applied sciences

MDPI

Article

PornNet: A Unified Deep Architecture for Pornographic Video Recognition

Zhikang Fu, Jun Li *, Guoqing Chen, Tianbao Yu and Tiansheng Deng

School of Computer and Electronic Information, Nanjing Normal University, Nanjing 210023, China; zhikangfu@yeah.net (Z.F.); 192235022@njnu.edu.cn (G.C.); 230119271@seu.edu.cn (T.Y.); 101101436@seu.edu.cn (T.D.)

* Correspondence: lijuncst@njnu.edu.cn

Abstract: In the era of big data, massive harmful multimedia resources publicly available on the Internet greatly threaten children and adolescents. In particular, recognizing pornographic videos is of great importance for protecting the mental and physical health of the underage. In contrast to the conventional methods which are only built on image classifier without considering audio clues in the video, we propose a unified deep architecture termed PornNet integrating dual sub-networks for pornographic video recognition. More specifically, with image frames and audio clues extracted from the pornographic videos from scratch, they are respectively delivered to two deep networks for pattern discrimination. For discriminating pornographic frames, we propose a local-context aware network that takes into account the image context in capturing the key contents, whilst leveraging an attention network which can capture temporal information for recognizing pornographic audios. Thus, we incorporate the recognition scores generated from the two sub-networks into a unified deep architecture, while making use of a pre-defined aggregation function to produce the whole video recognition result. The experiments on our newly-collected large dataset demonstrate that our proposed method exhibits a promising performance, achieving an accuracy at 93.4% on the dataset including 1 k pornographic samples along with 1 k normal videos and 1 k sexy videos.

Keywords: pornographic video recognition; local-context aware network; attention network; unified deep architecture

Citation: Fu, Z.; Li, J.; Chen, G.; Yu, T.; Deng, T. PornNet: A Unified Deep Architecture for Pornographic Video Recognition. *Appl. Sci.* **2021**, *11*, 3066. https://doi.org/10.3390/app11073066

Academic Editor: Hugo Pedro Proença

Received: 9 March 2021
Accepted: 26 March 2021
Published: 30 March 2021

1. Introduction

With the rapid development of the Internet, substantial short videos are uploaded freely onto the Internet by personal users every day. Among these videos publicly available, those with harmful or illegal contents are not only detrimental to personal mental health but also threaten social security and stability [1]. In particular, short pornographic videos seriously affect the mental growth of children and adolescents, since the underage have easy access to these harmful videos with the help of the Internet [2,3]. Therefore, pornographic video recognition is extremely important for preventing the current Internet environment from being contaminated, and thus plays a crucial role in protecting the mental health of the underage [4].

Although the last two decades have witnessed massive research devoted to recognizing pornographic images [5–8], pornographic video recognition is still an open problem. In general, the key information contained in the pornographic videos manifests itself in image frames and audio cues. Thus, these two modalities are usually extracted from the videos in the first place, and then handled separately for recognizing pornographic contents. On the one hand, the pornographic images exhibit significant intra-class variances when scenario, scale, and background change. In particular, the private part that distinguishes a pornographic image from normal images often accounts for a small local region, whereas the image background irrelevant to pornographic contents may consist of a large portion

of the image area. Previous research [2,9–16] focuses on searching a naked person and detecting skin regions by using low-level patterns [17] such as texture, color, and geometrical features, whereas it is unreasonable to assume that all images with large skin areas are pornographic. With the success of deep convolutional neural network (DCNN) in image classification and detection [18], more attention has been paid to the application of DCNN in porn detection [3,19,20]. However, distinguishing the difference between sexy photos and porn images is still a challenging task. On the other hand, massive pornography videos can just be accurately recognized by its audio information alone, while some videos with periodic screaming and moaning cannot be understood by image content, since no pornographic images are shown explicitly. Therefore, fusing pornographic audio and visual information is significantly beneficial for improving recognition accuracy. Despite massive efforts devoted to environmental sound classification (ESC) [21–25] in the audio domain, it is well known that few research focuses on recognizing pornographic audio. The most popular strategy aims to perform the feature embedding obtained by pre-processing an audio signal and subsequently delivers the resulting features to DCNNs.

To address the above challenges and difficulties, in this paper, we propose a unified deep architecture termed PornNet in which the two heterogeneous modalities of image and audio in the videos are separately handled for accurate pornographic video recognition. In terms of porn image detection, a detection-classification network (DCNet) based on an anchor free [24] and bidirectional feature pyramid network (BiFPN) [26] is proposed to capture the global and local information of images. More specifically, porn images are classified into three categories according to global information, i.e., normal, sexy, and porn images, whilst six categories which are breast_porn, vagina_porn, penis_porn, buttock_porn, breast_sexy, and buttock_sexy are defined according to local information. The proposed DCNet allows the extraction of discriminative features of normal, sexy, and porn images for both global information and local information from sensitive body parts. On the other hand, DCNN is used for recognizing pornographic audios for the first time. In particular, log Mel-Spectrogram is calculated as feature embedding from input audio data, while an additional log Mel-Spectrogram is generated for an image-like representation of the spectrum of frequencies varying with time [27]. Meanwhile, a Resnet-Attention network (RANet), which is also used as temporal segmentation network is proposed to extract inter-context and extra-context of log Mel-Spectrograms. To summarize, our contributions are listed as follows:

1. In order to capture the global and local information of porn images, we propose the DCNet including two carefully designed branches, namely detection and classification. In the detection branch, particularly, our proposed detector is anchor box free as well as proposal free, and thus completely avoids the complicated computation process related to anchor boxes such as setting the rate of the anchors. Besides, a weighted bi-directional feature pyramid network (BiFPN) is used to achieve multi-scale feature fusion;
2. We propose a RANet based on audio feature embedding for pornographic audio detection. Specifically, the feature embedding termed log Mel-Spectrogram is an image-like representation, and the number of features is equal to the audio seconds. Furthermore, a frequency attention block is used to extract the inter-spatial relationship of a spectrogram, while the framework of Temporal Segment Networks (TSN) [28] is used for capturing the relationship of spectrograms along the temporal dimension in RANet. To the best of our knowledge, this is the first attempt to introduce DCNN to recognize pornographic audio;
3. For pornographic video recognition, we specially assemble a dataset including 1 k real-world pornographic videos merged with 1 k videos and 1 k normal videos. Due to the privacy and copyright issue, we only show some examples analogous to our simulated data as illustrated in Figure 1. Experiments show that our proposed method can achieve an accuracy of 93.4% on the real-world dataset, demonstrating superior performance over the other state-of-the-art networks.

Figure 1. Some example sexy images analogous to our simulated data.

2. Related Work

Generally, methods for recognizing pornographic videos can be classified into image-based recognition and audio-based recognition approaches. Particularly, audio classification has attracted much attention in recent years, and thus our pornographic audio recognition benefits from recent advances in audio classification.

2.1. Porn Image Recognition

In terms of image representation, the existing methods for pornographic image recognition can be classified as hand-crafted feature-based and DCNN-based schemes.

2.1.1. Hand-Crafted Feature-Based Approaches

Prior to the advent of DCNN, conventional porn image recognition approaches rely on various low-level hand-crafted features to classify adult images. Wang et al. [29] made use of wavelet for image representation, while the normalized central moments, the daubechies wavelet transformation, and the color histogram are used to generate semantic-mating vectors for image classification. Zhao and Cai [2] combined the edge, color, and texture features along with SIFT descriptors for enhancing the recognition performance. Although hand-crafted features allow straightforward image representation, their limited discriminative power fails to capture the essential content of a pornographic image and thus leads to a degraded recognition performance.

2.1.2. DCNN-Based Approaches

With tremendous success achieved by DCNN in image classification, DCNN has been extensively used for recognizing pornographic images. Moustafa et al. [3] combined GoogLeNet [18] and AlextNet [30] to produce an ensemble model for porn image recognition. They show that the recognition accuracy of their model is slightly better than either one model. Mallmann et al. [31] considered the recognition of pornographic content as an object problem and used detection network for detecting pornographic private parts. Ou et al. [19] took full use of the complementarity of local context and global context information, and proposed a context-ensemble detection system with a fine-to-coarse strategy. Wang et al. [20] proposed GcNet and SpNet for capturing local and global context. Compared with the methods based on hand-crafted features, the major advantages of DCNN-based methods are two-fold. With sufficient descriptive and discriminating power, DCNN is capable of capturing the most sensitive features in porn images. Meanwhile, with the help of DCNN, those methods can effectively distinguish between sexy photos and porn images by combining local and global contents. Our proposed porn image recognition scheme falls into the group of DCNN-based methods.

2.2. Porn Audio Recognition

Different from image signals, an audio signal has distinct characteristics, and thus many methods are specifically tailored towards the audio domain. In general, the existing porn audio recognition methods can be roughly divided into two categories as follows:

2.2.1. Raw Waveform and 1D-CNN

In the 1D-CNN architecture, the raw waveform of an audio example is usually used as the input fed to the network. Tokozume and Harada [32] proposed a one-dimensional CNN architecture termed EnvNet which shows a promising performance using raw waveform data as input. Zhu et al. [33] used raw waveform data at different time scales as the input of 1D-CNN for improving performance. Abdoli et al. [34] used a gammatone filter bank for the initialization model which revealed an improved performance compared with the other random weight initialization methods. Note that these methods avoid the procedure of pre-processing the raw waveform data.

2.2.2. Time-Frequency Representation and 2D-CNN

In terms of 2D-CNN, raw waveform of audio data should be transformed into a two-dimensional representation, such as Mel-scaled spectrograms [35], Mel-frequency cepstral coefficients (MFCC) [36], and log-power Mel-Spectrogram [37]. In [38], 2D CNN is imposed on Mel-scaled spectrograms for environmental sound classification. Mydlarz et al. [39] proposed a 2D CNN architecture with five layers using the augmented data as new training samples. Guzhov et al. [40] proposed a 2D CNN with attention block termed EsResNet for Environmental Sound Classification. The EsRestNet uses log-pow SIFT spectrograms as input and achieves the state-of-the-art results on ESC-10/-50 [41] and UrbamSound8K [42]. In our framework, we adopt the EsRestNet-like structure which abandons the time block as our pornographic audio recognition network, whilst using log Mel-spectrogram as the input audio representation instead of log-pow Spectrograms. The framework of Temporal Segment Networks (TSN) [28] is employed for capturing the temporal context of audio examples.

3. Our Proposed Methods

In our framework, a video sample is decomposed into massive image frames and an audio file, each of which is handled by the corresponding network. The framework is illustrated in Figure 2. The DCNet is proposed to recognize pornographic frames and generate the results of images. The video-frames result is calculated through simple voting. Audio feature embeddings which are log Mel-spectrograms and image-like representation of the audio are produced by VGGish [37]. The RANet is used to recognize audio feature embeddings and generate video-audio result [43]. In the video-frames and video-audio fusion algorithm, a well-designed function is pre-defined to aggregate the recognition the result from video-frames and video-audio.

3.1. Detection-Classification Network

Figure 3 illustrates the architecture of DCNet for distinguishing the pornographic images in the video clips. To be specific, the network can be divided into four modules: The backbone, the bidirectional feature pyramid network (BiFPN) [26], the detection network, and global classification network.

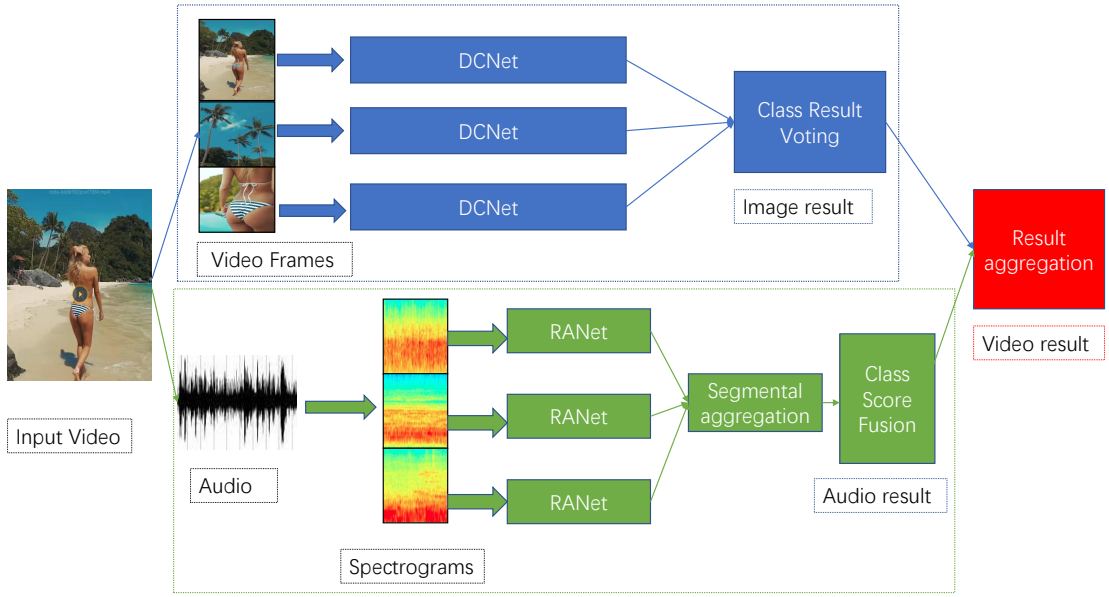

Figure 2. The framework of our proposed method. A video is decomposed into massive image frames and an audio file from scratch. Then, the detection-classification network (DCNet) is used to recognize video frames for porn image recognition, whilst the Resnet-Attention network (RANet) is proposed for porn audio recognition. The final score of the whole video being pornographic is obtained by fusing the image and audio recognition results.

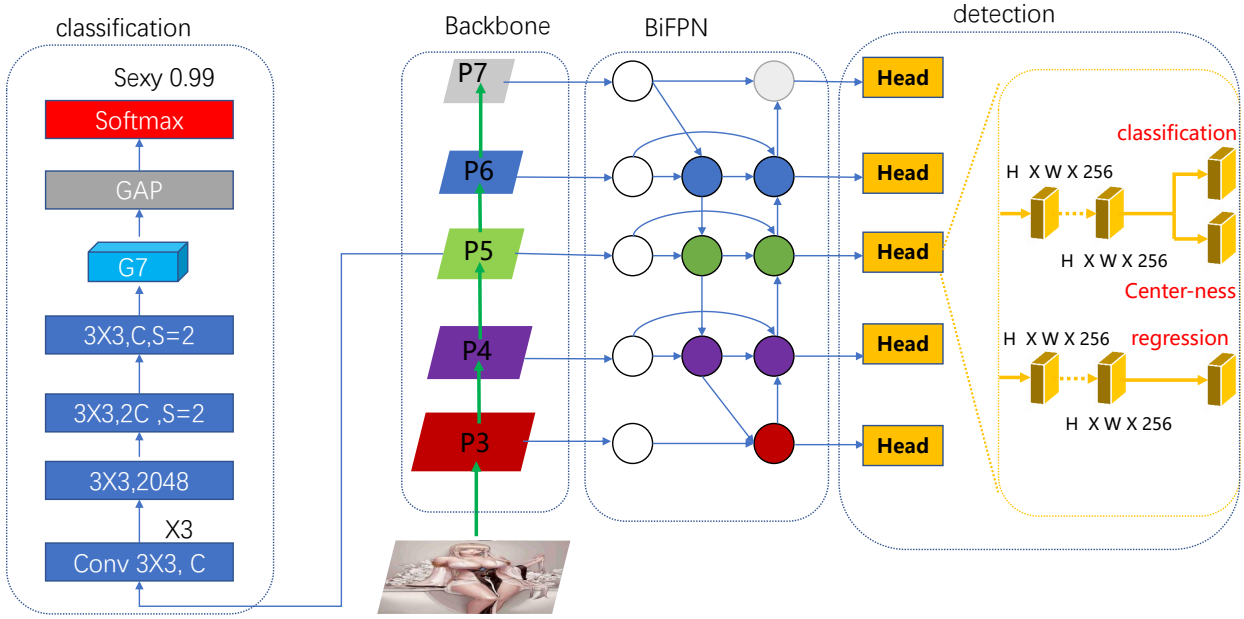

Figure 3. The DCNet architecture for porn image recognition. In DCNet, the bidirectional feature pyramid network (BiFPN) is used to achieve multi-scale feature fusion, while the classification network generates the recognition score of the video falling into any of the three categories: Normal, sexy, and porn. Besides, the detection branch is capable of capturing local information. Note that GAP denotes global average pooling and C is the channel of feature maps. In addition, S is the stride of convolutional kennel while $H \times W$ is the height and the width of feature maps.

With the ResNet-50 used as our backbone in our DCNet, BiFPN is built on the top layers at each stage of the ResNet-50. More specifically, activations from the 3rd layer to 7th layer $\tilde{P}^{in} = (P_3^{in}, ..., P_7^{in})$ are used as input features delivered to the subsequent BiFPN. P_i^{in} represents a feature level with $1/2^i$ resolutions of the input images. Here, $(P_3^{in}, P_4^{in}, P_5^{in})$ are computed from top-down and lateral connections to the output of the

convolutional layers at each residual stage of backbone network. P_6^{in} and P_7^{in} are obtained by imposing one convolutional layer on P_5^{in} and P_6^{in} separately with the stride at 2. Considering cross-scale connections, a bidirectional path, top-down, and bottom-up, works as one feature layer imposed on the same layer multiple times as shown in Figure 2. To balance among different input features at different resolution scales with different contributions, an additional weight for each input layer is used such that the network is capable of learning the importance of each input feature. Considering the computational efficiency, fast normalized fusion strategy is used as the weighted fusion approach:

$$O = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_j} I_i \tag{1}$$

where $\omega \geq 0$ by applying a Relu layer after each ω_i, while $\varepsilon = 0.001$ indicates a small value for avoiding numerical instability. In a nutshell, BiFPN integrates both the fast normalized feature fusion and the bidirectional cross-scale connections. Mathematically, the two fused features at level 5 for BiFPN are formulated as follows:

$$P_5^{td} = Conv\Big(\frac{\omega_1 P_5^{in} + \omega_2 Resize(P_6^{in})}{\omega_1 + \omega_2 + \varepsilon}\Big) \tag{2}$$

$$P_5^{out} = Conv\Big(\frac{\omega_1' P_5^{in} + \omega_2' P_6^{td} + \omega_3' Resize(P_4^{out})}{\omega_1' + \omega_2' + \omega_3' + \varepsilon}\Big) \tag{3}$$

where $Resize(\cdot)$ usually denotes upsampling or downsampling operation for resolution matching. P_5^{td} is the intermediate feature at 5th level on the top-down pathway, while P_5^{out} is the output feature at the 5th level on the bottom-up pathway.

The ground-truth bounding boxes in an image are defined as $\{B_i\}_{i=1}^N$, where $B_i = (x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)}, c^{(i)}) \in \mathbf{R}^4 \times \{1, 2, \ldots, C\}$. Here $(x_0^{(i)}, y_0^{(i)})$ and $(x_1^{(i)}, y_1^{(i)})$ are the coordinates of the left-top and right-bottom corners of the bounding box. $c^{(i)}$ denotes the class that the object in the bounding box belongs to while C is the number of classes. Inspired by [26], the detection network is anchor box free as well as proposal free. For each location (x, y) on the feature map P_i^{td}, the corresponding mapping location in the input image is $(xs + \frac{s}{2}, ys + \frac{s}{2})$ (s denotes the stride) near the center of the receptive field. Similar to the FCNS for semantic segmentation [44], our detection network directly uses the location-specific image regions as training samples instead of anchor boxes in anchor-based detectors. Specifically, when a location (x, y) falls into any ground-truth box, it is considered as a positive sample with the ground-truth label C^*. Otherwise, it is viewed as a negative sample. We use a 4D vector $t^* = (l^*, t^*, r^*, b^*)$ to denote the regression targets for the location. Here l^*, t^*, r^*, and b^* are the distance from the location to four sides of the bounding box. Thus, two scenarios usually occur. Firstly, if a location falls into multiple bounding boxes, we simply choose the one with minimal area used as its regression target. Secondly, unlike the anchor-based detectors assigning anchor boxes with different sizes to different levels of feature map, the range of bounding box regression for each level is limited. If a location satisfies $max(l^*, t^*, r^*, b^*) > m_i$ or $max(l^*, t^*, r^*, b^*) < m_{i-1}$, the location is defined as a negative sample. Here, m_i is the maximum distance that i^{th}-level feature map needs regression. In our work, m_2, m_3, m_4, m_5, m_6, and m_7 are set as 0, 64, 128, 256, 512, and ∞, respectively. Otherwise, to suppress the detected low-quality bounding boxes which are far away from the center of an object, a single-layer branch, in parallel with the classification branch, is used for predicting the "centernes" of the location. Mathematically, the centerness target is defined as:

$$centerness = \sqrt{\frac{min(l^*, r^*)}{max(l^*, r^*)} \times \frac{min(t^*, b^*)}{max(t^*, b^*).}} \tag{4}$$

In addition to the above-mentioned detection network, the global classification network classifies an image into three categories: normal, sexy, and porn with the ground-truth label of an image defined as g^*. Note that it is built on the last stage of backbone network. To generate high-level feature map G7, we make use of P_5^{in} for the input features of the six convolutional layers, followed by a global average pooling layer, a fully connected layer with softmax activation used for classification.

Mathematically, the training loss function of our DCNet is formulated as follows:

$$L = \lambda_1 L_{global\_cls} + \lambda_2 L_{detect} \tag{5}$$

where;

$$\begin{aligned} L_{detect} = & \frac{1}{N_{pos}} \sum_{x,y} L_{cls}(p_{(x,y)_{cls}}, c^*_{x,y}) \\ & + \frac{1}{N_{pos}} \sum_{x,y} L_{cent}(p_{(x,y)_{cent}}, centerness^*) \\ & + \frac{1}{N_{pos}} \sum_{x,y} I_{\{c^*_{x,y}>0\}} L_{reg}(t_{x,y}, t^*_{x,y}) \end{aligned} \tag{6}$$

where L_{global_cls}, L_{cls}, L_{cent}, and L_{reg} are cross entropy loss, focal loss [45], binary cross entropy loss and IOU loss [46]. Besides, N_{pos} denotes the number of positive samples, whilst λ_1 and λ_2 are tradeoff weights both of which are empirically set as 0.5.

3.2. ResNet-Attention Network

Figure 4 illustrates the architecture of RANet for pornographic audio detection. With the feature of audio data fed to RANet, the network can be divided into three modules: the backbone, the frequency attention block, and the temporal attention block. Consistent with DCNet, the backbone of RANet is used as the ResNet architecture.

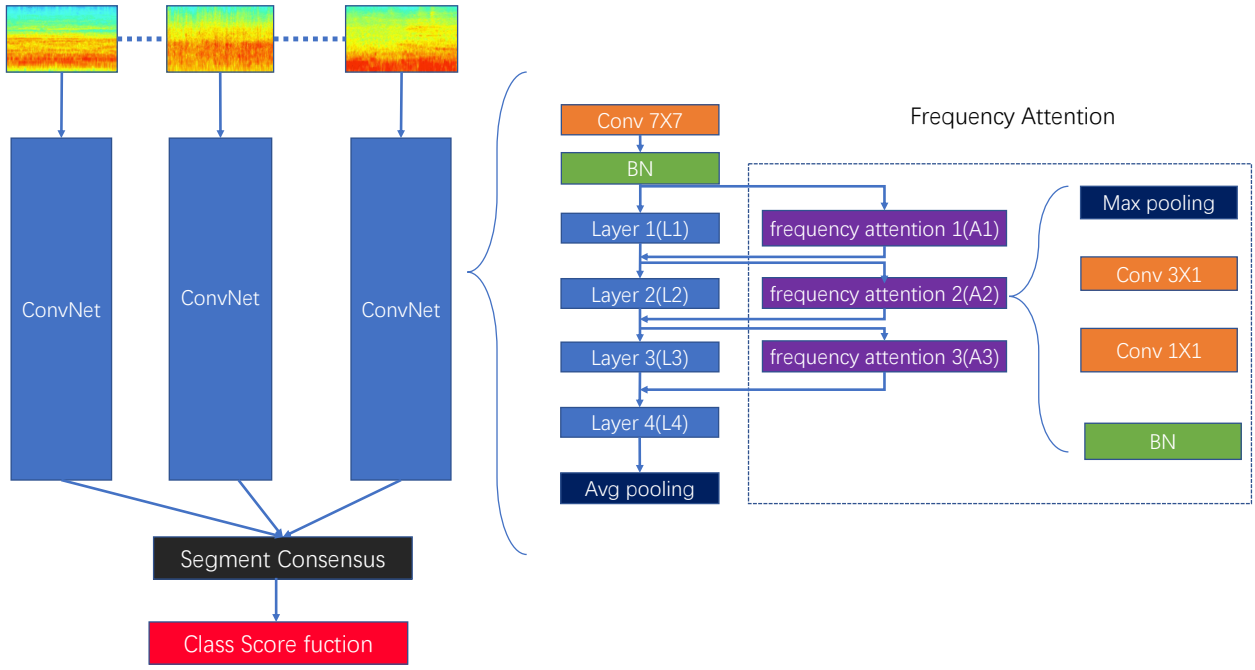

Figure 4. The RANet architecture for pornographic audio detection. Inspired by Temporal Segment Networks (TSN) [28], K log Mel-spectrograms are generated from the audio data and delivered into the RANet. In addition, the frequency attention which contains three attention blocks is used for capturing the most important information in frequency domain. Thus, the recognition scores of K log Mel-Spectrograms are fused by the segmental consensus function for porn audio recognition.

In our work, we use VGGish [37] to generate audio feature embedding from audio samples. In the pre-processing procedure, an input audio is first resampled to 16 KHZ, and then we compute log Mel-spectrogram $M \in \mathbf{R}^{H \times W}$ from every one second of the transformed audio data. Here, H is 96 and W is 64. The total number of log Mel-spectrograms generated from the audio sample is equal to the elapsed seconds. Formally, given a set of log Mel-spectrograms calculated by VGGish, we evenly divide them into K parts $\{S_1, S_2, ..., S_K\}$. Inspired by TSN [28], a log Mel-spectrogram T_K is randomly sampled from its corresponding segment S_K. Then, the RANet models a sequence of spectrograms $(T_1, T_2, ..., T_K)$ as follows:

$$RAN(T_1, T_2, ..., T_K) = H(g(F(T_1 : W), F(T_2 : W), ..., F(T_K : W))). \tag{7}$$

Here, $F(T_k : W)$ indicates a ConvNet with parameters W which operates on T_k and produces class scores. To achieve a consensus of class hypothesis among them, the segmental consensus function g which is defined as $g_i = \sum_{k=1}^{K} A(T_k) f_i^k$ combines the outputs from multiple spectrograms. $A(T_k)$ is the attention weight for T_k. Based on this consensus, the softmax function **H** predicts the probability of the whole audio being pornographic.

In our RANet, the frequency attention block enables capturing the most important information in frequency domain. To incorporate the frequency attention mechanism into our framework, we propose improving the Resnet network by adding a stack of attention blocks in parallel as shown in Figure 3. For instance, the first attention block frequency attention A_1 reconceives the same input x as the first layer L_1. Next, it processes x by frequency-dedicated convolutional filters and thus produces an output of the same shape as the one provided by L_1. At last, the input L^{att} of the second layer is constructed by the element-wise multiplication of A_i and L_1 blocks:

$$L_i^{att}(x) = L_i(x) \odot A_i(x) \tag{8}$$

3.3. Fusion of Pornographic Image and Audio Recognition Results

In our scenario, we extract image frames and audio data from the given video with a 1fps sampling rate and 16KHz sampling frequency respectively. Thus, the generated N images and the audio data are delivered to our DCNet and RANet for pornographic content recognition. Furthermore, with the classification result of each image frame $R_m^i \in \{0, 1, 2\}$ obtained, we aggregate the recognition results of all the images via voting strategy, leading to the aggregated result $R_m \in \{0, 1, 2\}$. Here, 0, 1, 2 represent three image classes, i.e., normal, porn, and sex. Analogously, the recognition result of the audio data can be computed as $R_a \in \{0, 1\}$. Thus, the following aggregation function is pre-defined to fuse the results of the porn image and audio recognition:

$$R = \begin{cases} 1(porn) & R_m = 1 \; or \; R_a = 1 \\ 0(normal) & R_m = R_a = 0 \\ 2(sexy) & R_m = 2 \; and \; R_a = 0 \end{cases} \tag{9}$$

Equation (9) can be interpreted as the following three cases: Firstly, the test video is identified as pornographic when either the image or audio data in the video is classified as pornographic. Secondly, the test video is normal when both of the two modalities are recognized as normal. Thirdly, the test video is classified as sexy when the audio data is normal whereas the image data is identified as sexy.

4. Experiments

4.1. Dataset

Since no public datasets are available for the task of pornographic video recognition, we have checked 100,000 videos on the Internet and collected a large-scale dataset from them. The newly assembled dataset consists of 10,000 pornographic videos, 10,000 videos, and 10,000 normal videos. Specifically, 8934 videos contain pornographic images and 8676 videos contain pornographic audios in the 10,000 pornographic videos. The average length of these videos is two minutes. All the pornographic videos involved in our dataset come from three pornographic web sites and are captured by personal mobile phone. Overall, they are categorized into two groups in terms of the video contents, namely nudity-typed and behavior-typed pornographic videos. The former type refers to videos revealing a human private part, such as a naked breast, vagina, penis, and buttock. Different from the nudity-typed videos, the latter video type represent those exhibiting pornographic behaviors, whereas the aforementioned human private part is not shown in videos. The pornographic videos are usually captured by personal users in an unprofessional way and thus they contain complex backgrounds with undesirable video quality.

In addition to the above-mentioned porn videos, the videos were downloaded from web sites. Similar to the pornographic videos in appearance, videos include bikini, a seductive posture, and man or baby with a bare upper body, demonstrating semi-exposed human private part, such as semi-exposed breast and buttock. The normal videos in our dataset are also downloaded from web sites, and can be categorized into two groups, namely normal-human type and no-human type. In the former type of normal videos, people in these videos are normally dressed, while the videos of no-human type cover a variety of topics including animals, natural, and living goods without humans contained.

The dataset is split into three partitions, the training set, validation set, and test set. Specifically, 80% of the data are used for training, while the rest are evenly divided for validation and test respectively. As aforementioned, we sampled image frames from videos with a 1fps sampling rate along with the audio data with 16 KHz frequency. In terms of DCNet training, we manually select 20,000 porn images, 20,000 sexy images, and 20,000 normal images from the sampled frames, and make use of bounding boxes to annotate the sensitive contents of the training images, including breast_porn, vagina_porn, penis_porn, buttock_porn, breast_sexy, and buttock_sexy. All audio data are also labeled as either normal or pornographic examples which encode the erotic voice. Similar to the training images, we randomly selected 20,000 pornographic and 20,000 normal audio files for training the RANet.

4.2. Experimental Setup

4.2.1. Training Setup

For training DCnet, the input training images were resized to maintain their short side being 768 and long side less or equal to 1333, since the input resolution must be dividable by $2^7 = 128$. Next, we used ResNet-50 as our backbone network and initialized it with the weights pre-trained on ImageNet [30]. Meanwhile, the newly added layers were initialized as in [45]. Our network was trained with mini-batch stochastic gradient descent (SGD) for 50 K iterations with the initial learning rate set as 0.001 and a mini-batch size of 32. The learning rate was reduced by a factor of 10 at iteration 20 k and 40 k, respectively. Momentum and weight decay are set as 0.9 and 0.0001, respectively.

Analogous to the training setup on DCNet, the mini-batch size, momentum, and weight were respectively set as 256, 0.9, and 0.0001 for training RANet. The learning rate was initialized as 0.001 and decreased by 0.1 every 150,00 iterations. The whole training procedure takes 35,000 iterations. Moreover, the most important parameter in training RANet was the number of segments K. In particular, RANet was reduced to the plain ConvNets when K equals to 1. With the increase in K, further performance improvement is expected. Inspired by [28], we evaluated the performance with varying values of K ranging from 1 to 9 by using the same test approach. The results are shown in Figure 5 . We observe

that increasing K leads to better performance. The highest accuracy is reported at 85.3% when K grows up to 5, while further increasing K does not improve the performance. Thus, we set $K = 5$ in the following experiments.

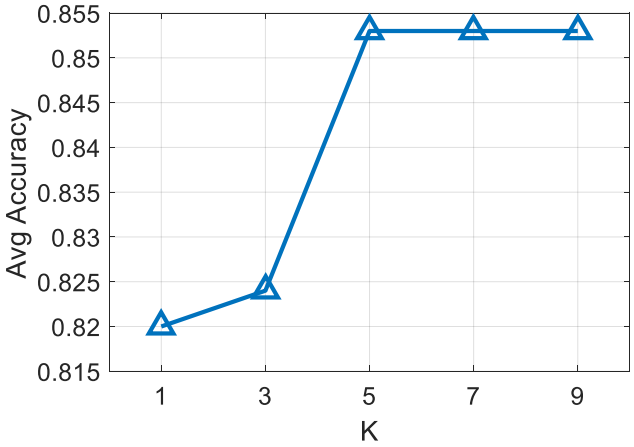

Figure 5. Performance of our RANet on validation set with varying values of K. The results indicate the average accuracy of recognizing porn and normal videos.

4.2.2. On-the-fly Inference

For the on-the-fly inference, given a test video, we firstly derive N image frames and a sequence of audio data from the video. Then, N log Mel-spectrograms are produced by VGGish for video representation. For a specific image frame, the detection result and classification result are obtained by the detection and the global classification network. For porn image recognition, we only use the classification results of the N images and employ the voting strategy to aggregate the scores. In addition, following TSN [28], K is set as 20 when feeding the audio data to the trained RANet model, leading to the audio recognition results. The final result is calculated by both image and audio classification scores.

4.3. Ablation Studies

To evaluate the performance of our proposed method, we conduct a set of ablation studies on the respective DCNet and RANet. In all the ablation experiments, we report the validation accuracy of the 1 k pornographic videos, 1 k videos, and 1 k normal videos respectively.

4.3.1. Ablation Studies on DCNet

We use Resnet-50 architecture as a single classification network without being combined with a detection network as our baseline. Apart from the baseline, we compare four different detection-classification frameworks: RCNet using RetinaNet as the detection net, FCNet which is bidirectional feature pyramid network, A-RCNet using anchor-free detection network, and DCNet using anchor-free bidirectional feature pyramid network.

It is shown in Table 1 that compared with the baseline, all four detection-classification architectures have boosted the validation performance to some extent. Specifically, the accuracy gain of 1.4% is achieved when a detection network works as an auxiliary branch. For porn video validation, BiFPN can boost the accuracy from 90.6% to 91.4%, while anchor-free detector improves the accuracy from 90.6% to 91.5%. Particularly, DCNet improves the ResNet-50 baseline from 89.2% to 92.5%. This is attributed to the obvious difference between breast_porn and breast_sexy that accelerates the function of detection network branch. On the contrary, the performance gain for the validation of normal videos is relatively limited. This implies that no sensitive information in normal videos weaken the function of the detection network branch.

Table 1. Comparison of different detection-classification networks. P, S, and N denote videos of a porn, sexy, and normal class, respectively.

Networks	Precision (%)			Recall (%)			Acc (%)
	P	S	N	P	S	N	
ResNet50	85.0	90.3	93.5	90.1	85.0	92.4	89.2
RCNet	87.5	92.3	94.2	91.6	86.3	94.0	90.6
FCNet	88.3	93.3	94.8	92.3	87.2	94.6	91.4
A-RCNet	88.5	93.5	94.7	92.4	87.3	94.8	91.5
DCNet	90.0	95.5	95.1	93.5	88.5	95.5	92.5

4.3.2. Ablation Studies on ResNet-Attention Network

To produce effective image-like feature embedding of audio data, we impose four features on our RANet: Mel-frequency cepstral coefficients (MFCC), gammatone frequency cepstral coefficients (GFCC), log-power short time fourier transform (STFT) spectrograms, and log Mel-spectrograms. As shown in Table 2, log Mel-Spectrograms can achieve a better accuracy 86.3% in recognizing pornographic audios, and thus it is used as the feature embedding of the audio data in the following experiments.

Table 2. The performance of our RANet with different features. Mel-frequency cepstral coefficients (MFCC), gammatone frequency cepstral coefficients (GFCC), and short time fourier transform (STFT).

Features	Precision (%)		Recall (%)		Acc (%)
	P	N	P	N	
MFCC	82.9	76.0	73.1	85.0	79.1
GFCC	86.2	78.1	75.6	87.9	81.8
log-power SIFT Spectrograms	82.3	82.1	82.1	82.4	82.3
Log Mel-Spectrograms	85.5	87.0	87.3	85.2	86.3

In addition, we use a VGG network excluding attention module as our baseline. We compare the performance of five networks: VGGNet-16, ResNet-18, A-ResNet18 (ResNet-18 with attention module), ResNet-50, and RANet (ResNet-50 with attention module). Since audio data are only categorized into porn and normal type, 1 k porn videos and 1 k normal videos for validation are used for evaluating the performance of these architectures. As illustrated in Table 3, deeper network architecture tends to achieve better results. Compared with the backbone, more specifically, Resnet-50 obtains a performance improvement from 81.6% to 83.8%. A-Resnet-18 embedded achieves the accuracy of 85.0%, outperforming Resnet-50 by 0.7%. Our RANet which embeds frequency attention module into the ResNet-50 achieves the best accuracy at 86.3%.

Table 3. Comparison of different audio-classification networks.

Networks	Precision (%)		Recall (%)		Acc (%)
	P	N	P	N	
VGG	80.6	82.6	83.1	80.0	81.6
ResNet-18	82.5	85.0	85.6	81.9	83.8
A-ResNet-18	83.8	86.1	86.6	83.3	85.0
ResNet-50	83.0	85.6	86.1	82.4	84.3
RANet	85.5	87.0	87.3	85.2	86.3

4.3.3. Combining DCNet and RANet

As discussed in the ablation studies above, we fuse the image and audios recognition results for pornographic video recognition. More specifically, we make use of Equation (9) to produce the final decision. In practice, we conduct two groups of experiments. First, A-ResNet-18 is used to produce the audio recognition results while it is combined with

five different detection-classification networks presented in Table 1. Second, we replace A-ResNet-18 with RANet for performing porn audio recognition. The results are illustrated in Tables 4 and 5 respectively. By comparing Tables 1 and 4, we can observe that A-ResNet-18 obviously increases the performance from 92.5% to 93.1% obtained by DCNet, along with the accuracy gains of 0.9%, 0.7%, 0.7%, and 0.9% achieved by ResNet50, RCNet, FCNet, and A-RCNet respectively. Particularly, significant performance improvement manifests itself into the precision and recall of porn video recognition. Furthermore, combining A-ResNet-18 and DCNet improves the precision from 85.0% to 93.2% and the recall from 90.1% to 95.0%. This sufficiently demonstrates the beneficial effect of RANet in further performance boost. Thus, the best accuracy at 93.4% is achieved by combining DCNet and RANet.

Table 4. Comparison of different frameworks in which A-ResNet-18 is combined with varying detection-classification networks.

Frameworks	Precision (%)			Recall (%)			Acc (%)
	P	S	N	P	S	N	
A-ResNet-18 + ResNet50	88.7	90.3	93.7	92.2	85.0	92.6	89.9
A-ResNet-18 + RCNet	90.1	92.3	94.3	93.6	86.3	94.1	91.3
A-ResNet-18 + FCNet	92.4	93.3	94.9	94.2	87.2	94.8	92.1
A-ResNet-18 + A-RCNet	92.3	93.5	94.8	94.8	87.3	95.0	92.4
A-ResNet-18 + DCNet	93.2	95.5	95.3	95.0	88.5	95.8	93.1

Table 5. Comparison of different frameworks in which RANet is combined with varying detection-classification networks.

Frameworks	Precision (%)			Recall (%)			Acc (%)
	P	S	N	P	S	N	
RANet + ResNet50	90.3	90.3	93.9	92.6	85.0	92.7	90.1
RANet + RCNet	93.3	92.3	94.4	94.0	86.3	94.3	91.5
RANet + FCNet	94.1	93.3	94.9	94.6	87.2	94.9	92.2
RANet + A-RCNet	94.3	93.5	95.1	95.1	87.3	95.3	92.6
RANet + DCNet	95.6	95.5	95.6	95.8	88.5	96.0	93.4

5. Conclusions

In this paper, we proposed a unified deep architecture termed PornNet integrating dual sub-networks for pornographic video recognition. Specifically, a local-context aware network is proposed for discriminating pornographic image frames, whilst an attention network which is also used as temporal segment networks is used to recognize pornographic audios. The results generated from the two sub-networks were aggregated for generating the whole video recognition result. Since no audio labels were available in the exiting porn video recognition datasets, we collected a large-scale dataset with both image and audio label annotated. Experiments on our newly-collected large dataset demonstrated the effectiveness of our proposed method, achieving an average accuracy with 93.4%, tested on 1 k pornographic videos, 1 k videos, and 1 k normal videos.

Author Contributions: Conceptualization, Z.F. and T.D.; methodology, Z.F.; software, Z.F.; validation, Z.F. and J.L.; formal analysis, Z.F.; writing—original draft preparation, Z.F.; writing—review and editing, J.L.; supervision, J.L.; project administration, G.C. and T.Y.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61703096).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and copyright issue.

Acknowledgments: The authors greatly appreciate all the reviewers for their positive and constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Levy, S. *Good Intentions, Bad Outcomes: Social Policy, Informality, and Economic Growth in Mexico*; Brookings Institution Press: Washington, DC, USA, 2010.
2. Zhao, Z.; Cai, A. Combining multiple SVM classifiers for adult image recognition. In Proceedings of the IEEE International Conference on Network Infrastructure and Digital Content, Beijing, China, 24–26 September 2010; pp. 149–153.
3. Moustafa, M. Applying deep learning to classify pornographic images and videos. *arXiv* **2015**, arXiv:1511.08899.
4. Tamburlini, G.; Ehrenstein, O.; Bertollini, R. *Children's Health and Environment: A Review of Evidence: A Joint Report from the European Environment Agency and the WHO Regional Office for Europe*; World Health Organization, Regional Office for Europe, EE Agency: Copenhagen, Denmark, 2002.
5. Bosson, A. Non-retrieval: Blocking pornographic images. In Proceedings of the International Conference on Image and Video Retrieval, London, UK, 18–19 July 2002; pp. 50–60.
6. Zheng, Q.-F.; Zeng, W.; Wang, W.-Q.; Gao, W. Shape-based adult image detection. *Inter. J. Image Graph.* **2006**, *6*, 115–124. [CrossRef]
7. Jang, S.-W.; Park, Y.-J.; Kim, G.-Y.; Choi, H.-I.; Hong, M.-C. An adult image identification system based on robust skin segmentation. *J. Imaging Sci. Technol.* **2011**, *55*, 20508–20601. [CrossRef]
8. Deselaers, T.; Pimenidis, L.; Ney, H. Bag-of-visual-words models for adult image classification and filtering. In Proceedings of the Pattern Recogn, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
9. Jiao, F.; Gao, W.; Duan, L.; Cui, G. Detecting adult image using multiple features. In Proceedings of the International Conferences on Info-Tech and Info-Net., Beijing, China, 29 October–1 November 2001; pp. 378–383.
10. Shih, J.-L.; Lee, C.-H.; Yang, C.-S. An adult image identification system employing image retrieval technique. *Pattern Recogn. Lett.* **2007**, *28*, 2367–2374. [CrossRef]
11. Yin, H.; Xu, X.; Ye, L. Big Skin Regions Detection for Adult Image Identification. In Proceedings of the 2011 Workshop on Digital Media and Digital Content Management, Hangzhou, China, 15–16 May 2011; pp. 242–247.
12. Zhu, Q.; Wu, C.-T.; Cheng, K.-T.; Wu, Y.-L. An adaptive skin model and its application to objectionable image filtering. In Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, 10–15 October 2004; pp. 56–63.
13. Smith, D.; Harvey, R.; Chan, Y.; Bangham, J. Classifying Web Pages by Content. In Proceedings of the IEE European Workshop Distributed Imaging, London, UK, 18 November 1999; pp. 1–7.
14. Chan, Y.; Harvey, R.; Bangham, J. Using Colour Features to Block Dubious Images. In Proceedings of the European Signal Processing Conference, Tampere, Finland, 4–8 September 2000; pp. 1–4.
15. Garcia, C.; Tziritas, G. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Trans. Multimed.* **1999**, *1*, 264–277. [CrossRef]
16. Fleck, M.M.; Forsyth, D.A.; Bregler, C. Finding naked people. In Proceedings of the European Conference on Computer Vision(ECCV), Cambridge, UK, 14–18 April 1996; pp. 593–602.
17. Lopes, A.P.; de Avila, S.E.; Peixoto, A.N.; Oliveira, R.S.; Araujo, A.A. A bag-of-features approach based on hue-sift descriptor for nude detection. In Proceedings of the European Signal Processing Conference, Scotland, UK, 24–28 August 2009; pp. 1552–1556.
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
19. Ou, X.; Ling, H.; Yu, H.; Li, P.; Zou, F.; Liu, S. Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-Coarse Strategy. *ACM T. Intel. Syst. Technol.* **2017**, *8*, 1–25. [CrossRef]
20. Wang, X.; Cheng, F.; Wang, S. Adult image classification by a local-context aware network. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 2989–2993.
21. Kong, Q.; Xu, Y.; Wang, W.; Plumbley, M.D. Audio Set classification with attention model: A probabilistic perspective. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 316–320.
22. Yu, C.; Barsim, K.S.; Kong, Q.; Yang, B. Multi-level attention model for weakly supervised audio classification. *arXiv* **2018**, arXiv:1803.02353.
23. Chou, S.-Y.; Jang, J.-S.R.; Yang, Y.-H. Learning to recognize transient sound events using attentional supervision. In Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 3336–3342.
24. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.

25. Wang, Y.; Li, J.; Metze, F. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 31–35.
26. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
27. Bozkurt, B.; Germanakis, I.; Stylianou, Y. A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection. *Comput. Biol. Med.* **2018**, *100*, 132–143. [CrossRef]
28. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal.* **2018**, *41*, 2740–2755. [CrossRef] [PubMed]
29. Wang, J.Z.; Li, J.; Wiederhold, G.; Firschein, O. System for Screening Objectionable Images. *Comput. Commun.* **1998**, *21*, 1355–1360. [CrossRef]
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, 3–6 December 2012; pp. 1097–1105.
31. Mallmann, J.; Santin, A.O.; Viegas, E.K. PPCensor: Architecture for real-time pornography detection in video streaming. *Future Gener. Comp. Syst.* **2020**, *112*, 945–955. [CrossRef]
32. Tokozume, Y.; Harada, T. Learning environmental sounds with end- to-end convolutional neural network. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2721–2725.
33. Zhu, B.; Xu, K.; Wang, D.; Zhang, L.; Li, B.; Peng, Y. Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features. In Proceedings of the Pacific Rim Conference on Multimedia, Hefei, China, 21–22 September 2018; pp. 528–537
34. Abdoli, S.; Cardinal, P.; Koerich, A.L. End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Syst. Appl.* **2019**, *136*, 252–263. [CrossRef]
35. Volkmann, J.; Stevens, S.S.; Newman, E.B. A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am.* **1937**, *8*, 208. [CrossRef]
36. Logan, B. Mel frequency cepstral coefficients for music modeling. In Proceedings of the International Symposium on Music Information Retrieval (ISMIR), Plymouth, MA, USA, 23–25 October 2000; Volume 270.
37. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W. CNN architectures for large-scale audio classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
38. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
39. Mydlarz, C.; Salamon, J.; Bello, J.P. The implementation of low-cost urban acoustic monitoring devices. *Appl. Acoust.* **2017**, *117*, 207–218. [CrossRef]
40. Guzhov, A.; Raue, F.; Hees, J.; Dengel, A. ESResNet: Environmental Sound Classification Based on Visual Domain Models. *arXiv* **2020**, arXiv:2004.07301.
41. Piczak, K.J. Esc: Dataset for environmental sound classification. In Proceedings of the ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1015–1018.
42. Salamon, J.; Jacoby, C.; Bello, J.P. A dataset and taxonomy for urban sound research. In Proceedings of the ACM International Conference on Multimedia, Mountain View, CA, USA, 18–19 June 2014; pp. 1041–1044.
43. Riaz, H.; Park, J.; Choi, H.; Kim, H.; Kim, J. Deep and Densely Connected Networks for Classification of Diabetic Retinopathy. *Diagnostics* **2020**, *100*, 24. [CrossRef] [PubMed]
44. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
45. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
46. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.