Look Through Masks: Towards Masked Face Recognition with De-Occlusion Distillation

Chenyu Li
Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Shiming Ge*
Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Daichi Zhang
Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Jia Li
State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, Beijing 100191, China
Peng Cheng Laboratory, Shenzhen 518055, China

ABSTRACT

Many real-world applications today like video surveillance and urban governance need to address the recognition of masked faces, where content replacement by diverse masks often brings in incomplete appearance and ambiguous representation, leading to a sharp drop in accuracy. Inspired by recent progress on amodal perception, we propose to migrate the mechanism of amodal completion for the task of masked face recognition with an end-to-end de-occlusion distillation framework, which consists of two modules. The *de-occlusion* module applies a generative adversarial network to perform face completion, which recovers the content under the mask and eliminates appearance ambiguity. The *distillation* module takes a pre-trained general face recognition model as the teacher and transfers its knowledge to train a student for completed faces using massive online synthesized face pairs. Especially, the teacher knowledge is represented with structural relations among instances in multiple orders, which serves as a posterior regularization to enable the adaptation. In this way, the knowledge can be fully distilled and transferred to identify masked faces. Experiments on synthetic and realistic datasets show the efficacy of the proposed approach.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Object recognition*; *Knowledge representation and reasoning*.

KEYWORDS

Masked Face Recognition; Amodal Completion; Generative Adversarial Networks(GANs)

*Shiming Ge is the corresponding author (geshiming@iie.ac.cn).

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413960

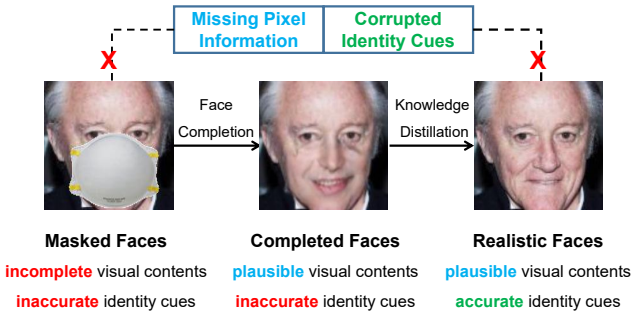

Figure 1: Inspired by the mechanism of amodal perception, we propose to solve masked face recognition via de-occlusion distillation that first enforces face completion, then inherits rich knowledge from pre-trained recognizor via distillation. In this way, both incomplete visual contents and inaccurate identity cues can be well recovered.

ACM Reference Format:
Chenyu Li, Shiming Ge, Daichi Zhang, and Jia Li. 2020. Look Through Masks: Towards Masked Face Recognition with De-Occlusion Distillation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413960

1 INTRODUCTION

Human faces in the wild are often occluded by masks, intentionally or unintentionally. The ability to handle recognition towards masked faces is essential for many visual applications, *e.g.* video surveillance [19] and urban governance [16]. In the last few years, the deep-learning-based face recognition models [3, 47, 49, 50, 52, 61] have been able to achieve or even exceed human-level performance on public benchmarks. This has much attributed to our growing understanding of how our brain may be solving the identity recognition task. However, face recognition under more challenging conditions, such as masked faces, is less characterized. The question is how to represent these masked parts of perceived faces: this is the problem of amodal perception [37].

To facilitate recognition towards masked faces, intuitive methods [54, 56] that seek to extract credible features from visible regions permit direct mapping between the input partial observation in the source domain and expected output in the target domain. They usually split the input image into several local parts and predict the possibilities of each being masked, on which the allocated weights for local parts are based. These approaches are supported by biological neural science conclusions that masked objects are selectively perceived as disconnected elements [5, 22].

Recent researches have proved amodal completion indispensable in perceiving partly-observed objects. In [5] Chen *et al.* revealed that amodal completion is first manifested in face-selective areas, then implemented through feedback and recurrent processing among different cortical areas. Via attention mechanism and context restraints, many effective algorithms for occluded object recognition have been proposed [20, 52]. However, different from general object or shape recognition, facial identity relies on both shape/view and appearance [4]. The latter involves more complexity and ambiguity when being masked. Given an example, with an image of a man with a face mask, it's rather easy to tell from isolated parts like eyes that there is a man, while difficult to tell his identity. Amodal perception for masked faces is hard to achieve by conventional Deep Neural Networks (DNNs), where completion is done implicitly.

We propose to explicitly regularize the amodal completion process. The most related task is face completion. Recent approaches [18, 26, 41, 57] based on Generative Adversarial Networks (GANs) [11], which formulate completion as a conditional image generation problem, have evolved as compelling tools that generate photo-realistic results. However, recently Joe *et al.* [36] looked into the question that whether generative face completion helps recognition. Experimental results showed that, though completion greatly please biometric systems, the benefit for recognition is limited. We suspect that it arises from the innate selectivity-invariance pattern of CNNs, where higher layers of representation amplify shared aspects and suppress irrelevant variations [23]. General constraints on efficient generative completion naturally lead to an axis representation with strong appearance bias. There exists a clear domain gap between the inexact visual imagery [38] and realistic faces.

In this work, inspired by the amodal completion mechanism in the human brain, we propose a novel de-occlusion distillation framework to deal with the task of masked face recognition, as shown in Fig. 1. The model consists of two main modules, *de-occlusion* and *distillation*. The de-occlusion module applies a GAN-based face completion network to eliminate the appearance ambiguity and enables the masked face to be perceived as a whole. The attention mechanism is introduced to teach the model to "look" at informative areas. Then to subsequently benefit recognition, the distillation module takes a pre-trained general face recognition model as the teacher and adapts its knowledge to completed faces through knowledge distillation. Recently Ge *et al.* [9] employed identity-centered regularization and gained an effective accuracy boost, which inspired us to exploit in deep generative models rich problem structures and domain knowledge. Assuming the distribution of unmasked faces could provide essential guidance, we represent the teacher knowledge with structural relations among instances. Via enforcing various orders of structural similarities to provide a posterior regularization, the student learns to perform accurate recognition towards completed faces. We evaluate the proposed method on both synthetic masked face datasets (Celeb-A [32] and LFW [17]) and realistic masked face datasets (AR [35]), both showing compelling improvements on recognition accuracy.

Our main contributions can be summarized as three folds: 1) We propose a novel end-to-end framework for masked face recognition, which first enforces face completion explicitly and then transfer rich domain knowledge from pre-trained general face recognition model via knowledge distillation; 2) We introduce the theory of amodal perception to shed light on the masked face recognition task, and our empirical results echo the theory and 3) We conduct extensive experiments to demonstrate the efficacy of our approach.

2 RELATED WORKS

2.1 Amodal Perception and Face Completion

Humans are able to recognize objects even when they are partially occluded by another pattern, so easily that one is usually not even aware of the occlusion. The phenomena of completion of partly occluded shape have been termed "amodal perception" [37], since the occluded contours are not seen. Kovacs *et al.* [22] found that single IT units remain selective for shape outlines under a variety of partial occlusion conditions, physiologically locating where amodal perception happens for the first time. So one last question we care about is: how the occluded contents are represented. In [22] the discrimination performance was found much better when they were familiar with the subjects. This suggests that amodal perception relies heavily on our background knowledge of how the occluded parts of the object (may) look. They also find that the IT cells only respond to selective fragments, and conclude that amodal completion doesn't happen. Chen *et al.* [5] delves into the time course of amodal completion in face perception. Their results suggest amodal completion is first manifested in face-selective areas, then implemented through feedback and recurrent processing among different cortical areas. Therefore, amodal completion plays an indispensable role in perceiving partly-observed objects.

Face completion, or inpainting, aims to recover masked or missing regions on faces with visually plausible contents. Traditional exemplar-based approaches [1, 12] searched similar patches as reference for the synthesis of missing regions. While this non-parametric manner achieves good results when similar content is available, the mechanism is not scalable for objects with unique textures, *e.g.* faces. Recently, the GAN-based architecture has been widely adopted in completion with visually satisfactory results [18, 26, 41, 44, 58]. They usually train an auto-encoder to predict the missing region using a combination of reconstruction loss and adversarial loss. Despite their capacity in recovering high-quality visual patterns, the recognition accuracy gain is still limited [36].

2.2 Occluded Object Recognition

Partial occlusions are one of the greatest challenges for many vision tasks, *e.g.* classification [20], recognition [61], and person re-ID [27]. Various approaches have been proposed to solve the problem, following "representation" or "representation" idea. The "representation" idea seeks to obtain robust representations for occluded objects by decreasing or excluding the influence of missing regions and tapping the useful information. Some methods first segmented

a face image into several local parts and then described the face using the ordered property of facial parts [52] or extracting discriminative components [20, 25]. Other methods directly take the whole face image as input instead, and represent it with a good descriptor, such as sparse representation [56] and low-rank regularization [42].

Different from the "representation" idea, the "reconstruction" idea utilizes the redundancy of images and performs information recovery before recognizing. Deng *et al.* [8] proposed an exemplar-based Graph Laplace algorithm to complete masked faces. In this way, the approach can use the completed faces to boost the recognition accuracy. It performs well when a similar appearance and expression can be found in the library. However, the type and shape of the occlusions are innumerable and unpredictable in real scenarios, which limits its applications. More recently, GAN-based face completion approaches [18, 26, 41, 44, 55, 57, 58] have achieved remarkable improvement in extracting high-level contextual representations and generating photo-realistic results. However, the identity consistency during completion is less considered. [60, 61] enforce identity preservation through perceptual loss. To exploit structural domain knowledge, [45] proposes a structural loss to constrain the structure of the generated image. Alternatively Ge *et al.* [9] exploit structural domain knowledge in feature space and employ identity-centered regularization.

2.3 Knowledge Distillation and Transfer

Transfer learning aims to mitigate the burden of manual labeling for machine learning by transferring information between different domains or tasks. The most common approach is to fine-tune models pre-trained on public datasets like ImageNet [6] for specific tasks with labeled data. Recently, as a special branch in transfer learning, knowledge distillation has gained much interest and exhibited remarkable capability in knowledge transfer. Knowledge distillation was first introduced by [2] and [15] presented a more general approach within the scope of a feed-forward neural network. By using the softmax output of the teacher network as soft labels instead of hard class labels, the student model can learn how the teacher network studied given tasks in a compressed form. Romero *et al.* [46] improved the method by using not only the final output but also intermediate hidden layer values of the teacher network to train the student network. To encourage the diversity of learning, Luo *et al.* [34] utilized the ensemble of multiple networks as the teacher to train a compact student network for face recognition. All these methods assumed that the input data of the teacher and student model are from the same domain. To boost the domain adaptation task, Su and Maji [48] proposed cross quality distillation to learn models for recognizing low-resolution images, non-localized objects, and line-drawings by using soft labels of high-resolution images, localized objects, and color images, respectively. Radosavovic *et al.* [43] proposed data distillation to ensemble predictions from multiple transformations of unlabeled data to automatically generate new training annotations.

3 DE-OCCLUSION DISTILLATION

In this section, we first provide an overview of our proposed approach, then describe the details of each network component as well as the loss functions.

3.1 Problem Formulation

Masked faces are faces that not fully observed. In this section, we dissect the problem of masked face recognition and try to provide simple yet solid explanations for two questions: i) What help does data recovery do? ii) What needs to do after data recovery?

Here we describe the generative process for partly observed data, following the setting of missing data processing [28]. Let $\mathbf{X} \in \mathbb{R}^n$ be a data vector and $\mathbf{M} \in \{0,1\}^n$ is a binary mask indicating which entries in $\mathbf{X}$ to reveal: x_d is observed if $m_d = 1$ and vise versa.

$$\mathbf{X} \sim p_\theta(\mathbf{X}),\ \mathbf{M} \sim p_\varphi(\mathbf{M}|\mathbf{X}), \tag{1}$$

where θ denotes the parameters of data distribution and φ denotes the parameters of the mask distribution. The mask distribution is usually assumed to depend on the data $\mathbf{X}$. Let $\mathbf{X}_o$ denote the observed elements of $\mathbf{X}$, and $\mathbf{X}_m$ denote the missing elements according to the mask $\mathbf{M}$. We define the target attribute as $\mathbf{a}_t$. In the standard maximum likelihood setting, the unknown parameters are estimated by maximizing the following marginal likelihood, integrating over the unknown missing data values:

$$p(\mathbf{a}_t, \mathbf{X}_o) = \int p_\theta(\mathbf{X}_o, \mathbf{X}_m) \cdot p_\varphi(\mathbf{M}|\mathbf{X}_o, \mathbf{X}_m) \cdot p_\psi(\mathbf{a}_t|\mathbf{X}_o, \mathbf{X}_m)\, d\mathbf{X}_m, \tag{2}$$

where $p_\psi(\mathbf{a}_t|\mathbf{X}_o, \mathbf{X}_m)$ is a recognizor that gives prediction with $\mathbf{X}_m$ replacing the missing region. Due to the ambiguity introduced by masks, the optimization process involves integration over literally infinite possible $\mathbf{X}_m$. Even with the remarkable capacity of well-crafted DNNs, it is difficult to reach convergence. One simple technique is to multiply with an impulse function $\delta(\mathbf{X}_m - \hat{\mathbf{X}}_m)$:

$$\delta(\mathbf{X} - \mathbf{X}_0) = \begin{cases} \infty, & \mathbf{X} = \mathbf{X}_0 \\ 0, & \mathbf{X} \neq \mathbf{X}_0 \end{cases} \tag{3}$$

The physical meaning for this operation is to select the best restoration $\hat{\mathbf{X}}_m$ for the missing data based on certain criterions, *e.g.* the coherence with the observed data regarding the wearing mask, measured by $p_\theta(\mathbf{X}_o, \mathbf{X}_m)p_\varphi(\mathbf{m}|\mathbf{X}_o, \hat{\mathbf{X}}_m)$. In this way, the optimization problem in Eq. 2 can be simplified as:

$$\begin{aligned} p(\mathbf{a}_t, \mathbf{X}_o) = \int\ & p_\theta(\mathbf{X}_o, \mathbf{X}_m) \cdot p_\varphi(\mathbf{m}|\mathbf{X}_o, \mathbf{X}_m) \\ & p_\psi(\mathbf{a}_t|\mathbf{X}_o, \mathbf{X}_m) \cdot \delta(\mathbf{X}_m - \hat{\mathbf{X}}_m)\, d\mathbf{X}_m \\ =\ & p_\theta(\mathbf{X}_o, \hat{\mathbf{X}}_m) \cdot p_\varphi(\mathbf{m}|\mathbf{X}_o, \hat{\mathbf{X}}_m) \cdot p_\psi(\mathbf{a}_t|\mathbf{X}_o, \hat{\mathbf{X}}_m), \end{aligned} \tag{4}$$

$$\hat{\mathbf{X}}_m = \arg\max_{\mathbf{X}_m} p_\theta(\mathbf{X}_o, \mathbf{X}_m) \cdot p_\varphi(\mathbf{m}|\mathbf{X}_o, \hat{\mathbf{X}}_m), \tag{5}$$

which turns out to be proportional to the prediction towards the completed data $p_\psi(\mathbf{a}_t|\mathbf{X}_o, \hat{\mathbf{X}}_m)$, as the former two terms can be seen as constants once $\hat{\mathbf{X}}_m$ is decided. Via data recovery, we turn the intricate problem in Eq. 2 into two sub-problem: finding the best restoration $\hat{\mathbf{X}}_m$ and acquiring accurate prediction $p_\psi(\mathbf{a}_t|\mathbf{X}_o, \hat{\mathbf{X}}_m)$. This answers for our first question.

Naturally we wonder if $p_\psi(\mathbf{a}_t|\mathbf{X}_o, \hat{\mathbf{X}}_m)$ could adopt a pre-trained state-of-the-art recognition model for faces in the wild. Following our former analysis, the optima $\hat{\mathbf{X}}_m$ is obtained by solving a maximum optimization. Numerical solutions for these high-order space concerning optimizations are not available. In practice, we approach

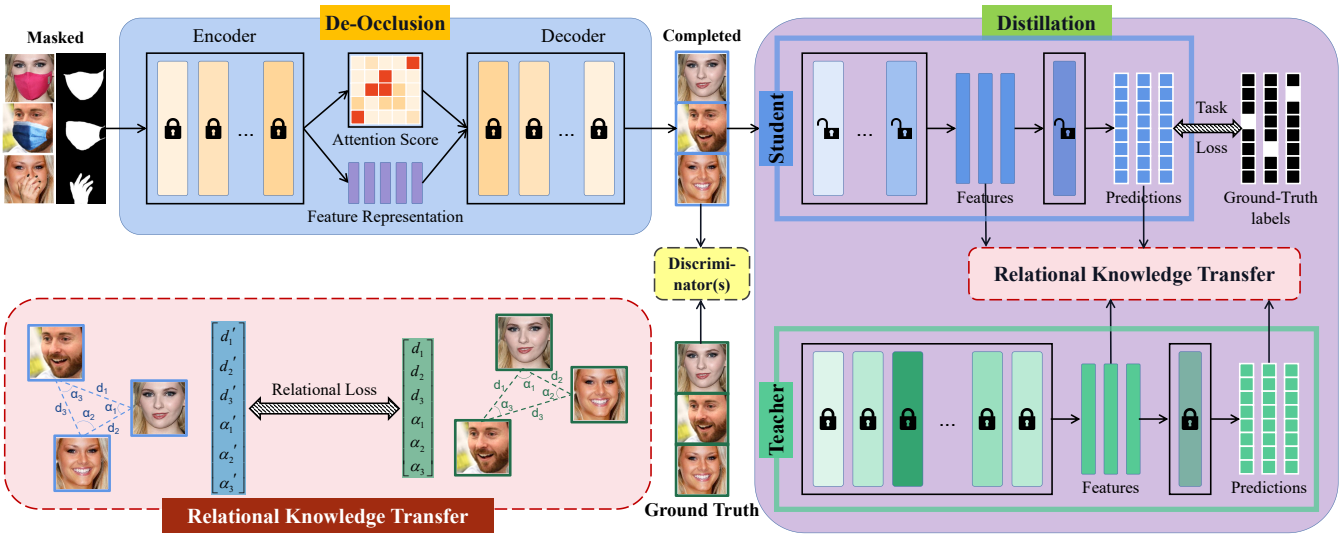

Figure 2: Overview of our proposed framework. The learning process consists of two stages. In the first stage, we initialize the input of the student model via an inpainting model. In the second stage, we do cross-quality knowledge distillation and transfer the knowledge contained in the teacher recognizor for normal faces into student recognizor by enforcing relational structure consistence. In this manner, the student network for recognizing masked faces learns representations for completed faces with the same clustering behaviors as the original ones, which could greatly benefit recognition accuracy.

the optima in a data-driven way. This leads to a recovery process that is logically tractable while not accurate enough. Experiments in [36] support our theory. As it claims, though face completion seems pleased for a biometric system, for recognition tasks, its contribution is limited. The tractability decreases the representation capacity of the model, making it hard to deal with the turbulence in feature space, brought by varied masks. The deficiency of accuracy also suggests the domain gap regarding latent identity features. To sum up, a general recognition model is not the best fit for recognizing completed faces, yet a knowledgeable adaptation source.

In this work, we propose to solve masked face recognition with **MFR** = $\{\mathcal{G};\mathcal{R}\}$, where $\mathcal{G}$ and $\mathcal{R}$ denote a de-occlusion module and a distillation module separately. The missing information is recovered first in image space via inpainting-based de-occlusion, then in feature space via knowledge distillation. We use a de-occlusion module $\mathcal{G}(\mathbf{X},\mathbf{M};\mathbb{W}_\mathcal{G})$ to approximate the maximum optimization in Eq. 5. It takes masked faces $\mathbf{X}$ as input, and aims to generate completed faces $\tilde{\mathbf{X}}$ by realistically approximating the ground-true faces $\mathbf{Y}$, where $\mathbf{M}$ denotes the binary masks labeling the masked regions with 1 inside Ω and $\mathbb{W}_\mathcal{G}$ is the set of model parameters. The distillation module adopts a teacher-student scheme to distill knowledge from a pre-trained teacher network $\mathcal{R}_t(\mathbf{Y};\mathbb{W}_t)$ for general faces $\mathbf{Y}$ into a simpler student network $\mathcal{R}_s(\tilde{\mathbf{X}};\mathbb{W}_s)$ for completed faces $\tilde{\mathbf{X}}$ by transferring structural relational knowledge. Here, $\mathbb{W}_t$ and $\mathbb{W}_s$ refer to the model parameters for the teacher and student, respectively. In this way, we formulate the final goal function as:

$$\max_{\mathbb{W}_\mathcal{G},\mathbb{W}_\mathcal{R}} \mathcal{R}(\mathbf{a}_t|\mathbf{X},\mathcal{G}(\mathbf{X},\mathbf{M};\mathbb{W}_\mathcal{G});\mathbb{W}_\mathcal{R}) - \mathbb{E}(\mathcal{R}_t(\mathbf{Y};\mathbb{W}_t),\mathcal{R}(\mathbf{X},\mathcal{G}(\mathbf{X},\mathbf{M};\mathbb{W}_\mathcal{G});\mathbb{W}_\mathcal{R})). \tag{6}$$

3.2 Appearance Recovery via Completion

In the de-occlusion module, we explicitly enforce amodal completion via a generative face completion model. First, it's important to emphasize that amodal perception is manifested in face-selective areas, and masked faces are perceived as disjointed segments. Attention plays a very important role here. We adopt the same architecture as in [57], which consists of a generator for inpainting and two auxiliary discriminators for regularizing from local and global views, with a contextual attention mechanism.

Generator Given an image of a masked face and a binary mask indicating the missing regions, the generator $\mathcal{G}(\mathbf{X},\mathbf{M};\mathbb{W}_\mathcal{G})$ aims to generate a photo-realistic result as similar with the ground-truth as possible. To achieve that, a pixel-wise reconstruction loss is employed to penalize the divergence, formulated as:

$$\mathcal{L}_\mathcal{G} = \ell_1(\tilde{\mathbf{X}},\mathbf{Y}) = \ell_1(\mathcal{G}(\mathbf{X},\mathbf{M};\mathbb{W}_\mathcal{G}),\mathbf{Y}). \tag{7}$$

Local and Global Discriminators Two discriminate networks are adopted to identify whether input images are real or fake from global and local views, respectively. The global discriminator takes the whole image as input, while the local one uses the completed region only. Contextual information from local and global views compensate each other, eventually reaching a balance between global consistency and local details. They regularize the generator via local and global adversary losses:

$$\mathcal{L}_{\mathcal{D}_i} = \min_\mathcal{G}\max_{\mathcal{D}_i} \mathbb{E}[\log \mathcal{D}_i(\mathbf{Y}) + \log(1-\mathcal{D}_i(\mathcal{G}(\mathbf{X},\mathbf{M});\mathbb{W}_\mathcal{G})],\ i\in\{g,l\}, \tag{8}$$

where $\mathcal{D}_i$ denotes the global discriminator when $i=g$ and the local discriminator when $i=l$.

Contextual Attention The contextual attention layer enables the generator to refer to features from the whole image and to learn long-distance semantic dependencies. It computes the similarity of patches centered in missing pixel (m, n) and observed pixel (p, q):

$$s_{p,q,m,n} = \langle \frac{x_{p,q}}{\|x_{p,q}\|}, \frac{\tilde{x}_{m,n}}{\|\tilde{x}_{m,n}\|} \rangle, \tag{9}$$

where x and $\tilde{x}$ denotes the masked and the completed face image, separately. The calculated similarities are then send through a softmax layer to obtain attention score for each pixel $s^*_{p,q,m,n} = softmax_{m,n}(\lambda \cdot s_{p,q,m,n})$, where λ is a constant value. Finally the image contents are reconstructed by performing de-convolution on attention score. The contextual attention layer is differentiable and fully convolutional. Implementation details refer to [57].

3.3 Identity Recovery via Distillation

The last stage has recovered missing visual contents via GAN-based face completion. Experiments suggest activations responsible for amodal completion happen in the same place where cells are activated when we visualize objects with our eyes closed [21]. It is easy to accept that between an actual visual stimulus and visual imagery, there exists a non-ignorable domain gap. We here raise our insight that, between the ground truth and the heuristic completion results, there also exists a non-ignorable domain gap. This is consistent with the unsatisfactory performance of generative face completion helping recognition applications. To bridge the gap, we propose to rearrange the identity features via knowledge distillation.

Knowledge distillation is a widely applied technology to transfer the knowledge from a cumbersome teacher network into a compact counterpart. To be general, the goal function for traditional knowledge distillation can be formulated as:

$$\mathcal{L}_\ell = \sum_{x_i \in \mathbf{X}} \ell(t_i, s_i), \tag{10}$$

where t_i and s_i denote the feature representation produced by the teacher and student respectively, with x_i as input; and l denotes specific loss function adopted to penalize the differences.

Traditional distillation usually focuses on classification tasks, trained with the Cross-Entropy loss. During training, the output class distribution generated by the student is forced to be close to that of the teacher. In this way, the student could obtain better results than directly trained with class labels. The main reason may lie in that probability distribution over classes provided by the teacher's output, reveals relevance information between classes, therefore providing richer knowledge than ground truth labels.

However, the present distillation methods remain limited. Existing distillation methods usually focus on the point-wise similarity between representations of teacher and student. Previous researches [7, 39, 49, 52] have verified that instance relationships can help reduce the intra-class variations and enlarge the inter-class divergences in the feature space. Nevertheless this is rarely considered in distillation. We assume that what constitutes the knowledge is better presented by relations of the learned representations than individuals of those, and the structural distribution of unmasked faces could provide essential guidance for the identity feature rearrangement of completed faces. Besides, point-wise distillation methods usually require the teacher and student to share similar network architecture and close data domains. Here in the masked face recognition scenario, we seek to distill the rich knowledge about feature distribution for unmasked faces and use them to guide the rearrangement of that of completed faces. The common characteristics we seek here should be the aggregation behaviors, in other words, the instance relationships, which are more robust to network changes and domain shifts.

Let $\hat{\phi}_t(\mathbf{Y}; \hat{\mathbb{W}}_t)$ and $\hat{\phi}_s(\tilde{\mathbf{X}}; \hat{\mathbb{W}}_s)$ be the sub-networks composed by the first several layers of the teacher network $\phi_t(\mathbf{Y}; \mathbb{W}_t)$ and the student network $\phi_s(\tilde{\mathbf{X}}; \mathbb{W}_s)$, respectively, where $\mathbf{Y}$ and $\tilde{\mathbf{X}}$ is the corresponding input. $\hat{\phi}_t(\mathbf{Y}; \hat{\mathbb{W}}_t)$ is the feature extraction backend before the softmax layer for extracting the identity features of unmasked faces, while $\hat{\phi}_s(\tilde{\mathbf{X}}; \hat{\mathbb{W}}_s)$ denotes the layers before the embedding layer, used to extract features of masked faces. The training process of the student network can be described as transferring the relational structure of the output representation $\hat{\phi}_t(\mathbf{Y}; \hat{\mathbb{W}}_t)$ to $\hat{\phi}_s(\tilde{\mathbf{X}}; \hat{\mathbb{W}}_s)$, to improve the final recognition ability of $\phi_s(\tilde{\mathbf{X}}; \mathbb{W}_s)$. Let $\mathbf{Y}^n$ and $\tilde{\mathbf{X}}^n$ denote a set of n-order tuple of unmasked and completed faces respectively, $s_i = \hat{\phi}_s(\tilde{x}_i; \hat{\mathbb{W}}_s)$ is the student knowledge gained from a completed face and $t_i = \hat{\phi}_t(y_i; \hat{\mathbb{W}}_t)$ is the teacher knowledge distilled from the corresponding ground-truth. The loss function for n-order distillation process can be formulated as

$$\mathcal{L}_n = \sum_{\substack{(y_1,\ldots,y_n)\in\mathbf{Y}^n,\\ (\tilde{x}_1,\ldots,\tilde{x}_n)\in\tilde{\mathbf{X}}^n}} \ell(\psi(t_1,\ldots,t_n), \psi(s_1,\ldots,s_n)), \tag{11}$$

where ψ is a relational potential function that measures relational similarity between given n-tuple of teacher and student models, and l is a loss that penalizes structural difference based on that.

To efficiently transfer the relational knowledge, we here introduce relational loss in three orders, enforcing structural similarity in instance-wise, pair-wise and triplet-wise fashion, respectively.

Instance-Wise Relational Loss Following the vanilla setting, we enforce instance-wise similarity via punishing the difference

$$\mathcal{L}_i = \sum_{y_i \in \mathbf{Y}, \tilde{x}_i \in \tilde{\mathbf{X}}} \ell_1(t_i, s_i), \tag{12}$$

where ℓ_1 loss is chosen instead of ℓ_2 loss because it deals better with abnormal points. In this task, besides the gap between teacher and student domain, there also exists great variance within the student domain. Despite better convergence and more robustness, ℓ_2 loss would bring unwanted smoothness.

Pair-Wise Relational Loss Recent several works have used pair-wise relational distillation loss in tasks such as image classification [30], image retrieval [59] and semantic segmentation [31]. It is used to transfer pair-wise relations, especially pair-wise similarities in our approach, among instances. We formulate the pair-wise relational knowledge distillation loss as follows:

$$\mathcal{L}_p = \sum_{\substack{(y_i,y_j)\in\mathbf{Y}^2,\\ (\tilde{x}_i,\tilde{x}_j)\in\tilde{\mathbf{X}}^2}} \ell_\delta(\psi_p(t_i,t_j), \psi_p(s_i,s_j)), \tag{13}$$

where ℓ_δ is Huber loss, and $\psi_p(t_i, t_j) = \frac{1}{\mu}\|t_i - t_j\|_2$ is the pair-wise potential function which measures the Euclidean distance between the two instances t_i and t_j in a mini-batch space. $\mu = \frac{1}{|\mathbf{Y}^2|}\sum_{(y_i,y_j)\in\mathbf{Y}^2}\|t_i - t_j\|_2$ is a normalization factor, which enables

relational structures transferring disregarding the difference in space dimensions between source and task field.

Triplet-Wise Relational Loss The structure within a triplet could provide more strict regularization than that of a pair. Inspired by this, [40] propose a triplet-wise relational distillation loss:

$$\mathcal{L}_t = \sum_{\substack{(y_i,y_j,y_k)\in\mathbf{Y}^3,\\(\tilde{x}_i,\tilde{x}_j,\tilde{x}_k)\in\tilde{\mathbf{X}}^3}} \ell_\delta(\psi_t(t_i,t_j,t_k),\psi_t(s_i,s_j,s_k)), \tag{14}$$

where l_δ is Huber loss, and the corresponding triplet-wise potential function which measures the angle formed by the three instances t_i, t_j and t_k in a mini-batch space is formulated as:

$$\psi_t(t_i,t_j,t_k) = \langle \frac{t_i - t_j}{\|t_i - t_j\|_2}, \frac{t_k - t_j}{\|t_k - t_j\|_2} \rangle. \tag{15}$$

The triplet-wise relational loss transfers relationships of instance embedding by penalizing angular differences. Compared with the pair-wise potential function, the triplet-wise potential function measures structural similarity in a higher-order space, enabling more effective relational knowledge transferring.

Total Loss The total loss is therefore formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_i\mathcal{L}_i + \lambda_p\mathcal{L}_p + \lambda_t\mathcal{L}_t, \tag{16}$$

where $\mathcal{L}_{CE}$ is the Cross-Entropy loss between outputs of the teacher and student network, as defined in Eq. 10 when ℓ adopts Cross-Entropy. $\mathcal{L}_i$, $\mathcal{L}_p$ and $\mathcal{L}_t$ are various orders of relational distillation loss defined in Eq. 12, Eq. 13 and Eq. 14, respectively. The λ_i, λ_p and λ_t are weighting hyper-parameters to balance the loss terms.

3.4 Implementation Details

We build the de-occlusion module with a generative inpainting network using the same architecture as in [57]. In the distillation module, we employ a pre-trained VGGFace2 [3] model as the teacher. It achieves a very high accuracy of 99.53% on the LFW dataset [17] after alignment. The student network is composed of a ResNet-18 [13] model with a single embedding layer on top.

Our end-to-end network is implemented based on the deep learning library Pytorch. In the experiments, we set $\lambda_p = 1.0, \lambda_t = 2.0$. All models were trained with a mini-batch size of 128. The initial learning rate is $lr = 0.1$ and decreases to 0.1 times every 24 epochs. When sampling tuples of instances for the relational losses, we simply use all the tuples (pairs or triplets) in the given mini-batch.

4 EXPERIMENTS

In this section, the proposed de-occlusion distillation framework is systemically evaluated on both synthesized and realistic masked face datasets. We first introduce the experiment setting, then present experimental results on two datasets, finally we conduct ablation studies and discuss the function paradigm of the proposed method.

4.1 Experiment Setting

Datasets Our experiments are carried out on three datasets: Celeb-A dataset [32], LFW dataset [17] and AR dataset [35].

The **Celeb-A** dataset consists of 202,599 face images covering 10,177 subjects. Each face image is cropped, roughly aligned by the position of two eyes, nose, and two mouth corners, and rescaled to $256\times256\times3$ pixels. We acquired synthetic masked faces via pasting collected masks onto these images, described later. We randomly split it into training and validation set with the ratio set as 6 : 1.

Figure 3: Examples of the masks adopted for synthetic masked faces.

The **LFW** dataset consists of 13,233 images of 5,749 identities. Same preprocessing as Celeb-A was performed to prepare the data. We used all 13,233 images on LFW to benchmark the results. $6K$ pairs (including $3K$ positive and $3K$ negative pairs) were selected to evaluate the performance of masked face recognition.

The **AR** dataset consists of face images with varying illumination conditions, expressions, and partial occlusions. Two variations of occlusions are available in the dataset, sunglasses and scarves, which makes 1,200 images in total. We followed the same protocol to prepare the data. For our study, we randomly took an unmasked face of the same subject, instead of its non-exist original, to send into the teacher network and provide guidance.

Synthesizing Protocols Considering the deficiency of masked face datasets, we synthesized masked face images by automatically pasting mask patterns into face images from Celeb-A [32] and LFW [17] Dataset. We collected transparent mask images online. To prevent over-fitting, we followed [10] and divided occlusions into four categories: Simple Mask (man-made objects with pure color), Complex Mask (man-made objects with complex textures or logos), Human Body (face covered by hand, hair, etc.) and Hybrid Mask (combinations of at least two of the aforementioned mask types, or one of the aforementioned mask types with eyes occluded by glasses), and select representative masks for each type. 45 mask images are employed. Several examples of the extracted masks are shown in Fig 3. All the masks were rescaled, covering an average of about $\frac{1}{5}$ of the face. To improve the generalization ability of the model, we did data augmentation including flipping and shift.

4.2 Results on Synthetic Masked Faces

In this subsection, we compare the recognition performance for synthetic masked faces of different models. We trained our end-to-end de-occlusion model on Celeb-A, then evaluate comparison accuracy on the LFW dataset. All models extract features of all 6000 face pairs and then computes the cosine similarities between the face pairs. The accuracy is the percentage of correct prediction, where the threshold is decided as the one with the highest accuracy.

We trained our network with total loss taking the form as Eq. 16. Several sotas are also presented and compared. **GFC** [26], **GA** [57] and **IDGAN** [9] are all state-of-the-art generative inpainting methods, especially **IDGAN** is designed and optimized for masked face recognition problem. We equip them with five high-performance recognizers, and the results are shown in Fig. 4. Our model trained with Eq. 16, denoted as **OUR-Hard**, surpass all combinations.

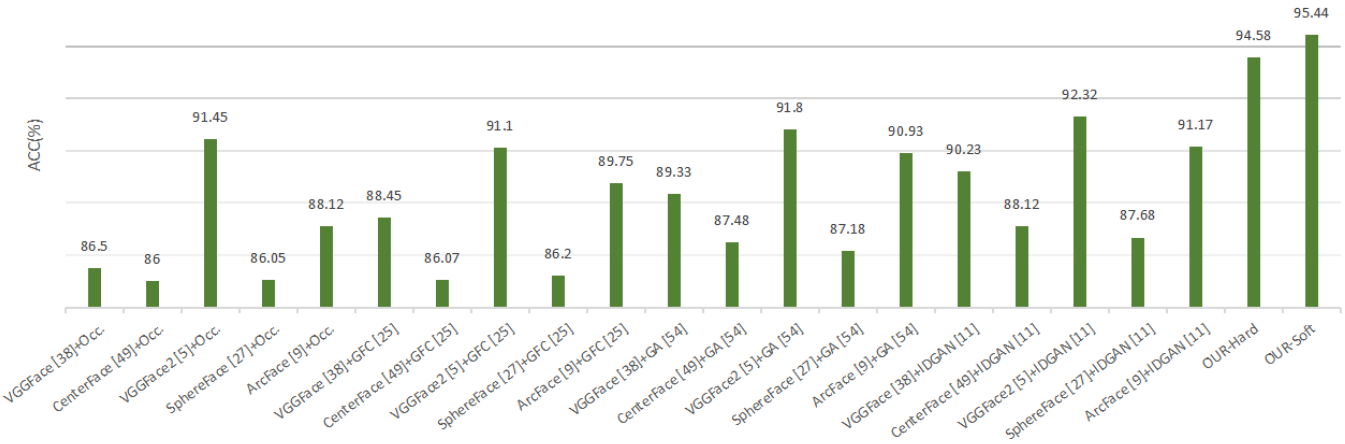

Figure 4: Evaluation accuracy of different models on LFW.

Table 1: Rank-1 recognition accuracy on AR Database by several existing recognizers.

	AR1: Recognition of faces with glasses (%)					**AR2**: Recognition of faces with scarfs (%)				
	Occ.	**RPCA** [53]	**GL** [8]	**GFC** [26]	**GA** [57]	**Occ.**	**RPCA** [53]	**GL** [8]	**GFC** [26]	**GA** [57]
PCA [51]	61.4	64.2	70.0	82.9	88.6	37.5	32.2	40.8	72.6	78.3
GPCA [24]	73.3	71.6	76.6	88.4	92.7	56.2	54.0	60.9	80.3	87.2
LPP [14]	45.7	61.4	59.0	83.5	90.0	43.0	38.3	47.1	75.9	80.1
SR [54]	59.2	57.3	60.6	85.7	90.3	51.8	47.7	56.7	79.8	86.4
VGGFace [39]	85.4	84.5	87.9	91.7	95.9	75.9	79.6	83.5	89.9	92.2
VGGFace2 [3]	**88.3**	86.7	89.0	94.2	97.0	**78.2**	81.4	85.7	93.1	93.3
SphereFace [29]	87.5	87.2	89.0	93.7	97.5	78.0	79.8	83.9	92.8	93.1
ArcFace [7]	85.5	85.2	87.6	92.3	95.5	76.4	79.2	82.6	90.2	91.9
OUR	-	**92.1**	**93.3**	**97.2**	**98.0**	-	**84.4**	**86.8**	**93.3**	**94.1**

During the training process, we noticed that the model converges stably at the early stages, then gets stuck soon, showing obvious over-fitting. We leave detailed illustration to Sec. 4.4 and go directly for refining strategy. With a large reduction in parameters and change on model structure, it is reasonable to suspect the student bear considerable instability, especially under perturbation settings like ours. Directly enforcing the feature to be the same could be too strict regularization. History researches indicate that soften knowledge is more efficient to learn [33]. Therefore, we reformulate the instance-wise relational loss $\mathcal{L}_i$ as:

$$\mathcal{L}_i^* = \sum_{y_i \in \mathbf{Y}, \tilde{x}_i \in \tilde{\mathbf{X}}} \ell_1(t_i - \mathbf{f}_{id}, s_i - \mathbf{f}_{id}) \tag{17}$$

where identity-centered feature $\mathbf{f}_{id}$ represents the centroid of identity features for training images with identity label id:

$$\mathbf{f}_{id} = \frac{\sum_{i=1}^{N} \delta(y_i = id) t_i}{\sum_{i=1}^{N} \delta(y_i - id)} \tag{18}$$

where N denotes the size of training datasets. In experiments, the identity-centered features are pre-computed off-line. In this way, we soften the knowledge of the teacher and enable a stabler knowledge transfer. The results are shown in Fig. 4 as **OUR-Soft**.

4.3 Results on Realistic Masked Faces

We then evaluate the proposed method on the AR dataset, where two variations of occlusions are available. For testing, the masked faces were divided into two subsets, denoted as AR1 and AR2, consisting of faces with sunglasses and scarfs, respectively.

We adopted four classic face recognition algorithms and four deep learning recognition models to test the recognition performances on the masked faces and the completed faces. Specifically, the four recognition algorithms are: 1) PCA [51], the typical statistic-based recognition algorithm; 2) Gabor wavelet-based recognition (GW+PCA) [24], using features in the transformed domain; 3) locality projection [14], a manifold-based recognition algorithm; and 4) SR [54], which is a branch of norm-based optimization. The four state-of-the-art deep learning recognition models include VGGFace [39], VGGFace2 [3], SphereFace [29] and ArcFace [7].

We completed faces by two tradition methods **RPCA** [53] and **GL** [8], as well as two sota generative inpainting methods **GFC** [26] and **GA** [57] respectively. All faces are then predicted by all recognizers above, as comparisons with our de-occlusion distillation model. The results are shown in Tab. 1, and our method achieves a higher recognition accuracy. The VGGFace2 and SphereFace model exhibit relatively milder degradation in masked scenarios among the baselines. We take this as evidence that data diversity and structural regularization are beneficial for model robustness.

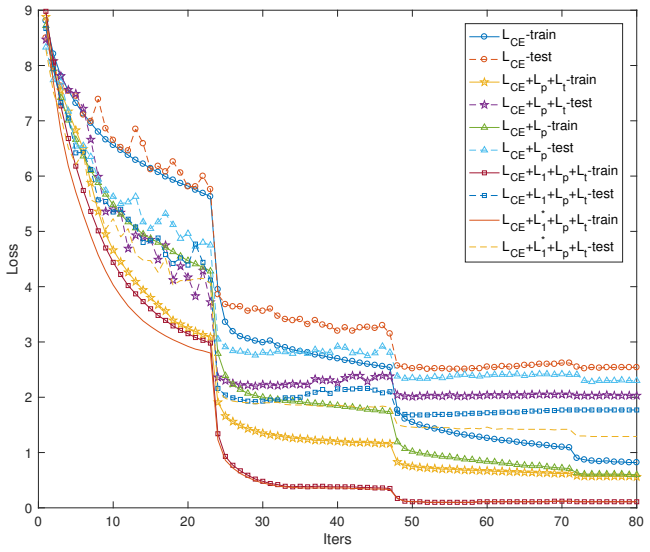

Figure 5: Loss changes with different loss settings.

4.4 Ablation Study

In this section, we conduct ablation studies to prove efficacy.

Contribution of each loss component. We trained our model with 1)Cross-Entropy(CE) loss only $\mathcal{L}_{CE}$, 2) CE loss and pair-wise relational loss $\mathcal{L}_{CE} + \lambda_p \mathcal{L}_p$, 3) CE loss, pair-wise and triplet-wise relational loss $\mathcal{L}_{CE} + \lambda_p \mathcal{L}_p + \lambda_t \mathcal{L}_t$, 4) CE loss, pair-wise, triplet-wise, and hard instance-wise relational loss $\mathcal{L}_{CE} + \lambda_i \mathcal{L}_i + \lambda_p \mathcal{L}_p + \lambda_t \mathcal{L}_t$ and 5)CE loss, pair-wise, triplet-wise, and soft instance-wise relational loss $\mathcal{L}_{CE} + \lambda_i \mathcal{L}_i^* + \lambda_p \mathcal{L}_p + \lambda_t \mathcal{L}_t$. Fig. 5 shows the trend of loss during training in both the training and evaluation set. Testing accuracy on Celeb-A and LFW share similar trends, and the trend on LFW is less aligned due to the domain gap. From the figure, we noticed that CE loss can well stabilize the training process. However, with CE loss only, the model fails to reach the optima. Our relational distillation loss in various orders exhibits good collaboration with CE loss and leads to the best performance eventually.

OUR-Hard vs OUR-Soft. It is worth to note in Fig. 5 that **OUR-Hard** trained with $\mathcal{L}_{CE} + \lambda_i \mathcal{L}_i + \lambda_p \mathcal{L}_p + \lambda_t \mathcal{L}_t$ shows apparent over-fitting in later stage, even worse than those trained without instance-wise loss. With a large reduction in parameters and change on model structure, it's reasonable to suspect the student bear considerable instability, especially under perturbation settings like ours. Directly enforcing the features to be exactly the same could be too strict. After we soften the instance-wise loss into identity-centered ensemble loss $\mathcal{L}_i^*$, as defined in Eq. 17, the resulting **OUR-Soft** shows more efficient convergence and finally reach the lowest loss. We believe this discovery is meaningful for more general adaptation tasks, especially with great domain gaps or severe perturbation.

Choice of completion methods. We then ask whether the choice of completion methods make a difference. We replace the face completion model in the de-occlusion module with 1)**GFC** [26], 2)**GA** [57] without contextual attention and 3)**IDGAN** [9], in comparison with the adopted 4)**GA** [57] with contextual attention. We trained model with Hard ($\mathcal{L}_{CE} + \lambda_i \mathcal{L}_i + \lambda_p \mathcal{L}_p + \lambda_t \mathcal{L}_t$) and Soft loss ($\mathcal{L}_{CE} + \lambda_i \mathcal{L}_i^* + \lambda_p \mathcal{L}_p + \lambda_t \mathcal{L}_t$) and Tab. 2 shows the results on LFW. Bold and underline denote the first and second highest in each column, separately. All models trained with Soft loss perform better than those with Hard loss, which verifies the efficacy of the softening mechanism. Besides, it's evident that the performance of **GA** [57] with attention stands out alone, while the others are close. We attribute it to the contextual attention mechanism which allows the model to focus on the most relevant and informative areas.

Table 2: Comparisons between models adopting different face completion methods and finetuned ArcFace [7] (%).

	GFC [26]	**GA** [57] w/o attention	**IDGAN** [9]	**GA** [57] w/ attention
ArcFace [7]-finetuned	90.15	91.28	92.52	92.17
OUR-Hard	<u>92.77</u>	<u>92.93</u>	<u>93.12</u>	<u>94.58</u>
OUR-Soft	**93.60**	**93.55**	**93.92**	**95.44**

Distillation vs Fine-tune. Finally, to prove the efficacy of the distillation module, we fine-tune the Arcface model [7], which has a similar size with our student model, on completed faces and make comparisons. The evaluation accuracy on the LFW dataset is reported in the first row of Tab. 2. Our **OUR-Soft** model surpasses the fine-tuned ArcFace model by 3.27%, which suggest that fine-tuning can hardly come across the semantic gap. Our models instead, learn to recover the identity features under the guidance of pre-trained recognizer, and effectively improve the masked face recognition.

5 CONCLUSION

Masked face recognition is a problem of wide prospects for applications. Despite great efforts and advancements made over the years, current methods are restrained by incomplete visual content and insufficient identity cues. In this work, we migrate the mechanism of amodal perception and propose a novel de-occlusion distillation framework for efficient masked face recognition. The model first recovers appearance information via a generative face completion based de-occlusion module, and then transfers rich structural knowledge from a high-performance pre-trained general recognizor to train a student model. In this way, the student model learns to recover the missing information both in appearance space and in identity space. By representing knowledge of existing high-performance recognition models with structural relations in various orders, the model is enforced to extract representations with similar aggregation behaviors with those of the teacher. Experimental results show that the amodal completion mechanism is also beneficial for deep neural networks, and our proposed de-occlusion distillation can deal with the masked face recognition task on both synthetic and realistic datasets. In the future, we will work on the establishment of amodal perception for computer vision and further investigation on suitable network architecture.

Acknowledgement. This research is supported in part by grants from the National Key Research and Development Program of China (2020AAA0140001), the National Natural Science Foundation of China (61772513 & 61922006), Beijing Natural Science Foundation (L192040) and Beijing Municipal Science and Technology Commission (Z191100007119002). Shiming Ge is also supported by the Youth Innovation Promotion Association, Chinese Academy of Sciences.

REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Transactions on Graphics (ToG)* 28, 3 (2009), 24.

[2] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model Compression. In *SIGKDD*. 535–541.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces Across Pose and Age. In *FG*. 67–74.

[4] Le Chang and Doris Y Tsao. 2017. The Code for Facial Identity in the Primate Brain. *Cell* 169, 6 (2017), 1013–1028.

[5] Juan Chen, Bingyun Liu, Bing Chen, and Fang Fang. 2009. Time Course of Amodal Completion in Face Perception. *Vision Research* 49, 7 (2009), 752–758.

[6] Jia Deng, Wei Dong, Richard Socher, Lijia Li, Kai Li, and Li Feifei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*. 248–255.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*. 4690–4699.

[8] Yue Deng, Qionghai Dai, and Zengke Zhang. 2011. Graph Laplace for Occluded Face Completion and Recognition. *IEEE Transactions on Image Processing (TIP)* 20, 8 (2011), 2329–2338.

[9] Shiming Ge, Chenyu Li, Shengwei Zhao, and Dan Zeng. 2020. Occluded Face Recognition in the Wild by Identity-Diversity Inpainting. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* (2020), 1–11.

[10] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. 2017. Detecting Masked Faces in the Wild with LLE-CNNs. In *CVPR*. 2682–2690.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*. 2672–2680.

[12] Kaiming He and Jian Sun. 2014. Image Completion Approaches using the Statistics of Similar Patches. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36, 12 (2014), 2423–2435.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[14] Xiaofei He and Partha Niyogi. 2003. Locality Preserving Projections. In *NeurIPS*. 153–160.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the Knowledge in a Neural Network. In *NIPS Workshop*.

[16] SK Alamgir Hossain, Abu Saleh Md Mahfujur Rahman, and Abdulmotaleb El Saddik. 2011. Fusion of Face Networks through the Surveillance of Public Spaces to Address Sociological Security Recommendations. In *ICME*. 1–6.

[17] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report 07-49. University of Massachusetts, Amherst.

[18] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 107:1–14.

[19] Jaywoo Kim, Younghun Sung, Sang Min Yoon, and Bo Gun Park. 2005. A New Video Surveillance System Employing Occluded Face Detection. In *IEA/AIE*. 65–68.

[20] Adam Kortylewski, Ju He, Qing Liu, and Alan L. Yuille. 2020. Compositional Convolutional Networks for Robust Object Classification under Occlusion. In *CVPR*. 8940–8949.

[21] Stephen M. Kosslyn, William L. Thompson, Irene J. Klm, and Nathaniel M. Alpert. 1995. Topographical Representations of Mental Images in Primary Visual Cortex. *Nature* 378 (1995), 496–498.

[22] Gyula Kovács, Rufin Vogels, and Guy A Orban. 1995. Selectivity of Macaque Inferior Temporal Neurons for Partially Occluded Shapes. *Journal of Neuroscience* 15, 3 (1995), 1984–1997.

[23] Y Lecun, Y Bengio, and G Hinton. 2015. Deep Learning. *Nature* 521 (2015), 436–444.

[24] Zhen Lei, Shengcai Liao, Ran He, Matti Pietikainen, and Stan Z Li. 2008. Gabor Volume based Local Binary Pattern for Face Representation and Recognition. In *FG*. 1–6.

[25] Hongjun Li and Ching Y Suen. 2016. Robust Face Recognition Based on Dynamic Rank Representation. *Pattern Recognition (PR)* 60 (2016), 13–24.

[26] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017. Generative Face Completion. In *CVPR*. 3911–3919.

[27] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. 2019. Recover and Identify: A Generative Dual Model for Cross-Resolution Person Re-Identification. In *ICCV*. 431–439.

[28] Roderick JA Little and Donald B Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

[29] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *CVPR*. 6738–6746.

[30] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. 2019. Knowledge Distillation via Instance Relationship Graph. In *CVPR*. 7096–7104.

[31] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. 2019. Structured Knowledge Distillation for Semantic Segmentation. In *CVPR*. 2604–2613.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*. 3730–3738.

[33] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards Natural and Accurate Future Motion Prediction of Humans and Animals. In *CVPR*. 10004–10012.

[34] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. 2016. Face Model Compression by Distilling Knowledge from Neurons. In *AAAI*. 3560–3566.

[35] Aleix Martínez and Robert Benavente. 1998. *The AR Face Database*. Technical Report 24. Computer Vision Center.

[36] Joe Mathai, Iacopo Masi, and Wael AbdAlmageed. 2019. Does Generative Face Completion Help Face Recognition?. In *ICB)*.

[37] Albert Michotte, Georges Thinès, and Geneviève Crabbé. 1964. Les Compléments Amodaux des Structures Perceptives. In *Michotte's Experimental Phenomenology of Perception*, G. Thinés, A. Costall, and G. Butterworth (Eds.). 140–167.

[38] Bence Nanay. 2007. Four Theories of Amodal Perception. In *CogSci*, Vol. 29. 1331–1336.

[39] M. Parkhi Omkar, Vedaldi Andrea, and Zisserman Andrew. 2015. Deep Face Recognition. In *BMVC*. 41.1–41.12.

[40] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational Knowledge Distillation. In *CVPR*. 3967–3976.

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context Encoders: Feature Learning by Inpainting. In *CVPR*. 2536–2544.

[42] Jianjun Qian, Jian Yang, Fanglong Zhang, and Zhouchen Lin. 2014. Robust Low-rank Regularized Regression for Face Recognition with Occlusion. In *CVPR Workshop*. 21–26.

[43] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. 2018. Data Distillation: Towards Omni-Supervised Learning. In *CVPR*. 4119–4128.

[44] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. 2019. StructureFlow: Image Inpainting via Structure-Aware Appearance Flow. In *ICCV*. 181–190.

[45] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. 2019. StructureFlow: Image Inpainting via Structure-Aware Appearance Flow. In *ICCV*. 181–190.

[46] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for Thin Deep Nets. In *ICLR*.

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*. 815–823.

[48] Jong-Chyi Su and Subhransu Maji. 2017. Adapting Models to Signal Degradation using Distillation. In *BMVC*. 21.1–21.14.

[49] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep Learning Face Representation by Joint Identification-Verification. In *NeurIPS*. 1988–1996.

[50] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *CVPR*. 1701–1708.

[51] Matthew Turk and Alex Pentland. 1991. Face Recognition using Eigenfaces. In *CVPR*. 586–591.

[52] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *ECCV*. 499–515.

[53] John Wright, Arvind Ganesh, Shankar R Rao, Yigang Peng, and Yi Ma. 2009. Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization. In *NeurIPS*. 2080–2088.

[54] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. 2008. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 31, 2 (2008), 210–227.

[55] Jianwen Xie, Yang Lu, Ruiqi Gao, and Ying Nian Wu. 2018. Cooperative Learning of Energy-Based Model and Latent Variable Model via MCMC Teaching. In *AAAI*. 4292–4301.

[56] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. 2011. Robust Sparse Coding for Face Recognition. In *CVPR*. 625–632.

[57] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative Image Inpainting with Contextual Attention. In *CVPR*. 5505–5514.

[58] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2019. Free-Form Image Inpainting with Gated Convolution. In *ICCV*. 4471–4480.

[59] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. 2019. Learning Metrics from Teachers: Compact Networks for Image Embedding. In *CVPR*. 2907–2916.

[60] Shu Zhang, Ran He, Zhenan Sun, and Tieniu Tan. 2017. DemeshNet: Blind Face Inpainting for Deep Meshface Verification. *IEEE Transactions on Information Forensics and Security (TIFS)* 13, 3 (2017), 637–647.

[61] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 2018. 3D-Aided Dual-Agent GANs for Unconstrained Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 10 (2018), 2380–2394.