Received: September 21, 2021. Revised: October 23, 2021. 198

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

LSPD: A Large-Scale Pornographic Dataset for Detection and Classification

Dinh Duy Phan[1,2] Thanh Thien Nguyen[1,2] Quang Huy Nguyen[1,2] Hoang Loc Tran[1,2]
Khac Ngoc Khoi Nguyen[1,2] Duc Lung Vu[1,2]*

[1]*Faculty of Computer Engineering, University of Information Technology, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University Ho Chi Minh City, Vietnam*
* Corresponding author's Email: lungvd@uit.edu.vn

Abstract: This paper introduces a new dataset named Large-Scale Pornographic Dataset for detection and classification (LSPD) that intends to advance the standard quality of pornographic visual content classification and sexual object detection tasks. As we recognize, the LSPD dataset is the first ever dataset for both object detection and image/video classification tasks in this area. The dataset gathers a large-scale corpus of pornographic/non-pornographic images and videos containing a rich diversity of context. The images and videos are not only labelled with their representative class but are also annotated by polygon masks of four private sexual objects (breasts, male and female genitals, and anuses). Our dataset contains 500,000 images and 4,000 videos, with more than 50,000 annotated images. To ensure fair use of the dataset, we present a detailed statistical analysis and provide baseline benchmarking scenarios for both image/video classification and instance detection/segmentation tasks. Finally, we evaluate the performance of four object detection algorithms and a Convolutional Neural Network (CNN) classifier on these scenarios.

Keywords: Dataset, Deep learning, Pornography image/video classification, Object detection.

1. Introduction

With the rapid development of technologies, the Internet has become an indispensable part of our lives. Within the environment of the internet, anybody can find and upload any digital content in the worldwide web. Although many of these contents are benign and useful, many others (such as pornography) are harmful. Censoring pornography is among the most challenging problems related to the Internet. Pornography spreading is illegal in many countries because it negatively impacts on people, especially on children. Moreover, many women have become victims of cyber-sexual crimes since releasing their private videos for sale via online chat rooms. Owing to these negative impacts, pornography detection and filtering has now become one of the most concerning problems. However, to another circumstance, people may access wide variety of content, including movies which consist of some adult scenes. Although these scenes are few and short, they can still be cut or even make the whole movie abandoned in some countries. However, cutting-off these adult contents may badly infer the user experiences because some of those is necessary for the whole content experience. Keeping the original content with least of changes while ensuring local regulation becomes a practical demand. To access this problem, we must clarify what constitutes pornography.

According to the oxford advanced learner dictionary, *"Pornography is magazines, DVDs, websites, etc. that describe or show naked people and sexual acts to make people feel sexually excited, especially in a way that many other people find offensive"*. From that definition, we can determine a pornographic image as one containing naked people. In other words, pornographic images explicitly illustrate sexual objects or organs i.e. bare breasts, genitals, and coitus. The problem now is not only detecting pornography but also handling erotic elements in a way that affects the user's experiences as little as possible. It means that erotic elements, such as breasts or genitals, have to be censored automatically in pornographic scene/image. This

problem directly involves object detection besides the image classification task.

Many researches [1-3] have proved that deep learning method is the state-of-the-art in the pornography classification and detection tasks. An appropriate dataset is essential to tackle these problems, especially for training deep CNN models which requires a large amount of data. Pornographic image classification requires a dataset of images divided into categories, while object detection and segmentation require object annotations. Although some datasets for pornography classification are available [4-6], their image quality is too low for today's applications. These datasets haven't been updated for a while, which makes them incapable to recognize recent types of pornography. In addition, there are not any dataset that includes these objects labelled for object detection task. This task might play a significant role in the film industry, e.g. automatic censoring, blurring, or sexual object replacement.

Motivated by these limitations, we propose a new benchmarking dataset named LSPD (large-scale pornographic dataset) for image/video classification and object detection tasks. The dataset is not only larger in quantity but also more variety in quality (resolution, content, duration...). According to our knowledge, this is the first visual pornography dataset that can be used for sexual object recognition. The dataset is also designed to be able to classify multiple types of pornographic content on images (such as hentai, sexy, porn...) rather than the convention porn/non-porn ones. This helps the classification task becomes more specific for practical implementation. With the dominance of short video social networks e.g. TikTok, Facebook Watch, our video dataset, diverse in short, middle, and long duration can be a useful resource helping models to deal with short porn clips or videos. Alongside this novel dataset, we propose a testing procedure with evaluation metrics and benchmarking scenarios and discuss the remaining challenges for further comparisons among future research.

The main contributions of this work are outlined below:

- We review the popular and significant methods for pornography classification along with human-sensitive object detection.
- We compile a novel and large-scale public dataset of 500,000 labelled images (including 50,212 annotated images with polygon instances on 93,810 sexual objects) and 4,000 labelled videos, which is probably the most comprehensive pornography dataset collected thus far. For detection tasks, the annotated set is labelled with four explicit sexual objects: male genitals, female genitals, breast, and anus. For classification tasks, the labelled set is divided into five main classes: sexy, hentai (porn drawing), porn, non-porn, and drawing.
- We develop benchmarking scenarios to evaluate classification and object detection tasks on image and video using the proposed LSPD dataset.
- Finally, we present some baseline performances on these benchmarking scenarios, with four object detection models You Only Look Once (YOLO), Single-Shot Multibox Detector (SSD), Mask R-CNN, and Cascade Mask R-CNN along with a CNN classifier. These results can be an initial point for further study on the LSPD dataset.

The remainder of this paper is organized as follows. Section 2 briefly reviews skin-based, handcrafted feature-based, deep learning-based, and object detection-based approaches for visual pornography classification. We also compare several related pornography datasets. Section 3 introduces our large-scale pornographic dataset for detection and classification, describing its quality, quantity, construction, and distribution in detail. Section 4 does not only provide some metrics and scenarios for evaluating methods on our LSPD dataset but also describes our experiments, results, and gives some further discussion overall. Finally, conclusions and suggestions for future works are given in section 5.

2. Related works

2.1 Pornography detection approach

The recognition and classification pornographic content have been a classic problem solved by various approaches and methods. Karamizadeh and arabsorkhi [7] divided porn-identification methods into six main categories: colour-based, shape-based, local and global feature-based, bag-of-words for filtering images, and deep learning methods. Alternatively, we recognize four main approaches to porn-identification: skin-based, handcrafted feature-based, deep learning-based, and sensitive-object based. Studies in each of these categories are briefly reviewed below.

2.1.1. Skin-based approaches

One of the earliest methods for pornography detection was the identification of naked people in

images or videos. These approaches try to identify the skin information of naked bodies based on vital factors such as skin proportion, histogram, and colour probabilities. Their focus is identifying whether a pixel is skin or not. Whether the image is pornographic or not is decided from the extent of nudity in the image. Combining shape and colour can improve the skin detection performance, and the upper features such as face and body parts can further strengthen the model accuracy. Some methods calculate the ratio of skin area on the face and body to distinguish a pornographic image when the nudity level exceeds a given threshold.

Moreira and fechine [8] proposed a skin-ratio detection method based on the red/blue/green (RBG) and the luma/blue-difference/red-difference (YCbCr) colour spaces. This model calculates the skin ratio in five major skin regions (R) of an image: the whole image (R1), the skin on both arms (R3, R4), belly skin (R5), and a rectangle R2 bounding R3, R4, and R5. The sensitivity of the image is determined from the skin ratios in these five regions. Balamurali and chandrasekar [9] combined face detection using the Viola-Jones algorithm with skin recognition on the YCbCr colour space. Their method calculates the skin ratio over the whole image. If the face pixels comprise less than 30 % of all skin pixels and the skin areas exceed 50 % of the whole image, the image is classified as pornographic; otherwise, it is classified as non-pornographic.

Zaidan [10] pointed out the advantages and disadvantages of real-life skin-based detection methods. The advantages of these methods are their quickness, uncostliness and effectiveness in classifying between obscene and benign images. However, one disadvantage of these approaches is dubious classification quality, as their performance depended on the quality and resolution of the input images likes lightning, illumination, or texture conditions. Furthermore, skin-based approaches can easily misclassify non-porn images containing skin-like objects or a vast amount of exposed body skin, such as images of swimming, wrestling, or bikini photo-shooting. This problem reduces the accuracy of skin-based pornography detection. The high false positive rate of these methods must be addressed.

2.1.2. Handcrafted feature-based approaches

Bag-of-visual-words (BoVW) approaches have recently been applied to image classification problems, including pornography recognition. A BoVW model extracts the key points from the low-level features in an image using various descriptors. Through mapping to a uniform representative vector, the extracted features are converted into a visual codebook representation. The codebook can be combined with a support vector machine (SVM) or some other classification algorithm to identify the pornography class of the image.

Various descriptors can extract handcrafted features in images. Lopes [11] added the colour information to the scale-invariant feature transform (SIFT) descriptor. They compared the performances of their so-called hue-SIFT and the original SIFT in the pornography problem. Because it combines the colour information and the local handcrafted features, hue-SIFT outperformed SIFT. Avila [5] presented an extension of the BoVW approach called BossaNova. This mid-level representation computes a histogram of the distances between the image features and descriptors in the codebook. This approach can represent the visual content-based concepts that distinguish pornographic from non-pornographic videos. The authors extracted videos from the public video dataset NPDI-800 as segmented shots called key-frames. The low-level features in the key-frames were extracted by a hue-SIFT descriptor, and their labels were detected by combining BossaNova (which encodes the local features) with a trained SVM classifier. The final result on each video was decided by a majority voting scheme. Moreira [4] introduced a spatio-temporal interest point detector and descriptor called temporal robust features (TRoF), and applied it to feature extraction from images. The features extracted by TRoF were aggregated into mid-level representations as a fisher vector, which is the state-of-the-art BoVW model. The TRoF method was evaluated on the NPDI-2k dataset (an expansion of the NPDI-800 dataset).

2.1.3. Deep Learning-based approaches

In recent years, pornography recognition (especially in visual contents) has been improved by robust deep learning-based methods. A particular deep learning model, which mainly is a multi-layer neural network capable for self-learning, is able to make classification decision based on learned features. Most of previous approaches involve deep learning on pornography detection apply transfer learning, which fine-tunes a pre-trained model to the target problem rather than training the neural network model from the beginning. Most of these methods employ a convolutional neural network (CNN) for image classification and object detection.

Nian [1] proposed a model that trains a deep neural network for pornographic image detection through two strategies: (1) model evaluation after fine tuning of a pre-trained mid-level representation,

and (2) adjusting the training data at an appropriate time based on the validation set performance. The data feeding to the training model for training were obtained by using an improved sliding-window method and were supported by data augmentation. mahadeokar [12] from Yahoo identified not-safe-for-work (NSFW) images after fine tuning a pre-trained thin ResNet-50 model on the ImageNet dataset. This open NSFW model scores the safety of each image. A low and high score denotes a safe and unsafe image, respectively. Similar to the NSFW classifier from Yahoo, many methods utilized an end-to-end CNN model (with different backbone networks or pre-trained models) for pornographic image classification [3, 13].

2.1.4. Sensitive object-based approaches

Sex organs and sensitive objects are known to carry rich information on the pornographic content of images and videos. Most pornographic visual content exposes sensitive sexual objects such as female breasts, genitals, and anuses, which raise the erotic level of the viewer. Some recent studies have identified sexually sensitive objects to determine the safety level of the visual content. Nugroho [14] proposed a model that detects nipples with a cascade classifier algorithm combined with haar-like features. To improve the true positive rate, they applied a face detection algorithm that distinguishes eyes from nipples. Wang [15] used a sliding window to generate multiple instances. Their model identifies possible sexual objects (female breasts, sex organs, and anuses) in an image. The center point of each sexual object is then marked as a keypoint using multiple instance learning. Finally, the model extracts the deep features by CaffeNet, and re-defines and fine-tunes those objects with GoogleNet. Tian [16] developed a sexual organ detector that detects female breasts, vulva, and male genitals. As part of the model development, they trained a colour-saliency preserved mixture deformable part model on the colour attributes and histogram-of-oriented gradient features, which reflect the shape and colour distributions of sexual objects in different poses. Tabone [13] proposed a classification system with seven sexual objects: buttocks, female breasts, female genitals (divided into two sub-classes: female genital posing and female genital active), male genitals, sex toys, and non-porn images. Eventually, they annotated these classes with five-set labelled points: one center point and four perpendicular offset points.

To identify explicit sexual objects in rectangle bounding boxes on an image, we employed four object-detection algorithms – mask R-CNN [17], YOLOv4 [18], SSD [19], and cascade mask R-CNN [20] – a pixel-level segmentation mask. Based on the detected information in the objects, we could determine if the content (image or video) was pornographic or not.

The above-mentioned approaches and their performances mostly on the NPDI datasets are presented (Table 1). The CNN-based method is the state-of-the-art method for pornographic content identification and classification.

Table 1. Summary of approaches in the existing literature

Method	Performance (Accuracy)	Evaluated Dataset
SIFT & Hue-SIFT Descriptor [11]	84.60%	Private Dataset
BossaNova + SVM [5]	89.50%	NPDI-800
Utilizing Deep CNN [1]	98.60%	Private dataset
Temporal Robust Features [4]	93.54%	NPDI-2k
Open NSFW + SVM [21]	88.40%	NPDI-2k
Open NSFW (thin ResNet50) * [3]	89.05%	NPDI-800
	82.18%	NPDI-2k
BossaNova Video Descriptor [22]	85.40%	NPDI-800
Multiple Instance DCNN-based Learning [15]	98.41%	Private dataset
	97.50%	NPDI-800
Deep Multicontext Network [2]	92.40%	NPDI-800
Shallow CNN (AlexNet) [3]	78.96%	NPDI-800
	79.26%	NPDI-2k
2-tiered SEIC Detector + SVM [21]	91.50%	NPDI-2k
Skin & Face detector + Random Forest [8]	96.96%	AIIA-PID4
End-to-end CNN (MobileNet) [13]	62.00%	Private dataset
Open NSFW + Mask R-CNN [23]	90.40%	NPDI-2k
Open NSFW + CNN Classifier (ResNet50) [23]	84.82%	NPDI-2k

* the original paper [3] did not work with the NPDI datasets.

2.2 Pornography dataset

The dataset crucially affects the detection and classification of visual content. Although pornography recognition is a long-standing problem, few datasets of pornographic visual content are publicly available owing to the sensitive and erotic nature of such material.

Algorithms and models based on those algorithms are usually evaluated on customized private datasets. Wang [15] tested their model on a dataset containing 155,000 pornographic images and 222,000 non-pornographic images.

Similarly, the AIC dataset introduced by xizi [24], includes 150,000 images divided into three classes (porn, normal, and sexy), but is also a private dataset. Consequently, these datasets cannot be used to benchmark other pornography detection methods. Connie [6] built a dataset for adult content recognition in images. This dataset, consisting of 41,154 pornographic images and 40,152 neutral images, is open but its images are of fixed size (128 x 128 pixels); therefore, they are generally unsuitable for fair evaluations. Karavarsamis [25] published the AIIA-PID pornography dataset, which contains 8,690 images in four classes: porn, bikini, skin, and non-skin. Unfortunately, we also were unable to access this dataset. Other authors [9], [13] also built their own datasets, but as these datasets contain a limited number of images, evaluations on them may not demonstrate the real performances of existing approaches.

At present, two main public pornography datasets containing videos are available: NPDI-800 [5] with 800 videos and NPDI-2k [4] with 2,000 videos. Although these NPDI datasets have been popularly used in experiments, they have a major drawback. Most of the videos are of lower quality and poorer resolution than today's pornographic videos. This drawback significantly affects the precision and accuracy of experimental evaluations. The constructions of the open pornography datasets and our LSPD dataset are compared (Tables 2 and 3).

Clearly, we require a standard dataset for training and evaluating different methods in this research field. This paper proposes a new dataset containing thousands of high-quality, high-resolution images, thus providing a standard dataset for pornographic visual content detection. The images and videos in the LSPD are of higher quality than those in the NPDI datasets, and better reflect today's pornography data. Our dataset is detailed in section 3.

3. LSPD dataset

As described above, although some datasets for pornography classification are available, their image quality is too low for today's applications. Moreover, there are not any public datasets with sensitive objects annotated for the object detection task. This study aims is to build a dataset, which can overcome the above limitations. One of the most useful applications can be used with our dataset is to use for detecting sensitive objects, blurring, hiding, or replacing these sensitive areas on images and video in social networks.

To provide the general intuitions beforehand, we present the specifications and purpose of the LSPD dataset (Table 4) along with the quantity distribution of data (Table 5). Then, the construction of LSPD is described in detail about the process from gathering data, filtering to annotating/labelling image and video for specific tasks. After that, statistical measurements are provided to provide the distribution of LSPD in categories for training, validating, and testing in quantity and quality. Furthermore, a comparison between the LSPD and other open pornographic datasets is made to provide some insight to our dataset qualification. Finally, metrics and evaluating scenarios – in term of two tasks: object detection and

Table 2. Comparison of pornographic image datasets

Dataset	Public	Porn	Non-porn	Total	# Class
Connie [6]	Yes	44,154	40,152	81,306	2 classes
Wang [15]	No	155,000	222,000	377,000	2 classes
Tabone [13]	No	16,033	1,656	17,689	2 classes
AIC [24]	No	50,000	100,000	150,000	3 classes
Karavarsamis [25]	No	1,891	6,799	8,690	4 classes
LSPD (ours)	**Yes**	**250,000**	**250,000**	**500,000**	**5 classes**

Table 3. Comparison of pornographic video datasets

Dataset	Public	Porn	Non-porn	Total	# Class
NPDI-800	Yes	400	400	800	2 classes
NPDI-2k	Yes	1,000	1,000	2,000	2 classes
LSPD (ours)	**Yes**	**2,000**	**2,000**	**4,000**	**2 classes**

Table 4. Specifications of the LSPD dataset

LSPD Specification	
Subject area	Visual content classification; object detection and instance segmentation; deep learning; convolutional neural network;
Specific subject area	Pornography image/video classification; sexual object detection;
Data format	2D-RGB image (.jpg, .jpeg) 2D-RGB video (.mp4)
Annotation file format	JSON file (.json) with polygon annotating structured described in VGG Image Annotator[1]
Data accessibility	The dataset is freely accessible for any research purposes[2]
LSPD value	
• The LSPD dataset can be used for training, validation, and benchmarking of algorithms for visual pornography classification and sexual object detection/instance segmentation; • Images in LSPD intentionally include irrelevant objects which may affect the accuracy of Deep CNNs trained on this dataset in real applications, e.g. skin-coloured sausage for male genital; • The LSPD can be used to develop new Deep CNN architectures or modify the existing ones, e.g. ResNet, YOLO, Mask R-CNN, in order to increase the efficiency of the network for pornographic recognition and sexual object detection; • The LSPD dataset can be used to help develop practical implementation to tackle real-world problems, e.g. automatic blurring or censoring inappropriate scenes, or clothing nude body;	

visual classification – of the LSPD are proposed for methods to benchmark their performances.

3.1 Data construction

As shown in section 2, high quality and large quantity datasets are highly demanded in pornography detection and classification tasks. To resolve this problem, we built our LSPD dataset by extracting a large number of images and videos containing both porn and non-porn contents from the internet.

For more than six months of the development process, we constructed the LSPD dataset through three steps: (1) image/video collection, (2) image/video filtering and labelling, and (3) image annotation. The whole image dataset is divided into two labelled and annotated subsets corresponding to two main sub-tasks: classification and detection. The images in the labelled subset are classified as non-porn or porn based on their main characteristic. Meanwhile, the annotated data consist of explicit pornographic images branded with polygon masks and their respective labels of specific sexual objects. The construction of the LSPD dataset, including data labelling and annotating, is described in the following subsections.

Table 5. Data distribution of the LSPD dataset

Data type	**Status**	**Class**	**Quantity**
Image	Labelled	Porn	200,000
		Non-porn	150,000
		Sexy	50,000
		Hentai	50,000
		Drawing	50,000
		Total	**500,000**
	Annotated*	4 sexual objects	50,212
Video	Labelled	Porn	2,000
		Non-porn	2,000
		Total	**4,000**

* subset of labelled porn images

3.1.1. Data collecting process

The porn images were mainly collected from adult content websites on the Internet. The non-porn images were obtained by searching for images on the Google search engine with approximately 250 keywords on several topics including people, nature, urban, rural, cartoon, art, transport, economics, and science. Then we chose approximately 1,000 images to download for each keyword's result.

On the other hand, the videos were collected from both adult and non-porn video websites. The videos contain varied scenes and are of various qualities, being produced by both amateur and professional users. Among the videos collected were hentai, cartoon, news, film, and music.

3.1.2. Data filtering and labelling for classification tasks

Image set: To ensure the quality of the dataset, we checked all downloaded images manually. Each image was retained only if it satisfied the following requirements:

- The width and height of the image must be at least 300 pixels.
- The content must be sufficiently clear to be recognized by normal people.

[1] http://www.robots.ox.ac.uk/~vgg/software/via

[2] http://uit.edu.vn/~LSPD

drawing hentai normal porn sexy

Figure. 1 Sample images in our dataset (left to right): drawing, hentai, non-porn, porn, and sexy

After double-checking by at least two authors, images that failed the above requirements were removed from the dataset.

The filtered images were classified into suitable categories for building the labelled set. Instead of dividing the dataset into porn and non-porn categories, we extended the number of categories to five: porn, hentai, drawing, sexy, and non-porn. All images in the porn class contain at least one pornographic characteristic, such as a sexual object or a sexual action. The hentai class contains pornographic drawings rather than photographs. This content type has grown in recent years but has received insufficient attention in previous works. The drawing class contains any comics, cartoons, or drawings with non-porn content. The sexy class is a set of non-porn photographs containing characteristics that might place them in the porn class, such as high skin ratio (bikini wearing, swimming activity) or sexual posture. Finally, the non-porn class contains all non-pornographic photographs that belong to neither the drawing nor the sexy class. Samples of these classes are represented in Figure. 1.

Finally, 6 co-authors voted 1 of 5 categories for each image, then we selected the final category for each image with the highest votes. Luckily, our consensus was so high that most members chose the same category for each image.

By developing five-category labels rather than the classic porn/non-porn classes, we hope to recognize various types of pornography in our further works. Moreover, these categories are easily combined into the conventional positive/negative classes for standard classification.

Video set: The video set was divided into only two classes: porn and non-porn. All videos with resolutions lower than 240p were discarded. We also removed videos that did not obviously fit into one of the categories. The filtered video dataset contained 4,000 videos including 2,000 porn, and 2,000 non-porn videos.

3.1.3. Image annotating for detection tasks

To build the annotated set for object detection tasks, we, 6 male authors with the ages ranging from 23 to under 50 years old, randomly selected 50,212 images from the porn and hentai classes, then branded four main sexual objects on these images: male genitals, female genitals, female breast, and anus. Using the VGG image annotator tool, we annotated every image with polygons and their respective labels describing sexual objects. This annotated information is easily converted into segment bitmaps for Mask R-CNN and Cascade Mask R-CNN training or rectangle bounding box coordinates for SSD and YOLOv4 training with polygon format structure, which is shown in Figure. 2. During the annotating process, the following rules were applied to ensure the quality of the annotated set:

- The annotated sexual objects must be clear and visible, and the annotations must not overlap with other annotations.
- The female breast is annotated only when its nipples are visible.
- The area of the annotated objects cannot be less than 20 x 20 pixels.

```
{ 'filename': 'image_name.jpg',
          'regions': {
              '0': {
                  'region_attributes': {},
                  'shape_attributes': {
                      'all_points_x': [...],
                      'all_points_y': [...],
                      'name': <class_name>}},
              ... more regions ...
          },
          'size': <file_size>
}
```

Figure. 2 Image annotation format structure

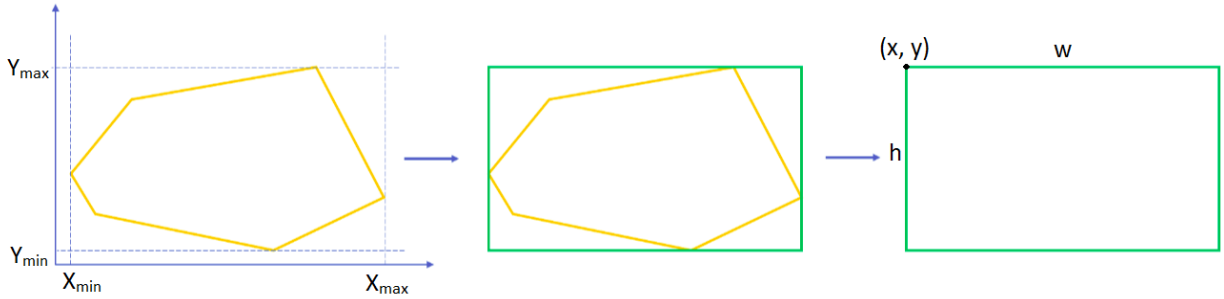

Figure. 3 Conversion from a polygon mask to a rectangular bounding box

The annotation process was performed in two phases:

- Phase 1: We annotated 30,000 explicit pornographic and hentai images from scratch. These annotations were then available for training object detection models.
- Phase 2: The Mask R-CNN model was trained on the annotated data from phase 1. The trained model then predicted the segmentation mask on the remaining 20,212 images to expand our annotated set. The annotations were re-checked by five authors to ensure their correctness. The checking was performed by (1) removing false predicted segments, (2) modifying incorrect segments, and (3) annotating wrongly detected objects.

To create the annotating information, we extracted the contour of the predicted segmentation bitmap of Mask R-CNN and converted it into coordinates. The converted annotations reduced the labelling time by approximately three quarters from that of manual annotation. However, this method has two drawbacks. First, the annotating results are highly influenced by the Mask R-CNN model's performance, necessitating the removal of wrongly annotated objects and the manual annotation of missed objects; nevertheless, correcting these errors was less time-consuming than annotating from scratch. Second, a number of images must be annotated from scratch for initial training of the Mask R-CNN model. Once trained, the model could predict and generate annotating information for the remaining large number of images.

Among the coordinate information extracted from a Mask R-CNN prediction of a single object, we can easily determine four values, X_{max}, X_{min}, Y_{max}, and Y_{min}, for constructing the rectangular bounding box (Figure. 3). Specifically, the coordinates of the bounding box are calculated as follows:

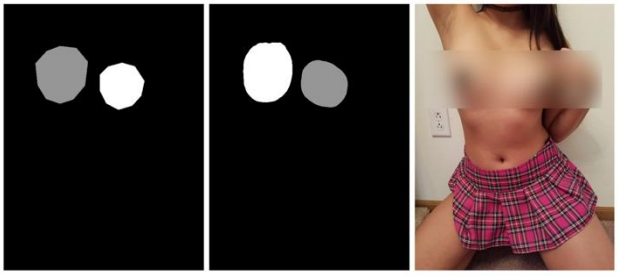

Figure. 4 Bitmaps showing the polygon masks annotated by hand (left) and predicted by Mask R-CNN's (center). The original image is shown at the right.

$$\begin{cases} x = X_{min} \\ y = Y_{max} \\ w = X_{max} - X_{min} \\ h = Y_{max} - Y_{min} \end{cases} \quad (1)$$

Besides reducing the annotating time, the automatic Mask R-CNN annotation generates a smoother polygon mask than handwork annotation while maintaining the localization accuracy. This advantage is depicted in Figure. 4.

To our knowledge, our dataset is the largest dataset for pornographic image classification. Unlike other pornographic datasets, it also contains annotations for the recognition and segmentation of sexual objects.

3.2 Data statistics

3.2.1. Image set

Table 6 shows the distributions of images in terms of resolution. More than half of the LSPD 500,000 images, overall, have the resolution < 0.5 Megapixel. Still, there are a lot of image data have the resolution more than 2 MP. We split the 500,000 images into three subsets: train, val, and test at ratios of 70 %, 20 %, and 10 %, respectively (Table 7). Similarly to the main dataset, the annotated set including 50,212 images with 93,810 instances mask was divided into three corresponding subsets, namely train-set, val-set, and test-set (Table 8).

Table 6. LSPD image resolution

Class	< 0.5 MP	0.5 – 1 MP	1 – 2 MP	> 2 MP	Total
Porn	99,439	30,354	24,054	46,153	200,000
Non-porn	130,975	5,341	4,061	9,623	150,000
Sexy	11,296	12,478	16,315	9,911	50,000
Hentai	33,948	9,709	4,210	2,133	50,000
Drawing	13,228	5,811	7,409	23,552	50,000
Total	**288,886**	**63,693**	**56,049**	**91,372**	**500,000**

Table 7. Distribution of the LSPD dataset

Class	Training	Validation	Testing	Total
Porn	140,000	40,000	20,000	200,000
Non-porn	105,000	30,000	15,000	150,000
Sexy	35,000	10,000	5,000	50,000
Hentai	35,000	10,000	5,000	50,000
Drawing	35,000	10,000	5,000	50,000
Total	**350,000**	**100,000**	**50,000**	**500,000**

Table 8. Distribution of annotated objects in the LSPD

Object	Training	Validation	Testing	Total
Breast	44,774	8,541	8,346	61,571
Male genital	8,525	2,791	2,649	13,965
Female genital	8,595	1,658	2,825	13,078
Anus	3,424	557	1,215	5,196
Total	**35,256**	**7,226**	**7,730**	**50,212**

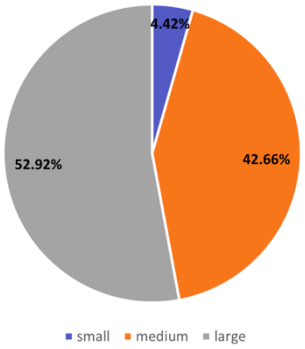

Figure. 5 Size distributions of objects in images

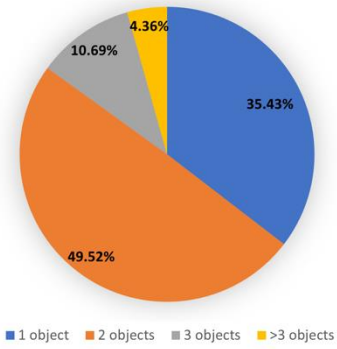

Figure. 6 Proportions of sexual objects in the images

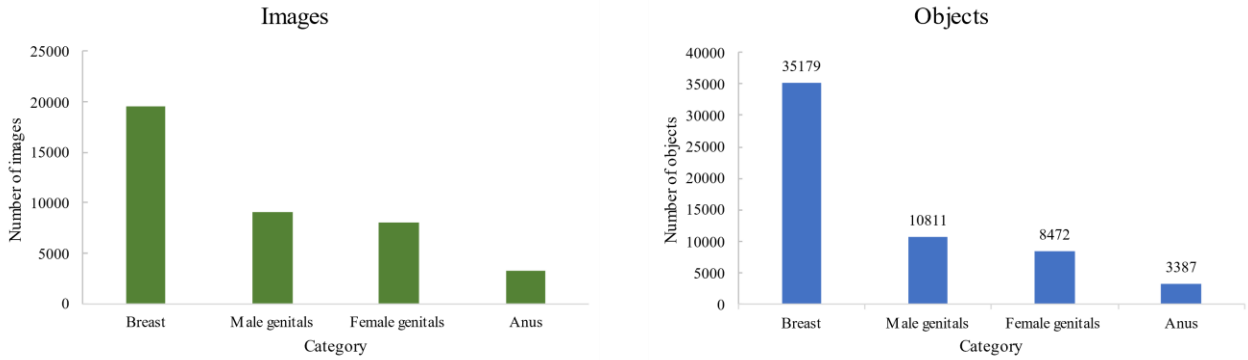

Figure. 7 Distributions of labelled and annotated sets

In the annotated 50,212 images set, small, medium, and large instance objects are defined as objects with areas smaller than 32 x 32 pixels, between 32 x 32 to 96 x 96 pixels, and larger than 96 x 96 pixels, respectively. Figure. 5. illustrates the distribution of small, medium, and large instances which is 4.42 %, 42.66 % and 52.92 %, respectively. The large percentage of small and medium objects creates

another challenge for object detection algorithms because the small objects are harder to detect. Moreover, the ratios of sexual objects in the images are illustrated in Figure. 6. According to the statistics, most of the images contain one or two sexual objects; images having more than two sexual objects comprise approximately by 15 % of the dataset.

Owing to the popularity of female breasts in explicit sexual images on the internet, our datasets are imbalanced. As shown in Figure. 7, the "female breast" object is the most abundant object in both the annotated and labelled sets.

3.2.2. Video set

Overall, the quantity of video duration and quality of our LSPD are more diverse than the NPDI datasets (Table 9, Table 10), in both pornographic and non-pornographic video. Not only the LSPD also focus on shot, low resolution clip, which is quite popular on the internet. The large differences pose a challenge for models working with the LSPD dataset.

3.3 Benchmarking scenarios

We now provide an evaluation script for further comparison of methods, including metrics and benchmarking procedures, on the labelled and annotated sets for classification and detection tasks of the LSPD. Furthermore, the experiments are made with four object detection algorithms to evaluate the effectiveness of the LSPD benchmarking scenarios, with results some challengers are then discussed. We expect that these experimental scenarios and the characteristics of the LSPD will serve as new standard evaluations in future studies with the dataset.

3.3.1. Detection task

To evaluate the detection, we selected the mean average precision (mAP), which is the primary metric of evaluate object detection problem. The main equation of mAP is calculated as follow:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (2)$$

Here, AP_i (average precision) is the area under the precision – recall curve of image i, determined by calculating the intersection over union (IoU) between the predicted bounding box and the ground truth box of the annotated image. The mAP is the most popular metric for measuring the accuracy of object detectors. When evaluating the object detection tasks, we calculated the percentage score of mAP over the whole annotated testing set to benchmark the detectors.

Based on the mAP metric, we proposed two evaluating standards for the object detection tasks: the PASCAL VOC metric and the COCO primary challenge metric[3]. With these two standard mAP metrics, the object detection task can be evaluated in the most comprehensive way.

Benchmarking scenarios: To benchmark object detection algorithms on the LSPD dataset, we selected 7,730 annotated images with 15,035 annotated instances for 4 sexual objects (Table 11).

On this testing set, we developed two benchmarking scenarios for evaluating the entire testing set, one using the mAP standard of pascal VOC, the other using the COCO primary AP metrics. Besides representing the performances of object detection models, these two standard metrics

Table 9. Comparison of video resolution

Label	Dataset	< 0.5 MP	0.5 – 1 MP	1 – 2 MP	> 2 MP
Porn	NPDI-800	400	0	0	0
	NPDI-2k	999	1	0	0
	LSPD	**1,150**	**565**	**81**	**204**
Non-porn	NPDI-800	399	1	0	0
	NPDI-2k	703	277	4	16
	LSPD	**1,243**	**596**	**28**	**133**

Table 10. Comparison of video duration

Label	Dataset	< 1 min	1 – 5 min	5 – 10 min	10 – 20 min	> 20 min
Porn	NPDI-800	10	181	83	78	48
	NPDI-2k	148	316	362	126	48
	LSPD	746	745	233	179	106
Non-porn	NPDI-800	97	232	65	5	1
	NPDI-2k	293	605	87	14	1
	LSPD	986	661	175	125	53

[3] https://cocodataset.org/#detection-eval

Table 11. Detection testing objects

Category	Quantity
breast	8,346
male_genital	2,649
female_genital	2,825
anus	1,215
Total	**15,035**

Table 12. Testing set for image classification

Multi-categories	Binary class	Quantity
Drawing	Non-porn	5,000
Hentai	Porn	5,000
Non-porn	Non-porn	15,000
Porn	Porn	20,000
Sexy	Non-porn	5,000
Total		**50,000**

determine whether the model can recognize small, medium, and large objects in the pornographic object detection task.

3.3.2. Classification task

Metrics: In binary classification tasks, visual content is most suitably evaluated through the confusion matrix, which divides the predicted data into four indexes: number of true positive (TP), number of true negative (TN), number of false positive (FP), and number of false negative (FN). Based on these four values, we define the accuracy, precision and recall in Eqs. (3)-(5):

$$Accuracy = \frac{TP+TN}{Total\ image} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

Here, the accuracy score represents the probability of predicting the right label with respect to the ground truth over the whole dataset, the precision score measures the probability of predicting the right label, and the recall score measures the probability of finding all positive data. For LSPD classification tasks, we use the accuracy metric to benchmark our proposed model's effectiveness on the labelled data.

Furthermore, one video is considered as a set of sequential image frames. Under this assumption, we split the video into key-frames per second throughout the video's length for video recognition. Based on the predictions in these key-frames, we propose three main scenario criteria for judging the pornographic-ness of the video: (1) the number of key-frames predicted as porn must exceed θ frames, (2) the number of key-frames predicted as porn must exceed ε percent of all key-frames, and (3) the number of continuous key-frames predicted as porn must exceed σ frames.

Benchmarking scenarios: For the image classification task, we proposed two scenarios based on the original five categories of the LSPD images and the contrasting porn/non-porn classes. The quantity and label of each image category are presented in detailed (Table 12).

This combination was chosen because it reflects the actual context of pornography classification. That is, the classification is primarily concerned with recognizing whether the input contents are obscene or not, rather than classifying input contents into various labels only. We then proposed two classification tasks for this combination: multi-class classification of the five original categories, and binary-class classification of the porn/non-porn classes.

The performance of the multi-class classification task was evaluated by the accuracy metric, whereas the binary-class classification task was evaluated by the accuracy, precision, and recall. Both classification tasks were performed on all images of the LSPD dataset.

On the other hand, for the video classification task, an 800-video set randomly chosen from the 4,000 videos was reserved as the video testing set. The performance of the video testing set was scored by the accuracy metric, based on the porn/non-porn video decisions of the three metrics described above.

4. Results and discussion

In this section, we present our baseline results on the LSPD object detection and image/video classification tasks, using object detection algorithms. The following parts describe in detail the configurations of our models and their performances on individual benchmark tasks that we proposed earlier. These results can be a starting point for further research in the future. Finally, we discuss some challenges due to the baseline results, along with the necessary and the potential of the LSPD dataset for future direction.

4.1 Environment and configuration

4.1.1. Experimental environment

All experimental scenarios were conducted on 64-bit ubuntu 18.04.2 LTS operating system powered

Table 13. Customization of mask R-CNN for training

Stage	Epoch	Layers	Learning rate	Augment-ation
1	0-20	head	lr	no
2	21-60	4+	lr	yes
3	61-90	3+	lr/10	yes
4	91-100	all	lr/100	yes

by an Intel(R) Core(TM) i7-4790 CPU@3.60 GHz, 8 GB RAM with a Nvidia GeForce GTX 1660 8 GB. The experimental models were implemented on Keras 2.5.5 and TensorFlow 1.14.0 under Python 3.6. We also used the Google Colab Pro with an Intel Xeon 2.3 GHz, RAM 25 GB, and Tesla P100 GPU.

4.1.2. Training configurations

Mask R-CNN: The Mask R-CNN model implements the mask R-CNN from Matterport[4] with ResNet101 + FPN backbones. During the training process, we combined transfer learning of the previously trained data with augmentation methods (randomly flipped left and right training images). The Mask R-CNN delivered its best results at a default learning rate of 0.001, image resizing to 512 x 512 pixels, and 1,000 learning steps per epoch (750 training steps and 250 valuating steps). To optimize the training process, we trained 200 regions of interest on each image and limited the number of predicted instances to 50. To ensure the high performance of Mask R-CNN, we also excluded all predicted objects scoring below 0.9. Prior to training the Mask R-CNN, we modified the network layers, learning rate, and image augmentation (Table 13).

YOLOv4: The YOLOv4 data were trained on a pre-trained CSPDarknet53 model with mish activation. The batch size, iteration number (number of batch), and learning rate of the training were set to 64; 35,000; and 0.001, respectively. The first 1,000 iterations were regarded as warm-up steps. The training images were resized to 608 x 608 pixels. During the training process, we randomly flipped the images and applied mosaic data augmentation.

SSD: To evaluate the effectiveness of SSD on sensitive object detection, we adapted the baseline provided by tensorflow. This model received no initial training but began as a pre-trained model obtained by training ResNet-50 + FPN backbone on the COCO dataset (this model is also known as RetinaNet). The batch size, number of iterations, and learning rate of the training were set to 64; 17,000 steps (including 2,000 initial warm-up steps); and 0.04, respectively. The data were augmented by horizontal flipping and random cropping.

Cascade Mask R-CNN: We adapted pre-trained Cascade Mask R-CNN models provided by Detectron2[5]. These models are provided with a ResNet-50 + FPN backbone in two scheduled configurations: training over 12 epochs (1X) and training over 37 epochs (3X) on the COCO dataset. In each case, we fine-tuned the model over 250,000 iterations at a learning rate of 0.02 and a batch size of 4. The training images were randomly flipped for data augmentation.

Table 14. Models' performances on the image sexual object detection task

Model	Pascal VOC mAP	COCO mAP
Mask R-CNN	85.24%	**56.65%**
YOLOv4	86.55%	47.90%
SSD	81.53%	44.32%
Cascade Mask R-CNN	**88.08%**	52.14%

4.2 Baseline results

4.2.1. Detection task

Experiment Results: Cascade Mask R-CNN achieved the highest performance with a pascal mAP metric of 88.08 % and YOLOv4 on the second place with 86.55 %, while Mask R-CNN and cascade Mask R-CNN achieved the highest and second-highest COCO mAP scores, 56.63 % and 52.14 % respectively (Table 14). The performance of the Mask R-CNN-based models show their effectiveness on the object detection approach.

Challenge: Several challenges should be considered in a detection task. Among the most notable is the presence of sexually linked objects in non-porn images and videos. These objects are benign objects with some visual similarity to explicit sexual objects, such as sausages (which resemble the male genital). Another challenge is the imbalanced distribution of sexual objects; in particular, annotated female breasts outnumber the combination of the other three sexual objects. Finally, the size differences among small, medium, and large objects, along with changes in illumination, scale, and viewpoint, might affect the quality of sexual object recognition, especially in hentai images which portray sexual objects in various drawing styles. Addressing these challenges would refine our benchmarking of the effectiveness of object detection models.

[4] https://github.com/matterport/Mask_RCNN

[5] https://github.com/facebookresearch/detectron2

4.2.2. Classification task

Experiment results: Because the object detection methods were adapted for porn/non-porn classification tasks, we re-defined the four metrics (TP, TN, FP, FN) in terms of the ground truth labels of the images and the predicted results of the object detection models. Specifically, a TP image was a porn image in the ground truth and predicted to have at least one sexual object, a TN image was a non-porn image in the ground truth and predicted to have no sexual objects, a FP image was a non-porn image in the ground truth and predicted to have at least one sexual object, and an FN image was a porn image in the ground truth and predicted to have no sexual objects.

Based on the sexual object prediction results as well as metric definitions described above, results for the four object detection models are presented (Table 15) for the binary classification task. The table also presents binary classifying results from the CNN classifier from our previous study [23] – which combines ResNet50 neural network with softmax classifier. Unlike the object detection approach, the end-to-end CNN classifier – with diverge backbones or pre-trained model – has long been applied for pornographic image/video classification (Table 1). This CNN model that we utilized was trained on 350,000 images and evaluated on 50,000 images for its ability to discriminate both two labels (porn, non-porn) and five labels (porn, non-porn, hentai, sexy, and drawing) as described in the previous section.

As can be observed, cascade Mask R-CNN achieved the highest performance with 92.62 % in accuracy overall. Clearly, cascade Mask R-CNN outperformed the other object detection algorithms in both detection task (Table 14) and classification task (Table 15). On the other hand, although the precision and recall scores were almost similar among the four methods, the Mask R-CNN-based methods outperformed both YOLOv4 and SSD, with a precision score of 98.33 % on Mask R-CNN and a recall score of 89.95 % on its cascade improvement. The object detection models, overall, overcome the performance of the CNN classifier except the recall metric.

On the other hand, as object detection methods can only be leveraged for the classify in binary class, the CNN classifier is solely used to evaluate the multi-class classification (Table 16). The classifier achieved higher predictions on the positive classes (porn, hentai) in multi-class classification than on negative classes (drawing, non-porn, and sexy). Among the five classes, pornographic images were most accurately recognized (91.27 % accuracy), whereas non-porn images were least accurately recognized (62.19 % accuracy).

In contrast to image classification, YOLOv4 achieved the best performance in this task (Table 17).

Table *17*When judging pornographic videos containing $\theta \geq 3$ key-frames recognized as sensitive, YOLOv4 achieved an accuracy score of 87.25 % on the 800-video testing set (Table 18). This model also achieved the accuracy of 87.75 % and 87.00 % on pornographic videos with $\varepsilon \geq 10\%$ and $\sigma \geq 3$ on two video evaluating criteria, respectively. This significant improvement over the other methods was attributed to the high reliability of YOLOv4 on pornographic video recognition. We did not experiment the ResNet50 CNN classifier for video classification, and that will be done in further research.

Over three evaluation methods, the third criterion (number of continuous key-frame recognized as

Table 15. Models' performances on the image binary classification task

Models	Accuracy	Precision	Recall
Mask-RCNN	86.70%	**98.33%**	88.00%
YOLOv4	92.59%	97.03%	87.86%
SSD	85.32%	94.11%	85.64%
Cascade Mask-RCNN	**92.62%**	95.01%	89.95%
CNN classifier	87.22%	84.86%	**90.59%**

Table 16. CNN classifier's performance on the image multi-class classification task

Class	Accuracy
Drawing	79.62%
Hentai	88.76%
Non-porn	62.19%
Porn	91.27%
Sexy	70.22%
Total	79.02%

Table 17. Models' performances on the video classification task

Model	Counting frames	Percent-frames	Porn-continuous frame
Mask R-CNN	85.63%	86.38%	85.63%
YOLOv4	87.25%	87.75%	87.00%
SSD	81.16%	83.48%	82.88%
Cascade Mask R-CNN	84.88%	86.63%	86.13%

porn) requires the lowest predicted number of porn frames recognized for models to achieve the peak performances (Table 18). However, the second one (percent of key-frame recognized as porn) achieved the highest performance, slightly better than the third (
Table *17*). During the experiment, we noticed that the first and third evaluating criteria do not require models to predict all key-frames beforehand to draw the conclusions – as they only have to calculate the frame prediction until the threshold θ or σ condition is satisfied – while the second criterion demands models to do so. Thus, we believe if using the third criterion to recognize pornographic video, the processing time can reduce significantly while the high accuracy still maintains in prediction.

Challenge: The characteristics of the LSPD dataset pose challenges on the classification task. The first challenge is the large quantity of porn and non-porn images and videos with a variety of characteristics and categories, which is difficult to draw a concrete boundary. Second is the style difference between drawings and realistic pornographic or non-pornographic images, which might affect the performance of the classifier. Finally, accounting for the large gap between short and long videos is expected to improve the judgment of video recognition and classification.

4.3 Discussion

The large-scale pornographic dataset (LSPD) provides data – both on images and videos distinctly – to deal with both pornographic classification and sexual object detection problems. Through detailed benchmarking scenarios, the LSPD provides metrics and methods in detail so models can be evaluated on two tasks. With five labels for images and two labels for videos, the LSPD can be used for binary or multi-class classification tasks. On the other hand, the sexual objects are annotated in sophisticated polygon masks to deal with both object detection and object instance segmentation recognition tasks. As we recognized, this is the first dataset that provides detailed polygon annotations on large scale – including more than 50,000 annotated images with nearly 94,000 instance masks – for the visual pornographic object recognition problem. With nearly 500,000 images and 4,000 videos – largest in quantity and more variety in quality comparing to existing pornographic datasets – the LSPD can be a valuable resource for further research in many aspects of computer vision. With the rising of short-clip social networks such as TikTok, a majority of LSPD video data are short clips within 5 minutes duration, which is very useful for training models to catch up with the trending. We also provided baseline methods and results as the starting point for other studies that work with LSPD dataset in the future.

Table 18. Best threshold criteria for pornographic video classification

Model	Counting frames θ	Percent-frames ε	Porn-continuous frame σ
Mask R-CNN	8	10	2
YOLOv4	3	10	3
SSD	6	14	3
Cascade Mask R-CNN	9	21	3

With θ, ε, and σ, respectively denote threshold of three criteria including the number of frames, percent of frames and number of continuous frames recognized as pornographic.

However, there are works to be done. As we mainly focus on object detection models and how to implement them for classifying pornographic visual data, we haven't worked deeply with the instance segmentation evaluating scenario, although adapted two models mask R-CNN and its cascade version. We will work with instance segmentation algorithm using the LSPD dataset. Moreover, the video data didn't build for multi-class classification, thus there is no corresponding scenario for video benchmarking comparing with the image data and that will be tackled soon.

Taken together, although exist challenges, our LSPD dataset not only demonstrates its usefulness in pornographic classification and sexual organ recognition tasks but also can be a new standard resource for training, evaluating, and benchmarking models in this very problem. With various tasks that we proposed, the dataset can help models to learn for practical implementation, from detecting adult images and videos in detail to automatically censoring sensitive parts within image/video or replacing sexual organs or body-part with clothing to make it safe for the general viewers.

5. Conclusion

We proposed a new large-scale pornographic dataset named LSPD, which provides numerous high-quality images for the classification of pornographic visual content and the detection and segmentation of sexual objects. The quality and quantity of the images and videos in LSPD are higher than in the existing datasets for pornographic

classification. The private sexual objects in the pornographic images are annotated by both class labels and instance segmentation masks. Moreover, we proposed baseline benchmarking scenarios for recognition/classification tasks and suggested several challenges that should be addressed in further research. Besides providing a comprehensive dataset for pornography detection, we developed benchmark procedures for fair comparisons among different studies. Four notable object detection algorithms and an end-to-end CNN classifier were adapted to provide initial evaluation results on the LSPD dataset. The baseline results demonstrate the effectiveness and potential of the LSPD dataset to access the automatic sexual object detection task besides the pornography classification problem.

In future work, we hope to adapt our study to practical implementations, such as preventing or bowdlerizing sensitive content on social media. We hope to adapt our study to practical implementations, such as preventing or bowdlerizing sensitive content on social media. Moreover, this study can be used to build a helpful tool for automatically detecting sensitive objects and blurring or hiding these sensitive areas on social networks. This dataset is intended to advance future research on the pornographic study and other fundamental classification or detection problems.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Dinh-Duy Phan and Duc-Lung Vu; methodology, Dinh-Duy Phan and Thanh-Thien Nguyen; validation, Quang-Huy Nguyen, Hoang-Loc Tran, and Khac-Ngoc-Khoi Nguyen; formal analysis, Dinh-Duy Phan, Thanh-Thien Nguyen, and Quang-Huy Nguyen; resources, Thanh-Thien Nguyen, Quang-Huy Nguyen, Hoang-Loc Tran, and Khac-Ngoc-Khoi Nguyen; data curation, Dinh-Duy Phan, Thanh-Thien Nguyen, Quang-Huy Nguyen, Hoang-Loc Tran, Khac-Ngoc-Khoi Nguyen, and Duc-Lung Vu; writing—original draft preparation, Quang-Huy Nguyen, Thanh-Thien Nguyen, Hoang-Loc Tran, and Dinh-Duy Phan; writing—review and editing, Hoang-Loc Tran and Duc-Lung Vu; visualization, Thanh-Thien Nguyen, Quang-Huy Nguyen, and Khac-Ngoc-Khoi Nguyen; supervision, Dinh-Duy Phan and Duc-Lung Vu; project administration, Dinh-Duy Phan; funding acquisition, Duc-Lung Vu.

Acknowledgments

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2019-26-02.

References

[1] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, “Pornographic image detection utilizing deep convolutional neural networks”, *Neurocomputing*, Vol. 210, pp. 283-293, 2016.

[2] X. Ou, H. Ling, H. Yu, P. Li, F. Zou, and S. Liu, “Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy”, *ACM Transactions on Intelligent Systems and Technology*, Vol. 8, No. 5, pp. 1-25, 2017.

[3] A. Gangwar, E. Fidalgo, E. Alegre, and V. G. Castro. “Pornography and child sexual abuse detection in image and video: A comparative evaluation”, In: *Proc. of 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017)*, Madrid, Spain, pp. 37-42, 2017.

[4] 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*, Vol. 268, pp. 46-61, 2016.

[5] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. A. AraúJo, “Pooling in image representation: The visual codeword point of view”, *Computer Vision and Image Understanding*, Vol. 117, No. 5, pp. 453-465, 2013.

[6] T. Connie, M. A. Shabi, and M. Goh. “Smart content recognition from images using a mixture of convolutional neural networks”, In: *Proc. of IT Convergence and Security 2017*, pp. 11-18, 2018.

[7] S. Karamizadeh and A. Arabsorkhi. “Methods of pornography detection”, In: *Proc. of the 10th International Conference on Computer Modeling and Simulation*, pp. 33-38, 2018.

[8] D. C. Moreira and J. M. Fechine. “A machine learning-based forensic discriminator of pornographic and bikini images”, In: *Proc. of 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2018.

[9] R. Balamurali and A. Chandrasekar, “Multiple parameter algorithm approach for adult image identification”, *Cluster Computing*, Vol. 22, No. 5, pp. 11909-11917, 2019.

[10] A. Zaidan, H. A. Karim, N. Ahmad, B. Zaidan, and A. Sali, “An automated anti-pornography

system using a skin detector based on artificial intelligence: A review", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 27, No. 04, pp. 1350012, 2013.

[11] A. P. Lopes, S. E. D. Avila, A. N. Peixoto, R. S. Oliveira, and A. D. A. Araújo. "A bag-of-features approach based on hue-sift descriptor for nude detection", In: *Proc. of 2009 17th European Signal Processing Conference*, pp. 1552-1556, 2009.

[12] J. Mahadeokar and G. Pesavento, "Open sourcing a deep learning solution for detecting NSFW images", *Retrieved August*, Vol. 24, pp. 2018, 2016.

[13] A. Tabone, A. Bonnici, S. Cristina, R. A. Farrugia, and K. P. Camilleri. "Private Body Part Detection using Deep Learning", In: *Proc. of ICPRAM*, pp. 205-211, 2020.

[14] H. A. Nugroho, D. Hardiyanto, and T. B. Adji. "Nipple detection to identify negative content on digital images", In: *Proc. of 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 43-48, 2016.

[15] Y. Wang, X. Jin, and X. Tan. "Pornographic image recognition by strongly-supervised deep multiple instance learning", In: *Proc. of 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 4418-4422, 2016.

[16] C. Tian, X. Zhang, W. Wei, and X. Gao, "Color pornographic image detection based on color-saliency preserved mixture deformable part model", *Multimedia Tools and Applications*, Vol. 77, No. 6, pp. 6629-6645, 2018.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask r-cnn", In: *Proc of the IEEE international conference on computer vision*, pp. 2961-2969, 2017.

[18] A. Bochkovskiy, C. Y. Wang, and H. Y.M. Liao, "Yolov4: Optimal speed and accuracy of object detection", *arXiv preprint arXiv:.10934*, 2020.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. "Ssd: Single shot multibox detector", In: *Proc. of European Conference on Computer Vision*, pp. 21-37, 2016.

[20] Z. Cai and N. Vasconcelos. "Cascade r-cnn: Delving into high quality object detection", In: *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 6154-6162, 2018.

[21] P. Vitorino, S. Avila, M. Perez, and A. Rocha, "Leveraging deep neural networks to fight child pornography in the age of social media", *Journal of Visual Communication and Image Representation*, Vol. 50, pp. 303-313, 2018.

[22] C. Caetano, S. Avila, W. R. Schwartz, S. J. F. Guimarães, and A. D. A. Araújo, "A mid-level video representation based on binary descriptors: A case study for pornography detection", *Neurocomputing*, Vol. 213, pp. 102-114, 2016.

[23] Q. H. Nguyen, H. L. Tran, T. T. Nguyen, D. D. Phan, and D. L. Vu. "Multi-level detector for pornographic content using CNN models", In: *Proc. of 2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 1-5, 2020.

[24] X. Wang, F. Cheng, S. Wang, H. Sun, G. Liu, and C. Zhou. "Adult image classification by a local-context aware network", In: *Proc. of 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2989-2993, 2018.

[25] S. Karavarsamis, N. Ntarmos, K. Blekas, and I. Pitas, "Detecting pornographic images by localizing skin ROIs", *International Journal of Digital Crime and Forensics*, Vol. 5, No. 1, pp. 39-53, 2013.