

Multitask ConvNet for Blind Face Inpainting with Application to Face Verification

Shu Zhang, Ran He*, Zhenan Sun and Tieniu Tan

National Laboratory of Pattern Recognition, CASIA

Center for Research on Intelligent Perception and Computing, CASIA

Center for Excellence in Brain Science and Intelligence Technology, CAS

University of Chinese Academy of Sciences, Beijing, 100049, China

{shu.zhang, rhe, znsun, tnt}@nlpr.ia.ac.cn

Abstract

Face verification between ID photos and life photos (FVBIL) is gaining traction with the rapid development of the Internet. However, ID photos provided by the Chinese administration center are often corrupted with wavy lines to prevent misuse, which poses great difficulty to accurate FVBIL. Therefore, this paper tries to improve the verification performance by studying a new problem, i.e. blind face inpainting, where we aim at restoring clean face images from the corrupted ID photos. The term blind indicates that the locations of corruptions are not known in advance.

We formulate blind face inpainting as a joint detection and reconstruction problem. A multi-task ConvNet is accordingly developed to facilitate end to end network training for accurate and fast inpainting. The ConvNet is used to (i) regress the residual values between the clean/corrupted ID photo pairs and (ii) predict the positions of residual regions. Moreover, to achieve better inpainting results, we employ a skip connection to fuse information in the intermediate layer. To enable training of our ConvNet, we collect a dataset of synthetic clean/corrupted ID photo pairs with 500 thousand samples from around 10 thousand individuals. Experiments demonstrate that our multi-task ConvNet achieves superior performance in terms of reconstruction errors, convergence speed and verification accuracy.

1. Introduction

Face recognition [8] has achieved significant advances in recent years benefiting from deep learning [2] methods, especially the ConvNet (convolutional neural network) [12, 18]. Since the groundbreaking work of DeepFace[17], ConvNets have continuously set new records on the LFW

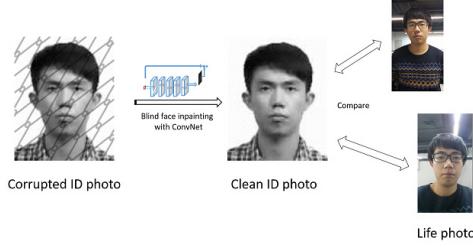


Figure 1. The face images corrupted by the *Chinese administration center* for ID photos and the flowchart of the application of FVBIL with blind face inpainting.

benchmark [9] and even achieved beyond human performance in [16]. These recent advances have promoted the development of a specific task, i.e. the face verification between ID photos and life photos [22]. In this task, face images acquired in unconstrained environment from daily life are compared with face images in the photos of the identity card. Challenging as it is, FVBIL has drawn much attention because it has various potential applications, e.g. clearance at an airport, opening bank account from remote and registration on a conference.

However, ID photos provided by the Chinese administration center are often corrupted with wavy lines or watermarks to prevent misuse. Fig 1 shows an example of this kind of ID photos. These imposed patterns would severely deteriorate the face detection and alignment performance and pose great challenges to face verification. Therefore, detecting those corrupted pixels and restoring them with ground truth face textures are crucial for FVBIL. The automatic detection and reconstruction process of corrupted images are often known as blind inpainting [3]. Since our inpainting problem for FVBIL specializes on the face images in ID photos, we denote our problem as the blind face inpainting problem, in which the term 'face' indicates the faces in ID photos. Fig 1 illustrates the blind face inpaint-

*Ran He is the corresponding author.

ing process for FVBIL.

To improve the performance of FVBIL, this paper studies a real time blind face inpainting approach to simultaneously detect the corrupted regions and reconstruct the corresponding face textures. To this end, we propose an multi-task ConvNet, which takes as input only the corrupted ID photos and as output the residuals between corrupted/clean photo pairs (both their values and positions). Specifically, by regressing the residuals between clean/corrupted pairs instead of the clean ID photos directly, an additional supervision, *i.e.* the positions of corruption, can be readily incorporated into the same ConvNet. This auxiliary task helps the network to converge faster and perform better than its single task alternatives, achieving better results in shorter time. Moreover, we introduce a skip connection to the traditional ConvNet architecture, which can provide accurate context information to the middle layers of the ConvNet. Although the performance gain is only marginal, this allows us to gain insight into the processing scheme of our proposed inpainting ConvNet.

Training the deep ConvNet requires a large set of clean/corrupted ID photo pairs. Therefore, we develop a technique to generate 500 thousand synthetic clean/corrupted pairs from 10 thousand clean ID photos. This amount of data would satisfy our demands to train the ConvNet. We also collect another dataset that consists of real corrupted ID photos and clean life photos to evaluate the model’s capability in FVBIL tasks.

In summary, our main contributions are:

1. We first address blind face inpainting problem and develop a real time approach to simultaneously detecting and reconstructing corrupted pixels. This approach is very fast at test time and achieves promising results on the FVBIL tasks.
2. We propose a multi-task ConvNet with residual learning to handle the blind face inpainting problem. The proposed model converges faster and performs better than its single task alternatives. It provides a potential solution of general blind or non-blind inpainting problems.
3. We visualize and analyze the ‘thinking process’ of the learned ConvNet, and propose a skip connection to further boost the performance.

2. Related Work

Blind inpainting [19] refers to the process of restoring the missing or corrupted areas of images or videos when their locations are not known in advance. It is more difficult compared with the traditional non-blind inpainting problem in that no information about the exact positions of the corrupted areas is provided. Therefore, blind inpainting, and blind face inpainting in particular, have seldom been addressed before. Authors in [19] resort to sparse auto encoder to learn a blind inpainting model for imposed text on

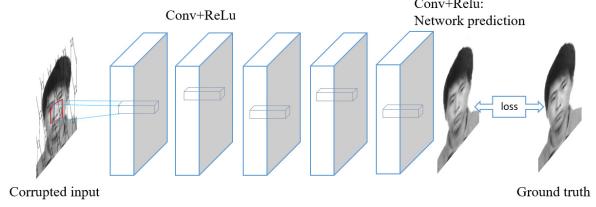


Figure 2. Network architecture of the TrivialNet. The Conv and ReLu layers are stacked to generate the predicted clean photo, the mean square loss is employed to measure the difference between network prediction and the ground truth.

natural images. A more recent work [15] also focuses on this general problem with a modified ConvNet which implements translation variant interpolation for non-blind inpainting. Rain removal, which is a special case of blind inpainting, is addressed in [6] with a specialized ConvNet. The analysis of the difference between patch based training and fully convolutional training is also presented on this paper. Besides the image inpainting task, ConvNet has also been shown to be effective in many other low-level computer vision problems, such as image denoising [4] and image super-resolution [5].

Multi-task ConvNet: By exploiting the commonality among related tasks, multi-task learning models can promote knowledge sharing among different tasks, thus boost the performance and improve the generalization ability. The feature sharing idea can be easily adapted to the ConvNet by concatenating multiple loss functions to a shared layer. Previous researches on various computer vision problems, including facial landmark detection [21], attribute prediction [1] and saliency detection [14] have demonstrated the effectiveness of the multi-task strategy in the ConvNet.

3. The Proposed Method

3.1. Multi-task Architecture

In this paper, we introduce a multi-task ConvNet that can be trained end to end for blind face inpainting. We begin the description of our proposed model with a trivial architecture. Intuitively, it is easy to develop a model that takes as input a entire corrupted photo and takes as output the prediction of the corresponding clean photo. An illustration of such network is shown on Fig 2. As implemented in this paper, all the six intermediate layers are convolutional layers with kernel size of 3×3 , filter numbers are set to 64 for all except the last layer. We pad zeros for all convolutional layers and do not introduce pooling layers to maintain the image size. In this work, we restrict our research to grayscale images, therefore, the input and output layer both have one channel.

This trivial ConvNet resembles the work of [19] which

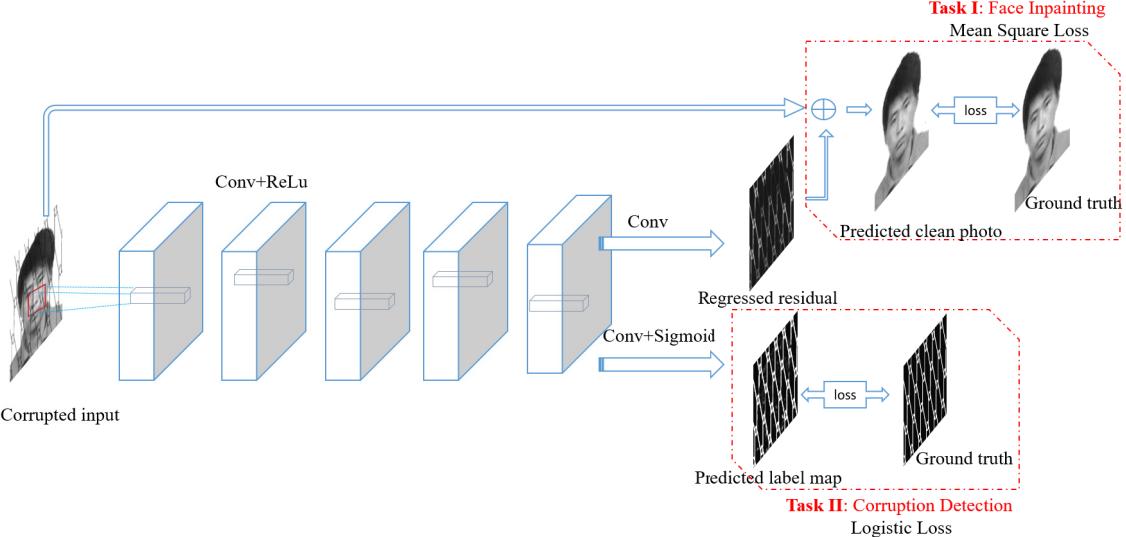


Figure 3. Network architecture of the multi-task ConvNet. The fifth Conv layer is followed by two separate Conv layers, with one outputting the predicted residual and the other outputting the predicted label map. The \oplus is an element-wise sum operation on the predicted residual and the input data. Mean squared loss and logistic loss are employed respectively for each task.



Figure 4. An example of the training sample. (a) is the corrupted photo, (b) is the clean photo and (c) is the corresponding label map.

uses a patch based sparse auto-encoder instead. However, this network only performs moderately well and it also suffers from the defect of slow convergence. Therefore, we turn to the multi-task learning paradigm to improve the convergence speed and the final performance. As emphasized in section 1, it is the purpose of this paper to simultaneously detect and reconstruct the corrupted areas. In that spirit, we introduce an auxiliary corruption detection task besides the original reconstruction task to our ConvNet to formulate a multi-task architecture.

However, these two tasks share little commonality under our current setting when the ConvNet is used to regress the entire clean image. The reason is that the reconstruction task focuses on the entire image whereas the corruption detection task only responds to the residual areas. Therefore, it is non-trivial to embed these two tasks to the current ConvNet architecture. We handle this problem with a major change in ways of solving this problem but need minor modifications in the network architecture. Specifically, instead of regressing the entire clean photos directly, we pro-

pose using the ConvNet to regress the residuals between the corrupted/clean ID photo pairs. By now, the network should attend to the residual areas directly in the main flow. Given this architecture, we can easily embed those two tasks in one shared ConvNet.

Our proposed multi-task ConvNet is illustrated in Fig 3. The detection and reconstruction task share the first five convolutional layers and have their respective sixth layer before the loss layers. By incorporating the auxiliary detection task into the ConvNet, the network could potentially learn how to distinguish between the corrupted and uncorrupted pixels and thus avoid an indiscriminative inpainting process on different types of pixel regions (inpainting on an uncorrupted region might cause blurry output as shown in Fig 9). Moreover, the multi-task paradigm is able to boost the overall performance and improve the generalization ability as indicated later in the experiments in section 5.

3.2. Learning Objectives

We denote the training example as (x, y, r, lr) , where x and y are the corrupted/clean photo pairs. r stands for the residual image $x - y$, and lr is a label map, which is a binary image with 1 indicating the corrupted areas and 0 indicating the non-corrupted ones. Fig 4 shows an example of such training data. We learn a regressor and a classifier at the same time with the ConvNet, which corresponds to the reconstruction and detection task respectively. Since the residual image $x - y$ may contain negative numbers, Conv layer without ReLu activations are employed to generate the predicted residual values in the residual regression task.

Although in the problem of blind face inpainting, the in-

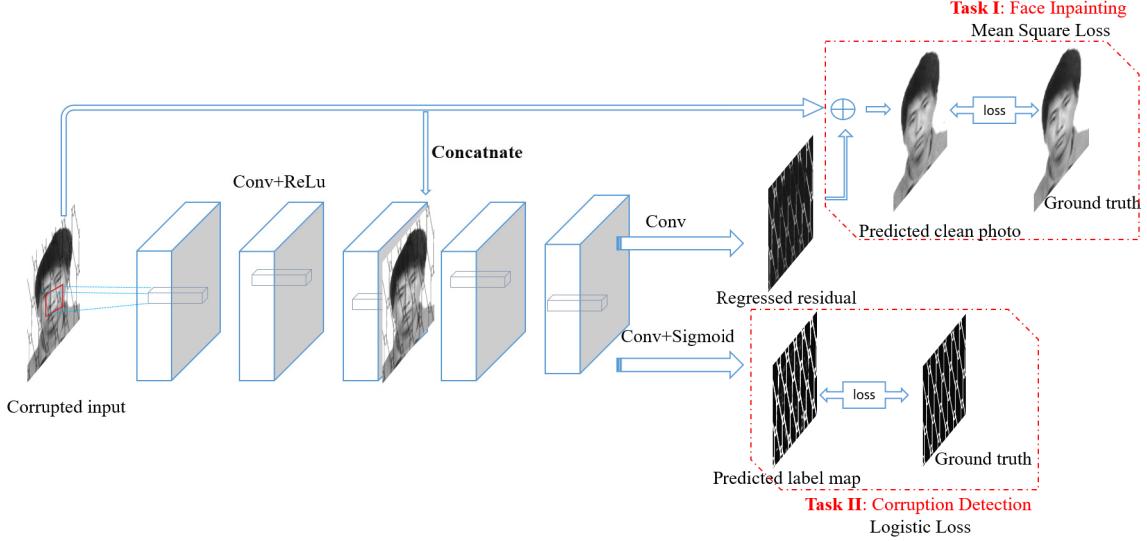


Figure 5. Network architecture of the multi-task ConvNet with skip connection. The input data layer are concatenated to the feature maps of the third ReLu layer as an extra channel.

formation on the positions of the corrupted areas cannot be provided as input to the ConvNet, it is still feasible to take the label map r as a supervision in the training phase. Because it is straightforward to get the residual image and the label map as long as the corrupted/clean photo pairs are provided.

Given the training set $\{x, y, r, lr\} \in N$, the ConvNet regressor φ and the network parameters w_1 , the training loss of the reconstruction task is the mean square error between the predicted residual and the ground truth $r = x - y$:

$$J(w_1) = \sum_{(x,y,r,lr) \in N} \sum_{i,j} \|r_{ij} - \varphi_{ij}(x)\|^2 \quad (1)$$

where φ_{ij} is the regressor that regresses the residual r_{ij} at position (i, j) .

The detection task can be regarded as training a ConvNet classifier that implement a pixel level dense prediction on whether it is corrupted. Given the ConvNet classifier ϕ and the network parameters w_2 , the logistic loss is exploited for the dense prediction problem at each position. The loss for all the data at all the positions is denoted as:

$$J(w_2) = \sum_{(x,y,r,lr) \in N} \sum_{i,j} [lr_{ij} \log(\phi_{ij}(x)) + (1 - lr_{ij}) \log(1 - \phi_{ij}(x))] \quad (2)$$

where lr_{ij} is the ground truth label indicating whether pixel at position (i, j) is corrupted or not, and ϕ_{ij} is the classifier that predicts whether pixel at position (i, j) is corrupted.

A weighting parameter α is further introduced to balance the relative importance of these two losses, and the total loss now writes $J(w_1, w_2) = J(w_1) + \alpha J(w_2)$. Note that w_1

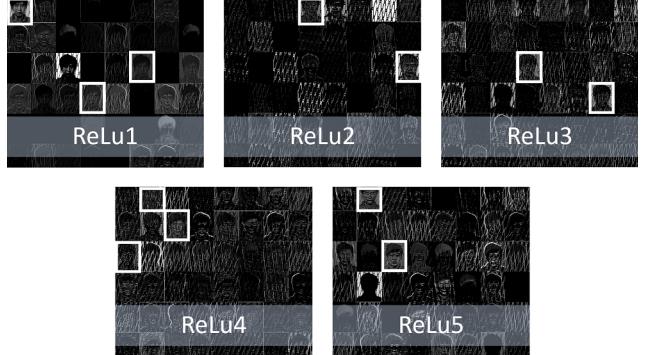


Figure 6. The 'thinking process' of our multi-task ConvNet. We visualize the responses of all the intermediate ReLU layers. The image contrast is adjusted for better visualization. Some feature maps (marked with white box) preserve the original image information.

and w_2 share most of the parameters. During the training process, we first give equal weight to each task and gradually decrease the weight for the auxiliary detection task to almost 0 while keeping the weight for the main reconstruction task fixed. We minimize the loss function with Stochastic Gradient Descent (SGD) using the Caffe [11] framework.

3.3. Skip Connection

Another feature of our proposed method is the introduction of a skip connection between the input and the intermediate feature maps. The term 'skip' refers to a structure that fuses feature maps from different layers, forming a



Figure 7. Examples of the synthetic corrupted data. Masks that exhibit different orientations, magnitudes, phases, line width and gray scale are imposed on the clean data to simulate the real data from the Chinese administration center.

inter-layer connection. When inpainting a corrupted image, traditional methods would consistently resort to neighboring pixels to restore the corrupted areas. Motivated by this scheme, we propose a skip connection that concatenates the input data as a feature map to an intermediate layer (output of ReLu3 in our case). The following layers will implement directly on this concatenated layer. One possible benefit of this architecture is that it provides more accurate context information to the subsequent layers. Without this skip connection, it is hypothesized that the ConvNet would have to learn identical mappings to preserve the context information in its intermediate layers. However, the learning of identical mapping is very difficult as depicted in the newly proposed residual learning model in [7]. By incorporating the skip connection, the context information would come at much lower cost.

As a proof of the necessity of the skip connection, we visualize the ‘thinking process’ of our multi-task ConvNet. In detail, we illustrate sample activations for each intermediate layers in Fig 6. As envisioned, most feature maps only respond to the residual areas. But it is also noted that a small portion of the feature maps (marked with white box) do maintain the original image information for further reconstruction. This reflects a coupled processing of detection and reconstruction in the ConvNet, and verifies our hypothesis that the ConvNet learns identical mapping to preserve the context information in its intermediate layers. By introducing the skip connection, we incorporate the original input to the intermediate layers, which can provide more accurate context information for further reconstruction.

3.4. Test-Time Evaluation

Our model processes the blind face inpainting problem with a single end to end ConvNet. At test time, we only need to calculate one pass of forward propagation on a corrupted ID photo of size 220×178 . Since the auxiliary task does not need to be computed at test time, the multi-task ConvNet will not introduce extra time consumption. With GPU acceleration on one NVIDIA GTX Titan, our approach runs at more than 100 frames per second. When running on a PC with core i7-3770 and 8G RAM, our network can still processes 1 frame in one second, whereas tradi-

tional methods would take minutes. It is totally viable to integrate our approach as a preprocess step in a real time FVBIL system with GPU acceleration. With recent advances in network compression [10], our ConvNet should be able to work much faster with only CPU in the future.

4. Dataset Collection

A large dataset is needed to enable the training of our multi-task ConvNet for real-life blind face inpainting problem. Since large scale corrupted/clean ID photo pairs are not accessible, we would have to come up with a procedure to generate the synthetic corrupted/clean ID photo pairs from a collection of clean ID photos.

We simulate the random patterns with sine and cosine waves with different magnitudes and phases, and random transformations are also introduced to increase the diversity of the patterns. Gaussian filtering is implemented on the binary pattern image (the grayscale of imposed patterns is 0 and that of other regions is 1) to generate random patterns with pixel values analogous to the actual distribution. The generated image is treated as a mask image. We add the mask to the clean ID photo to generate a corrupted ID photo with the following formula:

$$y = \begin{cases} \beta M + (1 - \beta)x & \text{if } M < 1 \\ x & \text{if } M = 1 \end{cases} \quad (3)$$

where, M is the mask image and β is a parameter drawn from a uniform distribution over $[0.1, 0.9]$ and controls the transparency of the patterns. The above equation means that only the corrupted areas are added to the clean photo, and the other regions remain the same during the procedure. We collect 11,648 clean ID photos to generate the synthetic dataset. 50 random corrupted images are synthesized for each individual. Finally, we get over 500,000 data pairs for the training process. Some examples of the synthetic data are shown in Fig 7.

For the FVBIL experiment, we collect another 1 : 1 synthetic corrupted ID photo/ life photo pairs of 500 individuals. We will call this dataset ‘SYN500’ in the next section. Due to the scarcity of test data, we conduct full comparison on this collected dataset. Since much more negative samples are tested than positive samples when conducting full comparison, the accuracy is not an ideal metric to evaluate the verification performance. Instead, we compare the TPR@FPR for fair evaluation. To further test the performance on unseen data, we also collect 300 ID photos corrupted by another institution (not the Chinese administration center, and due to copyright issues, we could not show these ID photos on this paper) and their corresponding life photos. Similarly, we call this dataset ‘US300’.

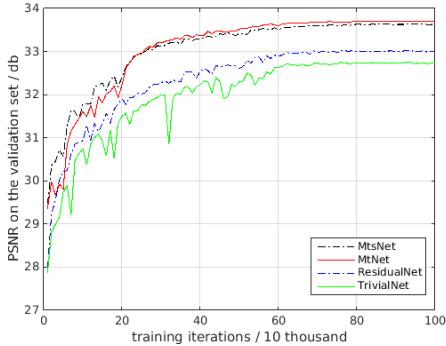


Figure 8. PSNR on the validation set as the function of training iterations (in ten thousand).

5. Experiments

5.1. Baseline Methods

Few algorithms have been proposed to address the inpainting problem, and none explicitly for the blind face inpainting problem. Therefore, in our work, we will compare our multi-task ConvNet with skip connection (MtsNet) and the multi-task ConvNet without the skip connection (MtNet) with (1) the trivial network (TrivialNet) introduced in section 3.1, (2) the trivial network that regress the residuals (ResidualNet) instead of the clean photos. Specifically, the TrivialNet can be seen as an variation of the compared method in [15], which has been shown to outperform the auto-encoder based method [19]. Since the method proposed in [15] focuses on non-blind inpainting which needs the corrupted regions to be provided in advance, we do not compare with their method (and other non-blind inpainting methods) in this work. The inpainting performance is evaluated with both quantitative metric (PSNR) and qualitative visual results.

5.2. Implementation Details

Training All the image pairs are resized to 220×178 . A total of 200 individuals are used for validation, and the rest are used for training. No crop, flip or rotation are implemented since the amount of the training data is abundant to train such a medium sized network. For all the compared network structure, the batch size is 30, momentum is set to 0.9 and weight decay is 0.0005 for all layers. The learning rate is set to 10^{-6} initially, and decreased by a factor of 10 to 10^{-8} after the loss in the validation set stops to decrease. For our multi-task ConvNet, the weight for the detection loss is set to 0.5 and decreased to 0.01 after 200K iterations.

ConvNet for facial feature extraction We train a deep ConvNet for the extraction of deep facial features. In detail, a model introduced by [18] is trained on the CASIA-WebFace dataset [20] and achieves an accuracy of 98.1%

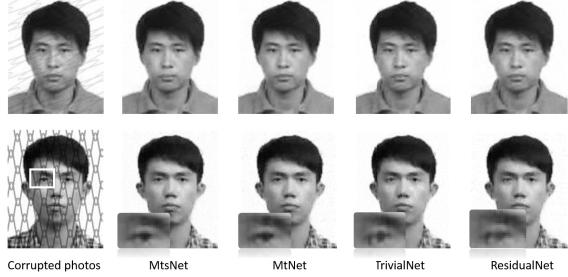


Figure 9. Visual results for different models. MtsNet can restore the corruption with more smooth textures. But TrivialNet and ResidualNet tend to make the entire image very blurry.

on the LFW [9] benchmark. We then finetune the network on a large ID/life photo dataset with triplet loss [16]. After that, this finetuned model is used to extract features for the FVBIL task. Cosine distance is employed to compare the similarity between face samples.

5.3. Inpainting Results

We compare the quantitative inpainting results for the synthetic dataset in this subsection. Specifically, PSNR for the validation set, which consists of 10,000 corrupted images from 200 individuals, are calculated. We report the PSNR of different training iterations in Fig 8. The weight for the detection task is set to 0.5 at first and reduced to 0.01 and 0.001 at 200K and 600K iterations respectively for both MtNet and MtsNet. The learning rate is decreased to 10^{-7} and 10^{-8} at 600K and 800K iterations respectively for all the compared ConvNets.

As illustrated in Fig 8, with the multi-task paradigm, the MtsNet and MtNet reach better PSNR after 10K iterations of training and continue to perform better than TrivialNet and ResidualNet. This suggests that the multi-task learning strategy speeds up the convergence and boosts the overall performance. At 200K iterations, we lower the loss weight for the detection task and observe a big leap in the performance and a relatively faster saturation soon after. This further verifies the multi-task learning strategy contributes to find a better local minimum that can benefit both the convergence speed.

The curves in Fig 8 also suggest that the skip connection indeed improves the performance before 200K iterations. But no evident improvement is observed after 200K iterations, and it even performs slightly worse than MtNet after 300K iterations. However, verification performance in the next subsection indicates that MtsNet still performs better than MtNet. Another phenomenon that is worth mentioning is that the ResidualNet consistently performs better than the TrivialNet. This can be considered as a validation of the theory proposed in [7] that learning an identical mapping is much harder than learning the residuals for the ConvNet.

Table 1. Evaluation of the FVBIL task.

	TPR@FPR=1%	TPR@FPR=0.1%	TPR@FPR=0.01%	PSNR
MtsNet	80.60	58.40	30.20	32.83
MtNet	81.00	57.80	29.20	32.89
TrivialNet	77.20	51.60	29.00	31.87
ResidualNet	79.60	55.00	29.20	32.24
Corrupted	49.00	30.80	16.40	22.66
Clean	91.20	75.20	50.20	-

5.3.1 Visual Results

We compare the qualitative visual results of different models in Fig 9. It is observed that both multi-task based ConvNets can remove more corruption areas than their single task alternatives. What's more, in the second row of Fig 9, the textures of the right eye is well preserved by both multi-task based ConvNets but are severely blurred by TrivialNet and ResidualNet. This suggests an indiscriminative inpainting process implemented by TrivialNet and ResidualNet. But the proposed MtsNet and MtNet can better identify the positions that needs to be inpainted and thus keep the uncorrupted regions intact. With a closer look at the first row of Fig 9, we can see that MtsNet can restore the corrupted areas with more smooth textures than MtNet. This is possibly because that the skip connection could provide more accurate context information for the reconstruction process.

As in [6], we also test our model on data with different statistics. To that end, we conduct experiments on the 'US300' dataset. As declared in [6], the quality of the results does depend on the statistics of test cases being similar to those of the training set. Therefore, the performance on this dataset is expected to be less appealing than that on our synthetic data. As anticipated, although most of the random patterns are removed and restored, more artifacts can be observed on this dataset. However, our trained network can well restore the background areas in most cases. This is crucial for increasing the detection rate of the corrupted ID photos and will in turn improve the verification performance. We will see more concrete results in the next subsection where the FVBIL task is considered. By expanding the diversity in patterns and distributions of the corrupted pixels of the training data, we would be able to further boost the performance on unseen data.

5.4. Verification Results

Due to the heterogeneous sources and different capturing conditions of ID photos and life photos, they exhibit large variations in poses, illumination conditions and facial expressions. These variations have made the FVBIL task one of the most challenging face recognition tasks. When we conduct full comparison on the 'SYN500' dataset with clean ID photo/life photo pairs, the TPR@FPR=1% (true positive rate when false positive rate is %1) is about 91.2%. But when we corrupt the ID photos with random wavy lines using our approach, the performance deteriorates severely and drops to 49.0%. This phenomenon has illustrated the

Table 2. Evaluation of the FVBIL task on unseen data.

	TPR@FPR=1%	TPR@FPR=0.1%	TPR@FPR=0.01%
MtsNet	68.67	46.67	27.33
MtNet	68.00	47.67	27.33
TrivialNet	67.67	42.00	26.67
ResidualNet	65.67	45.00	23.33
Corrupted	34.00	20.00	9.30

significance of our research on blind face inpainting.

Since the ultimate goal of blind face inpainting is to improve the verification performance, we will use all the compared ConvNets (trained after 1000K iterations) to inpaint the corrupted ID photos and conduct FVBIL task with the processed images. We report the verification performance in Table 1. All the compared methods can significantly improve the verification performance, among which our proposed MtsNet and MtNet performs the best. We also calculate the PSNR on this dataset and report them in the last column of Table 1. It is clear that there is a positive correlation between the PSNR and the verification performance.

It is also noted that although the verification performance has been significantly improved by blind face inpainting algorithms, there is still much room for improvement compared with the performance on clean photos. One possible approach to this is to finetune the ConvNet for feature extraction with the processed corrupted photos (by our proposed MtsNet, for instance), making it robust to subtle noise on the processed images. Another angle for dealing with this is to refine the detection [13] and alignment performance with the inpainted face photos.

As done in the last subsection, we also conduct FVBIL experiment on the unseen data 'US300' and report the results in Table 2. Even though the inpainting results is not as promising as that on our synthetic data, it is not surprising to see that our blind face inpainting approach can still significantly improve the verification performance. As mentioned above, our proposed methods remove most of the imposed patterns in the background areas, which is crucial to increase the face detection rate and finally improve the verification performance. The performance on unseen data can reflect the generalization ability of different models to a certain degree. Judging from the verification performance, we can see the multi-task based ConvNets generalize better than those two single task models. However, contrary to the results on synthetic data, ResidualNet seems to generalize worse than the TrivialNet on the unseen data.

6. Conclusion

In this work, to improve the face verification accuracy between the ID photo and life photo, we study the problem of blind face inpainting. Specifically, we have proposed a multi-task ConvNet with skip connection to inpaint the randomly corrupted ID photos. As indicated by both the inpainting performance and verification accuracy, our pro-

posed method has great advantages over its single task alternatives. However, the results on unseen data is not as promising as that on our synthetic data due to the difference in data distribution. This can be alleviated by generating more statistically similar training data and also remains to be an interesting problem for future work.

Although applied to blind face inpainting problem in this work, our proposed method is readily generalized to other blind and non-blind inpainting problems. To develop algorithms that specifically cater to blind face inpainting in the future, we will try to incorporate semantic information and take the facial structures into consideration.

Acknowledgements

This work is supported by the Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS) (Grant No. 2015190), the National Natural Science Foundation of China (Grant No. 61473289) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02070000).

References

- [1] A. Abdunabi, G. Wang, J. Lu, and K. Jia. Multi-task cnn model for attribute prediction. *Multimedia, IEEE Transactions on*, 17(11):1949–1959, Nov 2015. [2](#)
- [2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013. [1](#)
- [3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. [1](#)
- [4] H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2392–2399, 2012. [2](#)
- [5] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of European Conference on Computer Vision*, 2014. [2](#)
- [6] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 633–640, 2013. [2](#), [7](#)
- [7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. [5](#), [6](#)
- [8] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong. Two-stage nonnegative sparse representation for large-scale face recognition. *Neural Networks and Learning Systems, IEEE Transactions on*, 24(1):35–46, 2013. [1](#)
- [9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. [1](#), [6](#)
- [10] M. Jaderberg, A. Vedaldi, and A. Zisserman. Speeding up convolutional neural networks with low rank expansions. In *Proceedings of the British Machine Vision Conference. BMVA Press*, 2014. [5](#)
- [11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014. [4](#)
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. [1](#)
- [13] Q. Li, Z. Sun, R. He, and T. Tan. Learning symmetry features for face detection based on sparse group lasso. In *Chinese Conference on Biometric Recognition*, pages 162–169. Springer, 2013. [7](#)
- [14] X. Li, L. Zhao, L. Wei, M. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *arXiv preprint arXiv:1510.05484*, 2015. [2](#)
- [15] J. S. Ren, L. Xu, Q. Yan, and W. Sun. Shepard convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909, 2015. [2](#), [6](#)
- [16] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. [1](#), [6](#)
- [17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. [1](#)
- [18] X. Wu, R. He, and Z. Sun. A lightened cnn for deep face representation. *arXiv preprint arXiv:1511.02683*, 2015. [1](#), [6](#)
- [19] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012. [2](#), [6](#)
- [20] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. [6](#)
- [21] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proceedings of European Conference on Computer Vision*, 2014. [2](#)
- [22] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015. [1](#)