## 1. Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy

**Accession number:** 20173003968959
**Authors:** Ou, Xinyu (1, 2, 3); Ling, Hefei (1); Yu, Han (2); Li, Ping (1); Zou, Fuhao (1); Liu, Si (2)
**Author affiliation:** (1) School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China; (2) Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China; (3) College of Yunnan Province Cardre Online Learning, Yunnan Open University, Kunming, China
**Source title:** ACM Transactions on Intelligent Systems and Technology
**Abbreviated source title:** ACM Trans. Intell. Syst. Technolog.
**Volume:** 8
**Issue:** 5
**Issue date:** July 2017
**Publication year:** 2017
**Article number:** 68
**Language:** English
**ISSN:** 21576904
**E-ISSN:** 21576912
**Document type:** Journal article (JA)
**Publisher:** Association for Computing Machinery
**Abstract:** Adult image and video recognition is an important and challenging problem in the real world. Low-level feature cues do not produce good enough information, especially when the dataset is very large and has various data distributions. This issue raises a serious problem for conventional approaches. In this article, we tackle this problem by proposing a deep multicontext network with fine-to-coarse strategy for adult image and video recognition. We employ a deep convolution networks to model fusion features of sensitive objects in images. Global contexts and local contexts are both taken into consideration and are jointly modeled in a unified multicontext deep learning framework. To make the model more discriminative for diverse target objects, we investigate a novel hierarchical method, and a task-specific fine-to-coarse strategy is designed to make the multicontext modeling more suitable for adult object recognition. Furthermore, some recently proposed deep models are investigated. Our approach is extensively evaluated on four different datasets. One dataset is used for ablation experiments, whereas others are used for generalization experiments. Results show significant and consistent improvements over the state-of-the-art methods. © 2017 ACM.
**Number of references:** 79
**Main heading:** Deep learning
**Controlled terms:** Convolution - Object recognition
**Uncontrolled terms:** Ablation experiments - Conventional approach - Convolutional networks - Fine-to-coarse strategy - Hierarchical method - Learning frameworks - State-of-the-art methods - Video recognition
**Classification code:** 716.1 Information Theory and Signal Processing
**DOI:** 10.1145/3057733
**Compendex references:** YES
**Database:** Compendex

**Data Provider:** Engineering Village

# Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-Coarse Strategy

XINYU OU, Huazhong University of Science and Technology, Chinese Academy of Sciences, Yunnan Open University
HEFEI LING, Huazhong University of Science and Technology
HAN YU, Chinese Academy of Sciences
PING LI and FUHAO ZOU, Huazhong University of Science and Technology
SI LIU, Chinese Academy of Sciences

Adult image and video recognition is an important and challenging problem in the real world. Low-level feature cues do not produce good enough information, especially when the dataset is very large and has various data distributions. This issue raises a serious problem for conventional approaches. In this article, we tackle this problem by proposing a deep multicontext network with fine-to-coarse strategy for adult image and video recognition. We employ a deep convolution networks to model fusion features of sensitive objects in images. Global contexts and local contexts are both taken into consideration and are jointly modeled in a unified multicontext deep learning framework. To make the model more discriminative for diverse target objects, we investigate a novel hierarchical method, and a task-specific fine-to-coarse strategy is designed to make the multicontext modeling more suitable for adult object recognition. Furthermore, some recently proposed deep models are investigated. Our approach is extensively evaluated on four different datasets. One dataset is used for ablation experiments, whereas others are used for generalization experiments. Results show significant and consistent improvements over the state-of-the-art methods.

CCS Concepts: • **Computing methodologies** → **Object recognition**; *Image representations*; • **Security and privacy** → *Social network security and privacy;*

Additional Key Words and Phrases: Adult image and video recognition, multicontext modeling, fine-to-coarse strategy, deep convolutional network

**ACM Reference Format:**
Xinyu Ou, Hefei Ling, Han Yu, Ping Li, Fuhao Zou, and Si Liu. 2017. Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy. ACM Trans. Intell. Syst. Technol. 8, 5, Article 68 (July 2017), 25 pages.
DOI: http://dx.doi.org/10.1145/3057733



X. Ou was an intern at S-Lab of CASIIE at the time of our work.
This work was supported by the Natural Science Foundation of China (U1536203, 61572493, 61572214, 61672254), the Major Scientific and Technological Innovation Project of Hubei Province (2015AAA013), the Nature Science Foundation of the Open University of China (G16F3702Z, G16F2505Q), and the Major Scientific Research Project of Yunnan Provincial Education Department (2015Z169).
Authors' addresses: X. Ou, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, and College of Yunnan Province Cardre Online Learning, Yunnan Open University, Kunming, China; email: ouxinyu@hust.edu.cn; H. Ling, P. Li, and F. Zou, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China; emails:{lhefei, lpshome, fuhao_zou}@hust.edu.cn; H. Yu and S. Liu, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China; emails: {yuhan, liusi}@iie.ac.cn.

DOI: http://dx.doi.org/10.1145/3057733

## 1. INTRODUCTION

The Internet, as a global information center, allows people all over the world to browse, share, and exchange their resources and information. There are countless Web sites devoted to providing services, such as images and videos sharing sites (Flickr, YouTube, Qzone, etc.) and searching engines (Google, Baidu, Bing, etc.). Although the Internet brings convenience, Web pages with harmful or illegal content (e.g., adult content) are widely available. Recognition of adult content is important for the healthy development of Internet resources and culture but is an extremely challenging problem.

First, how to define *adult* is a major issue that we must face. Most countries define pornography (obscenity) as being "adult" in their laws (e.g., U.S. law [LegalDictionary 2015] and Chinese criminal law [ChinaPRC 2015]). They define the representation in books, magazines, photographs, films, and other media of scenes of sexual behavior that are erotic or lewd and are designed to arouse sexual interest. However, these legal documents have no official grading system to distinguish adult content. How to demarcate in detail and filter the sensitive images and videos is a headache to Webmasters and media administrators. Many sites ban adult files; however, it is difficult for the system to prevent users to upload adult files. Worse yet, some sites are proactive in offering adult images and videos. These phenomena cause the Internet to be inundated with many ungraceful images and videos.

Second, automatically detecting and recognizing adult content is a challenging problem because of the scale and diversity of image content on the Internet. With regard to scale, for any Web site or Web application, millions or even billions of images or videos will not be an amazement. With regard to diversity, there is a wide range in the content, size, resolution, and quality of images on the Internet. The nature of image content ranges from professionally created to phone photos, human action to organic close-up, fuzzy images to high-definition images, very small to very large sizes, binary and grayscale images to full-color images, cartoon and edited images to camera-captured images, and so on. Because of the large volume of media content on the Web, it is impossible to editorially screen the content. Therefore, automated techniques for identifying adult visual (image/video) content are required.

Considering the preceding challenges, this work proposes a deep multicontext network (DMCNet) with fine-to-coarse strategy that recognizes adult images and videos from mixed datasets. The specific contributions are the following:

—We present a simple but efficient unified deep learning framework for recognizing adult images and videos.
—A carefully designed multicontext strategy is proposed, which takes full account of the particularity of its task. Different from conventional feature fusion, we do not average different features directly but adopt a parameterized hierarchical fusion strategy. The purpose is to keep the accurate judgment of the global context and then use the local context to correct minor defects.
—Due to the diversity of adult images, a task-specific fine-to-coarse strategy is designed to make the deep modeling more suitable for adult object recognition.
—Modular design methodology allows improvement in overall performance by upgrading the global context or local context components.
—We collected three datasets with different styles for the adult recognition task. We plan to release these datasets via a formal application and authorization.

The article is organized as follows. Section 2 briefly reviews previous related works. We elaborate in detail our DMCNet with fine-to-coarse strategy methodology in Section 3. Section 4 demonstrates the experimental results, and we conclude in Section 5.

## 2. RELATED WORK

Automatic recognition of adult images has been studied by many researchers, which is discussed next. Current methods can be classified as human model based, handcraft based, and deep learning based. We also review some related work on deep convolutional networks, hierarchical learning, and multicontext modeling.

### 2.1. Recognition of Adult Images

*Human model–based recognition*. Human model–based methods emphasize the use of a human body model. The primary goal of these methods is to determine whether an image contains a naked human body. Skin color feature and human modeling are typically used. Forsyth and Ioffe [2001] employed a probabilistic framework to find people by drawing a human-like assembly. They also used the Monkatrol sampling method [Ioffe and Forsyth 1998; Forsyth and Ioffe 1999] to learn the geometric relationship between the different parts of the human body. Based on these methods, several works [Bregler et al. 1996; Forsyth and Fleck 1997; Fleck and Forsyth 1999; Forsyth et al. 1999] used skin color and texture information to segment image pixels and then connected skin regions to candidates for trunks and limbs. These skin regions were combined subject to constraints derived from a human body geometrical model, which was studied by Forsyth and Ioffe [Ioffe and Forsyth 1998; Forsyth and Ioffe 1999, 2001]. If the shape formed by the combination is a humanoid body, the image is treated as an adult image and otherwise is regarded as a normal image.

Human model–based methods are committed to utilizing a global perspective to achieve object recognition. However, due to a complicated structure, it is difficult to consider all possible relative positions of the parts of the body. Furthermore, these methods have a high computational complexity, rendering them unsuitable for widespread use.

*Handcraft-based recognition*. Handcraft-based methods emphasize feature selection and extraction in adult images. When we look back through the history, we can find that the most investigated feature is human skin color [Rehg and Jones 2002]. Once an image contains too many skin-colored pixels, it is taken as an indicator of nudity. Jiao et al. [2001] used the proportion of the largest connected skin region in the image to form a vector for recognition. Liang et al. [2004] extracted seven skin-based features to form a fusion feature vector, then fed it into a support vector machine (SVM) for classification. Zheng et al. [2004] constructed a maximum entropy model for the distribution of skin color to segment the skin regions in an image, and a multilayer perception was used to recognize adult images. Deselaers et al. [2008] introduced classification accuracy improvement by adopting the visual bag-of-words (BoW) model to extract the most common patches.

However, skin color is not reliable, as some images consisting of a great many skin pixels are not pornographic, such as face close-ups, boxing games, and baby photos. Some researchers augmented skin color with other constraints or shape features [Rowley et al. 2006]. Arentz and Olstad [2004] extracted a series of features from each connected skin region: color, texture, and shape, together with the normalized center and the area of the region. Caetano et al. [2014] utilized a binary BossaNova representation to classify adult videos, which depends on the HueSIFT descriptor representing both color and shape. Yu and Han [2014] used a simple operation in HSV color space and additional edge density postprocessing for identification. Zhao and Cai [2010] extracted RGB color, a global HSV correlogram, Gabor texture features, LBP, and a local SIFT feature to train multiple SVM classifiers for different erotic classes. Yin et al. [2011] proposed a hybrid multistage region detection approach to filter the non–skin color pixels. Then a texture-based coarse degree filter and a fractal dimension–based geometry filter were used to get more accurate results.

In addition, some techniques that do not refer to feature selection have been proposed. Wang et al. [1998] used Daubechies wavelet analysis to extract features. At the same time, the normalized central moments and the color histogram were used to form semantic-matching vectors to classify images. In Dong et al. [2014], based on the bag-of-visual-words model, image visual features such as texture and local shape were coalesced with text information extracted from the image file name, file header, or Web page, then the SVM was applied to accomplish the image classification. To better use the SIFT feature and color descriptor with a modified $k$-NN algorithm, a modification of the extreme learning machine was developed to improve the classifier by Akusok et al. [2015]. Face detection was used to filter out normal close-up images in Zhao and Cai [2010] and Sae-Bae et al. [2014]. Moreover, semantic understanding of a scene for object recognition and behavior recognition has positive significance, because specific objects appear only in specific scenarios. A saliency-guided unsupervised feature learning was proposed by Zhang et al. [2015] to use saliency information for finding interested objects from a particular scenario.

*Deep learning–based recognition*. In parallel with the traditional route of handcrafted features, another route is using deep learning. Moustafa [2015] proposed constructing a system based on one of the recently flourishing deep learning techniques—convolutional networks—to recognize adult images and videos. Their proposed *AGNet*, combining *AlexNet* [Krizhevsky et al. 2012] and *GoogLeNet* [Szegedy et al. 2015] features, achieved very competitive results on the NPDI [de Avila et al. 2013] benchmark dataset.

### 2.2. Deep Convolutional Networks

Since the 2012 ImageNet competition [Russakovsky et al. 2015] winning entry by Krizhevsky et al. [2012], their AlexNet network has been successfully applied to a large variety of computer vision tasks, such as object detection [Girshick et al. 2014; Li et al. 2015], segmentation [Long et al. 2015; Liu et al. 2014; Zheng et al. 2015], scene classification [Zhang et al. 2016a], human pose estimation [Tang et al. 2015], video classification [Karpathy et al. 2014], makeup transfer [Liu et al. 2016], action recognition [Pan et al. 2016], and retrieval [Liu et al. 2012a; Fu et al. 2016; Xia et al. 2016]. A very important research field is object detection. It focuses on finding and recognizing the objects from a given image. A series of works have been brought forward, such as Region CNN [Girshick et al. 2014], SPPnet [He et al. 2015], Fast R-CNN [Girshick 2015], Faster R-CNN [Ren et al. 2015], and Region FCN [Dai et al. 2016].

Deep learning–based features show its powerful representation ability; our work is also under this powerful framework.

### 2.3. Hierarchical Learning

Hierarchy is very useful and important for data understanding and management. Based on the hierarchical semantic structure in WordNet [Fellbaum 1998], a major project of the linguistics and natural language processing community, most of the existing datasets were organized into hierarchies [Griffin and Perona 2008], such as the large-scale images datasets *ImageNet* [Deng et al. 2009] and *TinyImages* [Torralba et al. 2008]. This line of work has begun to reveal the effects of hierarchical structure on classification accuracy [Branson et al. 2010; Deng et al. 2010], hints at the promise of exploiting such structure for classification when evaluated in terms of the hierarchy [Deng et al. 2010], and shows improved classification given very small amounts of training data [Fergus et al. 2010]. Hierarchical technology is also used in other image vision tasks [Yang et al. 2016; Liu et al. 2015]. In addition, the categories that are sensitive to the privacy information are easy to swamp in massive category space

in the real world, a hierarchical deep multitask learning algorithm [Yu et al. 2016b] was developed to jointly learn more representative deep convolutional neural networks (CNNs) and more discriminative tree classifiers over a visual tree for identifying sensitive objects. The coarse-to-fine method, a typical hierarchical method, is widely used in computer vision [Lin et al. 2015c; Figueroa et al. 2015; Li et al. 2016]. Lin et al. [2015c] adopted a coarse-to-fine search strategy for rapid and accurate image retrieval. They first picked up a set of candidates with similar high-level semantics, then filtered the images with deep mid-level image representations.

In contrast to these methods, we propose a novel and effective fine-to-coarse strategy for adult image recognition.

### 2.4. Multicontext Modeling

In recent years, many researchers [Ciresan et al. 2012; Zhao et al. 2015; Karpathy et al. 2014; Simonyan and Zisserman 2014; Szegedy et al. 2015; Liu et al. 2012b; Zhou et al. 2017] have considered using various contexts to improve performance. *Global-Global* fusion is the most common way. Ciresan et al. [2012] proposed a multicolumn deep neural network for image classification. The input image was preprocessed by different strategies, then trained independently to generate different features for the same image. The final predictions were calculated by averaging individual predictions of each deep neural network. In fact, the majority of participating teams [Simonyan and Zisserman 2014; Szegedy et al. 2015] in ILSVRC [Russakovsky et al. 2015] considered combining the output of several models by weighted averaging. *Global-Local* fusion is another common method. Zhao et al. [2015] concatenated the global context and local context into the finally fully connected layer, then produced a saliency map. Karpathy et al. [2014] separated the input frames into two context streams for gathering low-resolution and high-resolution features, respectively. Both streams converge to a fully connected layer to produce the final prediction.

In addition, a coupled CNN [Zhang et al. 2016b], multiview stochastic learning [Yu et al. 2014], multimodal distance metric learning [Yu et al. 2016a], and other multicontext methods have been suggested to learn the robust features. Zhang et al. [2016b] proposed a coupled CNN method, which combined a candidate region proposal network (RPN) and a localization network to extract the proposals and simultaneously locate the objects, which was more efficient and accurate. Due to the diverse data have its specific statistical property and physical interpretation, single view cannot get consistent discriminate information, Yu et al. [2014] proposed a high-order distance-based multiview stochastic learning method to learn features from different views. Deep-MDML, proposed by Yu et al. [2016a], is another method of solving the problem of diverse samplesthat employs multimodel learning and can reduce the semantic gap effectively.

The standard classification task aims to learn robust features from the whole image. This can be considered because it is more concerned with the global context information. Although the detection task is more concerned with regional features, it focuses more on local context information. In this work, we intend to take full advantage of the complementarity of these two kinds of context information with a task-specific multicontext decision strategy to replace the common methods of averaging or concatenation.

## 3. APPROACH

### 3.1. Fine-to-Coarse Strategy

Exploiting hierarchical relationships is very important for large-scale image recognition issues. The coarse-to-Fine inference procedure is the most commonly used method in many visual processing problems, such as image retrieval [Lin et al. 2015b] and

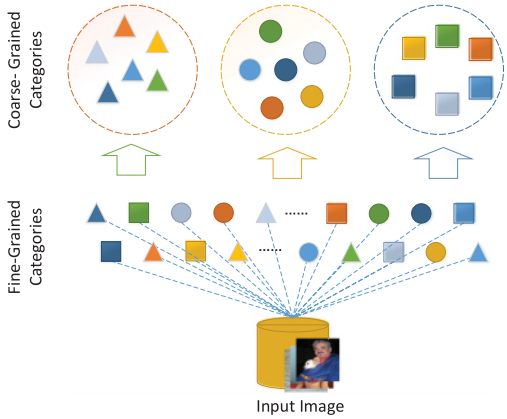


Fig. 1. Fine-to-coarse strategy. Given an input image $I$, it is first classified to a fine-grain category through a series inference and then is assigned to a coarser category space. It is noteworthy that the corresponding relationship between a fine-grain category and a coarse-grain category is allocated in advance.

object detection [Pedersoli et al. 2015]. The biggest advantage of this strategy is that it reduces search space dramatically for matching similar images from large-scale datasets, which further speeds up the process. In the adult image recognition task, the number of categories is usually limited, even with binary classification (porn vs. nonporn). However, there remain many different kinds of real images. For example, adult images may include nudity, oral sex, sexual organs, and sexual behaviors, and normal images may contain a variety of different objects, such as a cat, a person, a cake, or a car. To classify the complex concepts into two or three categories directly is difficult. Therefore, specifying a fine category for images and then combining them in a smaller class space is a simple method. We call this method a *fine-to-coarse strategy*. Figure 1 shows the fine-to-coarse strategy, and we formulate this strategy next.

Assume that $y = (y_c, y_f)$ is the label set of the input image $I$. We denote the coarse-grain category of an image $I$ as $y_c, (y_c \in C, c = (1, 2, \ldots, M))$ and the fine-grain category as $y_f, (y_f \in F, f = (1, 2, \ldots, N))$. Here, $M$, $N$ are the total numbers of categories of the coarse-grain category set $C$ and the fine-grain category set $F$, respectively. Hence, $f$ belongs to $c$ on a high-level semantic, and we can denote $y = (y_c, y_f)$, as image $I$ belongs to the coarse-grain category $y_c$ and the fine-grain category $y_f$ simultaneously. The corresponding relationship between the coarse-grain category $y_c$ and the fine-grain category $y_f$ is specified in advance. Specifically, we combine all normal images into one coarse-grain category that belongs to 997 fine-grain categories (995 classes are defined by ImageNet, and two categories are defined by our collections), nine remarkable eroticism and nudity fine-grain categories (e.g., nudity, oral sex, sexual organs, sexual behaviors) are combined into one coarse-grain category called *adult*, and three marginal categories (e.g., underwear, swimwear, leggy model) are combined and defined as being unsuitable for children. The inference problem is to find the highest score for the label $y$ of the input image $I$:

$$y_f = g(I), \tag{1}$$

$$y_c = T(y_f), \tag{2}$$

where $T : y_f \mapsto y_c$ is a mapping function, in which the symbol "$\mapsto$" means directly mapping the fine-grain category $y_f$ into the coarse-grain category $y_c$. $g(I)$ is the inference result of a deep network, such as CNN [Simonyan and Zisserman 2014] or Faster R-CNN [Ren et al. 2015]. In our method, $g(I)$ is generated by our DMCNet. This

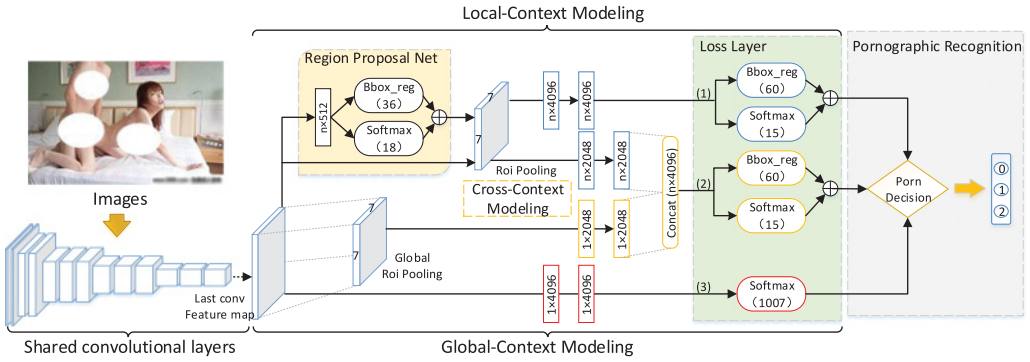


Fig. 2. Architecture of DMCNet. Upper branch (1): Deep Faster R-CNN–based [Ren et al. 2015] local context modeling for local object detection. Middle branch (2): Cross-context modeling for local object detection with global complementarity. Lower branch (3): Deep CNN–based global context modeling for a full-frame image. We visualize the fully connected layers with their corresponding dimensions. The convolutional layers can be easy to be replaced with contemporary deep CNN models (e.g., AlexNet [Krizhevsky et al. 2012], VGG16 [Simonyan and Zisserman 2014], and GoogLeNet [Szegedy et al. 2015]). All three branches share convolutional layers (with the same parameters and initialized using the ImageNet pretrained model and fine tuned on our Sensitive dataset), and a utility alternating training algorithm is used to learn the parameters (refer to Algorithm 1). In the test phase, the three branches are combined into a unified framework for adult recognition to produce a conclusive inference result. This figure is based on the VGG16 model and all hyperparameter settings (refer to Sections 3 and 4.1 (Sensitive)).

definition can be extended to the feature level, which is used to convert the fine-grain feature $\mathcal{F}_f$ to the coarse-grain feature $\mathcal{F}_c$. The extended edition is used in Section 3.5. Table II illustrates the fine-to-coarse strategy.

### 3.2. Global Context Modeling

As shown in Figure 2, the lower branch (global context modeling) of our full-frame recognition pipeline is a deep CNN architecture. A standard training procedure [Krizhevsky et al. 2012] is utilized to train the global context network. The input of this pipeline is a fixed-size ($227 \times 227$ for AlexNet [Krizhevsky et al. 2012], and $224 \times 224$ for VGG16 [Simonyan and Zisserman 2014] and GoogLeNet [Szegedy et al. 2015]) and mean-subtracted RGB image, and the output is a 1,007-Softmax corresponding to the categories of the Sensitive dataset. This network is initialized with an ImageNet-pretrained model and fine tuned end to end on the Sensitive dataset. This simple vanilla CNN model is used for initializing basic convolutional layers.

Thanks to the modular design method, it is flexible to incorporate any of the contemporary deep models into our framework, such as VGG16 [Simonyan and Zisserman 2014], VGG19 [Simonyan and Zisserman 2014], NIN [Lin et al. 2013], GoogLeNet [Szegedy et al. 2015], and ResNet [He et al. 2016a]. Some of these architectures are investigated in the experiments. As everyone knows, the deeper the model, the better the feature expression. This conclusion is further validated in our experiments, as shown in Table III.

### 3.3. Local Context Modeling

Whereas global context modeling at the lower branch aims to robustly model global features with few errors, local context modeling at the upper branch is designed to look at details. It focuses on a smaller context to refine the whole prediction, such as a local organ or a private part. In this work, we reimplement the previous work, Faster R-CNN [Ren et al. 2015] to achieve local context modeling. Local context modeling consists of two main parts with shared convolutional parts. One is an RPN, and the other is a

detection network. In this article, we investigate the AlexNet model [Krizhevsky et al. 2012], which has 5 shareable convolutional layers, and the VGG16 model [Simonyan and Zisserman 2014], which has 13 shareable convolutional layers to build shared convolutional layers. The upper part of Figure 2 illustrates the proposed local context modeling with a VGG16 architecture.

The RPN takes a feature map as input and outputs a set of rectangular object proposals with an objectness score. To generate regional proposals, we create a small network over the last convolutional feature map with a $3 \times 3$ convolutional kernel. Each sliding window is mapped to a lower-dimensional feature (256D for AlexNet and 512D for VGG16), then fed into two sibling fully connected layers, where one is a box regression layer (Bbox_reg, 36D) and the other is a binary classification layer (Softmax, 18D). Their dimensions are decided by the number of anchors (default = 9). Similar to RPN, the detection network is also separated into two sibling layers. But their dimension is decided by the number of categories, described in Section 4.2 (Sensitive).

We use a multitask loss $L$ to train the local context network jointly for classification and bounding box regression:

$$L(k, k^*, t, t^*) = \frac{1}{N_{cls}} L_{cls}(k, k^*) + \lambda \frac{1}{N_{reg}} k^* L_{reg}(t, t^*). \tag{3}$$

Here, the ground-truth label $k^*$ is 1 if the anchor is positive and is 0 if the anchor is negative. Moreover, $k = (k_0, k_1, \ldots, k_K)$ is a discrete probability distribution (per RoI) over $K + 1$ categories (the additional one is the background class), and $L_{cls}(k, k^*) = -\log p_{k^*}$ is the standard cross-entropy loss over two classes (object vs. nonobject). The second term, loss $k^* L_{reg}$, is defined over a tuple of true bounding box regression for class $k^*$; it is activated only for positive anchors ($k^* = 1$) and is shielded for others ($k^* = 0$). In addition, $t^* = (t_x^*, t_y^*, t_w^*, t_h^*)$ denotes the ground-truth bounding box and a predicted tuple $t = (t_x, t_y, t_w, t_h)$ again for class $k^*$. Finally, $L_{reg}(t, t^*) = \sum_{i \in x,y,w,h} R(t_i - t_i^*)$, where

$$R(*) = smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise,} \end{cases}$$

is a robust smoothed $L_1$ loss function defined in Girshick [2015] that is less sensitive to outliers than the $L_2$ loss used in R-CNN [Girshick et al. 2014]. The two terms are normalized by $N_{cls}$ and $N_{reg}$ and weighted by a balancing parameter $\lambda$. In our implementation, the *cls* term in Equation (3) is normalized by the mini-batch size (i.e., $N_{cls} = 64$ in VGG16), and the *reg* term is normalized by the number of anchor locations (i.e., $N_{reg} = 2{,}400$). By default, we set $\lambda = 10$ because the *reg* term is more important than *cls* in the regional proposal step. Although another value maybe more suitable for training a better RPN, this is not a decisive parameter.

After the RPN, an RoI pooling layer is used to map the region feature into a fixed-size feature map. This is configured by setting $H'$ and $W'$ to be compatible with the net's first full-connected layer (e.g., $H' = W' = 6$ for AlexNet and $H' = W' = 7$ for VGG16). An RoI pooling layer takes as input $N$ feature maps and a list of $n$ regions of interest, typically $n \gg N$. The $N$ feature maps are supplied by the last *conv* layer of the network, and each one of them is a multidimensional array of size $H \times W \times C$, with $H$ rows, $W$ columns, and $C$ channels. For each of the $n$ RoIs, RoI pooling layers output max pooled feature maps with spatial extent $H' \times W' \times C$ ($H' \leq H$ and $W' \leq W$).

### 3.4. Cross-Context Modeling

Since the global contexts are learned based on global appearance similarities, the local context concentrates on regional close-ups. Yet due to the diverse perspectives of portrait shooting, the local area or organ of people may be tricky to identify. For

Fig. 3. Some confusion examples. The first row illustrates the full images, and images in the second row are local regions with respect to the full images. The first and second columns are normal images, whereas the third and fourth columns should be considered as adult images.

example, in Figure 3, the images of the first and second columns maybe be judged as adult, and because the local area has great ambiguity, it will output high discriminant scores. However, the images in the third and fourth columns are always being judged as normal images due to the protagonist occupy a larger scope, and this will generate high discriminant scores through global modeling. Cross context is designed to estimate and weaken these confusion examples by embedding the cross-image region matching filters in a different perspective.

As the middle section of Figure 2 shows, after the last *conv* layer, the data flow is divided into two branches. One branch is with similar convolutional and RPN architecture as the local context modeling but with independent parameters in the *fc* layers. On the other branch, given the full-image *conv* feature map of the last convolutional layer, we pool the feature by global RoI pooling to obtain a global context feature that can be implemented with the existing RoI pooling layer proposed in He et al. [2015]. Similar to RoI pooling, global RoI pooling also needs an RoI region to generate the bounding box for feature extraction and feature mapping. In the detection network, the RoI regions generated by region regression in the whole image, and the RoI regions identify a local area or a local object. Instead, for global RoI pooling, the whole image is an RoI region. In other words, the whole image has only one RoI region, and the RoI region will cover the entire image. Global RoI pooling is also configured by setting $H'$ and $W'$ to be compatible with the first *fc* layer with the same parameters. Meanwhile, we set them equal to the value of local context. The cross-context feature is concatenated with the original global region feature and all proposal region features, then followed by a sibling classification and a box regression layer. It can be given by

$$\mathcal{F}_{cross}(n, D) = \mathcal{F}_{Local}(n, D_L) \oplus g(\mathcal{F}_{Global}(1, D_G), n). \tag{4}$$

Here, $n$ is the number of proposals generated by RPN, $D$ is the dimension of fusion feature, and $D_L, D_G$ are the dimension of the two *fc* layers ($D_L = D_G = 2{,}048$ in this article. Moreover, "$\oplus$" is a concatenation operator and can connect two small matrices into a large one. Function $g(A, n)$ achieves replicate matrix operation—that is, given a two-dimensional vector $\mathbf{V}$, $g(\mathbf{V}, n)$ transfer $\mathbf{V}$ to a three-dimensional $\mathbf{U}$. Here, the first two dimensions of $\mathbf{U}$ are the same to $\mathbf{V}$, but the third dimension is equal to $n$. In other words, $\mathbf{U}$ is an $n$ channel $\mathbf{V}$. We use this operation to make $\mathcal{F}_{Local}$ and $\mathcal{F}_{Global}$ have the same structures to further implement the cascade operation.

### 3.5. Multicontext Decision and Joint Training

Finally, we present the fusion and decision issues among multiple models. We discuss the following strategies for the multicontext decision:

—*Strategy 1*: One of the most straightforward and easiest methods is to integrate different-level context models into a unified framework and optimize for end-to-end learning to generate the final inference (unified optimization).
—*Strategy 2*: Incorporate multiple independent output probabilities of different-level context and generate a fusion probability to achieve recognition (avg-fusion).
—*Strategy 3*: Take full consideration of the dissimilarity and complementarity among different contexts of information, and employ a hierarchical selection algorithm to filter illegal samples (strategy-fusion).

End-to-end learning is a classic design principle in computer science, first explicitly articulated in Saltzer et al. [1981], and then introduced into the field of deep learning in recent years. It can back propagate the error from the output layer to the input layer and update all parameters at the same time. In addition, no disk storage is required for feature caching, which saves disk space and accelerates training. Nevertheless, it is not an easy job to design an end-to-end network. The most important limitation is as Ren et al. [2015] described, in which some gradients might be hard to take into account in the approximate joint training in a complex multifunction network, such as our DMCNet. Another real problem is that our design is a multipath concurrent network. Training these three branches simultaneously can produce a large number of parameters (about 350M), which is far beyond the processing capacity of our GPU equipment, NVIDIA Titan X. Unfortunately, Strategy 1 failed in our experiment.

In the past few years, several feature fusion methods have been used to improve performance. A simple but common method is to average the feature of each branch, such as was done by Karpathy et al. [2014] and Ciresan et al. [2012]. Additionally, Lin et al. [2015a] proposed a bilinear model that uses an outer product to generate a bilinear feature calculated from two different deep feature maps. In a more complicated method, Zhao et al. [2015] concatenate the two independent paths and feed into a binary classifier for saliency detection. Inspired by multicolumn CNN [Ciresan et al. 2012] of neurons in the cerebral cortex, given some input pattern, as Strategy 2 described, several paths are brought together to form a multicontext feature by simple averaging:

$$\mathcal{F}_{DMCN} = \frac{1}{N} \sum_{k=1}^{N=branches} \psi(\phi(\mathcal{F}_k))). \tag{5}$$

We denote $k = 1, 2, \ldots, N$ as DMCNet runs over $N$ branches, and the feature of the $k$-th branch is denoted by $\mathcal{F}_k$. Function $\phi(*)$ maps the fine-grain feature to the coarse-grain feature according to the fine-to-coarse strategy described in Section 3.1. Function $\psi(*)$ normalizes the feature value to [0, 1] for fair comparison. The branch number $N$ in this work is set at 3, and $\mathcal{F}_k, (k = 1, 2, 3)$ denotes the feature of the output of global context modeling, local context modeling, and cross-context modeling. The dimension of these features are 1,007, 15, 15—for instance, the numbers of classes in these branches are 1,007, 15, 15 (they will be introduced in the following section). Due to the difference of feature dimension, we cannot fuse the features directly. The mapping function $\phi(*)$ is crucial. The output dimension of $\phi(\mathcal{F}_k)$ is equal to the dimension of $\mathcal{F}_{DMCN}$. If the dimension equals 3, it means Level 3 recognition, whereas if it equals 2, it means Level 2 recognition (see Section 4.2).

In fact, there is potential trouble in Strategy 2. The observation probability of different branches not only has different dimensions but also has different connotations. The output of global context modeling is obtained by a standard 1,007-Softmax classifier, in which each node represents the probability of an input image to which category it belongs, which means that the sum of all probabilities is 1. Nevertheless, in local

Table I. Confusion Matrix of Global Context Modeling with Fine-to-Coarse Strategy on AlexNet and VGG16 Models

| | | S00 | S01 | S02 | S01+S02 | Sum |
|---|---|---|---|---|---|---|
| AlexNet | S00 | 56,126 | 97 | 28 | 125 | **56,251** |
| | S01 | 251 | 2,836 | 423 | — | **3,510** |
| | S02 | 422 | 735 | 2,353 | — | **3,510** |
| | S01+ S02 | 673 | — | — | 6,347 | **7,020** |
| VGG16 | S00 | 56,161 | 64 | 26 | 90 | **56,251** |
| | S01 | 139 | 3,085 | 286 | — | **3,510** |
| | S02 | 335 | 585 | 2,590 | — | **3,510** |
| | S01+S02 | 474 | — | — | 6,546 | **7,020** |

S00, normal; S01, adult; S02, unsuitable for children.
*Note*: The marks of the second column represent the ground truth class, and the first row represents predicted class. The numbers identify the classification performance. For example, "735" (column 4, row 4) indicates that 735 images (belong to class S02) are classified into class S01. The bold indicate the total number.

context modeling and cross-context modeling, the probability of an output node tells us whether the input image contains an object corresponding to a certain category or not. We can regard the value of each node as the output of a binary classifier. In other words, the output of global context modeling estimates to which category the input belongs, whereas the inference of local- and cross-context modeling predicts what objects the input contains.

As Table I shows, our global context has high recognition accuracy over normal images; the normal images mistaken as adult numbered only 125 (AlexNet) and 90 (VGG16) (i.e., 0.22% and 0.16%). Compared to the normal images, sensitive images (categories S01 and S02) have a higher failure rate. Even using VGG16 model, 4% and 9.5% of nonbenign images were wrongly identified as normal images, totaling 6.8% of the S01+S02 category. Inspired by these results, and considering the dissimilarity of the meaning between different inference probabilities, we proposed a novel hierarchical selection algorithm—Strategy 3—to recognize illegal samples. Our objective is to keep the right recognition results in a global context. Nevertheless, simply average output probability with different meanings may break existing advantages. Specifically, we set the *nms*_threshold of detection network at a high value (usually more than 0.95 on the Sensitive dataset). The purpose of this setting is to ensure that the detected objects have high confidence. Then we investigate the "exclusive gate," similar to Srivastava et al. [2015] and He et al. [2016b], to trade off the selections of different types of probability. The local context and cross-context paths are scaled by weight $w_1$, $w_2$, and the global context path is scaled by weight $1 - w_1 - w_2$. The final DMCNet feature can be formulated as follows:

$$\mathcal{F}_{DMCN} = \max((1 - w_1 - w_2) \cdot \tilde{\mathcal{F}}_{global}, w_1 \cdot \tilde{\mathcal{F}}_{local}, \quad w_2 \cdot \tilde{\mathcal{F}}_{cross}). \tag{6}$$

The feature $\tilde{\mathcal{F}}_{global} = \phi(\mathcal{F}_{global})$ is generated by the global context modeling with fine-to-coarse transform, and the feature $\tilde{\mathcal{F}}_{local} = \psi(\phi(\mathcal{F}_{local}, t_1))$ and $\tilde{\mathcal{F}}_{cross} = \psi(\phi(\mathcal{F}_{cross}, t_2))$ are the local context modeling and cross-context modeling with fine-to-coarse transform and normalization. As in Equation (5), function $\psi(*)$ and $\phi(*)$ are normalization and fine-to-coarse transformations, respectively. Parameters $t_1, t_2$ are the *nms*_threshold of detection network that controls the number of detected boxes. As Table III shows, local content modeling and cross-content modeling have similar results. Hence, we set $w = 2w_1 = 2w_2$ and $t = t_1 = t_2$ in this work, and we discuss the performance with different thresholds $w$ and $t$ in Section 4.4. For the sake of brevity, we use "local context" to describe the feature of $\tilde{\mathcal{F}}_{local-cross}$ in the following sections. Based on these settings,

the $\mathcal{F}_{DMCN}$ can be simplified as follows:

$$\mathcal{F}_{DMCN} = \max((1-w)\cdot\tilde{\mathcal{F}}_{global}, \quad w\cdot\tilde{\mathcal{F}}_{local-cross}). \tag{7}$$

Note that the physical implications of global context and local context are different. Although Equation (6) aims to compute the maximum among the features that is same as many maxout methods, the local context feature is very special in this article. As Table I shows, our global context has high recognition accuracy, which mainly benefits from the strong and robust CNN. However, compared to the global context model, the basic local context model is not credible. The recognition capacity of the local context is far worse than the global context, as shown in Table III. The fusion feature computed directly by Equation (6) (or another intuitive way) might hurt the overall performance. For this reason, it is our hope to retain the right recognition results produced by the global context, and we further recall some sensitive samples from the "normal images set" that are judged by the global context. More specifically, this recall processing is achieved through an accurate detection model. To obtain an accurate detection model, we adjust the *nms* threshold $t$ to increase the precision. It is common knowledge that a high *nms* value will reduce the recall of the detector, but in our multicontext decision Strategy 3, this is not a problem. First, our detector is designed to find sensitive objects or regions, not all objects. High detection accuracy means that the detected objects with a high degree confidence convicted as a sensitive object. Even if the detection system has missed many objects, the detected objects are credible. This improves the availability of the local context model as a component of the whole model. Second, from a logical perspective, the local context model is not designed to make key decisions but to make assistance decisions. Benefiting from the strong global context modeling, this design does not reduce the recall of the whole system, which will improve the recall rate of the overall system. Tables IV and V confirm this conclusion. Notably, our goal is to improve the recognition rate of pornographic objects. Certainly, a high degree of confidence does not guarantee correctness; joint decisions are necessary.

## 4. EXPERIMENTS

To better evaluate and compare our proposed method, four challenging benchmark datasets are used in this work, including Sensitive, NPDI, DMCV, and SPD. We first evaluate the effectiveness of the fine-to-coarse strategy. Then we discuss the impact of multicontext modeling. The hyperparameters $t$ and $w$ are set at different values on different datasets in this work. The assessment results are described in Section 4.4. We compare the results to further measure the generalization performance, we compare the results of our DMCNet and baseline methods with two state-of-the-arts methods, such as AGNet [Moustafa 2015] and Incremental Learning [Wang et al. 2015], on the other three datasets. All models in this work are only trained on the Sensitive dataset without training or fine tuning. In other words, only the Sensitive dataset includes training images, whereas the other three datasets only consist of validation images and testing images. The validation images are used to adjust hyperparameters, and the testing images are used to evaluate performance.

### 4.1. Implementation Details

We implemented our proposed DMCNet in the open source framework CAFFE [Jia et al. 2014]. We can recognize an image in less than 0.3 seconds on a PC with a Core E5 3.0GHz CPU and NVIDIA Titan X GPU. For training DMCNet with Strategy 3, we develop a five-step training process for joint optimization as shown in Algorithm 1. We adopt single-scale training and testing as in Girshick [2015] and Girshick et al. [2014] without using feature pyramids. An image is resized to $224 \times 224$. For RPN training, an anchor is considered as a positive example if it has an IoU ratio greater than 0.5 with

one ground truth box and otherwise is negative. We adopt the image-centric training scheme [Girshick 2015; Girshick et al. 2014], and each mini-batch consists of one image and 128 randomly sampled anchors for computing the loss. The ratio of positive and negative samples is 1:3 in a mini-batch. All models are employing a "step" learning rate policy and an initial learning rate of 0.001. We change the learning rate by factor 0.1 with step size 60,000 for RPN, 30,000 for local context and cross-context modeling, and 25,000 for global context modeling. We utilize momentum of 0.9 and weight decay of 0.0005. In training, we adopt nonmaximum suppression (*nms*) with a threshold of 0.7 to filter the proposal regions. In testing, the *nms* threshold $t$ is a tunable hyperparameter, as discussed in Section 4.4. The fine-to-coarse strategy can be considered as a postprocessing. As Figure 2 shows, the output dimensionalities of DMCNet are 15, 15, 1,007. For each branch, we group the dimensions into two or three for L2 or L3 evaluation. Specifically, in each group, we use the maxout operation to compute the maximum probability as the group output probability to achieve category transform from the fine-grain category to the coarse-grain category. The group definition and transformation rule are described in Section 4.2. Finally, a one-hot Softmax classifier is used to make the final decision.

**ALGORITHM 1:** Deep Multicontext Network Training Process

**Step 1:** Pretrain a deep CNN model with the ImageNet-pretrain [Simonyan and Zisserman 2014] model on our Sensitive dataset for initializing basic convolutional layers in Step 2 and Step 3, and obtain the global context model and a baseline model simultaneously. We could call the convolutional layers *shared* convolutional layers.
**Step 2:** Train the RPN, which is initialized with an extraordinary model pretrained in Step 1; the shared convolutional layers is fixed.
**Step 3:** Train a separate local context object detection network using the proposals generated by Step 2. This detection network is also initialized by the special pretrained model described in Step 1. The shared convolutional layers and the RPN are fixed.
**Step 4:** Keep the shared convolutional layers and the RPN fixed, combine local context and global context features, and fine tune the hybrid layers (Figure 2) to generate final cross context network. The proposal regions are also produced by Step 2.
**Step 5:** Output the unified DMCNet jointly trained in Step 1, Step 3, and Step 4 as the final DMCN model.

### 4.2. Datasets

*Sensitive* is a unique sensitive image dataset collected from the Internet and organized as in Imagenet; it contains roughly 30,000 ungraceful images. "Adult" images in this work are defined mainly as the images with naked men or women that involve various postures, sizes, actions, and close-ups. These images are divided into 14 classes to distinguish the content of porn, such as nudie, oral sex, private parts, sexual intercourse, underwear, swimsuit, and beautyleg. There are two exceptional classes that include facial portraits and formal dresses. To increase the diversity and complexity of this dataset, we extracted lots of images from the ImageNet challenge dataset [Russakovsky et al. 2015] and the Pascal VOC dataset [Everingham et al. 2010] to create the "normal" class. It is well known that ImageNet includes 1,000[1] object classes and can significantly increase the categories space, whereas each image in Pascal VOC contains multiple objects and can increase complexity. Hence, normal images include natural scenes, normal objects, and benign human and hybrid images. In addition,

[1]ImageNet actually contains seven similar classes to our collections, such as underwear and swimsuit, so we removed these images from the negative set.

most people in the entire dataset are Caucasian and Asian, and a small number of them are Black. Finally, the Sensitive dataset includes 1,413,765 images: 22,657 are adult images, and 1,391,108 are normal images. In total, there are 1,300,144 training images, 50,350 validation images, and 63,271 test images. For training DMCNet, we separated the training and validation images into two subdatasets: the classification dataset and the detection dataset. The latter only contains our collections—that is, it does not contain extra datasets. For training the global context model (i.e., classification task), there are 1,300,144 training images belonging to 1,007 categories (993 classes are defined by ImageNet, 14 classes are defined by our collection, and each category roughly has 1,300 images), and 50,350 validation images. For training the local context model (i.e., detection task), 17,637 images are used for training and 700 images are used for validation. In the detection dataset, each image contains multiple objects, and all objects are divided into 15 classes (14 classes are defined by our collections, and one for the background class). All experiments use the same test images, 7,020 are adult images, and 56,251 are normal images.

Furthermore, we defined two evaluation standards to assess performance: a two-level evaluation pattern (denoted as L2) and a three-level evaluation pattern (denoted as L3). To make dataset suitable for these evaluation standards, we combined all normal images into one category denoted as S00 and called *normal*; nine remarkable eroticism and nudity categories combined into one category denoted as S01 and called *adult*; and three marginal categories (e.g., underwear, swimwear, beautyleg) combined and defined as unsuitable for children, denoted as S02. We use categories S00, S01, and S02 to achieve the L3 evaluation and combine S00 and S01+S02 to finish the L2 evaluation. We denoted S0102 and S01+S02 to identify all sensitive images. In this work, all experiments are performed using this setting to train and test. The other three datasets are only assessed with the L2 evaluation.

The NPDI [de Avila et al. 2013] adult database contains nearly 80 hours of 400 adult and 400 nonadult videos. It has been collected by the NPDI group, Federal University of Minas Gerais (UFMG), Brazil. For the adult class, the database consists of several adult genres and depicts actors of various ethnicities, including multiethnic ones. The nonadult class includes two subcategories: 200 videos chosen at random (called *easy*) and 200 videos selected from textual search queries like "beach," "wrestling," and "swimming" that contain body skin but not porn (called *difficult*), which would be particularly challenging for the detector. In total, 16,727 key frames are extracted from all the videos. We use these key frames to estimate whether a video is adult or not. Different from the standard cross-validation protocol [de Avila et al. 2013], we divided the whole dataset into a validation set and a test set. There are 200 videos (100 adult and 100 nonadult) used for adjusting the hyperparameter, and there are 600 videos used for testing (300 adult and 300 nonadult). It is important to note that different from the original AGNet [Moustafa 2015], the result reported in this article over the reimplementation AGNet does not train or fine tune on the NPDI dataset; it trains on our Sensitive dataset by default.

The Dynamic Magnificent Colorful Video (DMCV) database is captured by a popular video software and contains 99 adult videos and 100 normal videos. The challenge is the large number of magnificent filters and special effects employed in shooting. The huge diversity of cases in both adult and nonadult videos makes this task very challenging. In this dataset, 40 videos are used for adjusting the hhyperparameter and 159 videos are used for testing. Adult videos and normal videos are balanced. We divided each video into 20 shots, and a key frame had been extracted from each video shot. Similar to the NPDI dataset, no sample was used to train.

The Sensitive Poster Dataset (SPD) is a small but difficult database that contains 1,074 adult images and 8,926 normal images. In total, there are 2,000 validation images


Fig. 4. Examples of four datasets. (a) *Sensitive*: The first three images are normal, the middle three images are adult, and the last three images belong to the unsuitable for children category. (b) *NPDI*: The "easy" and "difficult" nonadult cases are shown in the first and second columns, and the third image is an adult image. (c) *DMCV*: The first position is a normal image, and the other two are adult images; all three images used special filter to shoot. (d) *SPD*: All images are adult images; one is a cartoon-like sketch, and the other two are chaotic "Poster."

and 8,000 testing images. Different from other datasets, SPD includes many extremely complex images (called *Poster*), which have crowded screens and objects that usually are small.

To understand the details of these datasets better, some examples of selected images or key frames are shown in Figure 4. We can clearly see that each dataset has its own characteristics.

*Evaluation metrics*. We evaluated performance on four metrics: recall, precision, F1-score, and accuracy:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}}, \tag{11}$$

where TP, FN, TN, and FP are the true positive, false negative, true negative, and false positive, respectively.

### 4.3. Evaluation of the Fine-to-Coarse Strategy

According to the setting of Strategy 3, our multicontext method is implemented based on the global context model. Thus, we investigate the fine-to-coarse strategy depicted in Section 3.1 on global context. We separate the global context branch in our framework and set the output dimension equal to 2 or 3 for diverse datasets as the baseline model. We refer to them as the single context model since they are trained and tested under the same experimental setting. In this section, we designed two groups of different experiments to analyze our proposed fine-to-coarse strategy. On the one hand, we evaluate the fine-to-coarse strategy on the Sensitive dataset with three famous deep CNN models. The main purpose is to measure the adaptation of our proposed strategy on different CNN models. On the other hand, to verify that the strategy is robust to various datasets, we test the strategy on all four datasets. We are using the L2 evaluation pattern on the NPDI, DMCV, and SPD datasets, whereas the L2 and L3 evaluation patterns are both applied on the Sensitive dataset. For baseline training, we separate the datasets into two (normal and adult) or three (normal, adult, and

Table II. Global Context Modeling with (Ours) or without (Baseline) Fine-to-Coarse Strategy Performance (F1-Score) on Our Sensitive Dataset Using Contemporary Deep Networks, Including AlexNet [Krizhevsky et al. 2012], VGG16 [Simonyan and Zisserman 2014], and GoogLeNet [Szegedy et al. 2015]

| | | S00 | S01 | S02 | S0102 | Time (ms) |
|---|---|---|---|---|---|---|
| Baseline | AlexNet | 99.0 | 73.2 | 57.9 | 91.8 | 32 |
| | VGG16 | 99.3 | 80.0 | 72.9 | 93.9 | 145.4 |
| | GoogLeNet | 98.6 | 67.0 | 72.8 | 92.3 | 101.2 |
| Ours | AlexNet | 99.3 | 79.0 | 74.5 | 94.1 | 47 |
| | VGG16 | **99.5** | **85.2** | **80.8** | **95.9** | 158.5 |
| | GoogLeNet | 99.4 | 81.7 | 77.5 | 94.8 | 117.6 |

S00, normal; S01, adult; S02, unsuitable for children.

unsuitable for children) categories and run a binary or three classification task to evaluate the deep models.

First, our framework is flexible to incorporate many contemporary deep models. For simplicity, we replace the model structure in the global context model with other contemporary model structures for evaluation. Evaluated structures include AlexNet [Krizhevsky et al. 2012], VGG16 [Simonyan and Zisserman 2014], and GoogLeNet [Szegedy et al. 2015]. The fine-to-coarse strategy is not utilized in all of the baseline models. As shown in Table II, deeper models (VGG16 and GoogLeNet) slightly outperform the relatively shallow model (AlexNet), which shows that our strategy corresponds to the disputed basic rule of the convolutional network—deeper is better. It is particularly gratifying, after introducing the fine-to-coarse strategy, that the F1-score is increased on all baseline frameworks. We can find that the performance improvements mainly occur in the categories S01 and S02, and we believe that these occur for two major reasons. The first reason is that for normal images, the target objects usually have stronger semantic information. A good feature representation with a good classifier can complete the high-level semantics identification very well. However, influenced by the posture, scale, view, action, occlusion, object integrality, and many uncertain factors, semantics information of each adult image may be ambiguous. To treat them as a unified category, and recognize from a large-scale dataset is extremely difficult. The second reason is that for some adult images, especially local close-ups, may be very close to a normal image on the vision. If there is no fine-grain classification, the intraclass distance may be larger than the interclass distance for some images. Another piece of good news is that the time overhead increase is not obvious after introducing the fine-to-coarse strategy.

Second, as shown in Figure 5, our proposed fine-to-coarse strategy works well on all datasets. After converting the fine-grain categories to coarse-grain categories, the recognition performance is improved over all categories. For the same reasons, our proposed strategy for improving the recognition performance of adult images can be more significant. Especially on the DMCV dataset, our fine-to-coarse strategy increases the F1-score by around 15%.

To better understand why our proposed fine-to-coarse strategy can improve the classification performance in the adult recognition task, we visualize the classification probability distribution of each image on the Sensitive and DMCV datasets. As Figure 6 shows, each symbol represents an image, color is determined by the ground truth category, and position is decided by the class probability relative to each category. Assuming that image $I$ belongs to category $C$, the higher the probability relative to the category $C$ is closer in distance to the center $C$. The category center is denoted as a bigger symbol. If the classification probability of an image $I$ is 1, then it will overlap with the category center. Looking at the figure, it is not difficult to see that the denser the identical symbols are, the stronger the feature discriminative is. It is obvious that

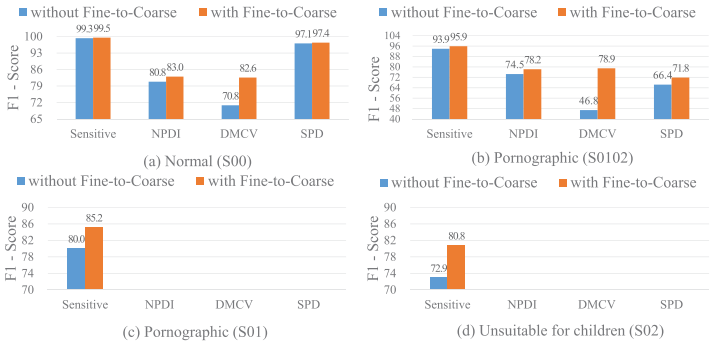


Fig. 5. F1-score on the four adult datasets for evaluating the fine-to-coarse strategy. (a-b) The results of normal (S00) and adult (S0102) images, respectively, are shown. Among them, category S0102 is the combination of categories S01 and S02 on the Sensitive dataset. (c-d) The results of adult (S01) and unsuitable for children (S02) images on the Sensitive dataset, respectively, are illustrated.

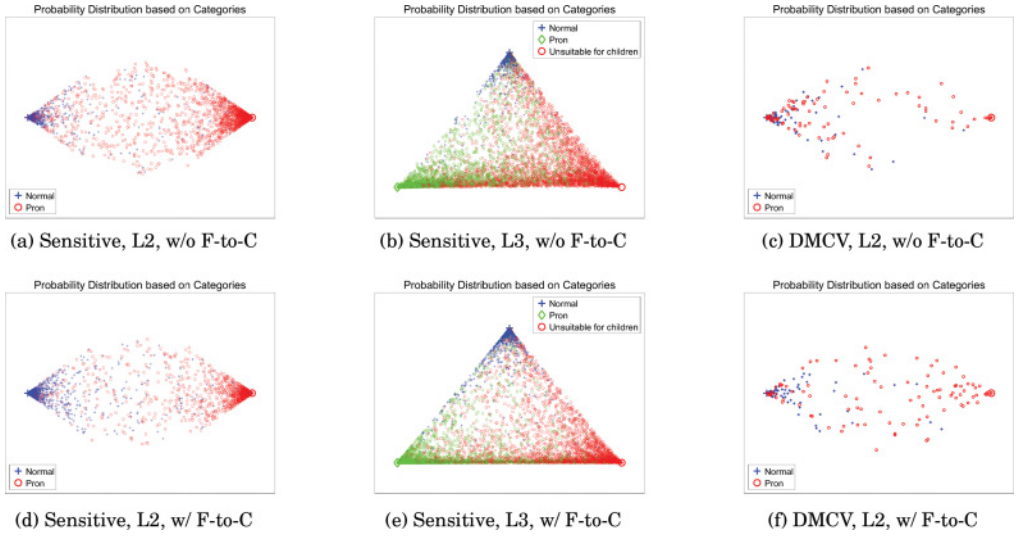


Fig. 6. Evaluation of the fine-to-coarse strategy on the Sensitive (the first and second columns) and DMCV (the third column) datasets. The probability distribution without the fine-to-coarse strategy is shown in (a) through (c), whereas the probability distribution with the fine-to-coarse strategy is presented in (d) through (f). Different colors represent different categories. The cluster center points are denoted by large symbols, whereas the images are denoted by small symbols. We first artificially specify the coordinates of the cluster center to identify a certain category, then plot the signature point of each image according to its determinable coordinates relative to the classification probability obtained by DMCNet inference. Parts (b) and (e) demonstrate the probability distribution of the L3 classification problem, and other four parts direct the probability distribution of the L2 classification problem.

after introducing the fine-to-coarse strategy, more symbols stream into the categories' center, especially in the adult images. In addition, we can see that the samples at the edge of the categories area is significantly reduced. For an example (please enlarge the figure for clear viewing if possible), compare Figure 6(a) to Figure 6(d), which demonstrate an L2 classification problem on the Sensitive dataset. We may discover that the density and quantity of red circles in Figure 6(d) gather around the right center significantly. This means that with the fine-to-coarse strategy, the adult images can be better allocated to the correct category. In Figure 6(e), as opposed to the Figure 6(b),

Table III. F1-Score on Four Datasets for Evaluating the Multicontext Modeling

| | | F1-Score | | | |
|---|---|---|---|---|---|
| | | S00 | S01 | S02 | S0102 |
| Sensitive | Global content | **99.5** | 85.2 | 80.8 | **95.9** |
| | Local content ($t$=0.99) | 94.6 | 19.8 | 9.8 | 15.2 |
| | Cross content ($t$=0.99) | 95.5 | 20.0 | 9.8 | 15.5 |
| | Multicontent ($w$=0.47) | **99.5** | **85.3** | **80.9** | **95.9** |
| NPDI | Global content | 83.0 | — | — | 78.2 |
| | Local content ($t$=0.3) | 77.7 | — | — | 79.2 |
| | Cross content ($t$=0.3) | 78.6 | — | — | 80.1 |
| | Multicontent ($w$=0.32) | **85.2** | — | — | **85.3** |
| DMCV | Global content | **82.6** | — | — | 78.9 |
| | Local content ($t$=0.2) | 66.0 | — | — | 70.4 |
| | Cross content ($t$=0.2) | 66.6 | — | — | 71.1 |
| | Multicontent ($w$=0.27) | 82.3 | — | — | **80.4** |
| SPD | Global content | 97.4 | — | — | 71.8 |
| | Local content ($t$=0.6) | 96.6 | — | — | 63.1 |
| | Cross content ($t$=0.6) | **97.5** | — | — | 63.7 |
| | Multicontent ($w$=0.48) | **97.5** | — | — | **74.7** |

S00, normal; S01, adult; S02, unsuitable for children; S0102, adult (both S01 and S02).

the green diamonds focus on the lower left corner, and the red circles focus on the lower right corner. In addition, one of the most obvious examples is shown in Figure 6(c) and Figure 6(f). We can see a large number of red circles around the left blue cross center in Figure 6(c), and we consider this a very poor classification result. By comparing the quantitative data, illustrated in Figure 5(b), we find that the F1-score of the DMCV dataset without thefine-to-coarse strategy is only 46.8%, which proves our conclusion. By comparison, under the fine-to-coarse strategy, the F1-score rises to 78.9%. These results can also be clearly seen in Figure 6(f), where a lot of red circles run to the right red circle center.

In conclusion, increasing the category space can significantly enhance discriminative capability and further improve adult image recognition performance. As Table II shows, the VGG16 model always has the best F1-score, so we use this model to achieve the following works.

### 4.4. Evaluation of Multicontext Modeling

In this section, we first compare the F1-score on four datasets using different context modeling methods. Then the effects of hyperparameters $t$ and $w$ are deeply analyzed. The fine-to-coarse strategy is used in all experiments.

As shown in Table III, most of the time our proposed multicontext model outperforms the single-context model on all four datasets. Unfortunately, multicontext modeling did not give us much pleasure over normal images. Even below baseline on the DMCV dataset (82.6 vs. 82.3 on DMCV). Nonetheless, for the adult images, the multicontext strategy performs well on the NPDI, DMCV, and SPD datasets. Unfortunately, our strategy is not outstanding on the Sensitive dataset; however, it consistently outperforms the baseline models. To the best of our knowledge, the preceding phenomenon results from three points. First, the deep convolutional network has very robust ability on feature learning. If the sample is obviously semantic, the discriminativeness of the feature will be very strong, especially for a single-object image or if the image has a main object, which is quite remarkable. Normal images are typical examples. Global

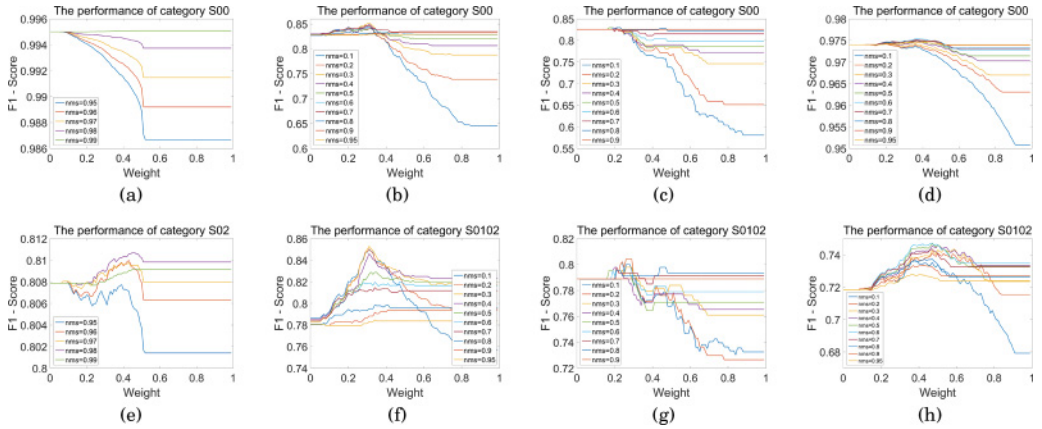


Fig. 7. Evaluation of multicontext modeling with different fusion weight $w$ and different boxes selected threshold $nms(t)$ on Sensitive, NPDI, DMCV, and SPD datasets (from left to right). The curves of the first row (a–d) correspond to category S00 (normal), and the results of the second row (e–h) correspond to category S0102 (adult). The only difference is that part (e) shows the result of category S02 instead of S0102 to be more clear.

context modeling has performed very well; therefore, adding local context information cannot effectively improve recognition performance. Second, adult images usually contain a large number of local close-ups, and the local context contributes to finding this information, which can make up for the deficits of global context. Hence, multicontext works well on the recognition of pornography. Third, restricted by the accuracy of the detection model, excessively depending on the local context information will damage overall performance, particularly when the global context is very powerful but with very little useful regional information. As with our Sensitive dataset, it contains a great quantity of various images, but the images with adult small objects are few. This is why the *nms* threshold $t$ is much bigger than other datasets on the Sensitive dataset, but the improvement is still very small. Some examples of local context modeling to help improve recognition capability are shown in Figure 3 (columns 3 and 4), and it is clearly shown that the multicontext model refines the erroneous predictions by the global context model since it combines both global context and local context.

In addition, the choice of hyperparameters is extremely important for fusion performance. We conduct a detailed analysis about hyperparameters $w$ and $t$ in this paragraph. To be fair and reliable, all parameters selected are achieved from the validation set, and the final results are reported on the testing set. Actually, the experimental results that run on validation and testing set are going to be very close, due to the dataset is divided randomly. As shown in Equation (6), $w$ is an exclusive gate parameter; if $w = 1$, $\mathcal{F}_{DMCN}$ is decided by the local context—that is, $\mathcal{F}_{DMCN} = \tilde{\mathcal{F}}_{local}$. When $w = 0$, $\mathcal{F}_{DMCN}$ is only decided by the global context—that is, $\mathcal{F}_{DMCN} = \tilde{\mathcal{F}}_{global}$. As shown in Figure 7, we can conclude the following. First, the F1-score extremum distribution usually lies between 0.2 and 0.5. From this, we can certainly infer that the global context has a crucial effect on overall performance. As well, adding the local context information into the global context will make recognition performance improve, especially for adult images. For images with many small objects or images with significant sexual organ close-ups, the effect of local context is more apparent. For example, weight $w$ on SPD is much bigger than in other datasets because it contains a lot of Poster images. Second, the choice of parameter $w$ (*nms*) is also crucial. Because the dependability of the detection network is inferior to the classification network, parameter $w$ becomes very sensitive. A smaller $w$ will increase the recall, but precision drops rapidly. In

Table IV. Comparison with Two State-of-the-Art Methods (AGNet [Moustafa 2015] and Incremental Learning [Wang et al. 2015]) on Our Global Context Baseline and the MultiContext Method on Our Sensitive Datasets

| Methods | Recall | | | Precision | | | F1-Score | | | Accuracy | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S01 | S02 | S0102 | S01 | S02 | S0102 | S01 | S02 | S0102 | L2 | L3 |
| AGNet | 80.2 | 56.4 | 86.0 | 55.9 | 92.9 | 98.8 | 65.9 | 70.2 | 92.0 | 98.3 | 96.3 |
| Incremental Learning | 58.7 | 26.5 | 68.5 | 14.3 | 30.0 | 35.8 | 23.0 | 28.1 | 47.0 | 75.1 | 72.3 |
| Baseline-VGG16 | 83.4 | 62.6 | 89.2 | 76.9 | 87.4 | **99.1** | 80.0 | 72.9 | 93.9 | 98.7 | 96.9 |
| DMCNet | **88.0** | **73.9** | **93.4** | **82.6** | **89.3** | 98.7 | **85.3** | **80.9** | **95.9** | **99.1** | **97.8** |

addition, a bigger $w$ can improve the precision, but there are very few images recalled, and most of these images are identified correctly via the global context. Note that the growth of the F1-score curve is limited on the Sensitive dataset and all normal images. Hence, how to weight the proportion of local context (by $w$) and the number of detections boxed (by $t$) becomes very important for the fusion feature. We have done many experiments on the validation set of four datasets to select appropriate hyperparameters. Third, some curves are insensitive to the weight $w$. For category S00, the expected performance improvement is mainly from those misrecognized porn images adjusted by the global context model ane re-recalled by the local context model. However, some normal images are incorrectly assigned to the pornography set at the same moment. In other words, the precision can be improved but the recall reduces. For category S02 of the Sensitive dataset, illustrated in Figure 7(e), the curve is also not notable. The Sensitive dataset is a very large dataset, and the images are usually of high resolution and are object centric (as illustrated in Figure 4(a)). This setting makes the global context (with the fine-to-coarse strategy) have strong discrimination. The improvement space left for the multicontext decision is very small. In fact, there are only dozens of pornographic images recalled correctly by the local context model, which are easy to disseminate in the data ocean. Therefore, the curve of S02 is tends to be stable in the Sensitive dataset. Fourth, according to the results, the hyperparameters are very sensitive for the images, which have various distribution in categories. Adjustment on the validation set is an effective way. However, a better approach is to improve the performance of detection network. A more robust local context model can make the parameters more stable and thus reduce the dependence of the prior knowledge of the dataset.

### 4.5. Comparison with the State of the Art

In Tables IV and V, we compare our approach to two of the latest state-of-the-art methods, including a deep learning method (AGNet [Moustafa 2015]) and a traditional method (Incremental Learning [Wang et al. 2015]). There are some differences between our reimplementation and the original version. For AGNet, we only train one network instead of five networks because our proposed approach does not use the five-fold method. For Incremental Learning, the covering center that we selected is only produced by initialization. We found that tuning coverage by the false samples is invalid and harmful in large-scale dataset such as our Sensitive dataset. The fine-to-coarse strategy and multicontext fusion strategy have been applied in the final testing phase. Results show that the deep learning method shows powerful performance. Apart from this, combined with our specially designed strategies for the adult recognition task, our proposed DMCNet significantly outperforms other state-of-the-art adult recognition algorithms. Our approach obtains a higher F1-score and accuracy on all four datasets. Except for the Sensitive dataset, the other datasets have not been used for training. This further validates the generalization ability of our proposed method (which is why the average accuracy reported in this article of 79% on the NPDI benchmark dataset by AGNet is much lower than the reported 94% of Moustafa [2015]). One example is the

Table V. Evaluation of General Performance of Our Proposed Methods and Two State-of-the-Art Methods (Wang et al. [2015] and Moustafa [2015]) on Three Other Datasets

| Datasets | Methods | Recall | | Precision | | F1-Score | | Accuracy |
|---|---|---|---|---|---|---|---|---|
| | | Normal | Porn | Normal | Porn | Normal | Porn | |
| NPDI | AGNet | 88.3 | 69.8 | 74.7 | **85.6** | 80.8 | 76.9 | 79.0 |
| | Incremental Learning | 76.2 | 51.6 | 71.8 | 57.2 | 73.9 | 54.3 | 63.9 |
| | Baseline-VGG16 | **92.2** | 64.0 | 71.9 | 89.2 | 80.8 | 74.5 | 78.1 |
| | DMCNet | 85.0 | **85.5** | **85.4** | 85.1 | **85.2** | **85.3** | **85.3** |
| DMCV | AGNet | 87.0 | 65.7 | 71.9 | 83.3 | 78.7 | 73.5 | 76.4 |
| | Incremental Learning | 23.0 | 86.9 | 63.9 | 52.8 | 33.8 | 65.7 | 54.8 |
| | Baseline-VGG16 | **91.0** | 33.3 | 58.0 | 78.6 | 70.8 | 46.8 | 62.3 |
| | DMCNet | 86.0 | **76.8** | **78.9** | **84.4** | **82.3** | **80.4** | **81.4** |
| SPD | AGNet | 99.8 | 49.2 | 94.4 | 96.9 | 97.0 | 65.8 | 94.4 |
| | Incremental Learning | 78.9 | 40.3 | 91.8 | 18.4 | 84.9 | 25.2 | 74.8 |
| | Baseline-VGG16 | **99.9** | 49.9 | 94.4 | **99.3** | 97.1 | 66.4 | 94.6 |
| | DMCNet | 99.3 | **63.1** | **95.8** | 91.7 | **97.5** | **74.7** | **95.4** |

exception: AGNet has the top precision on the NPDI dataset, which is slightly higher than our method (85.6 vs. 85.1). Nevertheless, the recall is much lower than ours (69.8 vs. 85.5).

## 5. CONCLUSION AND FUTURE WORK

In this article, we propose a multicontext deep learning framework for adult image and video recognition. First, we introduce a task-specific fine-to-coarse strategy to discover more detailed features and semantics to improve the recognition capability of the model. Second, a unique multicontext scheme is investigated to learn the robust fusion feature via deep convolutional networks. Global context, local context, and cross context are utilized and incorporated into a unified multicontext deep learning framework for feature learning. We take a utility five-step algorithm to learn shared features. Moreover, recently proposed contemporary deep models in ILSVRC are tested, and their effectiveness in adult recognition is investigated. Experiments validate the effectiveness of each component in our framework and show that our approach significantly and consistently outperforms all state-of-the-art methods. It is worth noting that this work focuses on a specific task to solve the problem of adult image and video recognition, but a promising future direction is to explore our proposed methods around fine-grain recognition and specific target recognition tasks.

## ACKNOWLEDGMENTS

The authors appreciate the valuable suggestions from the anonymous reviewers and the editors. Thanks to the NVIDIA Hardware Grant Project.

## REFERENCES

Anton Akusok, Yoan Miche, Juha Karhunen, Kaj-Mikael Bjork, Rui Nian, and Amaury Lendasse. 2015. Arbitrary category classification of Websites based on image content. *IEEE Computational Intelligence Magazine* 10, 2, 30–41.

Will Archer Arentz and Bjørn Olstad. 2004. Classifying offensive sites based on image content. *Computer Vision and Image Understanding* 94, 1–3, 295–310.

Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge J. Belongie. 2010. Visual recognition with humans in the loop. In *Proceedings of the 11th European Conference on Computer Vision (ECCV'10)*. 438–451.

Christoph Bregler, Margaret M. Fleck, and David A. Forsyth. 1996. Finding naked people. In *Proceedings of the 4th European Conference on Computer Vision (ECCV'96)*. 593–602.

Carlos Caetano, Sandra Eliza Fontes de Avila, Silvio Jamil Ferzoli Guimarães, and Arnaldo de Albuquerque Araújo. 2014. Pornography detection using BossaNova video descriptor. In *Proceedings of the 22nd European Signal Processing Conference*. 1681–1685.

ChinaPRC. 2015. Criminal Law of the People's Republic of China. Retrieved April 1, 2017, from http://www.lawtime.cn/faguizt/23.html.

Dan C. Ciresan, Ueli Meier, and Jürgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 3642–3649.

Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-FCN: Object detection via region-based fully convolutional networks. arXiv:1605.06409.

Sandra Eliza Fontes de Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de Albuquerque Araújo. 2013. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding* 117, 5, 453–465.

Jia Deng, Alexander C. Berg, Kai Li, and Fei-Fei Li. 2010. What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th European Conference on Computer Vision (ECCV'10)*. 71–84.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'09)*. 248–255.

Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. 2008. Bag-of-visual-words models for adult image classification and filtering. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08)*. 1–4.

Kaikun Dong, Li Guo, and Quansheng Fu. 2014. An adult image detection algorithm based on bag-of-visual-words and text information. In *Proceedings of the 2014 10th International Conference on Natural Computation (ICNC'14)*. IEEE, Los Alamitos, CA, 556–560.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision* 88, 2, 303–338.

C. Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Robert Fergus, Hector Bernal, Yair Weiss, and Antonio Torralba. 2010. Semantic label sharing for learning with many categories. In *Proceedings of the 11th European Conference on Computer Vision (ECCV'10)*. 762–775.

Nadia Figueroa, Haiwei Dong, and Abdulmotaleb El Saddik. 2015. A combined approach toward consistent reconstructions of indoor spaces based on 6D RGB-D odometry and KinectFusion. *ACM Transactions on Intelligent Systems and Technology* 6, 2, Article No. 14.

Margaret M. Fleck and David A. Forsyth. 1999. Automatic detection of human nudes. *International Journal of Computer Vision* 32, 1, 63–77.

David A. Forsyth and Margaret M. Fleck. 1997. Body plans. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR'97)*. 678–683.

David A. Forsyth, John A. Haddon, and Sergey Ioffe. 1999. Finding objects by grouping primitives. In *Proceedings of the 1999 Conference on Shape, Contour, and Grouping in Computer Vision*. 302–318.

David A. Forsyth and Sergey Ioffe. 1999. Finding people by sampling. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV'99)*. 1092–1097.

David A. Forsyth and Sergey Ioffe. 2001. Probabilistic methods for finding people. *International Journal of Computer Vision* 43, 1, 45–68.

Zhangjie Fu, Xingming Sun, Sai Ji, and Guowu Xie. 2016. Towards efficient content-aware search over encrypted outsourced data in cloud. In *Proceedings of the 2016 IEEE Conference on Computer Communications (INFOCOM'16)*. 1–9.

Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 1440–1448.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. 580–587.

Gregory Griffin and Pietro Perona. 2008. Learning and using taxonomies for fast visual categorization. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 9, 1904–1916.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Identity mappings in deep residual networks. arXiv:1603.05027.

Sergey Ioffe and David A. Forsyth. 1998. Learning to find pictures of people. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. 782–788.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 2014 ACM International Conference on Multimedia (MM'14)*. 675–678.

Feng Jiao, Wen Gao, Lijuan Duan, and Guoqin Cui. 2001. Detecting adult image using multiple features. In *Proceedings of the 2001 International Conferences on Info-Tech and Info-Net, Vol. 3*. 378–383.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. 1725–1732.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 2012 Conference on Advances in Neural Information Processing Systems*. 1106–1114.

LegalDictionary. 2015. Pornography. Retrieved April 1, 2017, from http://legal-dictionary.thefreedictionary.com/pornography.

Jian Li, Xiaolong Li, Bin Yang, and Xingming Sun. 2015. Segmentation-based image copy-move forgery detection scheme. *IEEE Transactions on Information Forensics and Security* 10, 3, 507–518.

Xiaoyan Li, Tongliang Liu, Jiankang Deng, and Dacheng Tao. 2016. Video face editing using temporal-spatial-smooth warping. *ACM Transactions on Intelligent Systems and Technology* 7, 3, 32.

K. M. Liang, S. D. Scott, and M. Waqas. 2004. Detecting pornographic images. In *Proceedings of the 2004 Asian Conference on Computer Vision.*

Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen. 2015b. Deep learning of binary hash codes for fast image retrieval. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops'15)*. 27–35.

Kevin Lin, Huei Fang Yang, Jen Hao Hsiao, and Chu Song Chen. 2015c. Deep learning of binary hash codes for fast image retrieval. In *Proceeeings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops'15)*. 27–35.

Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. arXiv:1312.4400.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015a. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*. 1449–1457.

Si Liu, Jiashi Feng, Csaba Domokos, and Hui Xu. 2014. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia* 16, 1, 253–265.

Si Liu, Xiaodan Liang, Luoqi Liu, Xiaohui Shen, Jianchao Yang, Changsheng Xu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. 2015. Matching-CNN meets KNN: Quasi-parametric human parsing. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1419–1427.

Si Liu, Tam V. Nguyen, Jiashi Feng, Meng Wang, and Shuicheng Yan. 2012a. Hi, magic closet, tell me what to wear! In *Proceedings of the 2012 Conference on ACM Multimedia*. 1333–1334.

Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. 2016. Makeup like a superstar: Deep localized makeup transfer network. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*. 2568–2575.

Si Liu, Zheng Song, Guangcan Liu, and Changsheng Xu. 2012b. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 3330–3337.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 3431–3440.

Mohamed Moustafa. 2015. Applying deep learning to classify pornographic images and videos. arXiv:1551.08899.

Zhaoqing Pan, Jianjun Lei, Yun Zhang, and Xingming Sun. 2016. Fast motion estimation based on content property for low-complexity H.265/HEVC encoder. *IEEE Transactions on Broadcasting* 62, 1–10.

Marco Pedersoli, Andrea Vedaldi, Jordi Gonzàlez, and F. Xavier Roca. 2015. A coarse-to-fine approach for fast deformable object detection. *Pattern Recognition* 48, 5, 1844–1853.

James M. Rehg and Michael J. Jones. 2002. Statistical color models with application to skin detection. *International Journal of Computer Vision* 46, 1, 81–96.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 2015 Conference on Advances in Neural Information Processing Systems*. 91–99.

Henry A. Rowley, Yushi Jing, and Shumeet Baluja. 2006. Large scale image-based adult-content filtering. In *Proceedings of the 1st International Conference on Computer Vision Theory and Applications (VISAPP'06)*. 290–296.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3, 211–252.

Napa Sae-Bae, Xiaoxi Sun, Husrev T. Sencar, and Nasir D. Memon. 2014. Towards automatic detection of child pornography. In *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP'14)*. IEEE, Los Alamitos, CA, 5332–5336.

Jerome H. Saltzer, David P. Reed, and David D. Clark. 1981. End-to-end arguments in system design. In *Proceedings of the 2nd International Conference on Distributed Computing Systems*. 509–512.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Rupesh Kumar Srivastava, Klaus Greff, and Jrgen Schmidhuber. 2015. Highway networks. arXiv:1505.00387.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1–9.

Ao Tang, Ke Lu, Yufei Wang, Jie Huang, and Houqiang Li. 2015. A real-time hand posture recognition system using deep neural networks. *ACM Transactions on Intelligent Systems and Technology* 6, 2, 21.

Antonio Torralba, Robert Fergus, and William T. Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 11, 1958–1970.

Chao Wang, Jing Zhang, Li Zhuo, and Xin Liu. 2015. Incremental learning for compressed pornographic image recognition. In *Proceedings of the 2015 IEEE International Conference on Multimedia Big Data (BigMM'15)*. 176–179.

James Ze Wang, Jia Li, Gio Wiederhold, and Oscar Firschein. 1998. System for screening objectionable images. *Computer Communications* 21, 15, 1355–1360.

Zhihua Xia, Xinhui Wang, Liangao Zhang, Zhan Qin, Xingming Sun, and Kui Ren. 2016. A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing. *IEEE Transactions on Information Forensics and Security* 11, 11, 2594–2608.

Cong Yang, Oliver Tiebe, Kimiaki Shirahama, and Marcin Grzegorzek. 2016. Object matching with hierarchical skeletons. *Pattern Recognition* 55, 183–197.

Haiming Yin, Xiaodong Xu, and Lihua Ye. 2011. Big skin regions detection for adult image identification. In *Proceedings of the 2011 Workshop on Digital Media and Digital Content Management (DMDCM'11)*. IEEE, Los Alamitos, CA, 242–247.

Jun Yu, Yong Rui, Yuan Yan Tang, and Dacheng Tao. 2014. High-order distance-based multiview stochastic learning in image classification. *IEEE Transactions on Cybernetics* 44, 12, 2431–2442.

Jun Yu, Xiaokang Yang, Fei Gao, and Dacheng Tao. 2016a. Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Transactions on Cybernetics* PP, 99, 1–11.

Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. 2016b. iPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security* 12, 5, 1005–1016.

Jung-Jae Yu and Seung-Wan Han. 2014. Skin detection for adult image identification. In *Proceedings of the 16th International Conference on Advanced Communication Technology*. IEEE, Los Alamitos, CA, 645–648.

Fan Zhang, Bo Du, and Liangpei Zhang. 2015. Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing* 53, 4, 2175–2184.

Fan Zhang, Bo Du, and Liangpei Zhang. 2016a. Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing* 54, 3, 1793–1802.

Fan Zhang, Bo Du, Liangpei Zhang, and Miaozhong Xu. 2016b. Weakly supervised learning based on coupled convolutional neural networks for aircraft detection. *IEEE Transactions on Geoscience and Remote Sensing* 54, 9, 5553–5563.

Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2015. Saliency detection by multi-context deep learning. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1265–1274.

Zhicheng Zhao and Anni Cai. 2010. Combining multiple SVM classifiers for adult image recognition. In *Proceedings of the 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content*. IEEE, Los Alamitos, CA, 149–153.

Huicheng Zheng, Hongmei Liu, and Mohamed Daoudi. 2004. Blocking objectionable images: Adult images and harmful symbols. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME'04)*. 1223–1226.

Yuhui Zheng, Byeungwoo Jeon, Danhua Xu, Q. M. Jonathan Wu, and Hui Zhang. 2015. Image segmentation by generalized hierarchical fuzzy C-means algorithm. *Journal of Intelligent and Fuzzy Systems* 28, 2, 961–973.

Zhili Zhou, Yunlong Wang, Q. M. Jonathan Wu, Ching-Nung Yang, and Xingming Sun. 2017. Effective and efficient global context verification for image copy detection. *IEEE Transactions on Information Forensics and Security* 12, 1, 48–63.

Received October 2016; revised January 2017; accepted February 2017