Journal of Visual Communication and Image Representation 50 (2018) 303–313

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

# Leveraging deep neural networks to fight child pornography in the age of social media☆

Paulo Vitorino[a,b], Sandra Avila[c,*], Mauricio Perez[d], Anderson Rocha[c,*]

[a] Department of Electrical Engineering, University of Brasilia, Brazil
[b] Brazilian Federal Police, Brazil
[c] Institute of Computing, University of Campinas, Brazil
[d] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

*Keywords:*
Child pornography
SEIC content
Deep learning
Transfer learning
Fine tuning

ABSTRACT

Over the past two decades, the nature of child pornography in terms of generation, distribution and possession of images drastically changed, evolving from basically covert and offline exchanges of content to a massive network of contacts and data sharing. Nowadays, the internet has become not only a transmission channel but, probably, a child pornography enabling factor by itself. As a consequence, most countries worldwide consider a crime to take, or permit to be taken, to store or to distribute images or videos depicting any child pornography grammar. But before action can even be taken, we must detect the very existence or presence of sexually exploitative imagery of children when gleaning over vast troves of data. With this backdrop, veering away from virtually all off-the-shelf solutions and existing methods in the literature, in this work, we leverage cutting-edge data-driven concepts and deep convolutional neural networks (CNNs) to harness enough characterization aspects from a wide range of images and point out the presence of child pornography content in an image. We explore different transfer-learning strategies for CNN modeling. CNNs are first trained with problems for which we can gather more training examples and upon which there are no serious concerns regarding collection and storage and then fine-tuned with data from the target problem of interest. The learned networks outperform different existing solutions and seem to represent an important step forward when dealing with child pornography content detection. The proposed solutions are encapsulated in a sandbox virtual machine ready for deployment by experts and practitioners. Experimental results with tens of thousands of real cases show the effectiveness of the proposed methods.

## 1. Introduction

It is no wonder that virtually all modern societies condemn the sexual abuse of children. In the past two decades, awareness and recognition of such problems have grown systematically and now permeate the forefront of discussions in several governments, media outlets and society circles [1]. Unfortunately, some aspects of sexual abuse of children still lag behind in terms of public policies and immediate actions of eradication. It was only recently that child pornography started to be cast as a significant element in the lineup of activities related to sexual abuse. According to Taylor and Quayle [1], although child pornography has been a recognized problem for decades, it was until recently deemed as a "rather small and essentially specialist correlate of a much broader and more significant problem". Nonetheless, some recent studies report some staggering projections pointing out that every forth girl and sixth boy in the U.S. alone will experience some form of sexual abuse before turning 18 [2]. Equally alarming, we must also be aware that it is very likely that a significant fraction of these cases will be video taped for further distribution and sharing in online platforms and social networks.

In recent years, new communication and computing advancements along with the rise of social networks have prompted unremarkable societal advances in our world. However, at the same time, these advancements have also brought dishonest elements of our society closer to us. In this vein, as Taylor and Quayle [1] properly put it, since the mid-1990s, we have seen a significant change in the nature of child pornography in terms of generation, distribution and possession of images. Until some years ago, child pornography was mostly done offline and, thus, had less impact and it was more easily traceable. However, in the last few years, it evolved to a much more difficult

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding authors.
*E-mail addresses:* sandra@ic.unicamp.br (S. Avila), anderson.rocha@ic.unicamp.br (A. Rocha).

https://doi.org/10.1016/j.jvcir.2017.12.005
Received 16 May 2017; Received in revised form 22 September 2017; Accepted 16 December 2017
Available online 19 December 2017

problem with the advent of social networks, transforming the internet not only in a transmission channel but, probably, in a child pornography enabling factor by itself [1,3].

The problem turns even more complicated when we analyze it under a behavioral optics [4], with which we seek to learn as much as possible about the perpetrators, victims and the dynamics of an offense. Under this vantage point, pedophiles[1] now not only share contents online but also organize themselves in their own social networks sharing interests and experiences, blurring the edge between the virtual and real worlds [6]. Moreover, child pornography offenses seem not to have any specific boundary of class, income or profession [1]. Although controversial, some of these studies also suggest that child pornography offending is an indicator of pedophilia [7,8]. Finally, with more children also having access to uncontrolled materials and uncensored "friendships" online, they are also more easily approachable and encouraged to engage in dubious relationships with offenders. The Tech Innovation to Fight Child Sexual Exploitation (THORNE) foundation reports that up to 42% of "sextortion" victims met perpetrators online [9].

According to the International Criminal Police Organization (Interpol) [10,11], child pornography is defined as "…the consequence of the exploitation or sexual abuse perpetrated against a child. It can be defined as any means of depicting or promoting sexual abuse of a child, including print and/or audio, centered on sex acts or the genital organs of children." As a consequence, most countries worldwide consider a crime to take, or permit to be taken, to store or to distribute images or videos depicting any child pornography grammar [1,3]. According to Taylor et al. [12], in spite of the actual scene depicted in an image or video, whenever "an image of a child is accessed for a sexual purpose, it victimizes the individual concerned". With this backdrop, it is paramount that we devise and deploy proper mechanisms (technical and legal) to combat child pornography online. In this vein, in this paper, we aim at the automatic detection of child pornography from images. For a proper nomenclature, whenever we refer to images depicting child pornography content, we adopt the term *sexually exploitative imagery of children* (SEIC) [4].

In recent years, some researchers took aim at this problem by proposing a diverse set of solutions in the literature. As we shall discuss in Section 2, solutions range from nudity detection [13] and facial analytics [14,15] as proxies for child pornography classification to bags of visual words [16,17] and behavioral analytics [4] to network profiling [4,18–21] and sensitive hashing techniques [14,22,23]. Departing from virtually all existing methods, in this work, we leverage data-driven techniques and deep convolutional neural networks (CNNs) to harness enough characterization aspects from a wide range of images and point out the presence of child pornography content in an image. We propose a two-tiered CNN modeling, first trained with the source-related problem of general pornography detection – for which we can gather more training examples and upon which there are no serious concerns regarding collection and storage – and then fine-tuned, in a second refinement stage, with child pornography concepts properly controlled by a government agent in a secured setup. The learned network outperforms different existing solutions and seems to represent a major leap forward when it comes to dealing with this complex problem.

In a nutshell, our contributions in this paper are threefold:

- We introduce data-driven solutions able to distinguish sexually exploitative imagery of children from adult (related to normal porn) and seemingly innocuous (related to everyday imagery) content. The methods can pinpoint images depicting assault, gross assault and sadistic/bestiality involving children, which are considered to be of high importance in an international scale for combating child pornography (see Section 2 for more details on the scale and Section 3 on the method).
- We also introduce an adult content detector able to detect adult content not necessarily involving children, including the ones depicting secretive photographs showing underwear/nakedness with sexual intent, intentional posing suggesting sexual content and erotic posing (intentional sexual or provocative poses), which are ranked as of medium importance in the same international scale for combating child pornography (see Section 2 for more details on the scale and Section 3 on the method).
- Finally, we produce a self-contained virtual machine with our solutions, free of cost, and ready for initial deployments by different law-enforcement agents and practitioners for combating SEIC content nowadays.

We organize the remaining of this paper into four sections. Section 2 presents works related to child pornography. Section 3 introduces our solution to detecting SEIC contents. Section 4 describes the used datasets and experimental setup while Section 5 presents the experiments and results comparing the proposed method with different counterparts in the literature and with some off-the-shelf solutions. Finally, Section 6 concludes the paper and sheds some light on future research directions.

## 2. Related work

The legal definition of child pornography does not capture the entire nuance of the problem [1,12]. In a study of online content at the Combating Paedophile Information Networks in Europe Center (COPINE), Taylor and Quayle [1,12] identified 10 categories of pictures that may be sexualized by an adult and created the so-called COPINE scale as Table 1 shows.

Establishing the COPINE scale is pivotal to define to which extent someone caught red-handed is involved with sexually exploitative imagery of children. It also sheds some light on the research directions we should take in order to deal with this challenging problem [2]. Clearly, focus should be given to detecting activities involving levels L8–L10. However, many techniques in the literature, as we describe next, focus on simple models that are only able to detect nudity, or levels L2 and L3 above. Moreover, studies such as the COPINE one can also help driving some appropriate legislation on the theme [3] and defining a typology of online child pornography offending [24], especially because the law, itself, might not be well defined [25].

As the public awareness about child pornography has increased over the past years, so has the interest of researchers in presenting effective methods to block or, at least, cope with the problem. On one hand, there have been efforts toward working on the server side and on the network itself by developing appropriate filtering and blocking tools [4,20,21]. On the other, researchers have been developing detection solutions exploiting the intrinsic distributed nature of the internet, thus focusing on the users. While the former group of methods might be effective by actively detecting suspicious content and users trying to have access to it, it raises serious questions regarding censorship and high false positive rates, oftentimes blocking otherwise legitimate content [26,27]. The second group, in turn, works as a passive solution in which law-enforcement agents can sift through large amounts of data looking for inappropriate content in apprehended materials or as an active filtering solution in which families can protect their children and beloved ones from accessing inappropriate content by installing such solutions in their computing devices.

Spearheading the first group, we can see the works of [18,19], which focus on developing network profiling techniques to pinpoint abnormal behavior linked to child pornography. In the same vein, some researchers focused on specific active analyses of peer-to-peer networks [4,20,21], by and large, one of the most important channels for exchange of SEIC content online.

Veering away from the network profiling models, we have the

[1] According to Dorland [5], pedophilia refers to an abnormal fondness for children; sexual activity of adults with children.

**Table 1**
Taylor and Quayle's COPINE scale for sexually exploitative imagery of children [1,12].

| | |
|---|---|
| L1 | Indicative (non-erotic pictures) |
| L2 | Nudist (naked or semi-naked in legitimate settings) |
| L3 | Erotica (secretive photographs showing underwear/nakedness) |
| L4 | Posing (intentional posing suggesting sexual content) |
| L5 | Erotic Posing (intentional sexual or provocative poses) |
| L6 | Explicit Erotic Posing (emphasis on genital areas) |
| L7 | Explicit Sexual Activity (explicit activity with no adult involved) |
| L8 | Assault (sexual assault involving adult) |
| L9 | Gross Assault (penetrative assault involving adult) |
| L10 | Sadistic/Bestiality (imagery involving pain or animal) |

methods more focused on the user. In this regard, some authors have been fighting SEIC content using facial analytics [13–15]. Ricanek et al. [14] proposed Artemis, a forensic tool underpinned by facial analytics and other technologies to quickly scan computers and memory devices for SEIC content. Artemis relies on image hashes to compare new content to previously identified child pornography content as well as active mechanisms such as object recognition, identification of exposed skin, and facial analysis to determine whether the subjects depicted in an image are children. Sae-Bae et al. [13] also exploited skin exposition along with facial features to detect imagery depicting children. However, the authors fell short in their objective of training a solution for actually detecting SEIC content. Their solution is more suitable for the lower levels in the COPINE scale than the actual harmful ones (levels L7 and beyond). Recently, Chatzis et al. [15] presented a new geometrical feature based on iris geometry as a proxy for detecting children in images and, consequently, SEIC contents. The major limitation of the proposed feature, however, is that the weight and growth of the human eyeball is mostly pronounced in the first three years of age not varying significantly afterwards. Therefore the method is highly dependent on the quality of the acquired images and on the ability to properly detect and isolate the eyeballs in an image [28].

As a matter of fact, the idea of using color and texture image descriptors to detect SEIC content is dominant in this area although its results are not always very promising [13,29–32]. Grega et al. [33] have exploited MPEG descriptors to find traces of child pornography imagery and presented the INDECT advanced image cataloguing tool. According to the authors, their method consists of MD5 hashes and MPEG-7 descriptors. While the hashes account for a faster way of comparing new suspect images to previous tagged harmful ones, the descriptors seek to identify new images containing SEIC content. Microsoft has also introduced a hash-based tool, referred to as PhotoDNA, to fight child pornography content [23,34]. The main problem with hash-based techniques, however, is that they are passive methods and can only be effective at comparing similar (previously annotated) SEIC content, not being able to spot out new (unseen) content. For this very reason, many authors have invested their time in combining hash-based solutions with pro-active solutions such as the ones exploiting image visual characteristics (e.g., color and texture). Taking a different direction, Panchenko et al. [35] proposed a classification system based on patterns present in SEIC's filenames. Clearly, this method has serious limitations if never-seen filename patterns are used by the perpetrators.

Focusing on skin color distributions, Polastro et al. [29,31,32] proposed NuDetective, a tool for detecting SEIC content in seized computer systems. The authors later extended the method to also work with video contents [36]. Although in current use by the Brazilian Federal Police, NuDetective suffers from the same problems that afflict skin-based detection methods: high rates of false alarms, as innocuous and some adult content are often incorrectly tagged as SEIC. To make things worse, imagery depicting levels L8 through L10 in the COPINE scale frequently goes unnoticed.

In addition to efforts such as PhotoDNA, NuDetective, Artemis, and INDECT, industry players have also attempted to produce solutions for nudity and pornography detection. There exist some software solutions, mostly commercial with focus on blocking websites that contain inadequate content (e.g., CyberPatrol, CYBERsitter, NetNanny, K9 Web Protection, Profil Parental Filter) while others scan the hard drive in search for adult content (e.g., SurfRecon, Porn Detection Stick, PornSeer Pro). In the latter group, we can also include MediaDetective [37] and Snitch Plus [38], to which we compare our proposed method in this work. Notwithstanding, these solutions heavily rely on skin detection methods, and consequently suffer from the same problems mentioned above.

Departing from the low-level skin-based modeling, Ulges et al. [16] have relied on a more discriminative mid-level representation based on color image parts rather than on the color directly, called visual words. Their solution leads to a prioritization of SEIC content with an equal classification error varying from 11% to 24%. However, this result must be taken with a grain of salt as the actual discrimination is done in a much easier problem between seemingly safe content (natural images) and adult/SEIC content rather than on SEIC vs. adult contents.

Seeking to incorporate the best of several existing description methods, Schulze et al. [17] aggregated skin color, visual words, sentibank and audio words to improve the detection rate of SEIC content. The main novelty in Schulze et al.'s work, besides the aggregation policy, was the use of sentibank, a large-scale visual sentiment ontology to capture sentiment nuances in the analyzed images. For videos, the performance shows an equal classification error rate close to 8% when comparing SEIC content and normal adult content. However, that success is only attainable when it is possible to extract all features from the videos (audio, sentiment, color detection). For images, a more complex problem as it relies only on a single frame for evidence each time, the error rate reported was higher, approximately 16%.

After surveying many different efforts to solve the problem of automatically detecting sexually exploitative imagery of children, it is clear that to detect any new activity, we must rely on content-based methods rather than on hashing- or filename pattern-based ones. However, it is also clear that given the high number of false alarms and the apparent lack of a real link between seemingly normal and SEIC contents, we must depart definitely from skin-based modeling methods and exploit richer and more discriminative patterns. Although richer mid-level representations have had relative success in the literature for detecting adult [39–43] and SEIC content [16,17,44], they are still not the final answer for the problem as finding enough and suitable training sets for SEIC imagery has been a challenge. Some authors have exploited these richer mid-level representations jointly with visual features (such as color moments and edge histograms) and audio vocabularies (based on an energy envelope unit) [45]. Veering away from those methods, in this work, we exploit a completely different path as we shall explain in Section 3.

In order to deal with the scarcity of properly labeled SEIC content and also with its intrinsic restrictions of access, even for research purposes, we present a two-tiered solution with two different levels of training, one with adult content (for which we can easily collect and annotate data) allied with a refinement training step involving SEIC content. Our solution is totally data-driven in the sense that we do not need to specify to the algorithm if we are interested in patterns of color or texture or even mid-level representations. On the contrary, the discriminative patterns for properly separating normal and adult content from SEIC content naturally emerge from the existing training data after our method gleans over several thousands of examples using convolutional neural networks underpinned by recent advances in deep learning and learning-from-data technologies. It is important to mention that adult content detection through deep learning has already been studied [46–49]. Nonetheless, in those works, the authors only focused on adult (non-SEIC) content, therefore not being able to deal with SEIC grammar at the COPINE scale levels of L8 through L10. More specifically, none of them considered real-world data of sexually exploitative imagery of children. In our case, we consider images and videos of real-world apprehensions of the Brazilian Federal Police, the

Pornography-2 k dataset [50], and data augmentation using images from COCO dataset. Finally, all of those methods are one-tiered methods, which means they only deal with the data directly in one level. In our case, we develop a one and two-tiered data-driven solutions.

## 3. A deep learning approach to detecting child pornography in images

Anyone can imagine how difficult it is to capture all the nuances of child pornography imagery through color, shape, textures, and other image hand-crafted descriptors alike while, at the same time, ruling out innocuous and/or non-SEIC adult content. Therefore, in this section, we present solutions to detect SEIC content, which leverage concepts from deep convolutional neural networks to learn discriminative patterns directly from the available training data. However, at the same time that deep learning solutions allow us to learn such patterns from available data rather than hand-crafting descriptors to capture those patterns, it comes with a high price tag, the need for appropriate training data capturing the different perspectives and vantage points of the problem itself.

Capturing such variations for a problem such as object classification and detection in images is reasonably easy as large amounts of data can be downloaded from the internet and labeled through services such as Mechanical Turk [51]. This is the key reason deep learning has caused big waves in the computer vision literature in the past few years with ever-better image classification models [52–54]. Unfortunately, detecting sexually exploitative imagery of children is a more sensitive problem for which access to data is, to say the least, very well controlled. Furthermore, obtaining properly labeled data also seems a daunting task given that it would involve sifting through degrading and loathsome content.

In this vein, in this paper, we come up with solutions that allow us to take full advantage of deep learning concepts while, at the same time, not having enough data to train a data-driven solution for SEIC content detection from scratch. Instead of training a data-driven solution from scratch, which we are not able to given the scarcity of training for the target problem of SEIC detection, we implement transfer-learning methods for this task. In this way, we initialize the network weights with side data available for training (from a source problem) and then fine-tune such weights to the target problem with annotated samples capturing the nuances of such problem.

We present a hierarchical solution in which we first learn network weights relative to a general image-related task (object classification), which we call source problem, and transfer the acquired knowledge (network parameters and weights) to a target problem. The domain-transfer procedure requires less data (sometimes one order of magnitude less) than the first training given that the network weights have already been initialized for an image-related problem.

In addition to this 1-tiered solution (source → SEIC target problem), we also experiment with a 2-tiered solution in which we first transfer from the source problem to a second problem (intermediary target) of adult content detection. Upon transferring the acquired knowledge from the source to the intermediary target problem, we perform network fine-tuning, properly finding network weights best adapted to detect adult content. After the network refinement, we perform a second transfer learning, now considering the knowledge acquired by the network specialized in adult content detection to the problem of SEIC content detection (ultimate target) and perform fine tuning for that problem. This second refinement also allows us to train a solution with less SEIC content than it would be necessary to train a solution from scratch. Under this vantage point, we can view this second source/target adaptation evolving from a network specialized in object classification, to adult image classification (adult/porn vs. non-porn content) to child pornography classification (SEIC vs. adult & non-porn content).

The first network (source) requires an order of millions of images for training. The target networks (1- and 2-tiered) require about 40 thousand images comprising SEIC/non-SEIC content for fine-tuning. When using the intermediary network for adult-content detection, it can be trained with some 200 thousand examples, also an order of magnitude less than the source network. This is remarkable feat when we consider how data hungry deep learning solutions can be. If we set forth the objective of training a SEIC detection network from scratch (with initial weights randomly selected) it is likely we would, at least, one order of magnitude more images containing child pornography examples to capture important nuances of the problem.

At this point is important to highlight that we do not intend with the proposed solutions to replace the ones based on signature matching such as NetClean or PhotoDNA, the so-called hash-based solutions. Instead we aim at proposing methods that can be complementary to the similarity- or hash-based ones. Hash-based techniques are passive methods and are effective at comparing similar (previously annotated) SEIC content, not being able to spot out new (unseen) content. Content-based methods, on the other hand, are specially designed for active detection. This means that the hashed-based group excels when it can spot out similar content (previously annotated by any means) while the content-based group shines when spotting unseen content.

As in practice we deal with contents of all sorts, sometimes totally new, sometimes coming from social networks with some indications (denouncements) of SEIC content or even from peer-to-peer networks, having complementary solutions if very important. In this regard, we believe the content-based methods we propose in this paper can be complementary to hash-based solutions currently available and this complementarity will certainly add value to daily investigations by forensic analysts.

### 3.1. Network architecture for the source and target problems

Following our experience from other problems in computer vision-related tasks, we start by choosing a network architecture with well-known performance for the source task (general object classification) referred to in the literature as GoogLeNet [54]. Fig. 1 depicts an overview of the chosen architecture for the source problem of object classification, which is then refined and adapted to the target problem of interest. The source problem consists of classifying images within 1000 everyday categories from dogs and cats to household items to food and planes.

The network consists of an initialization module, sometimes referred to as stem, nine processing modules called *inceptions*, two auxiliary classifiers and one final classification module. Fig. 2 shows a breakdown of each of these modules.

Everything starts in the initialization module responsible for

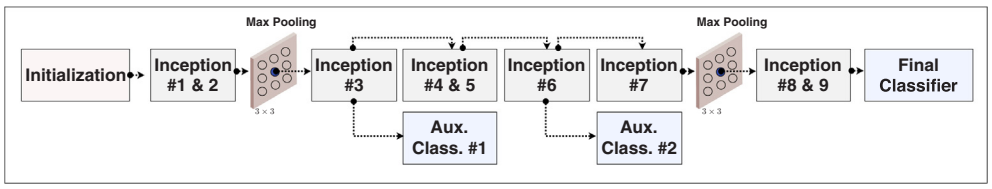


**Fig. 1.** Overview of the chosen architecture (source and target problems).

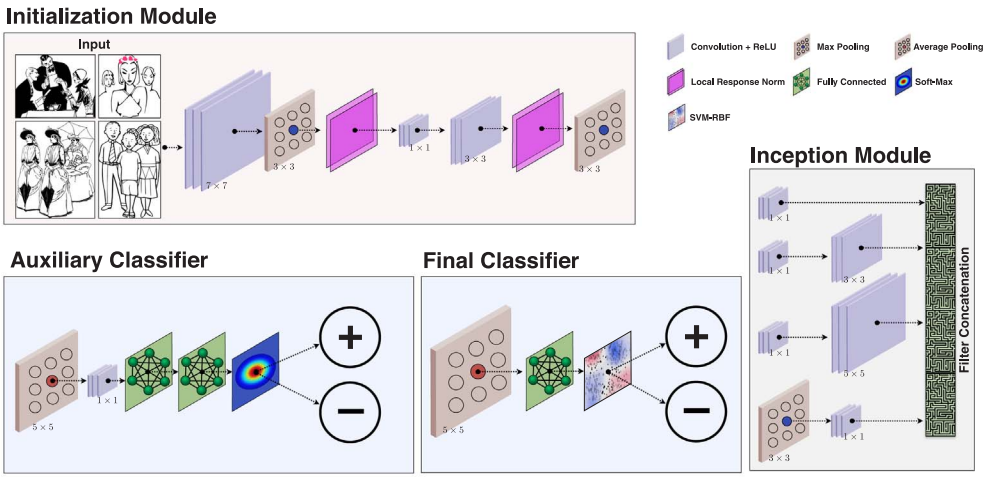


**Fig. 2.** Details of the different modules composing the chosen architecture (source and target problems).

receiving input images and transforming them with a series of convolution, pooling, and local response normalization operations, similarly to many other convolutional neural network models in the literature [52]. We start observing large object structures with $7 \times 7$ convolution kernels and then progressively apply additional convolutions, normalizations and data reductions (poolings). After this initialization, the processed data go through a series of nine feature-learning modules, referred to as *inception* modules, a key idea of this architecture with respect to other existing models in the literature. Each of these modules can be seen as a small network refining a set of features and properties of the problem. The name is derived from an allusion to the work of Lin et al. [55], network inside a network, and to Christopher Nolan's *Inception* movie.

GoogLeNet's atom, the inception module, is nothing more than a series of convolutions and data reductions/poolings at different scales, performed in parallel, ultimately concatenated together to feed the next stage of the network. In an inception, $1 \times 1$ convolutions are used to reduce the dimensionality of inputs to convolutions with larger, and more complex, filter sizes.

In addition to the network's nine stacked inception modules, we also have two auxiliary classifiers connected to the network in order to amplify the gradient signal back through the network, seeking to highlight the earlier representations of the data. Propagating the signal from the very last layer to the earlier layers in such a deep network may be very tricky and unstable. One common problem that appears is the vanishing gradient problem, in which the neurons in the earlier layers have a lower learning rate than neurons in later layers [56]. The auxiliary classifiers can then be used in some break points of the network to highlight the gradients. According to the authors [54], including the auxiliary classifiers encourages the discrimination in the lower stages of the network, increasing the gradient signal being back-propagated, also adding some regularization to the learning process. The two auxiliary classifiers are placed right after the third and sixth inception modules and are basically a small convolutional network with a softmax decision layer in the end. Each classifier consists of a data reduction layer through average pooling with a $5 \times 5$ kernel, two fully-connected layers with dropout [57] (70% dropout in the case of this architecture) and a linear layer with softmax loss as the 1000-way classifier. The auxiliary classifiers are used only during training and then removed during inference/testing time.

The final module is the classification. In the source problem, it consists of a data reduction layer through average pooling with a $7 \times 7$ kernel, a fully-connected layer and a linear layer with softmax loss as the 1000-way classifier. However, for the two target problems herein (adult content detection and SEIC detection afterwards), this final classifier is replaced with a two-way non-linear Support Vector Machine (SVM) with a radial basis function (RBF) kernel. The reason is that this classifier leads to a better discrimination when performing transfer learning considering the two classes of interest of each target problem. The final feature vector, collected right after the last fully-connected layer and before feeding the softmax classifier, has 1024 features.

### 3.2. Training for the source problem

The training of the source network is performed with about 1.2 million images of 1000 different object classes. This first training was performed by Szegedy et al. [54] and has used an asynchronous stochastic gradient descent with 0.9 momentum [58] and a fixed learning rate schedule, which decreases the learning rate by 4% every eight epochs. During the training of the network on the source problem, the losses measured on the two auxiliary classifiers for each image are added to the loss of the final classifier with a discount weight of 0.3. Differently from [54], to collect more training images, the initial pool is further augmented, on-the-fly, using some image transformations including rotations, mirroring and cropping but not scaling and photometric distortions [59]. The version we consider in this work uses a polynomial learning rate decay policy, instead of the step policy, because it leads to a $4\times$ faster training.

Given a set of inputs for training, all images are resized, maintaining the aspect ratio and having their smallest dimension as the network input dimension ($224 \times 224$ pixels). Then we perform center cropping, resulting in an image with the necessary shape for the convolutional network architecture chosen. This pre-processing of inputs is exactly the same for images in the training stage for the source problem as well as for the subsequent two stages of transfer learning and fine tuning to the two target problems of interest herein (see Section 3.3).

### 3.3. Transfer learning & fine tuning to the target problem

Once the network is fully trained and able to classify everyday object categories within the 1000 classes of interest, we first experiment with transferring the learned weights to a target network aimed at learning adult content detection (porn vs. non-porn content). We start by initializing the adult content detection network using the same architecture and weights previously learned for the object classification network, instead of training one from scratch. Then we adapt its last layer to a 2-way softmax one (adult content vs. non-adult content) and perform fine-tuning of the initial weights, receiving as input a series of images tagged as either adult and non-adult content and using back-propagation. Moreover, we also adapt the last layer of the two auxiliary classifiers to a 2-way softmax one.

Once this network converges with the new weights adapted to the intermediary target problem of adult content detection, we replace its very last layer with an SVM classifier with an RBF kernel. This means

that the learned network has weights specialized to adult content detection and will work now as a feature extractor for the problem of adult content classification and the SVM classifier will learn a discriminative model on top of those features. However, as the ultimate goal is to detect SEIC content, we can test how this network for adult content detection performs on the SEIC problem, by feeding it with SEIC and non-SEIC content examples, extracting the features and training an SVM classifier for SEIC detection.

Given an image set of SEIC and non-SEIC content, we extract the image features using this intermediary learned network and train a non-linear SVM with these examples. For training the SVM classifier, we perform 5-fold cross-validation within the available training set and perform grid-searching for finding the best classification parameters $C \in \{2^c: c \in [-5,-3,-1...,15]\}$ and $\gamma \in \{2^i: i \in [15,13,...,3]\}$. The obtained parameters for this first SVM were $\gamma = 0.0078125$ and $C = 8.0$. We refer to this first solution as a `1-tiered Adult Detector` as it uses one level of transfer learning, it can detect adult content and, to some extent, SEIC content but without having features totally specialized to this latter task yet.

After fine-tuning the initial network (object detection) to the problem of adult content classification (`1-tiered Adult Detector`), we turn our attention to fine-tuning the resulting network to our ultimate objective, SEIC content detection. To that purpose, we experiment with two options, a 1-tiered and a 2-tiered solution. For the 1-tiered SEIC solution, we start with the network trained for object classification (c.f., Section 3.2) and perform fine-tuning of its learned weights receiving as input a series of images tagged as either SEIC and non-SEIC content and using back-propagation. Just like before, once this network converges with the new weights adapted to the SEIC detection (target) problem, we replace its very last layer with an SVM classifier using an RBF kernel. We refer to this second solution as a `1-tiered SEIC Detector` as it uses one level of transfer learning from the source problem of object classification to the target problem of SEIC detection through fine-tuning with SEIC images (Object Classification → SEIC Content Classification).

For the 2-tiered solution, we start with network fine-tuned to the adult content detection problem and perform fine-tuning of its weights receiving as input a series of images tagged as either SEIC and non-SEIC content and using back-propagation once again. Just like before, once this network converges with the new weights adapted to the ultimate target problem, we replace its very last layer with an SVM classifier using an RBF kernel. We refer to this third solution as a `2-tiered SEIC Detector` as it uses two levels of transfer learning (Object Classification → Adult Content Classification → SEIC Content Classification).

Given an image set of SEIC and non-SEIC content, we extract the image features using the final 1- or 2-tiered learned networks and train non-linear SVMs with these examples. For the training, we also perform fivefold cross-validation within the available training set and perform grid-searching for finding the best classification parameters $C \in \{2^c: c \in [-5,-3,-1...,15]\}$ and $\gamma \in \{2^i: i \in [15,13,...,3]\}$. The obtained parameters for this first SVM were $\gamma = 0.0078125$ and $C = 0.5$.

Moreover, we select a dropout rate of 40% just like when the network was first trained for the source problem of general object recognition, a learning rate of 0.000009, a weight decay of 0.005, a polynomial power of 0.5, and a max number of 200 epochs. We also use the polynomial learning rate decay policy during fine-tuning as it is much faster than the original step optimization policy used by Szegedy et al. [54].

### 3.4. Data augmentation

Even when fine-tuning the network rather than training it from scratch, it might be heavily data hungry. Therefore, we also implement one extension of our 2-tiered solution to take into account data augmentation during the weight refinement to the target problems. For this augmentation, the initial training pool is further augmented for the non-SEIC content only – specifically, non-adult images –, since the dataset used for fine-tuning the 2-tiered network contains few actual negative examples. Thus, instead of generating additional images from the original ones (by rotating, mirroring, adjusting contrast, etc.), we include in the training set 20,000 images from Microsoft Common Objects in Context (COCO) dataset [60], which consists of 160,000 images labeled in 91 common object categories. For the 1-tiered solutions, we do not perform data augmentation. For the 2-tiered version, we experiment with one version comprising data augmentation and another one without it.

## 4. Experimental setup

In this section, we describe the adopted experimental setup in terms of datasets, experimental protocols, classification metrics, details on the available forensic and commercial tools, BoVW-based parametrization, and a deep-learning-based approach used in the experiments. The experimental setup designed for the evaluation of the proposed SEIC detectors – network parametrization and training details – was presented in Section 3.

### 4.1. Datasets

We validate the proposed deep learning-based SEIC detection methods with images of real-world apprehensions of the Brazilian Federal Police, i.e., the images come from a hard disk drive of a real forensic case involving child pornography. In total, the dataset comprises 58,974 images, with 33,723 depicting SEIC content. It is important to highlight that the non-SEIC class is composed of general content images, including nudity and pornographic content, which makes the task of detecting SEIC content very challenging. We split the dataset into training (33,646 images), validation (5938 images) and test (19,387 images) sets, with a proportion of 60%/40% images for SEIC and non-SEIC content, respectively. For pre-training the CNN model, all hyperparameters are optimized on the validation set to prevent overfitting. For non-deep-learning-based approaches, for instance, for the skin-detector- and BoVW-based techniques, the methods are trained on the training + validation.

Due to the illegal nature of child pornography, the data cannot be illustrated in this paper. Also, we emphasize that we – as research scientists – never possessed any of such material. Each iteration/version of our method was encapsulated in a sandbox virtual machine, sent to the Brazilian Federal Police, processed the data therein and returned to us in the form of feature vectors. The fine-tuning of our networks was performed using the prepared sandbox virtual machine in the premises of the Brazilian Federal Police as well.

For the task of adult content detection, as the source dataset to pre-train the CNN, we use the Pornography-2 k dataset [42], which comprises nearly 140 h of 1000 pornographic and 1000 non-pornographic videos, varying from six seconds to 33 min. For video annotation, the authors [42] adopted the definition of pornography proposed by Short et al. [61]: “any explicit sexual matter with the purpose of eliciting arousal”.

The official evaluation protocol for the Pornography-2 k dataset is a 5 × 2-fold cross-validation. Here we apply a similar protocol, which consists of randomly splitting the dataset – one time, because the main objective is to detect SEIC content, not pornography – into two same-size class-balanced folds. Also, to pre-train the CNN, we perform a 15% stratified split of the training set for validation. As we are concerned about detecting SEIC content in images, we adopt a sample rating of one frame per second, providing 214,171 frames from the training fold for training and validating our networks and 257,522 for testing them. Fig. 3 depicts some example frames from the Pornography-2 k dataset.

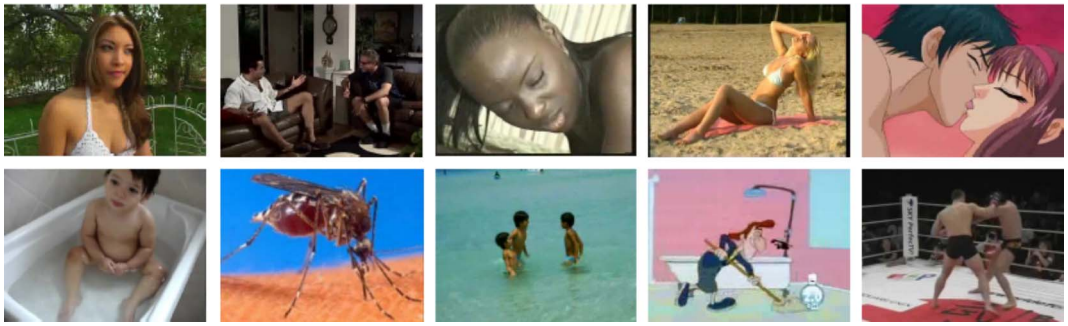

**Fig. 3.** Example frames from the pornography-2 k dataset, illustrating the diversity of pornographic (top row) and the non-pornographic content (bottom row), with high skin exposure.

### 4.2. Classification metrics

Following the default classification metrics of the Pornography-2 k dataset, we report the normalized accuracy (ACC), and the $F_2$ measure ($F_2$).

ACC corresponds to the percentage of correctly classified images. $F_2$, in turn, refers to the weighted harmonic mean of precision and recall, which gives twice the weight to recall ($\beta = 2$) than to precision. Note that high recall means a low number of false negatives, and high precision means a low number of false positives. $F^{\beta}$ measure is defined as:

$$F_{\beta} = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall} \quad (1)$$

where $\beta$ is a parameter denoting the importance of recall compared to precision.

### 4.3. Forensic tools

Despite finding several forensic tools for detecting child pornography content, only a very few of them are available on the Internet. Therefore, to evaluate the classification performance of the proposed solution, we also selected forensic tools that rely on visual content: NuDetective [31], MediaDetective [37] and SnitchPlus [38].

NuDetective is a tool, readily available for research purposes, developed by the criminal forensic experts of the Brazilian Federal Police for detecting SEIC content. MediaDetective and Snitch Plus, in turn, are both commercial solutions[2] with focus on detecting nudity and pornography content. All of these tools are underpinned by skin-based detectors to identify unsuitable material.

Furthermore, for MediaDetective and Snitch Plus, the image files are rated according to their suspicious (i.e., probability) for SEIC content. In those cases, we tag an image as SEIC if its returned probability is equal or greater than 50%. NuDetective, on the other hand, assigns binary labels to the image: positive (i.e., the image is SEIC) or negative (i.e., the image is non-SEIC).

Finally, MediaDetective and Snitch Plus have four predefined execution modes, which differ mostly on the rigorousness of the skin detector. In our experiments, we opted for the most rigorous execution mode. Regarding NuDetective, we employed its default settings.

### 4.4. Skin detector

Although a myriad of methods have been proposed to detect child pornography content, the simplicity of the human skin detection techniques has attracted many researchers [13,17,29,31,32], the most popular being the color-based techniques [62,63].

Kovac et al. [64] proposed a skin classifier by defining explicitly – through a number of rules – the skin region in the RGB color space. The skin model can be divided into three rules as follows:

Rule 1: $[(R > 95) \wedge (G > 40) \wedge (B > 20)] \wedge$
Rule 2: $[max(R,G,B) - min(R,G,B) > 15] \wedge$
Rule 3: $[|R - G| > 15 \wedge (R > G) \wedge (R > B)]$,

where R, G and B represent the pixel value in the RGB color space with values ranging from 0 to 255.

For comparison purposes, in this work, we chose Kovac et al.'s skin detector because it is widely used and easily implemented. In addition, we classified an image as SEIC if the percentage of skin pixels in the image is equal or greater than 13%. We performed cross-validation in the training set to select the best threshold value.

### 4.5. Bag of visual words-based approach

Before the introduction of Deep Learning techniques, the Bag of Visual Words (BoVW) modeling was the most successful approach to describe the content of images. In a nutshell, it characterizes an image as a histogram of the occurrence rate of visual words in a visual dictionary induced by quantizing the space of a local descriptor (e.g., Speeded-Up Robust Features (SURF) [65]). The dictionary of $k$ visual words is usually obtained by unsupervised learning over a sample of local descriptors.

In this paper, we evaluate the classical BoVW [66]: hard coding and average pooling. We first preprocess the dataset by resizing the images to an area of up to 100 thousand pixels if larger. We then extract SURF descriptors [65] on a dense spatial grid at five scales. More specifically, we use patch sizes of $24 \times 24$, $32 \times 32$, $48 \times 48$, $68 \times 68$ and $96 \times 96$ pixels, with step sizes of 4, 6, 8, 11 and 16 pixels, respectively. The dimensionality of the SURF is reduced down to 32 using principal component analysis (PCA). The application of BoVW to SEIC detection was proposed by [44] and this is the reference work we consider herein.

We learn the visual dictionary applying the $k$-means clustering algorithm with Euclidean distance over one million randomly sampled PCA-descriptors. We extract 2048 visual words, as suggested by [44].

Classification is performed by Support Vector Machine (SVM) classifiers, using the LIBSVM library [67]. By default, we use a non-linear SVM with RBF kernel. SVM parameters are estimated performing a grid search in the training set, considering $C \in \{2^c: c \in [-5,-3,-1...,15]\}$ and $\gamma \in \{2^i: i \in [15,13,...,3]\}$. The best obtained parameter were: $\gamma = 0.0078125$ and $C = 32$.

### 4.6. Yahoo! Porn detection approach

Very recently, Yahoo! open-sourced a CNN model for classifying adult content in images [49]. The Yahoo!'s network is based upon the ResNet model proposed by He et al. [68], a residual learning framework with a depth of up to 152 layers. According to Mahadeokar and Pesavento [49], their model replicates the ResNet paper's 50-layer network, using half of the number of filters in each layer.

For training residual networks, pre-trained on the ImageNet dataset, the authors [49] applied scale augmentation as proposed in [68]. They replaced the last layer by a 2-node fully-connected layer and performed

[2] We have purchased MediaDetective v3.1 and Snitch Plus v3.1 for our research.

fine-tuning with an image dataset of adult vs. non-adult content. According to Mahadeokar and Pesavento [49], the training images or other details are not available due to the nature of the data. Also, to optimize the classification performance, the hyperparameters (step size, base learning rate) were properly fine-tuned [49].

Moreover, they used the CaffeOnSpark deep learning framework, which was developed by Yahoo! for large-scale distributed deep learning on Hadoop clusters. Precisely, they benefited from 16 GPUs.

For the sake of comparison, in this paper, we use the Yahoo!'s model as a feature extractor and employ an SVM classifier with RBF kernel to make the decision for the SEIC detection task. For training the SVM classifier, we perform 5-fold cross-validation and apply grid-searching for finding the best classification parameters $C \in \{2^c: c \in [-5,-3,-1...,15]\}$ and $\gamma \in \{2^i: i \in [15,13,...,3]\}$. The best obtained parameters were $\gamma = 0.03125$ and $C = 2.0$. It must be noted, however, that any comparison with such method must be taken with a grain of salt as it has used a different training set.

Finally, although Yahoo!'s method has used a different image set for defining the network weights, it is validated here in the same test set as any other method. In addition, two important differences with regard to the methods we propose in this paper are that Yahoo!'s model considered a cluster of 16 GPUs while ours use a single GPU. A second, and important aspect, is that our solutions are trained with a public dataset while Yahoo's is proprietary.

## 5. Experiments and validation

In the following, we present the experimental results for detecting SEIC content, comparing our solutions with state-of-the-art methods, forensic tools and commercial softwares.

### 5.1. Adult content detection

At first, we analyze the performance for detecting adult content (porn vs. non-porn content) on the Pornography-2 k dataset (257,522 frames from the test data). As one might observe in Table 2, our `1-tiered Adult Detector` provides superior performance over `Yahoo! Detector` [49], for ACC and $F_2$ measures. This comparison is particularly relevant because both CNN models are trained on images of adult content. Our solution, besides being more accurate (i.e., an error reduction of 26.7% in ACC when compared to the Yahoo!'s one), ends up being less expensive in terms of computational resources: while the `Yahoo! Detector` uses of a 16 GPU-cluster for training the CNNs, we work with a single GPU machine. Although both networks have similar performance for actually detecting adult content (similar TPR), our method is much better at rejecting innocuous content (higher TNR). Concerning the COPINE scale, both detectors reasonably cover depictions of L1 through L7 levels.

### 5.2. SEIC detection

Turning the attention to the ultimate goal of detecting SEIC content, in Table 3, we have the results for the different methods for this more challenging task. Without surprise, we observe a considerable improvement of performance from skin-detector-based systems (Kovac et al. and different off-the-shelf forensic tools) to non-skin-detector-based ones (BoVW- and Deep Learning-based approaches), showing the relevance of exploiting richer and more discriminative patterns. This might indicate that human skin detection is of circumstantial – or no importance – for detecting SEIC content. The strength of non-skin-detector-based techniques is further highlighted when comparing the proposed `1-tiered` and `2-tiered SEIC Detectors` to the best skin-detector-based solution (`NuDetective`). Our solutions provide remarkable improvements (e.g., 29.4 percentage points for ACC and 13.9 percentage points for $F_2$ when considering the `2-tiered SEIC Detector`).

**Table 2**
Results for **adult** content detection.

| | TPR (%) | TNR (%) | $F_2$ (%) | ACC (%) |
|---|---|---|---|---|
| Yahoo! detector [49] | 94.5 | 82.3 | 94.1 | 88.4 |
| 1-Tiered adult detector | 94.8 | 88.2 | 94.8 | 91.5 |

TPR: true positive rate – TNR: true negative rate – ACC: accuracy – $F_2$: $F_2$ measure.

**Table 3**
Results for **SEIC** content detection.

| | TPR (%) | TNR (%) | $F_2$ (%) | ACC (%) |
|---|---|---|---|---|
| Kovac et al. [64] | 89.6 | 16.5 | 84.4 | 53.1 |
| MediaDetective [37] | 71.3 | 42.0 | 70.0 | 56.6 |
| Snitch plus [38] | 68.4 | 45.2 | 67.8 | 56.8 |
| NuDetective [31] | 76.4 | 37.0 | 73.8 | 56.7 |
| SURF & BoVW | 69.9 | 66.6 | 71.1 | 68.3 |
| Yahoo! detector [49] | 80.1 | 74.9 | 80.7 | 77.5 |
| SEIC detector (random initial weights) | 79.5 | 72.3 | 79.9 | 75.9 |
| 1-Tiered adult detector | 83.0 | 77.8 | 83.4 | 80.4 |
| 1-Tiered SEIC detector | 86.8 | 83.7 | 87.2 | 85.2 |
| 2-Tiered SEIC detector | 87.2 | 85.0 | 87.7 | 86.1 |
| 2-Tiered SEIC detector (ext.) | 86.2 | 86.7 | 87.1 | 86.5 |

TPR: true positive rate – TNR: true negative rate – ACC: accuracy – $F_2$: $F_2$ measure.

It should be mentioned that, with respect to $F_2$ measure, higher numbers do not mean necessarily higher performance (see Kovac et al.' results), since $F_2$ does not take into account the true negatives. Therefore, both classification metrics, ACC and $F_2$, must be considered together for evaluating the performance.

Regarding our proposed SEIC detectors, we observe that fine-tuning improves detection, both when transferring network weights and configurations from the original object detection network and fine-tuning it to SEIC content detection (`1-tiered SEIC Detector`), and when transferring weights and network configurations from the `1-tiered Adult Detector` to the SEIC content detection network (`2-tiered SEIC Detector`). With two different levels of fine-tuning and training, our solution is completely data-driven with the advantage of dealing with the scarcity of labeled SEIC content.

Table 3 also shows the comparison with the `Yahoo! Detector`. Our `1-tiered SEIC Detector` outperforms their detector by 7.7 and 6.5 percentage points, for ACC and $F_2$, respectively. If we consider the proposed best-performing solution (`2-tiered SEIC Detector`), we reach a notable improvement over `Yahoo! Detector` of 8.6 percentage points for ACC and 7.0 percentage points for $F_2$.

Fig. 4 illustrates the error reduction relative to the `2-tiered SEIC Detector`. Comparing our `2-tiered SEIC Detector` to skin-detector-based systems, we observe an error reduction of up to 70.4% (from 46.9% to 13.9%). In the case of BoVW-based method, the error reduction is of 56.2%, which agrees with current literature in Computer Vision, which shows outstanding results of deep convolution neural networks over BoVW-based techniques over various applications.

Comparing the strategies of transfer learning both `1-tiered` and `2-tiered SEIC Detectors` to the `1-tiered Adult Detector` and the `SEIC Detector (random initial weights)`, it is clear the importance of the transfer with both `1-tiered` and `2-tiered SEIC Detectors` outperforming the other solutions. This positive result is a major step forward for combating SEIC content.

Finally, comparing the `1-tiered SEIC Detector` and the `2-tiered SEIC Detector` directly, we can see the latter outperforms the former by one percentage point. One important aspect here is that as the 1- and 2-tiered solutions have different initializations policies, one with one level of knowledge transfer and the other with two levels, the two methods are somewhat complementary. An additional analysis of these two solutions show they are indeed complementary and when combined with a simple max probability fusion strategy they lead to an

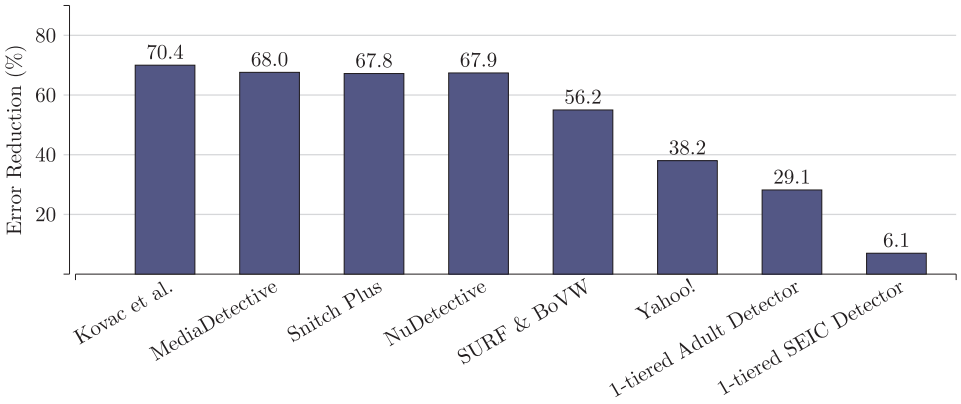


**Fig. 4.** Classification error reduction (%) with respect to our `2-tiered SEIC Detector` considering the metric ACC in Table 3.

$F_2$ of about 89.6%, which is three percentage points better than the best solution in isolation.

We also report the result for the extension of our 2-tiered solution. By augmenting the data for the non-SEIC class, we achieved slightly better results for ACC (86.5%); for $F_2$ measure, however, our result (87.1%) is lower than expected. Still, this experiment indicates that our results can be improved even further by training the network with more data, non-SEIC and SEIC content. In the next section, we discuss one possible future branch of research for automatically incorporating more training examples once the current solution is deployed in some investigation sites.

### 5.3. Detecting top-levels of the COPINE scale

As we discuss in the introduction, our solutions presented in this work are data-driven, which means they can automatically derive discriminative features directly from the available training data. When providing the model with adult content-type only, the one-tiered solution detects general pornography, which means it deals with COPINE scale levels of L1 through L7, mostly.

When we incorporate SEIC examples into the training, the methods (one- and two-tiered) improve the classification rate in at least four percentage points, being now capable of detecting some more images falling into the L8 through L10 levels of the COPINE scale. Of course, there is still much research to be done as the best available solution still misses about 13 percentage points of the evaluated cases. It is important to mention, however, that available off-the-shelf solutions perform with 77% at best (see NuDetective's performance) and it is currently in use by the Brazilian Federal Police. Equally important, a powerful model developed by Yahoo! [49] achieves about 80%. Compared to those solutions, proposed methods lead to a remarkable improvement of 10 and seven percentage points, respectively.

### 5.4. Variability under different training configurations

Sometimes, the order of examples with which a CNN-based solution is trained might affect the created model. To verify this, we performed some experiments randomizing the training dataset to assess any possible discrepancy when testing methods trained with different training sets but tested with the same testing data. For that, we first fixed the test set (19,387 images), the same used in all the experiments in this paper. Next, we randomly created five variations of training data. In each variation, we separated 29,442 images for training and the remaining 10,142 images for validation. We fine-tuned five CNNs based on such sets. Afterwards, we tested the CNNs with the fixed test set.

As one might observe in Table 4, the results are reasonably uniform over all five variations of training sets, for both ACC and $F_2$ measures. Comparing the results of the `2-tiered SEIC Detector`, in which the training dataset consists of 33,646 images, in the previous section to the results involving variations of the training dataset, which contains 29,442 images (12.5% decrease from the original training data), we observe a small impact on performance accuracy when changing the order of training sets/elements. Therefore, we can safely conclude, at least for the training data that we used in this work, the detector is reasonably robust when training with different training sets considering different/random order in the used batches.

**Table 4**
Results for **SEIC** content detection.

| | $F_2$ (%) | ACC (%) |
|---|---|---|
| `2-Tiered SEIC detector` | 87.7 | 86.1 |
| Variation #1 | 88.0 | 85.1 |
| Variation #2 | 88.2 | 85.0 |
| Variation #3 | 88.3 | 85.2 |
| Variation #4 | 88.3 | 85.1 |
| Variation #5 | 88.2 | 85.1 |

ACC: accuracy – $F_2$: $F_2$ measure.

### 5.5. First take on video classification

Although in this paper we focus on detecting SEIC content in images, we performed a small experiment with 100 random videos present in a real-case apprehension (50 normal videos and 50 with SEIC content at different parts of the streams) using our detector for images.

For the sake of simplicity, we have used the typical video classification pipeline: (i) split the video into frames; (ii) pool/aggregate the extracted features; and (iii) train a classifier. By using a frame sampling rate of five frames per second, we apply the `2-tiered SEIC Detector` model described in Section 3 and trained with images as our feature extractor. After that, we average pool the features to obtain a single description of the video, and we employ the SVM model trained for the `2-tiered SEIC Detector` to predict the video label.

The obtained classification accuracy is of 88.0% and an $F_2$ measure of 79.8%, in which all the negative samples – without child pornography content – are classified as negative (true negative rate is 100%). Regarding the positive samples, the analysis of the hardest false negatives revealed that the method has some difficulties when the videos are of very poor quality (often captured from webcams) or when the videos have few SEIC explicit elements. From this initial work, we intend to improve the detector by including temporal features and also to include audio, two important aspects present in videos that could boost the classification results significantly when compared to still images.

## 6. Conclusions and future work

Deep convolution neural networks are the state of the art for image classification tasks [53,54,68], but their use for sexually exploitative imagery of children is challenging, since those models require large amounts of training data. To bypass that problem, in this paper, we

proposed a data-driven solution in which: (1) we first transfer the network parameters and configurations trained on ImageNet (with 1.2 million images) to the target problem SEIC content detection (1-tiered solution) and (2) we also perform a 2-tiered transfer learning procedure, in which we first transfer knowledge from the network trained over ImageNet (with 1.2 million images) to the problem of adult content detection (with 200 thousand images) and fine-tune the network for SEIC content detection problem.

The evaluation of our solution shows remarkable improvements not only over current scientific state of the art, but also over off-the-shelf forensic and commercial tools. Additionally, to the best of our knowledge, this is the first time that the deep convolutional neural networks are evaluated to cope with the problem of SEIC content detection. Most importantly, our work is one of the very few solutions, as far as we know, able to deal with SEIC grammar at the COPINE scale levels of L8 through L10.

We believe our solution can potentially aid forensic experts in their daily routines of automatically examining vast troves of data in a process that heretofore was chiefly manual [14] or that relied on solutions producing a high number of false positives. In view of that, we produced a self-contained sandbox virtual machine with our solution, freely available upon request for direct deployment. We hope it will be used by different law-enforcement agents for combating SEIC content. To access the sandbox, an interested reader can contact the authors or access http://www.recod.ic.unicamp.br/.

In addition, in this paper we showed that it is possible to perform different levels of transfer learning and network adaptations for the problem of SEIC content detection, which often suffers with lack of proper training data. In our case, it was possible to perform different levels of knowledge transfer, reducing in one order of magnitude the requirements of training size. This approach might be the key for other related and complicated problems such as violence detection and grammar refinement for each class of the COPINE scale, for instance.

As future work, it is only natural to envision extending our method to detect child pornography in videos. Video files have important features that make automatic detection a challenge task [36]: they are usually large, have different formats and encodings, and may have low resolution. For further reducing the error rate, we intend to improve the 1- and 2-tiered SEIC models by using better pre-trained models and by adjusting the architecture in addition to using more training data. Another branch of research could focus on refining the solution for classifying the content among the different COPINE scale levels. This would be important to further focus the training on failing cases and improving the solution for those cases with a guided-training process. Finally, we also envision an online automatic classification and self-tagging method in which the current solution, once deployed at some forensic sites, could collect the most reliable classified images (higher classification confidence) and automatically integrate those examples into the fine-tuning and training stages re-running such stages periodically. A similar process was recently proposed by Guo and Aarabi [69] for the problem of hair segmentation and holds promise.

## Acknowledgments

This work was supported in part by Microsoft Research, São Paulo Research Foundation (Fapesp) under the grant #2017/12646-3 (DéjàVu project), CAPES DeepEyes project, PNPD/CAPES, Google Research Awards for Latin America 2016, and Brazil's National Program for Public Security with Citizenship (PRONASCI). We acknowledge the support of NVIDIA Corporation with the donation of two GPUs used in this research.

## References

[1] M. Taylor, E. Quayle, Child Pornography: An Internet Crime, Brunner-Routledge, 2003.
[2] The National Child Traumatic Stress Network (NCTSN), Child Sexual Abuse Fact Sheet, Tech. Rep., NCTSN Fact Sheet, The National Child Traumatic Stress Network (NCTSN), 2009.
[3] International Centre for Missing & Exploited Children (ICMEC), Child Pornography: Model Legislation & Global Review, Tech. Rep., eighth edition, International Centre for Missing & Exploited Children (ICMEC), 2016.
[4] N. Al Mutawa, J. Bryce, V.N. Franqueira, A. Marrington, Behavioural evidence analysis applied to digital forensics: An empirical analysis of child pornography cases using p2p networks, in: IEEE International Conference on Availability, Reliability and Security (ARES), 2015, pp. 293–302.
[5] W.A. Dorland, Dorland's Illustrated Medical Dictionary, 28th ed., Elsevier Saunders, 1994 p. 1248 <http://www.dorlands.com/wsearch.jsp>.
[6] K. Eichenwald, On the Web, Pedophiles Extend their Reach, p. A1, The New York Times, August 21, 2006.
[7] M.C. Seto, J.M. Cantor, R. Blanchard, Child pornography offenses are a valid diagnostic indicator of pedophilia, J. Abnorm. Psychol. 115 (3) (2006) 610.
[8] J. Sher, B. Carey, Debate on Child Pornography's Link to Molesting, The New York Times, July 19, 2007, p. A20.
[9] Tech Innovation to Fight Child Sexual Exploitation (THORNE) < https://www.wearethorn.org/ > (last accessed on November 2nd, 2016).
[10] International Criminal Police Organization (Interpol), Interpol Recommendations on Offences Against Minors, Tech. Rep., Interpol 61st General Assembly, International Criminal Police Organization (Interpol), 1995.
[11] F. Madsen, Transnational Organized Crime, Routledge, 2009 (p. 146).
[12] M. Taylor, E. Quayle, G. Holland, Child pornography, the internet and offending, Can. J. Pol. Res. 2 (2) (2001) 94–100.
[13] N. Sae-Bae, X. Sun, H.T. Sencar, N.D. Memon, Towards automatic detection of child pornography, in: IEEE International Conference on Image Processing, 2014, pp. 5332–5336.
[14] K. Ricanek Jr, C. Boehnen, Facial analytics: from big data to law enforcement, IEEE Comp. 9 (45) (2012) 95–97.
[15] V. Chatzis, F. Panagiotopoulos, V. Mardiris, Face to iris area ratio as a feature for children detection in digital forensics applications, in: IEEE Digital Media Industry & Academic Forum, 2016, pp. 121–124.
[16] A. Ulges, A. Stahl, Automatic detection of child pornography using color visual words, in: IEEE International Conference on Multimedia and Expo, 2011, pp. 1–6.
[17] C. Schulze, D. Henter, D. Borth, A. Dengel, Automatic detection of CSA media by multi-modal feature fusion for law enforcement support, in: ACM International Conference on Multimedia Retrieval, 2014, p. 353.
[18] M. Chopra, M.V. Martin, L. Rueda, et al., Toward new paradigms to combating internet child pornography, in: IEEE Canadian Conference on Electrical and Computer Engineering, 2006, pp. 1012–1015.
[19] A. Shupo, M.V. Martin, L. Rueda, A. Bulkan, Y. Chen, P.C. Hung, Toward efficient detection of child pornography in the network infrastructure, IADIS Int. J. Comp. Sci. Inf. Syst. 1 (2) (2006) 15–31.
[20] G. Bissias, B. Levine, M. Liberatore, B. Lynn, J. Moore, H. Wallach, J. Wolak, Characterization of contact offenders and child exploitation material trafficking on five peer-to-peer networks, Child Abuse Neglect 52 (2016) 185–199.
[21] C. Peersman, C. Schulze, A. Rashid, M. Brennan, C. Fischer, iCOP: automatically identifying new child abuse media in P2P networks, IEEE Security and Privacy Workshops, IEEE, 2014, pp. 124–131.
[22] M. Grega, D. Bryk, M. Napora, INACT-INDECT advanced image cataloguing tool, Multim. Tools Appl. 68 (1) (2014) 95–110.
[23] M. Inc., PhotoDNA Cloud Services, 2016 < https://www.microsoft.com/en-us/PhotoDNA > .
[24] T. Krone, A typology of online child pornography offending, Trends Issues Crime Crim Just. (July) (2004) 1–6.
[25] A. Adler, The perverse law of child pornography, JSTOR Columbia Law Rev. (2001) 209–273.
[26] L. Edwards, Content filtering and the new censorship, in: IEEE International Conference on Digital Society, 2010, pp. 317–322.
[27] D.J. Weitzner, Free speech and child protection on the web, IEEE Internet Comput. 11 (3) (2007) 86–89.
[28] T.W. Todd, H. Beecher, G.H. Williams, A.W. Todd, The weight and growth of the human eyeball, JSTOR Hum. Biol. 12 (1) (1940) 1–20.
[29] P. Eleuterio, M. Polastro, Identification of high-resolution images of child and adolescent pornography at crime scenes, Int. J. Foren. Comp. Sci. 5 (1) (2010).
[30] M. Islam, P.A. Watters, J. Yearwood, Real-time detection of children's skin on social networking sites using markov random field modeling, Inf. Sec. Tech. Rep. 16 (2) (2011) 51–58.
[31] M. Polastro, P. Eleuterio, NuDetective: A forensic tool to help combat child pornography through automatic nudity detection, in: IEEE International Workshops on Database and Expert Systems Applications, 2010, pp. 349–353.
[32] M. Polastro, P. Eleuterio, A statistical approach for identifying videos of child pornography at crime scenes, in: IEEE International Conference on Availability, Reliability and Security, 2012, pp. 604–612.
[33] M. Grega, D. Bryk, M. Napora, M. Gusta, Inact-indect advanced image cataloguing tool, in: Springer International Conference on Multimedia Communications, Services and Security, 2011, pp. 28–36.
[34] S. Lohr, Microsoft Tackles the Child Pornography Problem, The New York Times < http://bits.blogs.nytimes.com/2009/12/16/microsoft-tackles-the-child-pornography-problem/ > (December 16, 2009).
[35] A. Panchenko, R. Beaufort, H. Naets, C. Fairon, Towards detection of child sexual abuse media: categorization of the associated filenames, in: Advances in Information Retrieval: European Conference on IR Research, 2013, pp. 776–779.
[36] M. Polastro, P. Eleuterio, Quick identification of child pornography in digital

videos, Int. J. Foren. Comp. Sci. 2 (2012) 21–32.
[37] Media Detective < www.mediadetective.com > .
[38] Snitch Plus < www.hyperdynesoftware.com > .
[39] T. Deselaers, L. Pimenidis, H. Ney, Bag-of-visual-words models for adult image classification and filtering, in: International Conference on Pattern Recognition, 2008, pp. 1–4.
[40] S. Avila, N. Thome, M. Cord, E. Valle, A. Araújo, Pooling in image representation: the visual codeword point of view, Comp. Vis. Image Understand. 117 (5) (2013) 453–465.
[41] C.C. Yan, Y. Liu, H. Xie, Z. Liao, J. Yin, Extracting salient region for pornographic image detection, J. Vis. Commun. Image Represent. (JVCIR) 25 (5) (2014) 1130–1135.
[42] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, Pornography classification: the hidden clues in video space-time, Foren. Sci. Int. 268 (2016) 46–61.
[43] C. Caetano, S. Avila, W.R. Schwartz, S.J.F. Guimarães, A. Araújo, A mid-level video representation based on binary descriptors: a case study for pornography detection, Neurocomputing 213 (2016) 102–114.
[44] P. Vitorino, S. Avila, A. Rocha, A two-tier image representation approach to detecting child pornography, in: XII Workshop de Visão Computational, 2016, pp. 129–134.
[45] Y. Liu, Y. Yang, H. Xie, S. Tang, Fusing audio vocabulary with visual features for pornographic video detection, Fut. Gener. Comp. Syst. (FGCS) 31 (2014) 69–76.
[46] M. Moustafa, Applying deep learning to classify pornographic images and videos, in: Pacific-Rim Symposium on Image and Video Technology, 2015, pp. 1–10.
[47] F. Nian, T. Li, Y. Wang, M. Xu, J. Wu, Pornographic image detection utilizing deep convolutional neural networks, Neurocomputing 210 (2016) 283–293.
[48] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, Video pornography detection through deep learning techniques and motion information, Neurocomputing 230 (2017) 279–293.
[49] J. Mahadeokar, G. Pesavento, Open Sourcing a Deep Learning Solution for Detecting NSFW Images, 2016 < https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for > .
[50] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, Pornography classification: the hidden clues in video space-time, Foren. Sci. Int. 268 (2016) 46–61.
[51] A. Sorokin, D. Forsyth, Utility data annotation with Amazon Mechanical Turk, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–8.
[52] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[53] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 512–519.
[54] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[55] M. Lin, Q. Chen, S. Yan, Network in Network, CoRR, 2013, abs/1312.4400.
[56] S. Hochreiter, F. F. Informatik, Y. Bengio, P. Frasconi, J. Schmidhuber, Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies, IEEE Press Field Guide to Dynamical Recurrent Networks, 2000, pp. 237–243.
[57] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors, arXiv preprint arXiv:1207.0580.
[58] I. Sutskever, J. Martens, G.E. Dahl, G.E. Hinton, On the importance of initialization and momentum in deep learning, in: International Conference on Machine Learning, 2013, pp. 1139–1147.
[59] A.G. Howard, Some Improvements on Deep Convolutional Neural Network Based Image Classification, arXiv preprint arXiv:1312.5402.
[60] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, L. Zitnick, Microsoft COCO: Common Objects in Context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[61] M. Short, L. Black, A. Smith, C. Wetterneck, D. Wells, A review of internet pornography use research: methodology and content from the past 10 years, Behav., Soc. Network. 15 (1) (2012) 13–23.
[62] P. Kakumanu, S. Makrogiannis, N. Bourbakis, A survey of skin-color modeling and detection methods, Pattern Recog. 40 (3) (2007) 1106–1122.
[63] W. Kelly, A. Donnellan, D. Molloy, Screening for objectionable images: A review of skin detection techniques, in: IEEE International Machine Vision and Image Processing Conference, 2008, pp. 151–158.
[64] J. Kovac, P. Peer, F. Solina, Human skin color clustering for face detection, in: IEEE International Conference on Computer as a Tool, 2003, pp. 144–148.
[65] H. Bay, A. Ess, T. Tuytelaars, L.V. Gool, SURF: speeded up robust features, Comp. Vis. Image Understand. 110 (3) (2008) 346–359.
[66] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: International Conference on Computer Vision, vol. 2, 2003, pp. 1470–1477.
[67] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27 <http://www.csie.ntu.edu.tw/cjlin/libsvm>.
[68] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv preprint arXiv:1512.03385.
[69] W. Guo, P. Aarabi, Hair segmentation using heuristically-trained neural networks, IEEE Trans. Neural Netw. Learn. Syst. (99) (2016) 1–12.