


Covert photo classification by deep convolutional neural networks

Haiqiang Zuo^{1,2}  · Haitao Lang³ · Erik Blasch⁴ · Haibin Ling²

Received: 16 September 2016 / Revised: 11 June 2017 / Accepted: 24 June 2017 / Published online: 17 July 2017
© Springer-Verlag GmbH Germany 2017

Abstract The increasing presence of image/video capture devices such as camera phones and surveillance cameras has become a ubiquitous element of providing convenience and improving security in modern life. On the other hand, the pervasiveness of such image/video capture devices raises growing privacy concerns. In this paper, we concentrate on a new visual privacy protection problem—covert photo classification. Covert photography means that the subject being photographed is purposely made unaware that he or she is photographed. A covert photo often contains information that is inherently sensitive and private to a person. If such photos are released on the public without approval, it may lead to serious negative consequences. We explore deep convolutional neural networks (DCNNs) to discover intricate structures of covert photos and automatically learn the representations for covert photo classification. Experimental results demonstrate that DCNN-based architectures which are fully end-to-end trained reach beyond previous experience-dependent hand-

engineered feature methods in covert photo classification. The fusion of three DCNN-based architectures (AlexNet, VGGs, and GoogleNet) shows enhanced performance over individual networks on the Covert-2500 dataset and achieves an average classification rate (1-EER) of 0.925 which significantly outperforms the result (1-EER) of 0.8940 of hand-engineered feature methods.

Keywords Privacy protection · Covert photography · Image classification · Visual attribute · Deep convolutional neural networks

1 Introduction

Modern technology has brought a revolution in photograph acquisition and publication. Image/video capture devices, such as digital cameras and camera phones, have become cheaper, ubiquitous, and easier for owners of these devices to record images anywhere and keep a daily visual journal of their lives. Meanwhile, to prevent crime and insure public security, surveillance cameras have proliferated and can be found in both public and private spaces watching our movements around the clock. If connected to Internet, it is also effortless to share captured photographs within seconds by posting them to social-networking sites (e.g., Facebook, Twitter, LinkedIn), photo/video-sharing sites (e.g., Instagram, Flickr, YouTube), and personal web sites and blogs. According to the KPCB's (Kleiner Perkins Caufield & Byers) 2016 Internet Trends Report [1], about 2 billion photos are shared on Facebook-owned websites every day. These modern techniques provide a great convenience and security to users. Despite these many advantages, however, challenges do exist. The pervasiveness of such images/video capture devices creates a growing concern for invasions of privacy

✉ Haibin Ling
hbling@temple.edu

Haiqiang Zuo
zhqupc@upc.edu.cn

Haitao Lang
langht@mail.buct.edu.cn

Erik Blasch
erik.blasch@rl.af.mil

¹ Department of Chemical Equipment and Control Engineering, China University of Petroleum, Qingdao 266580, China

² Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA

³ Department of Physics and Electronics, Beijing University of Chemical Technology, Beijing 100029, China

⁴ Air Force Research Lab, 525 Brooks Rd, Rome, NY 13441, USA

to unassuming entities. There is always a trade-off between availability and privacy, and therefore, it is important to look for ways that technology can be used to protect privacy.

In this paper, we concentrate on an emerging visual privacy protection problem—covert photo classification. *Covert photography* (also called secret photography or unauthorized photography) refers to the use of an image or video recording device to photograph or film a person who is unaware that they are being intentionally photographed or filmed [2]. Photos taken in this manner are called covert photographs. These photos may be collected by normal cameras, while the method of acquisition from the device or observer is concealed. Usually covert photos contain information that is inherently special or sensitive and is private to an unaware person. If these photos are shared on the Internet, it may lead to serious negative consequences [3–6].

Covert photo classification was first studied in [7]. In the work, an algorithm is proposed to fuse heterogeneous image features and visual attributes in a multiple kernel learning framework, and it typically relies on manual feature design. On the other hand, recent progress in machine learning, especially deep learning technology, has achieved remarkable successes in many tasks such as image recognition [8–10], speech recognition [11–13], natural language understanding and translation [14, 15], achieving performance that equals or even beats humans assessment [16]. These successes suggest a promising trend of using deep neural networks trained end-to-end (from the raw image pixels to class scores) to improve performance of covert photo classification.

This paper explores the capability of applying deep neural networks to covert photo classification. In particular, three DCNN-based (AlexNet [17], GoogleNet [18], and VGGs [19]) architectures are developed for covert photo classification, with the following contributions:

- DCNN-based methods are first introduced to solve the emerging covert photo classification problem. Experimental results demonstrate that the performance of all the explored DCNN-based architectures, after transferring parameters from ImageNet pre-trained models, significantly surpasses hand-engineered feature methods.
- Activation maps demonstrate what intrinsic characteristics DCNN has learned to discriminate covert photos from non-covert ones. Though different DCNN architectures show different discriminative regions, most of the dark occlusion parts due to a hidden camera are highlighted which is consistent with our human instincts.
- With the aid of auxiliary attributes, a two-stage parameter transferring method is exploited to enhance the performance of the algorithm.
- The final fusion of three DCNN-based architectures further boosts the classification performance and significantly outperforms hand-crafted feature methods.

In the rest of the paper, Sect. 2 summarizes related work. The proposed covert photo classification with DCNN algorithm is described in Sect. 3, with detailed experimentation and analysis. Section 4 explores leveraging auxiliary attributes to improve the final covert photo classification performance. Finally, the paper is concluded in Sect. 5.

2 Related work

2.1 Privacy protection in visual data

Covert photo classification is highly related to studies of privacy protection, which is an increasing concern in modern society. Almost all countries have laws to protect privacy, although the boundaries and content of what is considered private differ among cultures and individuals. Many researchers and groups have proposed various algorithms and systems to protect privacy in visual media such as images and videos [20–32].

Previous studies on visual privacy protection mainly focus on privacy information detection and privacy information hiding. Martin et al. [33] developed an algorithm that implements a specific identification filter on video sequences of a driver from naturalistic driving data to protect the identity and preserve the behavior of the driver. Nakashima et al. [34] proposed a method for intended human object detection and developed a system for obscuring human object regions in videos taken for mobile video surveillance that contain privacy sensitive information. Elhadad et al. [35] developed a high capacity hiding technique which embeds the video captured by the surveillance camera into another processed video where the private information was removed. Ross and Othman [36] explored using visual cryptography to preserve the privacy of biometric data (such as face images, fingerprint images, and iris codes) by decomposing the original image into two images that were stored in two separate database servers. The original private image can be revealed only when both images were simultaneously available and the individual component images did not reveal any identity of the original private image.

Our work is most related to that in [7], which addresses the problem of classifying covert photos and establishes a covert image dataset with 2500 covert photos and 10,000 non-covert photos. The photos were collected from varying sources, e.g., web, surveillance system, voyeurism publishing, and real covert photography on street. Each sample image in the dataset was verified rigorously and carefully, by checking its source from which the final dataset was adjusted to reduce the potential bias toward specific topics or content. Eight hand-crafted low-level image features and 13 mid-level image attributes were fused for image representation using a multiple kernel learning framework for covert photo classi-

fication. The experimental results showed that the approach achieved an average classification rate (1-EER) of 0.8940, which significantly outperforms other contemporary algorithms as well as human performance.

In contrast to the prior efforts that use hand-designed models based on user-defined features, we propose using a deep learning architecture to discover the intricate structure in the training set and automatically learn the representations for covert photo classification.

2.2 Deep convolutional neural networks for image classification

Conventional computer vision strategies were constrained in their capacity to process pixel values of an image. Building an artificial intelligence or machine learning framework required domain experts’ careful design to extract feature vectors from the raw image data and followed a classifier that maps input vectors to different categories [37]. By contrast, deep convolutional neural networks [17–19] allow a machine to be fed with raw image pixels and learns representations automatically. Therefore, DCNN depends less on prior knowledge and human effort in features design.

DCNN is comprised of multiple convolutional and sub-sampling layers optionally followed by fully connected layers and is therefore said to be deep (in contrast, classical representation will be referred to as shallow) [38]. A DCNN architecture can take full advantage of the 2D structure of an input image, its local connectivity, and shared weighting properties to help dramatically reduce the number of free parameters to estimate and at the same time improve generalization. DCNN can also be easily trained with standard back

propagation algorithm. DCNN has achieved leading performances on a variety of vision recognition task [8–10,39]. In recent ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [40] competitions, almost all the highly ranked teams used DCNN as their basic framework.

3 Covert photo classification with DCNN

3.1 Problem formulation

We develop three DCNN-based (AlexNet, GoogleNet, and VGGs) architectures for covert photo classification. All the three networks are first trained on the source dataset—ImageNet, which contains 1.2 million images with 1000 categories, and then all the pre-trained DCNN parameters of the internal layers of the network are then transferred to the target task—covert photo classification. The framework is illustrated in Fig. 1.

A DCNN architecture usually contains millions of parameters and directly learning so many parameters from only a few thousand training images is problematic [41–43]. However, in our task only a small amount of training data is available. As illustrated in [7], the process of covert dataset collection needs a rigorous verification and bias reduction for which the final dataset contains 2500 covert photos.

A common technique to resolve the limited dataset problem is *transfer learning*, which aims to transfer knowledge from related source to target domains [44]. In this paper, we directly use the pre-trained models which are shared on Caffe Model Zoo [45] to get the source parameters. These models usually need 2–3 weeks to train on ImageNet. We ini-

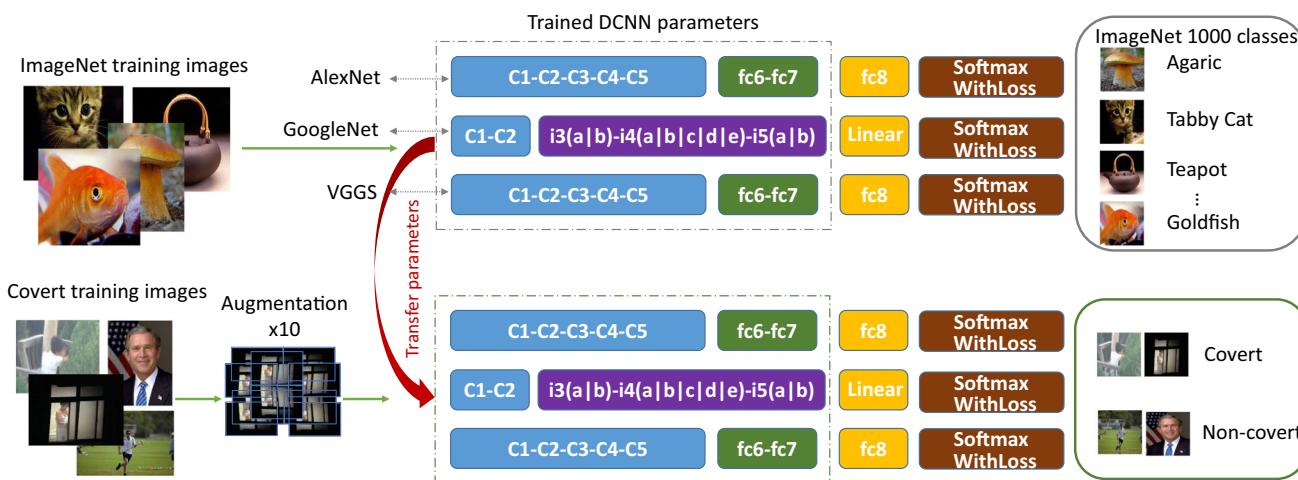


Fig. 1 Framework of covert photo classification by DCNN. Three DCNN-based (AlexNet, GoogleNet, and VGGs) architectures are first trained on the source dataset—ImageNet and then all the pre-trained DCNN parameters of the internal layers of the network are then trans-

ferred to the target task—covert photo classification. Data augmentation by extracting 10 different subcrops is employed to make up the deficiency of training set and help to reduce overfitting

tialize networks parameters by transferring from pre-trained models on ImageNet and keep the earlier layer parameters fixed (these parameters are not specific to a particular object category or dataset and usually appear like Gabor filters or color blobs) and then fine-tune higher layer parameters (these parameters are more specific to the classes contained in the training dataset) on our Covert-2500 dataset. This is based on the work of Yosinski et al. [41] which demonstrates transferring features even from distant tasks can be better than using random features. Recently, similar strategy is also adopted by Banerjee et al. [46] which transfer ImageNet pre-trained AlexNet parameters to classify medical images and Ghazi et al. [47] which transfer ImageNet pre-trained GoogleNet, AlexNet and VGGNet parameters to identify plant species.

For comparison, experiments of training the models from scratch, i.e., initializing parameters with random numbers instead of transferring them from pre-trained models are conducted.

Our problem is to distinguish covert and non-covert photos, and this is a typical two-class classification problem. Here the covert photos are denoted by the positive class, and the non-covert photos by negative class, respectively.

3.2 Experimental protocol

The classification performance is evaluated using two measures: area under an receiver operating characteristic (ROC) curve (AUC) and equal error rate (EER). The two measures are consistent with those used in [7] and derived from the ROC curve which plots the true positive rate against the false positive rate as the decision threshold varies along the score range [48, 49]. The larger the AUC, the better the ROC. The EER identifies where the false positive rate and the false negative rate are equal, where the smaller the EER, the better of the system. The EER point marked with a ‘*’ in the ROC figures locates at the intersection of ROC curve and straight line through (1, 0) and (0, 1).

All models are trained and tested with Caffe [50] on a NVIDIA GeForce GT640 2GB GPU.

3.3 Dataset

The Covert-2500 [7] dataset includes 2500 covert photos and 10,000 non-covert photos. The training and testing sets are the same with [7]. The training set contains 2000 covert photos and 8000 non-covert photos, and the testing set contains 500 covert photos and 2000 non-covert photos.

Each input image is preprocessed by resizing to 256×256 and subtracting the per-pixel mean across all training images. The system employs data augmentation which consists of generating image translations and horizontal reflections. The method extracts 10 different subcrops (4 corners, center and their horizontal flips) from the resized 256×256 images.

The subcrops are of size 227×227 (AlexNet) or 224×224 (GoogleNet and VGGs), and the networks are then trained on these extracted subcrops. This increases the size of our training set by a factor of 10.

3.4 Covert photo classification

3.4.1 AlexNet-based covert photo classification architecture

AlexNet was the first popularized DCNN architecture in Computer Vision, proposed by Krizhevsky et al. [17]. It was the winner of ImageNet ILSVRC 2012, and its performance significantly outperformed the second runner-up. In this paper, we use a Caffe [50] version of AlexNet, which is slightly different from original one, where pooling is done before normalization. Figure 2a describes the architecture of AlexNet-based covert photo classification. A 227×227 crop of an image (with 3 RGB color channels) is taken as the input. In the first layer, it is convolved with 96 different filters, each is of size 11×11 , using a larger stride of 4 pixels, which enable fast processing. The resulting 96 feature maps of size 55×55 are firstly passed through a rectified linear unit (ReLU [51]) and then are subsampled to 27×27 with 3×3 max-pooling (using stride 2) operation and finally normalized by local input regions. Similar operations are repeated in layers 2, 3, 4, and 5. The last three layers (fc6, fc7, and fc8) are fully connected, taking all neurons in the previous layer as inputs and connecting them to every single neuron available. The fully connected fc6 and fc7 layers have 4096 neurons each, and a drop-out [52] probability of 0.5 is followed to avoid overfitting. The number of neurons of the last fully connected layer (fc8) is equal to the number of classes, i.e., 1000 for ImageNet and two for covert photo classification. The last fully connected layer is followed by a softmax with loss layer which represents the class scores.

We train the AlexNet-based model with stochastic gradient descent with momentum. The batch size is set to 50, and the momentum is fixed to 0.9, and the multiplicative weight decay is set to 5×10^{-4} per iteration. The learning rate starts at 0.001 and anneals over the course of training by dropping by a factor of 10 when the validation error rate stops decreasing with the current learning rate. In our experiments, best performance of AUC:97.29% and EER:8.65% is reached after 40 epochs when transferring parameters from the ImageNet pre-trained model. By contrast, best performance of AUC:93.09% and EER:14.04% is observed after 68 epochs when training without transfer parameters.

3.4.2 GoogleNet-based covert photo classification architecture

GoogleNet [18] was the winner of the ILSVRC 2014. The network is 22 layers deep when counting only lay-

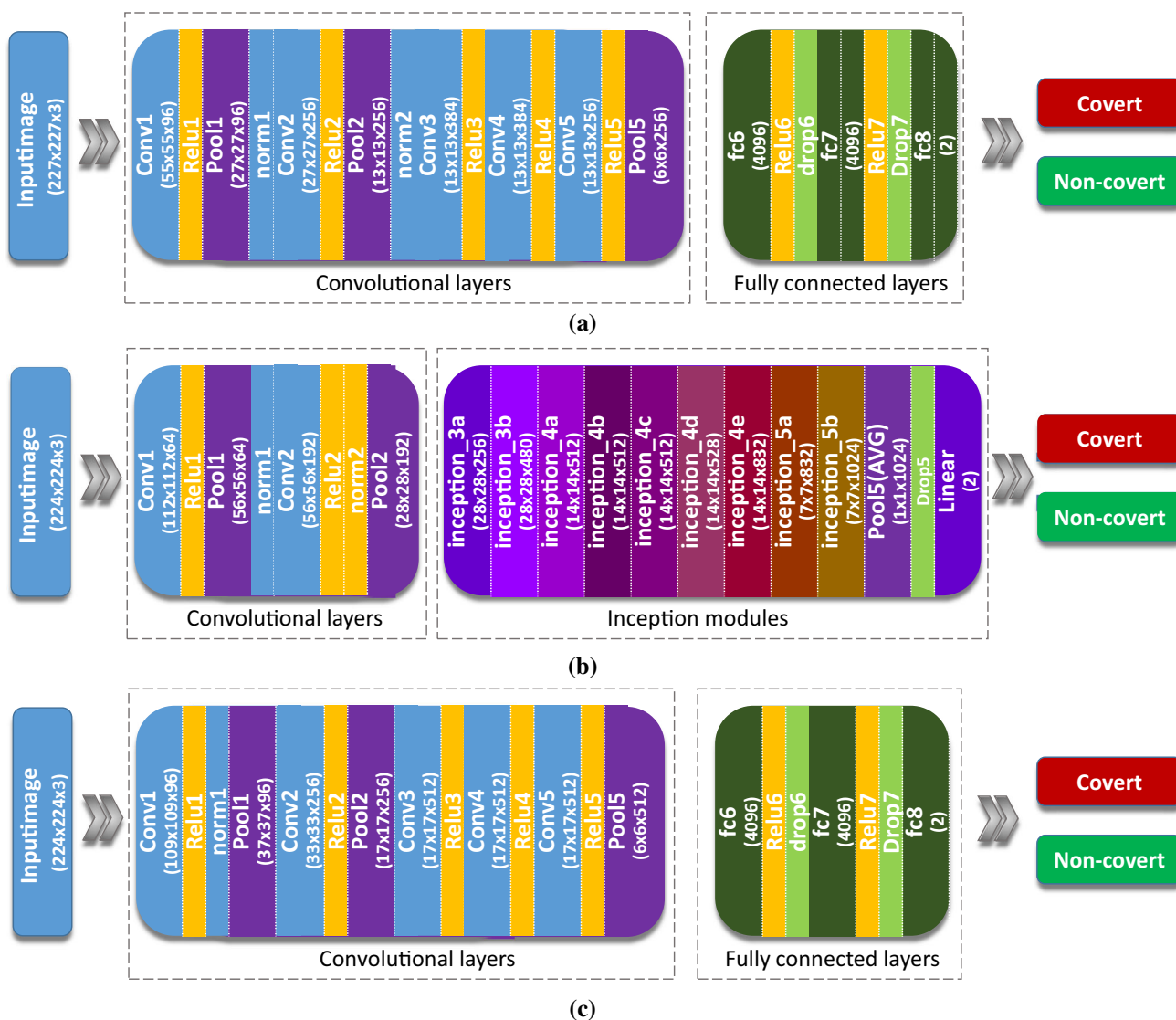


Fig. 2 An illustration of the three DCNN-based covert photo classification architectures. The source images are first resized to a fix size of 256×256 and then are subcropped to different input size (227×227 for AlexNet and 224×224 for GoogleNet and VGGs networks) and after passing through a number of convolutional, subsampling, and optional

fully connected layers, the networks finally output the class scores of covert and non-covert. The DCNN-based networks are trained end-to-end (from the raw image pixels on one end to class scores at the other)

ers with parameters. Figure 2b describes the architecture of GoogleNet, and we omit the details of the network, which is available from [18]. The input image size is of 224×224 , and after passed through two convolutional layers, the resulting feature maps are then fed to a series of inception modules. At the top of the network, average pooling instead of fully connected layers is performed. The inception module is a combination of multiple convolution layers and a parallel pooling with their output filter banks concatenated into a single output vector forming the input of the next stage. In these layers, the filter size is restricted to 1×1 , 3×3 and

5×5 . This architecture leads to a dramatically reduced number of parameters in the network, which is 12 times fewer parameters than AlexNet.

We train GoogleNet-based models using stochastic gradient descent with a batch size of 8 (due to memory limitation, this is the largest value we can set in our 2GB GPU) examples, momentum of 0.9, and weight decay of 2×10^{-4} . The base learning rate starts at 0.001 and is decreased by a factor of 10, until the test set accuracy stops improving. Best performance of AUC:97.71% and EER:7.85% is reached after 13.6 epochs when transferring parameters from Ima-

geNet pre-trained model. By contrast, the best performance of AUC:90.82% and EER:17.40% is observed after 15.5 epochs when training from scratch.

3.4.3 VGGs-based covert photo classification architecture

The VGG network [38] was proposed by Chatfield et al. from the Visual Geometry Group at the University of Oxford, and it includes three versions (the fast version VGG-F, the medium version VGG-M, and the slow version VGG-S) with different accuracy/speed comparisons. Here we use the slow but more accurate version VGG-S (denoted VGGs for simplicity) network. Similar to AlexNet [17], VGGs network contains five convolutional layers and three fully connected layers, and its architecture is described in Fig. 2c. The input is a 224×224 RGB image, and in conv1 layer it uses 7×7 filters with smaller stride 2. In conv3, conv4, and conv5 layers, more filters (512 instead of 384 and 256) are used than AlexNet. The configuration of fully connected layers (fc6, fc7, and fc8) is the same with AlexNet: the first two have 4096 neurons each, the third performs two-way covert classification and has only two outputs (covert and non-covert). All hidden layers are equipped with ReLU activation unit, and the final layer is the softmax layer.

The VGGs training procedure follows that of [38], learning on the Covert-2500 dataset using stochastic gradient descent with momentum. A mini-batch size of 10 is used to update the parameters, starting with a learning rate of 10^{-4} , in conjunction with a momentum term of 0.9. The training is regularized by weight decay, and the L2 penalty multiplier is set to 5×10^{-4} . Best performance of AUC:97.76% and EER:8.05% is reached after 7.6 epochs when transferring parameters from ImageNet pre-trained model. By contrast, best performance of AUC:79.32% and EER:29.20% is observed after 20 epochs when training from scratch.

The ROC curves of the three DCNN-based architectures with parameters transferred from ImageNet and trained from scratch are illustrated in Fig. 3. The figure shows that the results of all DCNN architectures with training from scratch (with random initialization) on the Covert-2500 dataset show drastically decreased performance. This is understandable since the training set is insufficient in size. However, when transferring parameters from ImageNet pre-trained models, VGGs-based architecture reaches best performance. It is interesting to see that all the three DCNN-based architectures, which are fully trained, outperform the hand-crafted feature methods.

3.5 Detailed analysis of what DCNN learned

Although DCNN has demonstrated excellent performance on a variety of challenging machine learning tasks, it has long been thought of a series of black boxes because it is difficult

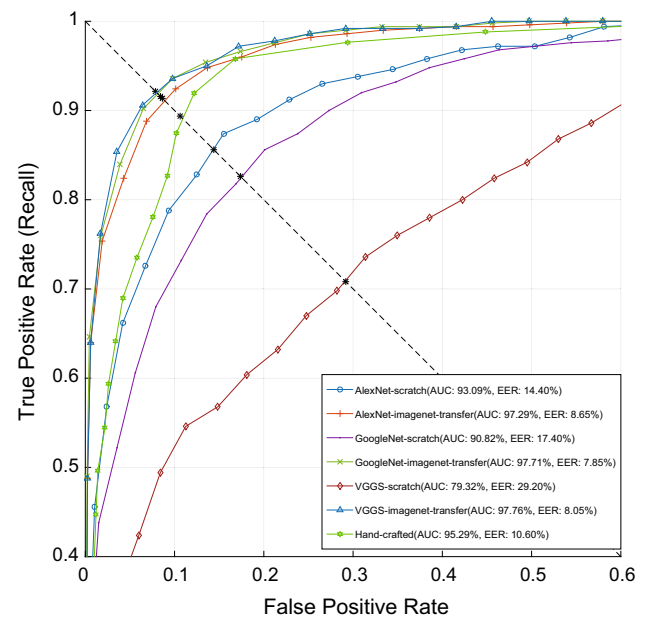


Fig. 3 ROC curves of three DCNN-based covert photo classification algorithms trained from scratch and transferring parameters from pre-trained ImageNet models, respectively. The ROC curve of hand-crafted method in [7] is also listed for comparison

to exactly understand their inner workings. In recent years, researchers have developed a series of algorithms [53–57] to expose the settings inside these black boxes and visualize the internal structure in an attempt to better understand exactly what DCNN has learned.

Figure 4 visualizes the learned filters of the first convolutional layer. All the networks parameters are initialized by transferring from pre-trained models on ImageNet and then fine-tuned on the Covert-2500 dataset. The first-layer filters exhibit human interpretable values, similar to Gabor filters or edge detectors and color blobs used in Computer Vision. This phenomenon appears in many datasets and tasks [17,41].

Figure 5 demonstrates the activation maps which followed the procedure of [55]. In Fig. 5, all the nine input photos are covert, and the discriminative image regions used by DCNN to identify covert photos from non-covert ones are highlighted. As we can see from the activation maps, different networks use different discriminative regions which means they have learned different internal representations. Most of the dark occlusion regions in the photos caused by the hidden camera are highlighted, and this infers that DCNN actually learns the intrinsic characteristics similar to human instincts. When the predicted score is mapped back to the convolutional layer, fully connected layers will lose the ability to localize objects. When we compute these activation maps, we remove the fully connected layers before the final output and replace them with a global average pooling layer for AlexNet and VGGs, while GoogleNet remains unchanged.

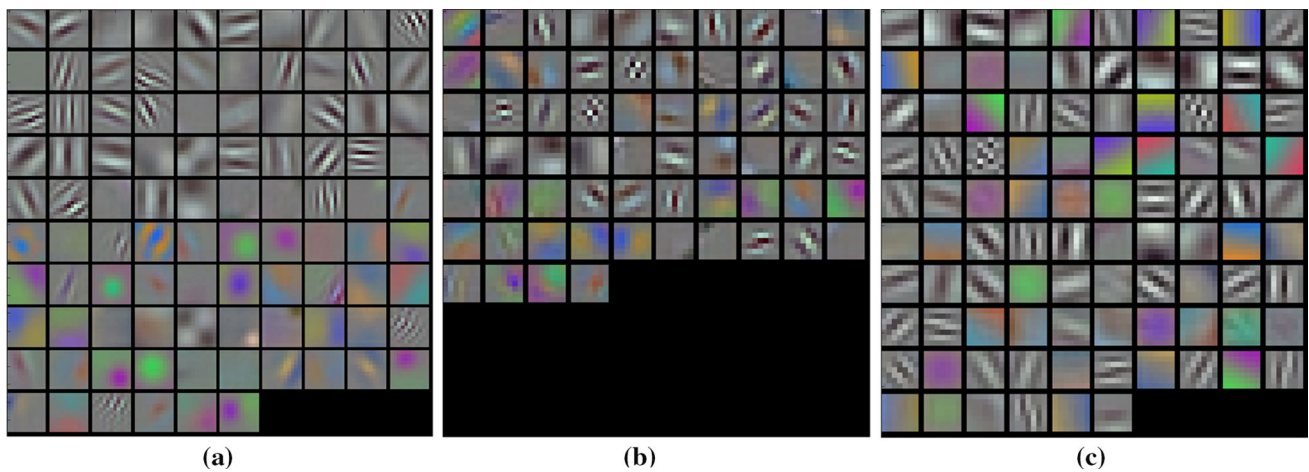


Fig. 4 Visualization of learned filters of the first convolution layers. All the three networks are pre-trained on ImageNet and then fine-tuned on the Covert-2500 dataset. Each single patch corresponds to one filter,

and all the filters resemble either Gabor filters or *color blobs*. **a** AlexNet (96 learned filters, size 11×11). **b** GoogleNet (64 learned filters, size 7×7). **c** VGGs (96 learned filters, size 7×7)

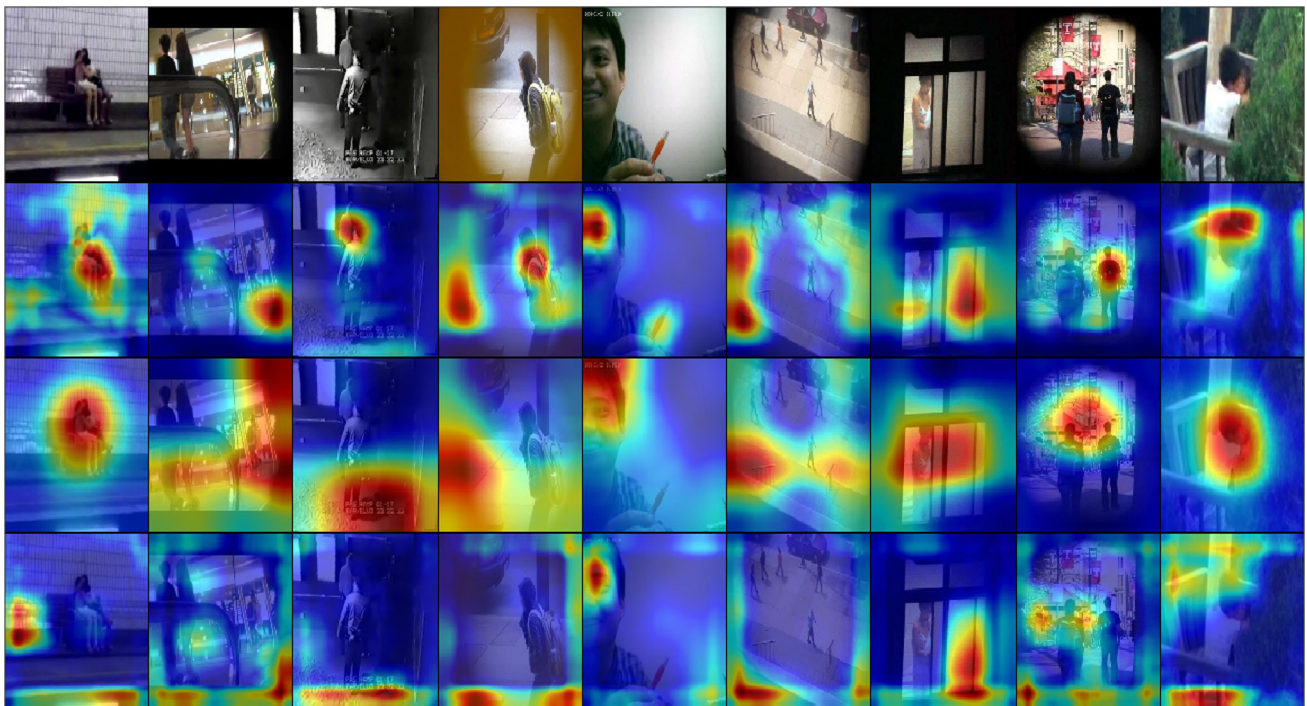


Fig. 5 Activation maps. The discriminative regions which DCNN used for covert photo classification are highlighted. The *first row* is the source covert example images; the *second row* is the activation maps of AlexNet; the *third row* is the activation maps of GoogleNet; the *last row* is the activation maps of VGGs

4 Covert-related attributes classification

Inspired by work of Zhang et al. [58] which exploits auxiliary attribute to improve the landmark detection or face alignment task, we also investigated the possibility of leveraging covert-related attributes to improve the final covert photo classification performance.

Apart from its final category, an image also has many other attributes. An attribute within the context of computer vision is defined as some semantic or abstract quality which different categories share [59]. Automatic learning and recognition of attributes can complement category-level classification and therefore improve the degree to which machines perceive visual content [60–65].

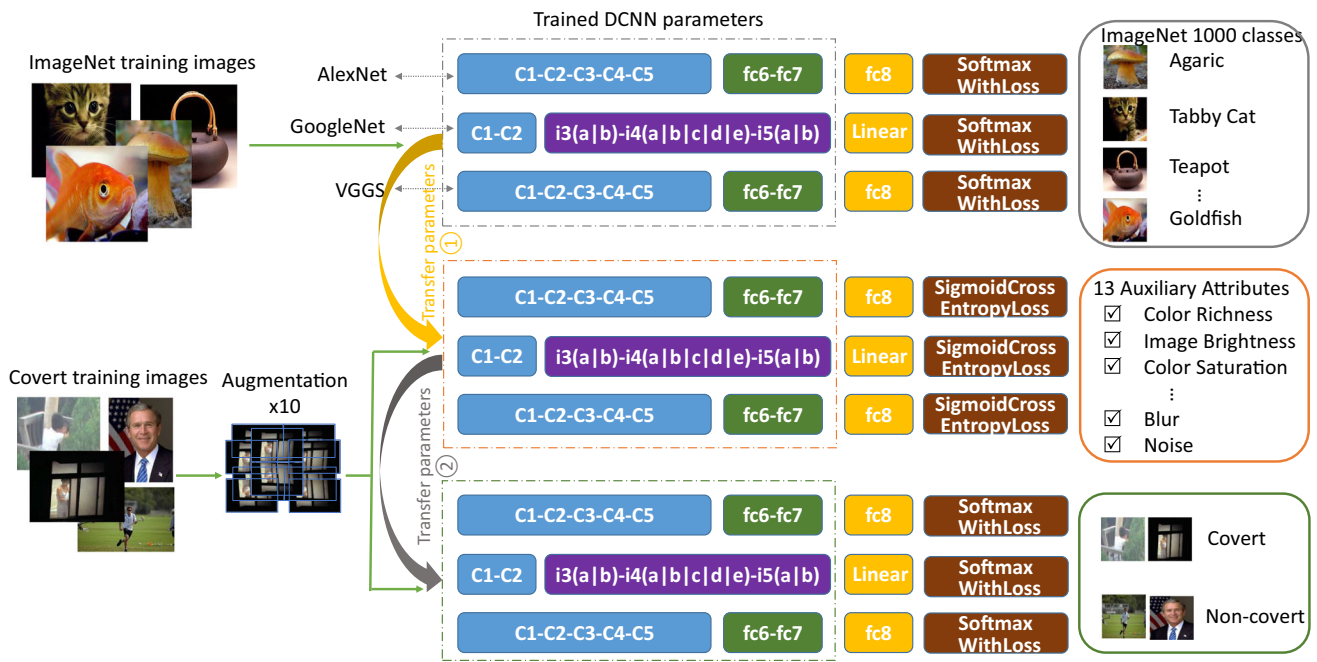


Fig. 6 Framework of covert-related attributes classification and a two-stage parameter transferring method leveraging these auxiliary attributes to improve the final covert photo classification performance. First, the pre-trained DCNN parameters on the ImageNet are transferred to the intermediate task of covert-related attributes classification

and fine-tuned, and then the fine-tuned parameters on the intermediate task are transferred to the final covert photo classification task. In the multi-label attributes classification task, sigmoid cross entropy loss layer is used

Table 1 Covert-related attributes

Attributes	Positive (1)	Negative (0)
A1-Color richness	Rich	Otherwise
A2-Image brightness	Bright	Otherwise
A3-Color saturation	High	Otherwise
A4-Image contrast	High	Otherwise
A5-Body wholeness	Whole body	Otherwise
A6-Dressing	Naked	Otherwise
A7-Pornographic	Porn	Otherwise
A8-Depth of focus	Short	Otherwise
A9-View angle	Good	Otherwise
A10-Image composition	Good	Otherwise
A11-Capturing distance	Small	Otherwise
A12-Blur	High	Otherwise
A13-Noise	High	Otherwise

In this section, a two-stage parameter transfer method is exploited. In the first stage, the pre-trained DCNN parameters on ImageNet are transferred to an intermediate task of covert-related attributes classification and fine-tuned, see Fig. 6 transfer parameters ①. Then in the second stage, the fine-tuned parameters on the intermediate task are transferred to the final covert photo classification task, see Fig. 6 transfer parameters ②.

As stated in [7], some visual attributes play important roles for human decision making of photo covertness, and 13 hand-

engineered attributes are used for covert photo classification. These 13 covert-related attributes, denoted as A1, A2, ..., A13, are listed in Table 1. In this paper, we first explore DCNN architectures for these covert-related attributes classification. We use the three DCNN architectures in the previous section, i.e., AlexNet, GoogleNet, and VGGs. The attributes classification is a multi-label task, where each input image has multiple binary labels. The DCNN networks are trained using a sigmoid cross entropy loss layer to replace the softmax with a loss layer. The multi-label vector of each input image is feed to Caffe in HDF5 format. The label is a 13-dimension binary vector, and the 13 attributes are listed in Table 1. If an attribute is positive, it is labeled as 1; otherwise, it is labeled as 0. Figure 7 demonstrates the ROC curves of 13 covert-related attributes. Figure 8 shows the accuracies (1-EER) of attributes classification by three DCNN architectures and the hand-crafted method used in [7] which exploits eight features together to estimate each visual attribute. Only 11 attributes are estimated in [7], as the last two attributes A12 (“blur”) and A13 (“noise”) are directly defined on the input image by the blind image quality index (BIQI) detector [66]. The DCNN-based architectures demonstrate approximately equivalent results compared with hand-crafted method for covert-related attributes classification.

At the second stage, the fine-tuned parameters of covert-related attributes classification are transferred to the covert photo classification. Figure 9 demonstrates the ROC curves,

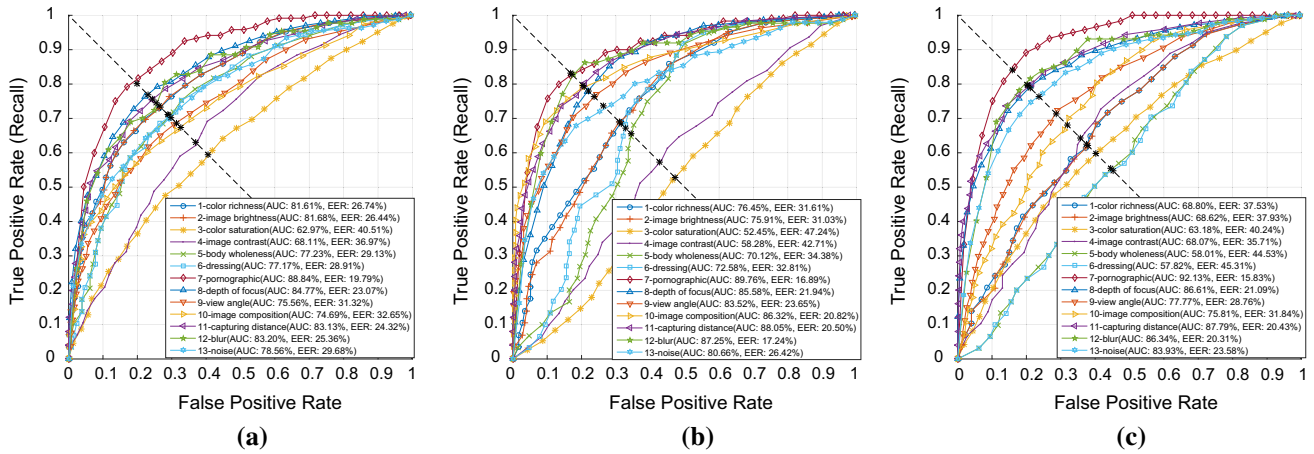


Fig. 7 ROC curves of covert attributes classification by DCNN-based architectures. a AlexNet. b GoogleNet. c VGGs

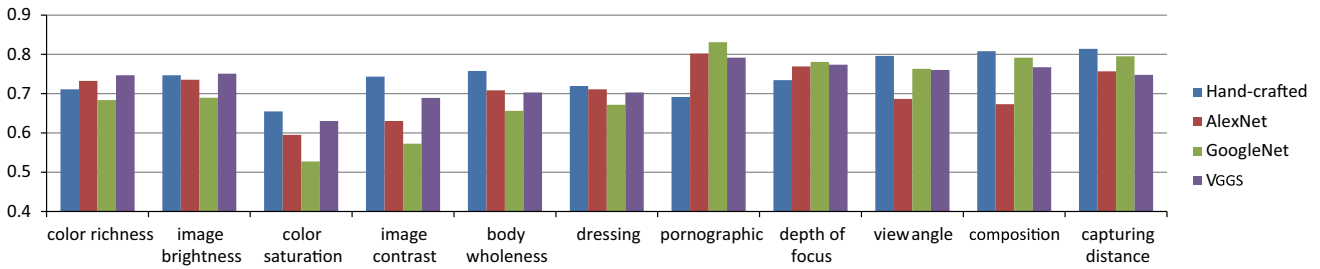


Fig. 8 Accuracies (1-EER) of DCNN-based covert attributes classification and hand-crafted method used in [7]

and the curves show that all the DCNN-based architectures with transferred parameters are within close proximity and surpassing hand-crafted method in covert photo classification. In order to further enhance the classification effect, we fuse the best models of AlexNet, GoogleNet, and VGGs (i.e., AlexNet-ImageNet-transfer, GoogleNet-ImageNet-transfer, and VGGs-attribute-transfer). Fusion is performed at the final softmax layers, and the softmax scores of individual network are combined by a weighted sum rule to produce the final score fusion result. For AlexNet, GoogleNet, and VGGs, the weights are set to be 0.2, 0.3, and 0.5, respectively. The final fusion of three DCNN-based architectures takes 41 ms to deal with an image (about 24 fps) with GPU acceleration.

At last, all the experimental results are summarized in Table 2. It shows that among all the explored individual DCNN-based architectures, the VGGs-based architecture which transferred from auxiliary attributes model reaches the best result. The fusion of three DCNN-based architectures further boosts the classification performance and achieves an average classification rate (1-EER) of 0.925

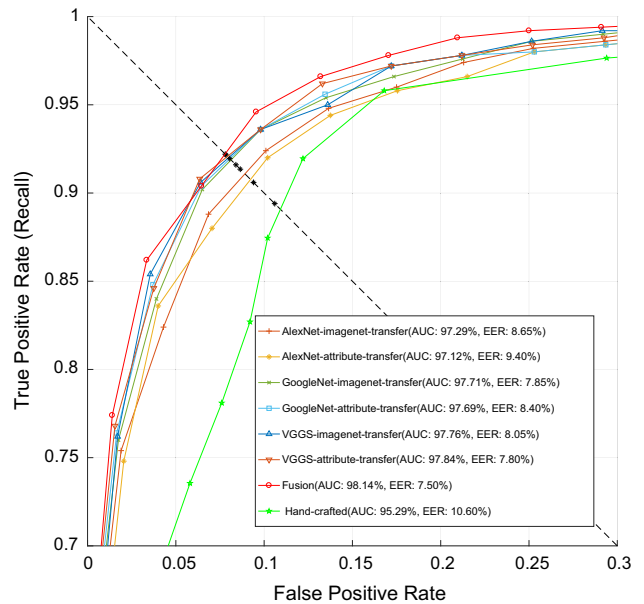


Fig. 9 ROC curves of covert photo classification transferred from ImageNet and attributes models. The ROC curve of hand-crafted method in [7] is also listed for comparison

Table 2 Summarization of experimental results of different DCNN-based architectures

Network	AUC (%)	EER (%)	1-EER (%)	Epochs
AlexNet-scratch	93.09	14.40	85.96	68
AlexNet-ImageNet-transfer	97.29	8.65	91.35	40
AlexNet-attribute-transfer	97.12	9.40	90.60	34
GoogleNet-scratch	90.82	17.40	82.60	15.5
GoogleNet-ImageNet-transfer	97.71	7.85	92.15	13.6
GoogleNet-attribute-transfer	97.69	8.40	91.60	12.8
VGGs-scratch	79.32	29.20	70.80	20
VGGs-ImageNet-transfer	97.76	8.05	91.95	7.6
VGGs-attribute-transfer	97.84	7.80	92.20	7.6
Fusion of AlexNet, GoogleNet, and VGGs	98.14	7.50	92.50	–
Hand-crafted	95.29	10.60	89.40	–

AUC area under a receiver operating characteristic curve, *EER* equal error rate
 Bold values indicate the best result

which significantly outperforms the result (1-EER) of 0.8940 of hand-crafted feature methods.

5 Conclusion

Instead of exploiting experience-dependent hand-crafted manual features, we have introduced DCNN-based architectures to automatically discover intricate structure and learn the representations for covert photo classification. We have demonstrated that the performance of DCNN-based architectures (AlexNet, GoogleNet, and VGGs) when transferring parameters from ImageNet pre-trained models significantly surpass those training from scratch. We also investigate leveraging auxiliary attributes to improve the final covert photo classification performance. A two-stage parameter transferring method is exploited. Firstly, the pre-trained DCNN parameters on the ImageNet are transferred to an intermediate attributes classification task, and then the fine-tuned parameters on the intermediate task are transferred to the final covert photo classification task. Experimental results demonstrate that all the fully trained DCNN-based architectures are within close proximity and surpassing hand-crafted method in covert photo classification. The fusion of three DCNN-based architectures shows enhanced performance over individual networks.

Acknowledgements This work was supported in part by the U.S. National Science Foundation under Grants 1618398 and 1350521 and in part by the National Natural Science Foundation of China under Grant 61103056 and in part by China Scholarship Council.

References

1. Kpcb 2016 internet trends report. <http://www.kpcb.com/internet-trends>. Accessed 1 July 2016

2. Wikipedia. <https://en.wikipedia.org/wiki/Secret-photography>. Accessed 1 July 2016
3. Zimmer, M.: Privacy on planet Google: using the theory of contextual integrity to clarify the privacy threats of Google's quest for the perfect search engine. *J. Bus. & Tech. L.* **3**, 109 (2008)
4. Hargrave, A.M., Livingstone, S.M.: Harm and Offence in Media Content: A Review of the Evidence. Intellect Books, Bristol (2009)
5. Lemire, J., Feeney, M.J., Mcshane, L.: He Wanted Roomie Out Rutgers Suicide Complained of Video Voyeur Before Fatal Fall, p. 2. Daily News, New York (2010)
6. Oulasvirta, A., Suomalainen, T., Hamari, J., Lampinen, A., Karvonen, K.: Transparency of intentions decreases privacy concerns in ubiquitous surveillance. *Cyberpsychol. Behav. Soc. Netw.* **17**(10), 633–638 (2014)
7. Lang, H., Ling, H.: Covert photo classification by fusing image features and visual attributes. *IEEE Trans. Image Process.* **24**(10), 2996–3008 (2015)
8. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: *Advances in Neural Information Processing Systems*, pp. 487–495 (2014). <http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf?spm=5176.100239.blogcont55892.31.pm8zm1&file=5349-learning-deep-features-for-scene-recognition-using-places-database.pdf>
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440 (2015)
10. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732 (2014)
11. Mikolov, T., Deoras, A., Povey, D., Burget, L., Černocký, J.: Strategies for training large scale neural network language models. In: *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 196–201. IEEE (2011)
12. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**(6), 82–97 (2012)
13. Sainath, T.N., Mohamed, A.R., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for IVCSR. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8614–8618. IEEE (2013)

14. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
15. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems*, pp. 3104–3112 (2014). <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>
16. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. arXiv preprint [arXiv:1604.00289](https://arxiv.org/abs/1604.00289) (2016)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012). <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
18. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
20. Greiner, S., Yang, J.: Privacy protection in an electronic chronicle system. In: *Proceedings of 34th Annual IEEE Northeast Bioengineering Conference*, Citeseer (2008)
21. Senior, A., Senior, A.W.: *Protecting Privacy in Video Surveillance*, vol. 1, pp. 11–33. Springer, Berlin (2009)
22. Winkler, T., Rimmer, B.: A systematic approach towards user-centric privacy and security for smart camera networks. In: *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, pp. 133–141. ACM (2010)
23. Senior, A., Pankanti, S., Hampapur, A., Brown, L., Tian, Y.L., Ekin, A., Connell, J., Shu, C.F., Lu, M.: Enabling video privacy through computer vision. *IEEE Secur. Priv.* **3**, 50–57 (2005)
24. Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.: Efficient privacy preserving video surveillance. In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 1639–1646. IEEE (2009)
25. Chen, D., Chang, Y., Yan, R., Yang, J.: Tools for protecting the privacy of specific individuals in video. *EURASIP J. Appl. Signal Process.* **2007**, 075427 (2007)
26. Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., Vincent, L.: Large-scale privacy protection in Google street view. In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 2373–2380. IEEE (2009)
27. Agrawal, P., Narayanan, P.: Person de-identification in videos. *IEEE Trans. Circuits Syst. Video Technol.* **21**(3), 299–310 (2011). <https://cvit.iiit.ac.in/images/ConferencePapers/2011/deidentTCSVT2k11.pdf>
28. Martin, K., Plataniotis, K.N.: Privacy protected surveillance using secure visual object coding. *IEEE Trans. Circuits Syst. Video Technol.* **18**(8), 1152–1162 (2008)
29. Du, L., Ling, H.: Preservative license plate de-identification for privacy protection. In: *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 468–472. IEEE (2011)
30. Chattopadhyay, A., Boulton, T.E.: Privacycam: a privacy preserving camera using uclinux on the blackfin dsp. In: *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pp. 1–8. IEEE (2007)
31. Wilber, M.I., Boulton, T.E.: Secure remote matching with privacy: scrambled support vector vaulted verification (s 2 v 3). In: *2012 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 169–176. IEEE (2012)
32. Jain, A.K., Nandakumar, K.: Biometric authentication: system security and user privacy. *IEEE Comput.* **45**(11), 87–92 (2012)
33. Martin, S., Tawari, A., Trivedi, M.M.: Toward privacy-protecting safety systems for naturalistic driving videos. *IEEE Trans. Intell. Transp. Syst.* **15**(4), 1811–1822 (2014)
34. Nakashima, Y., Babaguchi, N., Fan, J.: Intended human object detection for automatically protecting privacy in mobile video surveillance. *Multimedia Syst.* **18**(2), 157–173 (2012)
35. Elhadad, A., Hamad, S., Khalifa, A., Ghareeb, A.: High capacity information hiding for privacy protection in digital video files. *Neural Comput. Appl.* (2016). doi:[10.1007/s00521-016-2323-7](https://doi.org/10.1007/s00521-016-2323-7)
36. Ross, A., Othman, A.: Visual cryptography for biometric privacy. *IEEE Trans. Inf. Forensics Secur.* **6**(1), 70–81 (2011)
37. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015). <http://pages.cs.wisc.edu/~dyer/cs540/handouts/deep-learning-nature2015.pdf>
38. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: *British Machine Vision Conference* (2014)
39. Wallach, I., Dzamba, M., Heifets, A.: Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint [arXiv:1510.02855](https://arxiv.org/abs/1510.02855) (2015)
40. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015)
41. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: *Advances in Neural Information Processing Systems*, pp. 3320–3328 (2014)
42. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724 (2014)
43. Maitra, D.S., Bhattacharya, U., Parui, S.K.: CNN based common approach to handwritten character recognition of multiple scripts. In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1021–1025. IEEE (2015)
44. Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**(10), 1345–1359 (2010)
45. Caffe model zoo. <https://github.com/BVLC/caffe/wiki/Model-Zoo>. Accessed 15 June 2016
46. Banerjee, I., Crawley, A., Bhethanabotla, M., Daldrup-Link, H.E., Rubin, D.L.: Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma. *Comput. Med. Imaging Graph.* (2017). doi:[10.1016/j.compmedimag.2017.05.002](https://doi.org/10.1016/j.compmedimag.2017.05.002)
47. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing* **235**, 228–235 (2017)
48. Cortes, C., Mohri, M.: Auc optimization vs. error rate minimization. *Adv. Neural Inf. Process. Syst.* **16**(16), 313–320 (2004). <http://papers.nips.cc/paper/2518-auc-optimization-vs-error-rate-minimization.pdf>
49. Tronci, R., Giacinto, G., Roli, F.: Dynamic score combination: a supervised and unsupervised score combination method. In: *Machine Learning and Data Mining in Pattern Recognition*, pp. 163–177. Springer, Berlin (2009)
50. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678. ACM (2014)
51. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814 (2010)
52. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)

53. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: *Computer vision—ECCV 2014*, pp. 818–833. Springer (2014)
54. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196. IEEE (2015)
55. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. *arXiv preprint arXiv:1512.04150* (2015)
56. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. In: *Deep Learning Workshop, International Conference on Machine Learning (ICML)* (2015)
57. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**(7), e0130140 (2015)
58. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(5), 918–930 (2016)
59. Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813 (2014)
60. Ferrari, V., Zisserman, A.: Learning visual attributes. In: *Advances in Neural Information Processing Systems*, pp. 433–440 (2007). <http://papers.nips.cc/paper/3217-learning-visual-attributes.pdf>
61. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, pp. 1778–1785. IEEE (2009)
62. Russakovsky, O., Fei-Fei, L.: Attribute learning in large-scale datasets. In: *Trends and Topics in Computer Vision*, pp. 1–14. Springer, Berlin (2010)
63. Shankar, S., Garg, V.K., Cipolla, R.: Deep-carving: discovering visual attributes by carving deep neural nets. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3403–3412 (2015)
64. Jourabloo, A., Yin, X., Liu, X.: Attribute preserved face de-identification. In: *2015 International Conference on Biometrics (ICB)*, pp. 278–285. IEEE (2015)
65. Klare, B.F., Klum, S., Klontz, J.C., Taborsky, E., Akgul, T., Jain, A.K.: Suspect identification based on descriptive facial attributes. In: *2014 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–8. IEEE (2014)
66. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. *IEEE Signal Process. Lett.* **17**(5), 513–516 (2010)

Haiqiang Zuo received B.S. and M.S. degrees from China University of Petroleum (UPC) in 1999 and 2002, respectively, and Ph.D. degree in Computer Science from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China, in 2010. Since 2002, he has been with China University of Petroleum, where he is currently an Associate Professor. From 2015 to 2016, he was a visiting scholar at Temple University in the United States. His research interests include computer vision and data mining.

Haitao Lang received the B.S. and M.S. degrees from the Ocean University of China, in 2000 and 2003, respectively, and the Ph.D. degree from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, in 2006. Since 2006, he has been an Associate Professor with the Beijing University of Chemical Technology. His research interests include computer vision, machine learning, and remote sensing imagery understanding.

Erik Blasch received the B.S. degree in mechanical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1992, the Masters degrees in mechanical engineering in 1994, health science in 1995, and industrial engineering (human factors) in 1995 from Georgia Tech, Atlanta, GA, USA, and the Ph.D. degree in electrical engineering in 1999 from Wright State University, Dayton, OH, USA. He is a Program Officer with the United States Air Force Office of Scientific Research, Arlington, VA, USA. He has authored more than 650 scientific papers and book chapters. His research interests include information fusion, target tracking, pattern recognition, and robotics. Dr. Blasch is a Fellow of SPIE and an Associate Fellow of AIAA.

Haibin Ling received the B.S. degree in mathematics and the M.S. degree in computer science from Peking University, China, in 1997 and 2000, respectively, and the PhD degree from the University of Maryland, College Park, in Computer Science in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia. From 2006 to 2007, he worked as a postdoctoral scientist at the University of California Los Angeles. After that, he joined Siemens Corporate Research as a research scientist. Since fall 2008, he has been with Temple University where he is now an Associate Professor. Dr. Ling's research interests include computer vision, augmented reality, medical image analysis, and human-computer interaction. He received the Best Student Paper Award at the ACM Symposium on User Interface Software and Technology (UIST) in 2003 and the NSF CAREER Award in 2014. He is an Associate Editor of *IEEE Trans. on Pattern Analysis and Machine Intelligence* and serves on the editorial board of *Pattern Recognition* and served as Area Chairs for CVPR 2014 and CVPR 2016.