Privacy-Preserving Age Estimation for Content Rating

Linwei Ye*
University of Manitoba
Winnipeg, Canada
yel3@cs.umanitoba.ca

Binglin Li*
Simon Fraser University
Burnaby, Canada
binglinl@sfu.ca

Noman Mohammed
University of Manitoba
Winnipeg, Canada
noman@cs.umanitoba.ca

Yang Wang
University of Manitoba
Winnipeg, Canada
ywang@cs.umanitoba.ca

Jie Liang
Simon Fraser University
Burnaby, Canada
jiel@sfu.ca

***Abstract*—Content rating (*aka.* maturity rating) rates the suitability of kinds of media (*e.g.,* movies and video games) to its audience. It is essential to prevent a specific age group of people such as children from inappropriate information. However, in practice the administration of content rating system is usually suggestion-based declaration by media sources or key-based password which can easily fail if someone ignores the suggestions or somehow knows the keys. In this paper, we propose to estimate user's age in a privacy-preserving manner for automatic content rating. Several privacy-preserving approaches on facial images with different degree of privacy are proposed and evaluated on a deep neural network architecture for age estimation accuracy. We also introduce an attention mechanism which can adaptively learn discriminative features from the processed facial images. Experiments show that the proposed attention-based model performs better than the baseline model and achieves a reasonable performance to that with raw images in testing.**

I. Introduction

Content rating is important to prevent a specific age group of people especially children from inappropriate contents. As the Internet spreads globally, parents are concerned about their children surrounded with improper materials such as violent, sexual abused, frightening scenes and coarse languages when using social networks and online entertainment softwares. Personally monitoring what the children are browsing all the time is not feasible. Different types of filters exist for content management, such as browser-based filters, client-side filters, DNS-based filtering and search-engine filters [1]. Recently many software products are deployed for parental control. However, most of them require people to manually turn on or off the filters, which is not convenient when parents and children share the same computer device. Moreover, the systems are usually password-based administration [2] which sometimes can be fooled if the children somehow know how to turn the filter off. Therefore, how to automatically recognize whether the user is a real adult or kid is still a challenge as well as significant problem.

Human face conveys rich perceptible information related to individual attributes including identity, age, gender, expression, ethnic, etc. Human age, as one of the important individual attributes, can be directly inferred based on facial appearance by human to identify the exact age or the age group. Automatic age estimation by machines is useful in scenarios where a system needs to specifically identify the age of the individual without massive inspections by human power. Age estimation from facial images plays a significant role in the real-world applications [3], [4]. For example, a well designed age specific human-computer interaction system can help prevent kids browsing adult web pages or purchasing age restricted materials.

On the other hand, privacy of facial images is a key part in the content rating. Simply estimating the age from one's facial image will result in privacy issues as the user's face is captured and so will be exposed to age estimation system before his/her age is identified. There are several papers [5]–[7] studying on how to preserve privacy in specific applications. However, they are not suitable to be merged into content rating management for automatic age estimation.

Contributions: Inspired by the above observations, the proposed approach is designed for both privacy-preserving and automatic content rating. By learning the discriminative features out of private regions of facial images with attention mechanism, the accuracy of the proposed age estimation approach outperforms the baseline models and achieves a comparable performance in the privacy preserved scenario. In summary, our contribution has three manifolds:

- We propose and evaluate several privacy-preserving approaches with different degrees of privacy. These approaches can be incorporated into age estimation model and meanwhile preserve privacy of users' facial images.
- We propose attention-based deep learning model for automatic age estimation. This attention-based model is capable of seeking for discriminative features in a processed facial image, which effectively improves the accuracy over baseline model.
- Experimental results demonstrate that the proposed attention-based model outperforms baseline model in privacy-preserving facial images with two methods (Mask and Mosaic on eyes) and achieves 90.0 % exact accuracy and 93.4 % 1-off accuracy performance compared to the model with original facial image.

II. Background

Privacy and Security: A system of preventing privacy leakage from photos in social networks is studied to control permis-

*Equal contribution

sion [5]. Erkin *et al.* [6] propose a strongly privacy-enhanced face recognition system to hide both the biometrics and server results. It considers a scenario where one party provides a face image, while another party has access to a database of facial templates. The proposed protocol tries to find a way that the first party cannot learn from the execution of the protocol more than basic parameters of the database, while the second party does not learn the input image or the result of the recognition process. Therefore, it allows to efficiently hide both the biometrics and the result from the server that performs the secure matching operation. Another way to remedy the privacy concern and maintain image quality such as context information is introduced by [7]. For images that contain appearance of people, it aims to visually unintrusive privacy protection for facial images with preserving facial expressions. Specifically, this method shares the same basic idea as morphing-based privacy protection because it modifies a person's face by mixing with other person's face.

Deep Learning: Deep neural network (DNN) is a current dominant approach for computer vision problems attributed to its powerful feature representation ability. Compared to traditional machine learning methods that use hand-carfted feature extraction, DNN can automatically learn the feature representation by itself during training. Most of existing DNN based approaches are using Alexnet [8], VGG [9] or Resnet [10] as feature extraction network. The feature extraction network contains a stacked convolution and pooling layers. It is followed by several fully-connected layers or global average pooling layer to convert feature channel to the exact number of labels for classification. Although DNN has achieved satisfying performance in many applications including Natural Language Processing [11] and computer vision [9], it has some intrinsic drawbacks. Because of its deep architecture, the number of parameters is usually much larger than that of traditional classifiers, which means it requires more training data in order to avoid overfitting problem.

Age Estimation: Automatic age estimation is a difficult and challenging problem even under the strong DNN framework, because one's appearance change along age growing depends on not only the gene, but also many external factors around, such as living style and environment, sociality, health conditions, and even facial expression variations that will possibly influence the estimation. Research on age estimation are in different setting from ours. Basically they do not consider users' privacy issues. Some work treats age estimation as regression problem [12], [13], [14] or multi-class classification problem [15], [16]. Many argue that aging progress is ordinal and continuous, so it can also be cast as an oridinal problem [17], [18], [19]. Most of them are modified to a series of binary-classification-based model [18], [19]. [20] considers each face image as an example associated with a label distribution and estimate facial ages by learning from those label distributions due to insufficient training data.

Convolutional neural network is used to extract features directly from the data and the generated feature maps obtained in different layers are combined. A manifold learning algorithm is incorporated in the proposed scheme with another regressor to estimate age [4] . The DEX method [21] introduces a larger biological age prediction dataset IMDB-WIKI [22] which is crawled internet facial images with available age as augmented data. The VGG-based [9] network is pretrained on ImageNet [23] and IMDB-WIKI dataset, and then finetune on a small dataset. They pose the age regression problem as a deep classification problem followed by a softmax expected value refinement, which results in substantial improvement over direct regression training of DNNs.

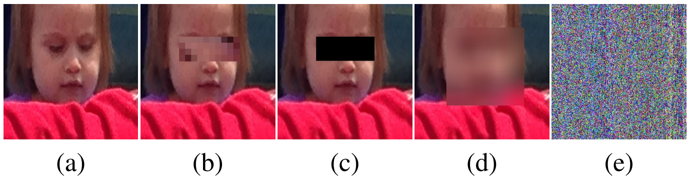

Fig. 1. Four proposed privacy-preserving approaches to address a facial image. From left to right: (a) original image, (b) mosaic on eyes, (c) mask on eyes, (d) noise on face, (e) encryption.

A large scale dataset is a crucial step to conduct massive deep neural network experiments. Recently some age datasets have been released to the public such as IMDB-WIKI [24] and Adience [16] where each dataset has more than 10k images, which gives rise to a few amount of work addressing age estimation based on deep learning architecture. However, to the best of our knowledge, our work is the first one to strive for improving age estimation performance using DNN in a private-preserving way which considers to protect users' privacy.

Attention Models: Attention models automatically select informative regions that can be adapted to learn age features at non-private regions. It is applied in various computer vision tasks such as fine-grained classification [25], object localization and image captioning [26]. Spatial Transformer Network (STN) [27] can also be viewed as one of attention approaches that spatially transform the input image to focus on relevant patches to the task.

Several attention-based works explore other related regions except the most discriminative part learned in training in order to drive the model to get ability to generalize features for testing. A random sampling [28] for hiding visible image patches in training is proposed to the weakly-supervised object localization. The model is forced to focus on other relevant object parts. A two-phase learning method [29] is also developed for object localization. It uses the first network to learn the most discriminative part as usual and adds another network to seek rest discriminative parts of objects. Wei *et al.* [30] exploits adversarial erasing framework to iteratively mine object regions to achieve semantic segmentation.

III. Approach and Models

In this section, we first describe several privacy-preserving strategies to process facial images to make their identity unrecognizable by human. Then we introduce the baseline model based on deep learning network and the proposed attention model in order to learn features of age adaptively.

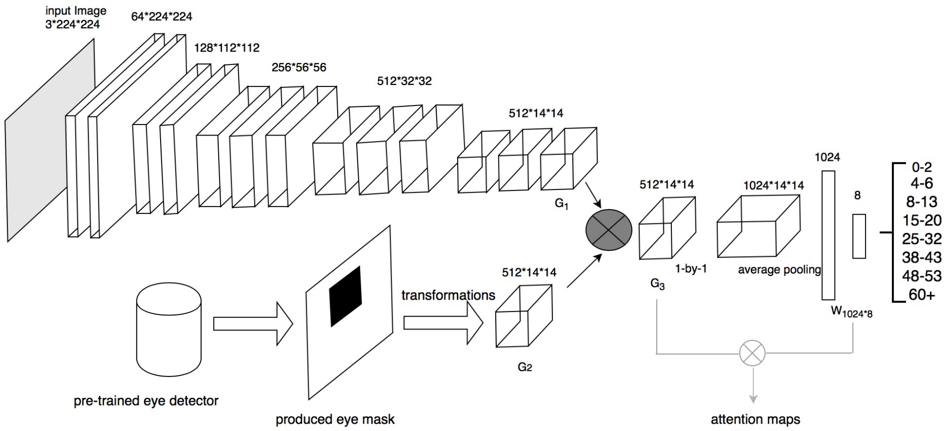

Fig. 2. Proposed attention-based model for age estimation. From the upper branch, DNN feature map G_1 is extracted from an input image. This image is also passed through the eye detector to locate eye regions to generate eye masks G_2. Then a new feature map G_3 is produced by element-wise multiplication on G_2 and G_1. G_3 is followed with a 1-by-1 convolutional layer and average pooling for final age classification. By hiding eye regions during training, the model sees different facial parts and is forced to focus on learning other relevantly discriminative regions for age estimation automatically.

A. Privacy-preserving Approaches on Facial Images

The automatic age estimation model from facial images is a complement to content management which is vulnerable to manual controls. It takes a person's facial image as input and automatically estimates this current user's age. However, one issue that people may be worried about is the leakage of privacy, since collecting their facial images is an essential part of age estimation. As shown in Fig. 1, an original facial image shows a user's full face and poses a risk of exposure due to curious database administrator (DBA) and arbitrary adversary attacks. Therefore, in this section, we propose four common but effective approaches to remove recognizable parts of a face for the sake of privacy. It should be noted that the proposed approaches also need to be unrecoverable in case anonymized algorithms are cracked, which will be discussed in Sec. IV-D later.

1) Mosaic on Eyes: Inspired by observations from life and TV interviews, eyes are quite important recognizing parts of human identification. In addition, eyes are also close to other distinct facial attributes such as ala nasi and eyebrow. So to the best of our knowledge, using a rectangle mosaic around eye areas is sufficient to anonymize the identification of a certain person. Specifically, we first detect landmarks of the face by [31], and then connect regions of eyes and expand this area by 10%, so that ala nasi and eyebrow are also be covered. In this specified area, for every subset of pixels which is with size of $10 * 10$, we randomly sample a pixel and use this value to fill the whole pixel in this subset. As shown in Fig. 1 (b), a rectangle mosaic is unrecoverable because of missing pixel values. Therefore it is able to anonymize identification of the person to some extent.

2) Mask on Eyes: This is similar to mosaic on eyes mentioned in above paragragh. Instead of using mosaic, we simply use black mask to cover around eye regions. An example image is shown in Fig. 1 (c), where all regions are filled with black pixels $(0, 0, 0)$ for anonymization.

3) Noise on Face: A more thorough way to anonymize a user's face is to add noise on the whole face. In that case, we instead detect the face in the image [31] and all facial attributes are masked and the privacy of the user is preserved naturally. As shown in Fig. 1 (d), the whole face is blurred by a gaussian filter with size of $25 * 25$. Note that due to some technical issues, some facial regions *e.g.*, hair, may not be fully covered. This is acceptable for our aim of privacy, because the hair style of a person varies in different time period.

4) Encryption: It uses Data Encryption Standard (DES) block cipher which is a symmetric-key algorithm for the encryption. A block cipher can operate on fixed-length groups of bits with an unvarying transformation by a symmetric key. We adapt the ImageEncryptor algorithm [32] on the whole image. It takes a 2-key permutation system to biject the pixels in a given image. Figure 1 (e) shows the visualized result of encrypted image, where the whole image becomes invisible to human vision system.

B. Age Estimation

We treat age estimation as image classification problem, where a group of ages is clustered as a certain classification label (clusters for different age groups are defined in Sec. IV). Then for predicting the age in each input facial image I, the output will be assigned to a certain classification label that means the age of the person is within this age cluster. This is more feasible than predicting an exact age value of a person. First, an age cluster greatly reduces the prediction distribution of age estimation problem, which also alleviates the workload of labelling ground truth for data-hungry deep learning method. Second, predicting an exact age value may confuse the learning of deep networks since the appearance change of different people is independent. But age groups are robust to this problem. The prediction can also be evaluated effectively and efficiently. Third, content rating is originally designed for different groups of ages rather than restricting to certain age value of people only.

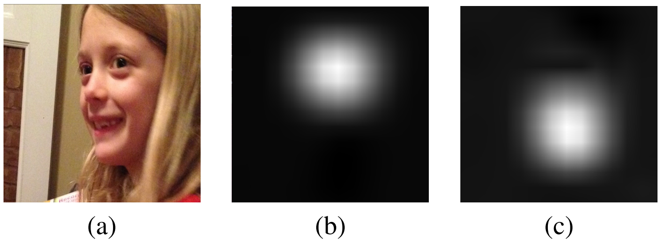

Fig. 3. An example shows how the attention model forces DNN to focus on other discriminative part out of eye areas in an image. (a) is an input image, (b) shows high response positions for the age using baseline model, and (b) shows high response positions for the age using attention model.

1) Baseline Model: In this paper, we use VGG16 [9] as a baseline architecture for age estimation. Original VGG16 contains 16 layers in total. The first 13 layers are considered as feature extraction, and consist of convolutional layers, max pooling layers and ReLU layers. Convolutional layers do computations with sliding filters. These filters are the weights that will be updated during training. Max pooling layers reduce the spatial size. Larger pooling size helps to get context information but at expenses of the resolutions. ReLU layers add non-linearity so that the architecture has large capacity to fit the model. The last three layers can be taken as a multi-layer classifier, and the number of output in the last layer is the number of categories according to applications. We remove the last pooling layer to keep feature maps with larger spatial size $512 \times 14 \times 14$ (we denote this feature map as G_1). In order to maintain rich feature representations, a 1-by-1 convolutional layer is added to convert the size to $1024 \times 14 \times 14$ and average pooling is applied to obtain global feature representation. At the end, a linear layer is applied to classify the 8 categories (age groups). For classification problem, it is commonly adopted to initialize the weights using pretrained model *e.g.* on ImageNet [23] because it has been trained on millions of images for 1000-category classification and the weights fine-tuned from it can lead to faster convergence as well as avoid local minima. IMDB-WIKI model [24] is trained on 500k+ celebrities' facial images for age estimation. Compared to ImageNet pretrained model, it learns more facial information concerning age. So here we fine-tune our model from the pretrained IMDB-WIKI model.

As shown in Fig. 3 (b), we visualize the response map after training by applying matrix multiplication to feature map G_1 and the weight $W_{1024\times 8}$ in last linear layer, where we can get a response map with size 14×14 for each corresponding category. We only visualize the response map of the category that has the highest probability.

2) Attention Model: From Fig. 3 (b), the baseline DNN recognizes the age of a person by focusing on the most discriminative parts of the image (*e.g.*, regions around eyes for baseline model). However, during testing since the most discriminative part of the person, i.e., eyes, is removed for the sake of identification privacy, age estimation is prone to be misclassified.

In our approach, testing images are pre-processed by one of the privacy-preserving approaches proposed in Sec. III-A. In order to address this case, we propose an attention-based model to force DNN to learn discriminative features out of eyes for age estimation as showed in Fig. 2.

We employ an eye detector [31] to locate eye regions. Based on these eye coordinates, an eye mask is estimated as showed at the bottom in Fig. 2. Then the eye mask is integrated into the attention-based model and we perform transformations (e.g., alignment and dimension expansion) on the eye mask to generate G_2. An element-wise multiplication on G_2 and the feature map G_1 generated from the baseline VGG16 model is used to produce the combined new feature map G_3 which is followed with a 1-by-1 convolutional layer similar to baseline model.

Again as showed in Fig. 3 (c), by applying matrix multiplication to feature map G_3 and the weight $W_{1024\times 8}$ in last linear layer, we can get a new response map for each corresponding category. We see that by hiding eye regions during training, the model sees different facial parts and is forced to focus on learning other relevantly discriminative regions (*e.g.*, regions around nose and mouth) for age estimation in Fig. 3 (c). After using this mask-out learning approach during training, the model is able to move its attention to other discriminative parts and achieve better age estimation results in testing.

IV. Experiment

In this section, we experiment and evaluate the age estimation with different privacy-preserving processing approaches on the baseline model and attention-based model.

A. Setup

1) Dataset: We use the age group estimation dataset Adience [16] in our experiment. This dataset contains 26,580 images with 8 age groups (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-years), in which 12242 and 4406 images are used for training and testing, respectively. Note that we omit some images that eye and face detectors fail to localize correct regions. All of them are cropped and aligned facial images. We tackle the age estimation as a 8-class classification problem and learn to predict one of the 8 age groups.

2) Evaluation Metrics: Two evaluation metrics (exact accuracy and 1-off accuracy) are applied. Exact accuracy is the standard classification accuracy which is the ratio of the number of images classified as correct age groups over the total number of testing images, while 1-off accuracy is the ratio of the number of images classified as correct age groups or two adjacent age groups over the total number of testing images. For example, if the true age group is 8-13, if the predicted age group is 4-6 or 15-20, it will count as correct classification. This metric is reasonable if we only want to know the rough age group of a person and can tolerate some extend of errors.

B. Baseline Model Result

We evaluate the proposed privacy-preserving approaches to figure out the appropriate tradeoff between privacy and utility.

Method	Exact accuracy (%)	1-off accuracy (%)
Random guess	12.50	37.50
VGG-encryption	36.79	53.20
VGG-face	41.42	61.85
VGG-mask	54.60	83.77
VGG-mosaic	56.96	85.34
VGG original	66.50	94.98

TABLE I
COMPARISON OF AGE ESTIMATION OF DIFFERENT VGG-BASED APPROACHES FOR FACIAL IMAGES ON THE ADIENCE DATASET.

Method	Exact accuracy (%)	1-off accuracy (%)
VGG-mask	54.60	83.77
VGG-mask atten	58.78	88.66
VGG-mosaic	56.96	85.34
VGG-mosaic atten	59.79	88.17
VGG original	66.50	94.98

TABLE II
COMPARISON OF AGE ESTIMATION OF DIFFERENT VGG-BASED APPROACHES WITH ATTENTION-BASED MODEL AND BASELINE MODEL FOR FACIAL IMAGES ON THE ADIENCE DATASET.

Tab. I shows age estimation results by using different privacy-preserving approaches proposed in Sec. III-A to process test images and the last row is VGG model tested on the original images without any processing for privacy consideration. Here, all approaches except for the random guess are tested on the same VGG model which is trained with complete facial images, as we assume the training facial images are publicly available or agreement-based available. In the first row random guess means that we randomly guess the age out of the 8 categories without looking at the user's image.

It can be seen that global encryption and global noise on face (VGG-face) are the strongest approaches for privacy but since the whole face are noised or encrypted, they perform significantly worse compared to the other eye-based approaches. Compared to random guess, they maintain certain amount of information that helps for prediction. Nonetheless, it is not suitable for privacy-preserving applications.

VGG-mosaic and VGG-mask which both anonymize eye regions show similar accuracy in testing. But they are about 10% worse than unanonymized baseline VGG (VGG original in Tab. I), because some discriminative features of eyes are learned from training but not available in testing.

C. Attention Model Result

We further compare the different VGG-based eye processing approaches using attention model and the baseline model both with original images as the input. From Tab. II, we can see that attention model achieves better performance (about 4% improvement for VGG-mask and 3% for VGG-mosaic) compared to that using baseline model. Meanwhile VGG-mask and mosaic guarantee intermediate level of privacy. In our proposed model, a compromise is reached between privacy and data utility. Fig. 4 shows some qualitative results for given images.

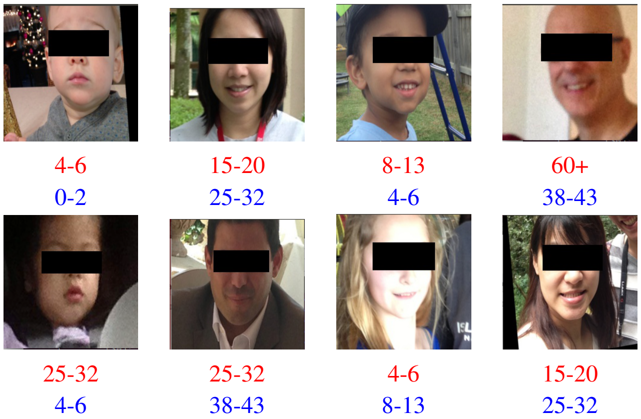

Fig. 4. Qualitative results.The first row are estimation results from VGG-mask (red) and the second row are results from VGG-mask attention model (blue). Age groups with red color are wrong predications and those with blue color are correct predications.

D. Discussions

We can see that encryption method is probably the most secure way to protect user's privacy. But when the model is applied to the system, each user might have their own ways to encrypt their facial images, which may results in inconsistency of all the testing images. Since the training images are raw facial images, the performance in content rating can be worse than that given in our experiment. On the other hand, if the user chooses some simple encryption method, it would be vulnerable to be attacked by adversaries.

Other ways such as blurring, eye mask or mosaic are different from the encryption technique. From the perspective of information theory, the processed facial images have lost some information permanently, so it is impossible to recover these areas exactly. Blurring the whole face loses the most information and it leads to poor age estimation performance. In terms of eye mask and mosaic, provided that adversary hacked the cloud and applies trained model to try to recover the user's face, it is difficult to get good reconstruction because the user's original facial image is not in the training set [33]. The recovered face is more likely to average all the faces in the training set to make up the missed eye regions of the user's face. In addition, when we do not have prior information about a person, it is more difficult to recognize him/her.

V. CONCLUSION

We present multiple ways to process facial images to preserve privacy and experiment the processed images for age estimation with baseline models as well as the improved attention-based models. The proposed attention-based model improves the age estimation performance close to baseline and meanwhile preserves intermediate level of privacy. We also discuss the privacy and security issues of the different privacy-preserving approaches at the end. Our proposed model can be tuned to real-world application for content rating and helps to effectively protect people from inappropriate materials.

References

[1] Wikipedia, "Content-control software," 2017, [accessed 30-October-2017]. [Online]. Available: https://en.wikipedia.org/wiki/Content-control_software

[2] A. Rockley, P. Kostur, and S. Manning, *Managing enterprise content: A unified content strategy*. New Riders, 2003.

[3] G. Guo, G. Mu, Y. Fu, and T. Huang, "Human age estimation using bio-inspired features," in *IEEE Computer Vision and Pattern Recognition*, 2009.

[4] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *IEEE Winter Conference on Applications of Computer Vision*, 2015.

[5] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis, "Face/off: Preventing privacy leakage from photos in social networks," in *ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[6] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in *International Symposium on Privacy Enhancing Technologies Symposium*, 2009.

[7] Y. Nakashima, T. Koyama, N. Yokoya, and N. Babaguchi, "Facial expression preserving privacy protection using image melding," in *IEEE International Conference on Multimedia and Expo*, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[12] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *IEEE Computer Vision and Pattern Recognition*, 2011.

[13] Y. Zhang and D. Y. Yeung, "Multi-task warped gaussian process for personalized age estimation," in *IEEE Computer Vision and Pattern Recognition*, 2010, pp. 2622–2629.

[14] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *IEEE Computer Vision and Pattern Recognition*, 2013.

[15] X. Geng, Z. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, vol. 29, no. 12, pp. 2234–2240, 2007.

[16] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[17] C. Li, Q. Liu, J. Liu, and H. Lu, "Learning ordinal discriminative features for age estimation," in *IEEE Computer Vision and Pattern Recognition*, 2012.

[18] K. Y. Chang, C. S. Chen, and Y. P. Hung, "A ranking approach for human ages estimation based on face images," in *IEEE International Conference on Pattern Recognition*, 2010.

[19] K. Y. Change, C. S. Chen, and Y. P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *IEEE Computer Vision and Pattern Recognition*, 2011.

[20] X. Geng, C. Yin, and Z. Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[21] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, pp. 1–14, 2016.

[22] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10–15.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[24] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision (IJCV)*, July 2016.

[25] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *IEEE Computer Vision and Pattern Recognition*, 2015.

[26] K. Xu, J. Ba, R. Kiros, K. C. andA. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, sttend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015.

[27] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015.

[28] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," *CoRR*, vol. abs/1704.04232, 2017.

[29] D. Kim, D. Cho, D. Yoo, and I. So Kweon, "Two-phase learning for weakly supervised object localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[30] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," *arXiv:1703.08448*, 2017.

[31] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[32] "Imageencryptor algorithm." [Online]. Available: https://github.com/AtheMathmo/ImageEncryptor

[33] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *ACM SIGSAC Conference on Computer and Communications Security*, 2015.