中图法分类号:TP391　**文献标识码**:A　**文章编号**:1006-8961(年) -

论文引用格式:

面向大姿态人脸识别的由粗到细人脸正面化

胡蓝青[1,2]，阚美娜[1,2]，山世光[1,2*]，陈熙霖[1,2]

1.中国科学院计算技术研究所，北京 100190； 2.中国科学院大学，北京 100049

摘　要：**目的** 尽管人脸识别已经得到了广泛的应用，但大姿态人脸识别问题仍未得到完美解决。已有方法或提取姿态鲁棒特征，或进行人脸姿态的正面化。其中主流的人脸正面化方法包括2D回归生成和3D模型形变建模，前者能够生成相对自然真实的人脸，但会引入额外的噪声导致图像信息的扭曲；后者能够保持原始的人脸结构信息，但生成过程是基于物理模型的，不够自然灵活。**方法** 为结合2D和3D方法的优势，本文提出了基于由粗到细形变场的人脸正面化方法。该形变场由深度网络以2D回归方式学得，反映的是不同视角人脸图像像素之间的语义级对应关系，可被利用以类3D的方式实现非正面人脸图像的正面化，因此该方法兼具了2D正面化方法的灵活性与3D正面化方法的保真性。且借鉴分步渐进的思路，本文中提出了由粗到细的形变场学习框架，以获得更加准确鲁棒的形变场。**结果** 本文采用大姿态人脸识别实验来验证本文方法的有效性，在MultiPIE、LFW、CFP、IJB-A等4个数据集上均取得了比已有方法更高的人脸识别精度。**结论** 本文所提出的基于由粗到细的形变场学习的人脸正面化方法，综合了2D和3D人脸正面化方法的优点，使人脸正面化结果的学习更加灵活、准确，保持更多有利于识别的身份信息。

关键词：大姿态人脸识别；人脸正面化；可学习形变场；由粗到细学习；全卷积网络

Face frontalization with coarse-to-fine morphing field learning for large pose face recognition

Hu Lanqing[1,2]，Kan Meina[1,2]，Shan Shiguang[1,2*]，Chen Xilin[1,2]

1. Institute of Computing Technology, Beijing 100190； 2. University of Chinese Academy of Sciences, Beijing 100049

Abstract: Objective Currently, face recognition is applied world widely. However, face recognition in the wild remains challenging due to large variations in pose, expression, aging, lighting, occlusion and so on. Among these factors, pose variations usually result in large non-planar face transformation which does severe harm to the performance of face recognition. To address the notorious pose variations, previous methods mainly attempt to extract pose invariant feature or frontalize non-frontal faces. Among them, the frontalization methods can relieve the difficulty of discriminative feature learning by eliminating pose variations, and they indeed achieve promising progress. There are two kinds of frontalization methods: 2D generative and 3D model based frontalization methods. 2D methods are flexible and can generate more natural frontal faces but it may lose facial structural information, which is important for identity discrimination. 3D methods can well preserve facial structural information but are not so flexible. In summary, both 3D methods and 2D methods have information loss in the frontalized faces especially in the case of large pose variations, for example, invisible pixels in 3D morphable model or pixel aberrance in 2D generative methods. **Method** To handle these problems, in this work we propose a new scheme, referred to as Coarse-to-fine Morphing Field Network (CFMF-Net), combining both 2D and 3D face

收稿日期:　　；　**修回日期**:

基金项目:国家重点研发计划课题(A1802)；国家自然科学基金(61772496)

Supported by: National Key Research and Development Program of China (A1802); National Natural Science Foundation of China (61772496)

transformation methods to frontalize a non-frontal face image via the coarse-to-fine optimized morphing field for shifting each pixel. It borrows the flexibility of 2D learning based methods and structure preservation of 3D morphable model-based methods. Besides, the proposed method learns the morphing field via a progressive and residual manner to make the learning process easier and reduce the probability of overfitting. Specifically, a coarse morphing field is learned firstly to capture the major structure variation of a face image. Then a residual module extracting detailed facial information is designed to complement the coarse morphing field, whose output is concatenated with the coarse morphing field to generate the final fine morphing field for frontalizing the input face image. The overall framework is for regressing the pixel correspondences but not pixel values as other frontalization works. The work ensures that all pixels in the frontalized face image are taken from the input non-frontal image, thus reducing information distortion to a large extent. Therefore, the identity information in the input non-frontal face images are well preserved with favorable visual results, thus further facilitating the subsequent face recognition task. All in all, the design of the coarse-to-fine morphing field learning assures the robustness of learned morphing field and the residual complementing branch achieves more accurate morphing field output. **Result** To verify the effectiveness of our proposed work, extensive experiments on MultiPIE, LFW, CFP and IJB-A datasets are conducted, and the results are compared with other newly proposed face transformation methods. Among these testing sets, MultiPIE, CFP and IJB-A datasets are all with full pose variation. Besides, IJB-A contains full pose variations as well as other complicated variations such as low resolution and occlusion. The experiments follow the same training and testing protocol with previous works, i.e., training with both original and frontalized face images. For fair comparison, the commonly used LightCNN-29 is exploited as the recognition model. Our method achieves the best performance among the related works on the large pose testing protocol of MultiPIE and CFP and comparable performance on LFW and IJB-A. In addition, our visualization results also show that our method can well preserve the identity information. Further, the ablation study implies the rationality of the coarse-to-fine framework in our CFMF-Net. In a word, the recognition accuracies and visualization results demonstrate that the proposed CFMF-Net can not only generate frontalized faces with identity information preserved but also achieve better face recognition accuracy especially in the case of large pose variations. **Conclusion** In conclusion, this work proposes a coarse-to-fine morphing field learning framework to frontalize face images by shifting pixels to ensure the flexible learnability and identity information preservation. Specifically, the flexible learnability helps the network to optimize according to face frontalization objective but not predefined 3D transformation rules, which can improve the accuracy. Moreover, the learned morphing field for each pixel makes the output frontal face shifted from only the input image, reducing the information loss. In addition, the design of coarse-to-fine and residual architecture further ensures more robust and accurate results. Taking all things into consideration, frontalizing faces by reorganizing pixels of input faces can preserve more identity information benefitting for face recognition and the coarse-to-fine framework helps improve the accuracy and robustness of the learned morphing field.

Key words： large pose face recognition; face frontalization; morphing field learning; coarse-to-fine learning; fully convolutional network

0 引 言

当今社会，人脸识别技术在各领域得到了广泛的应用，为人们的生活带来了巨大的便利。随着技术的发展，人脸识别的性能得到了极大的提升。非极端姿态的人脸识别已经取得了良好的效果。但是，大姿态下的人脸识别仍然面临很大的挑战。这是由于人脸在大姿态下会发生很强的非平面内形变，影响对人脸身份的判别。主流的针对大姿态人脸识别问题的方法分为两大类：第一类方法直接在原图上提取姿态鲁棒特征，第二类方法先将人脸进行正面化之后再提取特征。其中直接提取姿态鲁棒特征的方法在应用到极端姿态人脸识别时，可以提取的特征变得非常有限，人脸识别性能会明显降低，因此出现了先将人脸正面化再进行公共特征提取的方法，即人脸正面化方法，它们可以提取出更多有效的公共判别特征。正面化方法又可以被分为 2D 生成人脸方法和利用 3D 模型变换方法。2D 生成方法通过一个网络直接回归出正面人脸图像，3D 方法则是将人脸图像建模为 3D 模型，通过该模型算出原图与正面人脸的像素坐标对应关系，从而实现正面化。2D 生成方法比基于 3D 模型的方法更加灵活，生成的人脸也更加自然。然而 3D 方法得到的正面化人脸图像能够保留更多的人脸身份信息。一种自然的想法就是希望结合两种正面化方法的优点，摒弃它们的缺点，由此提出了本文的方法。

本文提出的方法为一种基于由粗到细形变场学习的人脸正面化方法，即 Coarse-to-fine Morphing Field Network (CFMF-Net)，如图 1 所示。此处形变场指正面人脸与输入人脸的像素点的位置对应关系，即非正面人脸图像的像素可以根据形变场进行重组得到对应的正面人脸图像。CFMF-Net 通过一个深度网络以由粗到细的优化策略学习形变场，来对输入人脸进行正面化。

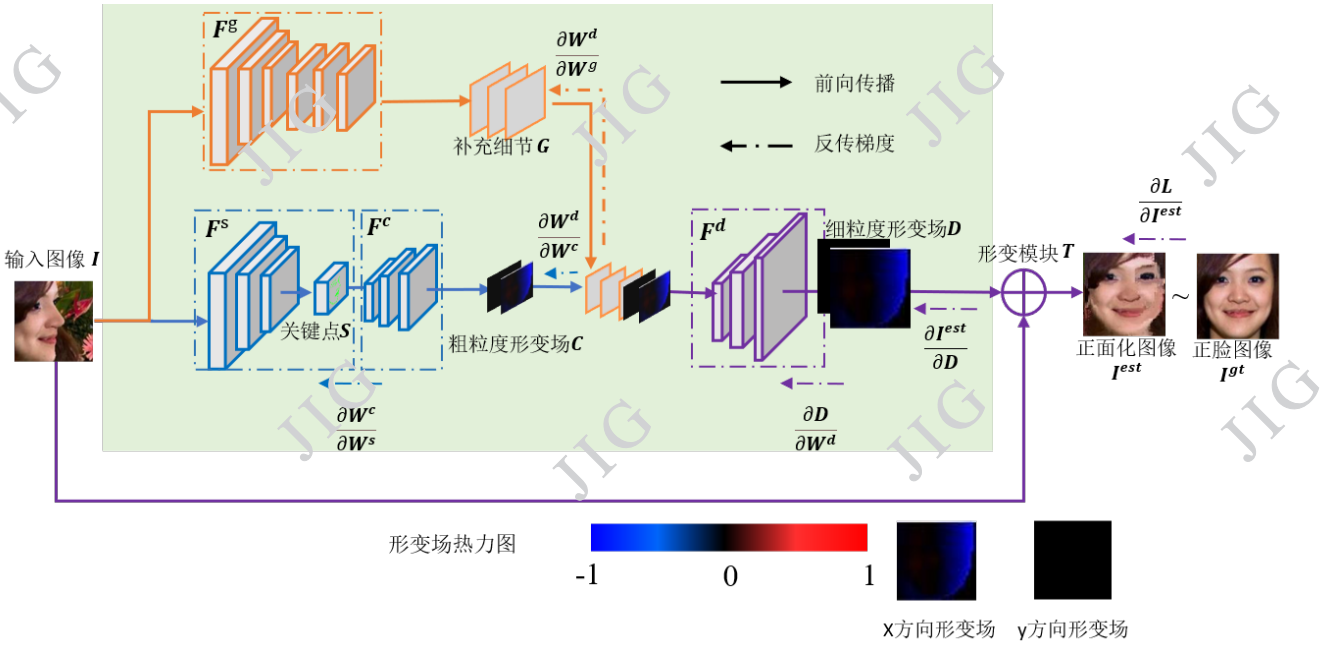


图 1 基于由粗到细形变场学习的方法 CFMF-Net 流程图。CFMF-Net 通过学习形变场将任意人脸图像$\boldsymbol{I}$正面化为图像$\boldsymbol{I^{est}}$。该网络首先通过$\boldsymbol{F^s}$提取人脸关键点特征$\boldsymbol{S}$，并将$\boldsymbol{S}$输入$\boldsymbol{F^c}$以得到粗粒度形变场$\boldsymbol{C}$。之后将$\boldsymbol{C}$和$\boldsymbol{F^g}$学到的细节特征$\boldsymbol{G}$拼接在一起输入$\boldsymbol{F^d}$以得到细粒度形变场$\boldsymbol{D}$。形变模块$\boldsymbol{T}$将形变场$\boldsymbol{D}$作用于输入图像$\boldsymbol{I}$得到$\boldsymbol{I^{est}}$。CFMF-Net 通过拉近$\boldsymbol{I^{est}}$与真实的正脸图像$\boldsymbol{I^{gt}}$的距离来进行优化。其中形变场的值由下方的热力图表示，红色表示该像素点上的形变场向左，蓝色表示该像素点上的形变场向右，颜色明度越高则移动距离越大。

Fig.1 Overview of our CFMF-Net. The overall objective is to transform an input non-frontal face image $\boldsymbol{I}$ into a frontal one $\boldsymbol{I^{est}}$ by applying an estimated morphing field with the same resolution as $\boldsymbol{I}$. The morphing field is predicted via a residual scheme. First, a coarse morphing field in low resolution $\boldsymbol{C}$ is decoded from a low-dimensional and robust representation $\boldsymbol{S}$, attempting to handle the major morphing variations. Later, a residual branch characterizing the detailed morphing variation $\boldsymbol{G}$ complements this coarse morphing field to further accurately predict the final morphing field $\boldsymbol{D}$. After obtaining the morphing field, the warping module $\boldsymbol{T}$

warps the input non-frontal face image $\boldsymbol{I}$ to the frontal one $\boldsymbol{I}^{est}$. The morphing field is illustrated via a heatmap, where blue points represent the leftward field, red ones are the rightward field, and brighter color stands for longer shifting distances.

本文采用的以形变场进行正面化的方式与利用3D人脸模型进行人脸正面化的方法类似，都能够通过像素点的移动来变换图像，保证正面化人脸图像中的像素点全部来源于原始图像。并且本文方法与2D回归方法类似，都是通过网络自动学习，而不是学自于人为设计的规则。因此该方法兼具了2D正面化方法的灵活性与3D正面化方法的保真性。

然而目标形变场来自高维空间，这给网络的优化带来了不小的难度。因此本文借鉴分步渐进的优化思路，进而提出了由粗到细的形变场学习框架，以获得更加准确鲁棒的形变场。然而在学习粗粒度形变信息时，模型只留意了人脸的主要形变，会导致细节信息的丢失，因而再增加一路细节补充分支网络，以进一步保证预测出的形变场的准确性。

总的来说，本文工作的主要贡献在于：1）采用2D回归的方式以类3D的行为对人脸进行正面化，结合了2D正面化方法的灵活性与3D正面化方法的保真性；2）用由粗到细的学习方式提升模型的易学习性。本文将在第2节方法部分对CFMF-Net的框架进行细节描述。

1 相关工作

上一章中大姿态人脸识别方法被分为两大类，本章将对它们进行更细致的介绍。

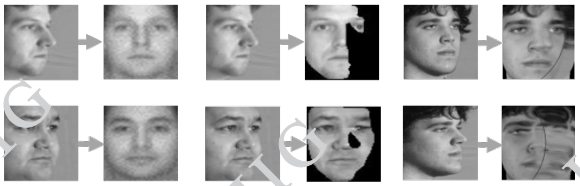
直接提取姿态鲁棒特征的方法主要是将不同姿态的人脸图像都映射到一个公共的特征空间中。典型相关分析（Li等人，2009）是直接提取姿态鲁棒特征的早期经典方法，它通过最大化两组不同姿态的图像的特征相关性，来将不同姿态的特征映射到统一的空间中。然而，该方法只保证了提取到的是不同姿态图像的公共特征，却忽略了特征的判别能力。其后，Sharma等人（2011）改进了典型相关分析，他们的方法通过偏最小二乘法最小化同一个人所有姿态的图像的特征距离，得到的特征不仅是姿态鲁棒的，且具有较好的判别能力。Zhang等人（2013）给训练集中同一人所有姿态的图像设定同一张随机的人脸作为映射目标，以得到姿态鲁棒的具有良好判别能力的特征。多视角判别网络（Kan等人，2016）针对不同姿态的图像采用不同的特征映射，将多姿态的图像映射到公共特征空间中，从而得到了更准确的公共判别特征。深度网络的提出与发展进一步赋予了模型更强大的特征学习能力。基于深度学习的特征解耦方法（Peng等人，2017）首先利用深度网络提取出更准确的人脸表示，之后通过特征解耦与交叉重组得到姿态鲁棒特征。

这些直接提取姿态鲁棒特征的方法对非极端姿态的人脸识别已经有了不错的效果，但对极端姿态的人脸识别却效果有限。这是因为，对这些姿态差异巨大的人脸图像，直接提取公共特征会丢失很多对识别有用的信息。因此，研究者们提出了先将人脸转正，再进行人脸识别的人脸正面化方法。这些方法又分为2D正面化方法和3D模型正面化方法。

2D人脸正面化方法直接通过一个编码器网络将不同姿态的人脸图像映射为正面姿态的图像。经典的方法（Zhu等人，2013）（Kan等人，2014）用渐进式学习的方式，对侧面人脸进行逐步的姿态调整，以映射到正面人脸。之后，随着生成对抗网络GAN（Goodfellow等人，2014）的提出，很多方法借助GAN强大的分布拟合能力来生成各种姿态的人脸，包括正脸。相比于通过回归生成人脸的方法，基于GAN的方法生成的人脸图像更加逼真。在Luan等人（2017）方法中，由特征提取器得到的身份特征和指定的姿态信息一起被输入GAN中，以生成多姿态的人脸图像。Yin等人提出了另一个更精细的基于GAN的方法（Yin等人，2017b），该方法给予了GAN更多的信息，即3D可变形模型的系数，从而得到了保留了更多原始信息的正面人脸图像。Huang等人（2017）同时兼顾了整张人脸和人脸局部图像块的逼真程度，从而使生成的人脸图像保留了更多的细节。Zhang等人（2019）认为更大姿态的人脸更难以识别与正面化，因此在通过GAN正面化人脸的训练过程中对难样本采用更大的训练权重。Rong等人（2020）通过特征级和图像级两种GAN判别器，加强GAN正面化人脸的效果。Luan等人（2020）在GAN判别器中加入自注意力机制保持人脸图像的几何结构，令人脸正面化更加真实。

3D人脸正面化方法通过建立人脸图像的3D模型将人脸映射到正面姿态。相比于2D方法，3D人脸正面化方法能保留更多的人脸结构信息。早期的经典方法，3D通用弹性模型（Prabhu等人，2011）和基于视角的主动外观模型（Asthana等人，2012）

等直接利用 3D 模型进行人脸姿态变换。这些方法通过将 2D 图像映射到 3D 坐标上，再投影到任意的角度，以生成相应姿态的人脸。而更直接的方法是计算侧面人脸图像到其正面人脸图像的像素点的位置对应关系，即形变场，再用该形变场来进行图像变换。Li 等人（2012）用从训练集得到的正面化形变场的线性组合来正面化测试集人脸图像。而这些 3D 方法都不能处理姿态变化引起的自遮挡，如图 2(b)所示。Ding 等人（2015）在 3D 模型变换的基础上，利用人脸的对称性来填补遮挡部分。但是该方法生成的人脸依然存在严重的失真，如图 2(c)所示。Hu 等人（2017）提出了一种利用全连接网络自动回归正面化形变场的方法，生成了更逼真并保留更多原始信息的正面人脸。Cao 等人（2018）提出了一种结合了 3D 模型和 GAN 的方法。该方法首先通过形变场得到一个初始的正脸图像，再通过 GAN 进行图像调整，最终得到足够逼真且身份保持的正面人脸。



(a) Kan 等人，2014　　(b) Li 等人，2012　　(c) Ding 等人，2015

图 2 几种经典方法在 MultiPIE 数据集上的正面化结果

Fig.2 Visualization results of several methods

((a) Kan et al., 2014；(b) Li et al., 2012；(c) Ding et al., 2015)

综上所述，人脸正面化方法相比于直接提取姿态鲁棒特征的方法能够提取出更多有效的公共判别特征。正面化方法中，2D 方法比 3D 方法更加灵活，生成的人脸也更加自然。3D 方法得到的正面化人脸图像能够保留更多的人脸身份信息。

2　方　法

如图 1 所示，本文提出的 CFMF-Net 主要由两部分组成：可学习形变场网络$\boldsymbol{F}$和用形变场进行正面化的模块$\boldsymbol{T}$。网络$\boldsymbol{F}$的输入为原始人脸图像$\boldsymbol{I}$，其输出为正面化$\boldsymbol{I}$的形变场$\boldsymbol{D}$。$\boldsymbol{T}$的输入为原始图像$\boldsymbol{I}$和形变场$\boldsymbol{D}$，其输出为正面化后的图像$\boldsymbol{I}^{est}$。

可学习形变场网络$\boldsymbol{F}$通过渐进式的方式学习形变场，即先学习粗粒度形变场以捕捉人脸结构的主要形变，在此基础上再学习细粒度形变场来精修细节上的形变。因此，网络$\boldsymbol{F}$主要包含两个部分：粗粒度形变场网络$\boldsymbol{F^c}$和细粒度形变场网络$\boldsymbol{F^d}$。具体来讲，$\boldsymbol{F^c}$首先学习人脸关键点，再解码出粗粒度形变场。$\boldsymbol{F^d}$进一步完善粗粒度形变场，得到与原图同分辨率的细粒度形变场，其输入包含$\boldsymbol{F^c}$的输出与一个分支网络$\boldsymbol{F^g}$从原图学到的补充细节两个部分。

CFMF-Net 通过学习到的形变场对图像进行变换，因而，其输出图像的像素值都是来自于原图，保留了更多的身份信息，减少了额外噪声的引入。相比于 2D 方法通过回归像素值来生成正脸图像，本文方法通过学习形变场来进行正面化，从而限制了正面化图像中的像素均来自于原图，更好地保持了原始信息。相比于 3D 方法基于 3D 模型规则计算形变场，本文方法得到的形变场是基于学习得到的，从而能够得到更逼真的正面化结果。

2.1 形式化

本文方法利用成对的人脸数据$\{(\boldsymbol{I}_1, \boldsymbol{I}_1^{\boldsymbol{gt}}), \ldots, (\boldsymbol{I}_k, \boldsymbol{I}_k^{\boldsymbol{gt}}), \ldots, (\boldsymbol{I}_n, \boldsymbol{I}_n^{\boldsymbol{gt}})\}$来进行训练，其中$\boldsymbol{I}_k \in \boldsymbol{R}^{h\times w}$与$\boldsymbol{I}_k^{\boldsymbol{gt}} \in \boldsymbol{R}^{h\times w}$分别代表输入人脸图像和它对应的正面人脸图像。CFMF-Net 通过一个深度网络学习形变场来正面化输入图像$\boldsymbol{I}_k$，得到估计的正面人脸图像$\boldsymbol{I}_k^{\boldsymbol{est}}$，其目标为最小化真实正脸图像和估计得到的正脸图像的差别，即：

$$\min_{\boldsymbol{W}} \frac{1}{n}\sum_{k=1}^{n} || \boldsymbol{I}_k^{\boldsymbol{est}} - \boldsymbol{I}_k^{\boldsymbol{gt}} ||_2^2, \tag{1}$$

式中$\boldsymbol{W}$为整个模型的可学习参数。

更具体地讲，$\boldsymbol{I}_k^{\boldsymbol{est}}$是由形变模块$\boldsymbol{T}$将形变场网络$\boldsymbol{F}$得到的形变场$\boldsymbol{F}(\boldsymbol{I}_k)$作用于输入图像$\boldsymbol{I}_k$形变而来，即：

$$\boldsymbol{I}_k^{\boldsymbol{est}} = \boldsymbol{T}\big(\boldsymbol{I}_k, \boldsymbol{F}(\boldsymbol{I}_k)\big). \tag{2}$$

下面依次阐述每个模块的细节。

2.1.1　形变场学习网络

根据形变场的定义，可以得到输出图像$\boldsymbol{I}_k^{\boldsymbol{est}}$上位于$(i, j)$位置的像素点$\boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}$的形变场为：

$$\boldsymbol{D}_{k,i,j} = (\Delta_{k,i,j}^{h}, \Delta_{k,i,j}^{w}), \tag{3}$$

其表示该像素点取自于输入图像$\boldsymbol{I}_k$的$(i + \Delta_{k,i,j}^{h},\ j + \Delta_{k,i,j}^{w})$位置。如果$i + \Delta_{k,i,j}^{h}$，$j + \Delta_{k,i,j}^{w}$超出图像空间范围，则分别对$\boldsymbol{h}$，$\boldsymbol{w}$取余。为了表示简单，后文仍将最终坐标写作$i + \Delta_{k,i,j}^{h}$，$j + \Delta_{k,i,j}^{w}$。

接下来依次介绍 CFMF-Net 的两个重要组成部分，即粗粒度形变场网络$\boldsymbol{F^c}$和细粒度的形变场$\boldsymbol{F^d}$。

$\boldsymbol{F^c}$的目标为学习输入到输出图像的人脸结构的

主要变化，其目标为得到一个大小为$\frac{h}{4}\times\frac{w}{4}$的粗粒度形变场$\boldsymbol{C}_k$，即：

$$\boldsymbol{S}_k = \boldsymbol{F}^{\boldsymbol{s}}(\boldsymbol{I}_k), \tag{4}$$

$$\boldsymbol{C}_k = \boldsymbol{F}^{\boldsymbol{c}}(\boldsymbol{S}_k) \tag{5}$$

式中$\boldsymbol{F}^{\boldsymbol{s}}, \boldsymbol{F}^{\boldsymbol{c}}$是两个连接在一起的卷积网络，它们的参数分别为$\boldsymbol{W}^{\boldsymbol{s}}, \boldsymbol{W}^{\boldsymbol{c}}$。$\boldsymbol{S}_k \in \boldsymbol{R}^{68\times 2}$为 68 个稀疏人脸关键点的位置表示，作为人脸结构鲁棒特征表示用来指导粗粒度形变场的学习。而学得的形变场$\boldsymbol{C}_k$将作为学习大小为$\boldsymbol{h}\times\boldsymbol{w}$的细粒度形变场的中间表示，以打下细粒度形变场学习的良好基础。

$\boldsymbol{C}_k$建模了输入到输出人脸图像的主要形变，但$\boldsymbol{C}_k$忽略了细节的变化，因此还需要进一步细化。在 CFMF-Net 中，一个分支网络$\boldsymbol{F}^{\boldsymbol{g}}$被用来提取原始图片$\boldsymbol{I}_k$的细节特征$\boldsymbol{G}_k = \boldsymbol{F}^{\boldsymbol{g}}(\boldsymbol{I}_k)$，其中$\boldsymbol{F}^{\boldsymbol{g}}$的参数为$\boldsymbol{W}^{\boldsymbol{g}}$。之后，将$\boldsymbol{C}_k$与$\boldsymbol{G}_k$拼接在一起，输入到细粒度形变场网络$\boldsymbol{F}^{\boldsymbol{d}}$中，得到与原图分辨率大小相同的细粒度形变场$\boldsymbol{D}_{\boldsymbol{k}} \in \boldsymbol{R}^{h\times w}$：

$$\boldsymbol{D}_k = \boldsymbol{F}^{\boldsymbol{d}}([\boldsymbol{C}_k, \boldsymbol{G}_k]), \tag{6}$$

式中$\boldsymbol{F}^{\boldsymbol{d}}$为反卷积网络，可以对粗粒度形变场进行上采样，其参数为$\boldsymbol{W}^{\boldsymbol{d}}$。

2.1.2　形变模块

当得到形变场$\boldsymbol{D}_k$之后，一个形变模块$\boldsymbol{T}$将$\boldsymbol{D}_k$作用于原图$\boldsymbol{I}_k$以得到正面化后的图像$\boldsymbol{I}_k^{\boldsymbol{est}}$，即：

$$\boldsymbol{I}_k^{\boldsymbol{est}} = \boldsymbol{T}(\boldsymbol{I}_k, \boldsymbol{D}_k). \tag{7}$$

$\boldsymbol{T}$通过形变场$\boldsymbol{D}_k$将原始图像$\boldsymbol{I}_k$的像素点进行重组得到$\boldsymbol{I}_k^{\boldsymbol{est}}$，这个过程没有可学习参数。如果$\boldsymbol{D}_{k,i,j} = (\Delta_{k,i,j}^h, \Delta_{k,i,j}^w)$是整数，$\boldsymbol{I}_k^{\boldsymbol{est}}$中位于坐标$(i,j)$的像素就直接取自于原图$\boldsymbol{I}_k$中位于坐标$(i+\Delta_{k,i,j}^h, j+\Delta_{k,i,j}^w)$上的像素，即：

$$\boldsymbol{I}_{k,i,j}^{\boldsymbol{est}} = \boldsymbol{I}_{k,i+\Delta_{k,i,j}^h, j+\Delta_{k,i,j}^w}. \tag{8}$$

一般情况下，$\boldsymbol{D}_{k,i,j}$是实数（为了方便求导，本文中并不会限制它是整数）。此时，$\boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}$像素值为$\boldsymbol{I}_k$$(i+\Delta_{k,i,j}^h, j+\Delta_{k,i,j}^w)$邻近四个像素点像素值的双线性插值。令$\tilde{i} = i + \Delta_{k,i,j}^h$，$\tilde{j} = j + \Delta_{k,i,j}^w$，则：

$$\begin{aligned}\boldsymbol{I}_{k,i,j}^{\boldsymbol{est}} = {} & (1-|\lfloor\tilde{i}\rfloor - \tilde{i}|)\times(1-|\lfloor\tilde{j}\rfloor - \tilde{j}|)\times \boldsymbol{I}_{k,\lfloor\tilde{i}\rfloor,\lfloor\tilde{j}\rfloor} \\ & + (1-|\lfloor\tilde{i}\rfloor - \tilde{i}|)\times(1-|\lceil\tilde{j}\rceil - \tilde{j}|)\times \boldsymbol{I}_{k,\lfloor\tilde{i}\rfloor,\lceil\tilde{j}\rceil} \\ & + (1-|\lceil\tilde{i}\rceil - \tilde{i}|)\times(1-|\lfloor\tilde{j}\rfloor - \tilde{j}|)\times \boldsymbol{I}_{k,\lceil\tilde{i}\rceil,\lfloor\tilde{j}\rfloor} \\ & + (1-|\lceil\tilde{i}\rceil - \tilde{i}|)\times(1-|\lceil\tilde{j}\rceil - \tilde{j}|)\times \boldsymbol{I}_{k,\lceil\tilde{i}\rceil,\lceil\tilde{j}\rceil}.\end{aligned} \tag{9}$$

容易看出，等式(8)是等式(9)的特殊情况。从等式(8)和(9)可以看出，正面化图像$\boldsymbol{I}_k^{\boldsymbol{est}}$中所有的像素点都来自原图$\boldsymbol{I}_k$某一个像素点或者由四个临近点加权得到。因此，$\boldsymbol{I}_k^{\boldsymbol{est}}$极大地保留了$\boldsymbol{I}_k$中的原始信息。

2.1.3　整体训练目标

本文方法通过端到端的学习方式来优化整个形变场网络 CFMF-Net=$\{\boldsymbol{F}^{\boldsymbol{s}}, \boldsymbol{F}^{\boldsymbol{c}}, \boldsymbol{F}^{\boldsymbol{d}}, \boldsymbol{F}^{\boldsymbol{g}}\}$。其目标为正面化后的人脸图像$\boldsymbol{I}_k^{\boldsymbol{est}}$与真实的正脸图像$\boldsymbol{I}_k^{\boldsymbol{gt}}$尽量相同，即：

$$\begin{aligned}L &= \min_{\boldsymbol{W}} \frac{1}{n}\sum_{k=1}^{n} ||\boldsymbol{I}_k^{\boldsymbol{est}} - \boldsymbol{I}_k^{\boldsymbol{gt}}||_2^2 \\ &= \min_{\substack{\boldsymbol{W}^{\boldsymbol{s}},\boldsymbol{W}^{\boldsymbol{c}},\\ \boldsymbol{W}^{\boldsymbol{d}},\boldsymbol{W}^{\boldsymbol{g}}}} \frac{1}{n}\sum_{k=1}^{n} ||\boldsymbol{T}\left(\boldsymbol{I}_k, \boldsymbol{F}^{\boldsymbol{d}}([\boldsymbol{F}^{\boldsymbol{c}}(\boldsymbol{F}^{\boldsymbol{s}}(\boldsymbol{I}_k)), \boldsymbol{F}^{\boldsymbol{g}}(\boldsymbol{I}_k)])\right) - \boldsymbol{I}_k^{\boldsymbol{gt}}||_2^2.\end{aligned} \tag{10}$$

2.2　优化过程

为了加快 CFMF-Net 的收敛，首先预训练 CFMF-Net 每个模块，得到一个较好的初始化参数，再以等式(10)为目标进行端到端的训练。

2.2.1 预训练

如前所述，粗粒度形变场学习中的$\boldsymbol{F}^{\boldsymbol{s}}$用来学习人脸关键点位置$\boldsymbol{S}_k$。此处用人脸关键点对$\boldsymbol{F}^{\boldsymbol{s}}$进行优化（如果没有标定的关键点，这一步也可以省略），即：

$$L^S = \min_{\boldsymbol{W}^{\boldsymbol{s}}} \frac{1}{n}\sum_{k=1}^{n} ||\boldsymbol{F}^{\boldsymbol{s}}(\boldsymbol{I}_k) - \boldsymbol{S}_k^{\boldsymbol{gt}}||_2^2, \tag{11}$$

式中，$\boldsymbol{S}_k^{\boldsymbol{gt}}$是人工标注的人脸关键点，如图 3 所示。这里，$\boldsymbol{F}^{\boldsymbol{s}}$通过梯度下降进行优化，梯度为$\frac{\partial L^S}{\partial \boldsymbol{W}^{\boldsymbol{s}}}$。

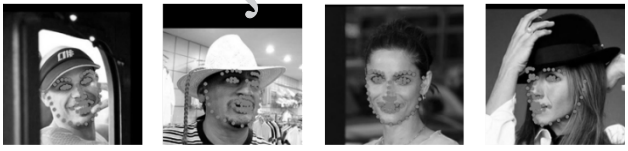


图 3　人脸关键点示例

Fig.3 Exemplars of facial landmarks

粗粒度形变场网络$\boldsymbol{F}^{\boldsymbol{c}}$以人脸关键点位置$\boldsymbol{S}_k$为输入，学习粗粒度形变场$\boldsymbol{C}_k$。本文方法借助事先计算得到的粗粒度形变场$\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$对$\boldsymbol{F}^{\boldsymbol{c}}$进行初始化，即：

$$\begin{aligned}L^c &= \min_{\boldsymbol{W}^{\boldsymbol{c}}} \frac{1}{n}\sum_{k=1}^{n} ||\boldsymbol{C}_k - \widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}||_2^2 \\ &= \min_{\boldsymbol{W}^{\boldsymbol{c}}} \frac{1}{n}\sum_{k=1}^{n} ||\boldsymbol{F}^{\boldsymbol{c}}(\boldsymbol{S}_k) - \widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}||_2^2,\end{aligned} \tag{12}$$

这里$\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$的大小为$\frac{h}{4}\times\frac{w}{4}$。为了加速预处理过程，对于

同一姿态的所有人脸图像，只取出一张代表性的人脸图像，计算得到一个统一的$\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$。给定人脸图像$\boldsymbol{I}_k$和其对应正面人脸$\boldsymbol{I}_k^{\boldsymbol{est}}$，借助这对图像的关键点，利用薄板样条插值 TPS（Bookstein，2002）粗略地估算出一个正面化形变场。但 TPS 无法填补自遮挡部分，因此，为了解决自遮挡问题，本文方法进一步利用人脸对称部分补齐被遮挡的像素点，从而得到最终估算的$\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$，具体的实现过程由图 4 说明。同样地，$\boldsymbol{F}^{\boldsymbol{c}}$也通过梯度下降进行优化，梯度为$\frac{\partial L^c}{\partial \boldsymbol{W}^{\boldsymbol{c}}}$。

类似地，借助事先计算得到的细粒度正面化形变场$\widehat{\boldsymbol{D}}_k^{\boldsymbol{gt}}$可以对细粒度形变场网络$\boldsymbol{F}^{\boldsymbol{d}}$和细节分支网络$\boldsymbol{F}^{\boldsymbol{g}}$一同进行初始化训练，即：

$$L^d = \min_{\boldsymbol{W}^{\boldsymbol{g}},\boldsymbol{W}^{\boldsymbol{d}}} \frac{1}{n}\sum_{k=1}^{n} ||\boldsymbol{D}_k - \widehat{\boldsymbol{D}}_k^{\boldsymbol{gt}}||_2^2 = \min_{\boldsymbol{W}^{\boldsymbol{g}},\boldsymbol{W}^{\boldsymbol{d}}} \frac{1}{n}\sum_{k=1}^{n} ||\boldsymbol{F}^{\boldsymbol{d}}([\boldsymbol{C}_k, \boldsymbol{F}^{\boldsymbol{g}}(\boldsymbol{I}_k)]) - \widehat{\boldsymbol{D}}_k^{\boldsymbol{gt}}||_2^2. \quad (13)$$

类似于$\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$，$\widehat{\boldsymbol{D}}_k^{\boldsymbol{gt}}$也同样由 TPS 得到，只是其大小为$\boldsymbol{h} * \boldsymbol{w}$，可以看作是$\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$的上采样。同样地，$\boldsymbol{F}^{\boldsymbol{d}}, \boldsymbol{F}^{\boldsymbol{g}}$也通过梯度下降进行优化，梯度分别为$\frac{\partial L^d}{\partial \boldsymbol{W}^{\boldsymbol{d}}}$，$\frac{\partial L^d}{\partial \boldsymbol{W}^{\boldsymbol{g}}}$。

2.2.2 端到端调优

在预训练的基础上，CFMF-Net 以等式(10)为目标对网络进行端到端的优化。

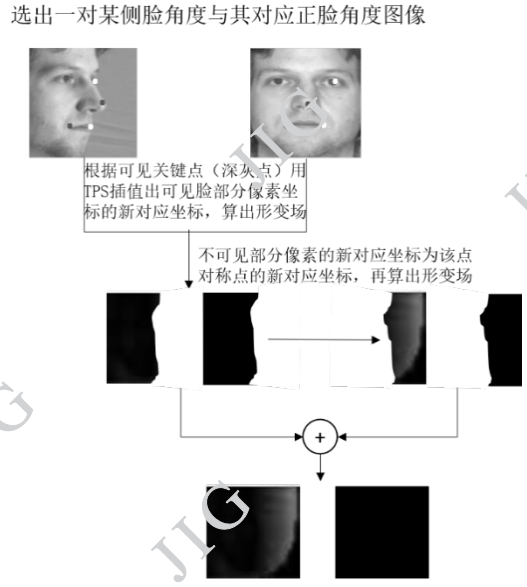


图 4 CFMF-Net 预训练时粗略估计$\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$过程

Fig.4 The estimating process of $\widehat{\boldsymbol{C}}_k^{\boldsymbol{gt}}$ during pretraining

由等式(10)，优化目标L关于$\boldsymbol{I}_k^{\boldsymbol{est}}$的导数为：

$$\frac{\partial L}{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}} = 2\left(\boldsymbol{I}_{k,i,j}^{\boldsymbol{est}} - \boldsymbol{I}_{k,i,j}^{\boldsymbol{gt}}\right). \quad (14)$$

由等式(9)，$\boldsymbol{I}_k^{\boldsymbol{est}}$关于形变场$\boldsymbol{D}_k$的导数为对每个像素点进行求导，分为长方向和宽方向上的形变场两个部分（$\frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \Delta_{k,i,j}^{h}}, \frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \Delta_{k,i,j}^{w}}$），具体而言：

$$\begin{aligned}\frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \Delta_{k,i,j}^{h}} &= \frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \tilde{i}} \times \frac{\partial \tilde{i}}{\partial \Delta_{k,i,j}^{h}} = \frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \tilde{i}} \\ &= (\lfloor \tilde{j} \rfloor - \tilde{j} + 1) \times (\boldsymbol{I}_{k,\lfloor \tilde{i} \rfloor,\lfloor \tilde{j} \rfloor} - \boldsymbol{I}_{k,\lfloor \tilde{i} \rfloor,\lceil \tilde{j} \rceil}) \\ &\quad + (\lceil \tilde{j} \rceil - \tilde{j} + 1) \times (\boldsymbol{I}_{k,\lceil \tilde{i} \rceil,\lfloor \tilde{j} \rfloor} - \boldsymbol{I}_{k,\lceil \tilde{i} \rceil,\lceil \tilde{j} \rceil}), \end{aligned} \quad (15)$$

$$\begin{aligned}\frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \Delta_{k,i,j}^{w}} &= \frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \tilde{j}} \times \frac{\partial \tilde{j}}{\partial \Delta_{k,i,j}^{w}} = \frac{\partial \boldsymbol{I}_{k,i,j}^{\boldsymbol{est}}}{\partial \tilde{j}} \\ &= (\lfloor \tilde{i} \rfloor - \tilde{i} + 1) \times (\boldsymbol{I}_{k,\lfloor \tilde{i} \rfloor,\lfloor \tilde{j} \rfloor} - \boldsymbol{I}_{k,\lceil \tilde{i} \rceil,\lfloor \tilde{j} \rfloor}) \\ &\quad + (\lceil \tilde{i} \rceil - \tilde{i} + 1) \times (\boldsymbol{I}_{k,\lceil \tilde{i} \rceil,\lfloor \tilde{j} \rfloor} - \boldsymbol{I}_{k,\lceil \tilde{i} \rceil,\lceil \tilde{j} \rceil}). \end{aligned} \quad (16)$$

整个 CFMF-Net 网络参数$\{\boldsymbol{W}^{\boldsymbol{d}}, \boldsymbol{W}^{\boldsymbol{g}}, \boldsymbol{W}^{\boldsymbol{c}}, \boldsymbol{W}^{\boldsymbol{s}}\}$通过梯度下降进行优化，对应每个模块的梯度如下：

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{W}^{\boldsymbol{d}}} &= \frac{\partial L}{\partial \boldsymbol{I}_k^{\boldsymbol{est}}} \times \frac{\partial \boldsymbol{I}_k^{\boldsymbol{est}}}{\partial \boldsymbol{D}_k} \times \frac{\partial \boldsymbol{D}_k}{\partial \boldsymbol{W}^{\boldsymbol{d}}}, \\ \frac{\partial L}{\partial \boldsymbol{W}^{\boldsymbol{g}}} &= \frac{\partial L}{\partial \boldsymbol{W}^{\boldsymbol{d}}} \times \frac{\partial \boldsymbol{W}^{\boldsymbol{d}}}{\partial \boldsymbol{W}^{\boldsymbol{g}}}, \\ \frac{\partial L}{\partial \boldsymbol{W}^{\boldsymbol{c}}} &= \frac{\partial L}{\partial \boldsymbol{W}^{\boldsymbol{d}}} \times \frac{\partial \boldsymbol{W}^{\boldsymbol{d}}}{\partial \boldsymbol{W}^{\boldsymbol{c}}}, \\ \frac{\partial L}{\partial \boldsymbol{W}^{\boldsymbol{s}}} &= \frac{\partial L}{\partial \boldsymbol{W}^{\boldsymbol{d}}} \times \frac{\partial \boldsymbol{W}^{\boldsymbol{d}}}{\partial \boldsymbol{W}^{\boldsymbol{c}}} \times \frac{\partial \boldsymbol{W}^{\boldsymbol{c}}}{\partial \boldsymbol{W}^{\boldsymbol{s}}}. \end{aligned} \quad (17)$$

3 实 验

为验证本文方法对大姿态人脸识别问题的有效性，本节在四个代表性大姿态人脸识别数据集上进行实验，包括通用人脸识别数据集 LFW、包含更多更极端姿态变化的数据集 MultiPIE 、CFP、IJB-A。本节将首先介绍大姿态人脸识别实验中使用的数据集，接着展示本文方法与其它方法在大姿态人脸识别上的对比实验结果，最后通过消融实验验证所提出的由粗到细学习策略的有效性。

3.1 数据集与实验设置

表1 训练集和测试集的设置说明

Table 1 Overview of training and testing datasets

测试集	条件是否可控	是否包含大姿态	训练集	
			正面化	识别
MultiPIE	是	是	MultiPIE	MultiPIE
LFW	否	否	300W-LP	Webface
CFP	否	是	300W-LP	Webface
IJB-A	否	是	300W-LP	Webface+IJB-A

本文在 MultiPIE 数据集（Sim 等人，2013）上进行可控场景下的大姿态人脸识别实验。在 300W-LP（Zhu 等人，2015）、Webface（Yi 等人，2014）、LFW（Huang 等人，2014）、CFP（Sengupta 等人，2016）和 IJB-A（Klare 等人，2015）上进行非可控场景下的大姿态人脸识别实验。在所有实验中，首先通过 CFMF-Net 进行人脸正面化，之后再通过一个人脸识别网络来进行人脸识别。其中，300W-LP（Zhu 等人，2015）为 CFMF-Net 网络的训练集，Webface（Yi 等人，2014）为人脸识别训练集，LFW（Huang 等人，2014）、CFP（Sengupta 等人，2016）和 IJB-A（Klare 等人，2015）为人脸识别测试集。数据集的主要情况参见表 1。更具体的细节将在 3.1.1 节中阐述。

通过裁剪缩放，所有的人脸图像都被调整至 128×128 大小，像素值被归一化至[-1,1]。图像坐标值被归一化到[0,1]，形变场被归一到到[-1,1]。不同实验的 CFMF-Net 网络结果在图 5 中展示。

3.1.1 数据集

MultiPIE 数据集（Sim 等人，2013）是最常用的可控场景下的大姿态人脸识别数据集，包含 337 个人在不同姿态、光照和表情下的照片。实验中采用与大姿态人脸识别的代表性工作 Cao 等人（2018）等相同的实验设置，即取前 200 个人的所有图像进行人脸正面化和识别的训练，剩下 137 个人的所有图像进行测试。在测试阶段，采用这 137 个人的正面姿态、光照和中性表情的照片作为 gallery，剩下 72000 张照片作为 probe。与大多数对比方法相同，在 MultiPIE 的实验中，本方法采用 LightCNN-29（Wu 等人，2018）作为识别网络。

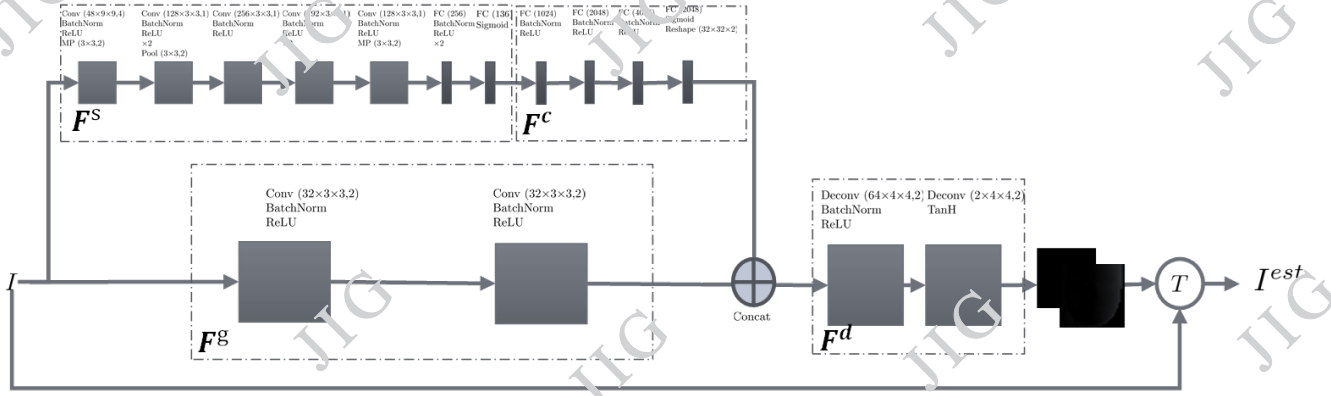


图 5 CFMF-Net 网络结构

Fig.5 Architecture of CFMF-Net

LFW（Huang 等人，2014）和 CFP（Sengupta 等人，2016）是两个经典的非可控场景下的人脸识别数据集，通常被用来测试人脸识别方法的性能。LFW 总共包含 13233 张采集自网络的人脸图像，其中通常用于人脸识别测试的部分为 3000 对来自于同一人的图像与 3000 对来自于不同人的图像。CFP 总共包含来自于 500 个人的 7000 张图像，其中每个人都有 10 张准正面（小于 10 度）图像与 4 张大姿态（大于 10 度）的图像。在本文的实验中， LFW 和 CFP 被用来进行人脸验证实验。在 LFW 上的测试指标为人脸验证准确率 ACC 与 ROC 曲线下的面积 AUC，在 CFP 上的测试包含正脸-正脸图像对（FF）和正脸-侧脸图像对（FP）两个部分，其测试指标为人脸验证准确率 ACC。同样地，在 LFW 和 CFP 的实验中，本方法用 LightCNN-29（Wu 等人，2018）作为识别网络。

IJB-A（Klare 等人，2015）是更大的不可控场景下的人脸识别数据集，它主要用来测试大姿态人脸识别方法的性能。因为 IJB-A 中包含很多极端姿态和光照条件下的人脸图像，相比于前面介绍的测试数据集，它更具有挑战性。IJB-A 总共包含来自 500 个人的 5396 张网络图片与 20412 张截取自网络视频的图片。其测试协议为十折交叉验证，每次划分出 333 人的图像作为训练集，剩余 167 人的图像作

为测试集，方法最终报告的准确率为十次实验的平均准确率。在多数方法中，首先它们在一个更大数据集（如 Webface）上训练一个识别模型，再用 IJB-A 的小训练集进行微调（Klare 等人，2015）。相比于之前介绍的数据集，IJB-A 上的测试不再是单一图片的对比，而是图片集合之间的对比。它的测试包含人脸验证和人脸识别两个部分。人脸验证的指标为在某个指定错误接受率（FAR）下的正确接受率（TAR）。人脸识别通常为闭集测试，其指标为第一名准确率和前五名准确率。在之前的方法中，IJB-A 上的测试没有统一的训练集和训练网络结构，为了与之前的方法公平比较，本方法采用了两个不同的人脸识别网络，分别为 Fast AlexNet 和 LightCNN-29（Wu 等人，2018）。其中，Fast AlexNet 是对 AlexNet 进行优化后得到的模型，与大多数已有方法的模型能力相当，但收敛速度更快，具体结构参见表 2。

表2 Fast AlexNet网络结构

Table 2 Architecture of our Fast AlexNet

模块	结构	模块	结构
1	Conv 48X9X9, S4, BatchNorm ReLU MAXPOOL 3X3, S2	6	Conv 192X3X3, S1, BatchNorm ReLU
2	Conv 128X3X3, S1, BatchNorm ReLU	7	Conv 128X3X3, S1, BatchNorm ReLU MAXPOOL 3X3, S2
3	Conv 128X3X3, S1, BatchNorm ReLU MAXPOOL 3X3,S2	8	FC 4096 BatchNorm ReLU
4	Conv 256X3X3, S1, BatchNorm ReLU	9	FC 2048 BatchNorm ReLU
5	Conv 192X3X3, S1, BatchNorm ReLU		

300W-LP 是人脸姿态增广方法（Zhu 等人，2015）对 300W 数据集（Sagonas 等人，2016）增广后构建的增广数据集，包含 122450 张图像。实验中每个正面图像与对应的所有增广侧面图像作为人脸正面化训练集，而该准正面图像就作为训练目标[1]。并且，300W-LP 每张人脸图像包含 68 个人脸关键点标注，可以作为$\boldsymbol{S}_k^{gt}$对$\boldsymbol{F}^s$进行预训练。

Webface（Yi 等人，2014）是一个通用人脸识别训练集，包含来自于 10575 个人的 494414 张图像。实验中使用 Webface 训练非可控条件下的人脸识别模型。

3.2 实验结果

CFMF-Net 在 MultiPIE、LFW 和 CFP 上的实验中同多种方法进行对比，包括多任务学习方法 Yin 等人（2017a），与本文方法同为图像生成类的基于 GAN 的方法 Luan 等人（2017）、Yin 等人（2017b）、Zhao 等人（2018a）、Zhao 等人（2018b）和 Cao 等人（2018）。其中 Luan 等人（2017）是一种直接基于 GAN 的 2D 人脸正面化方法。Yin 等人（2017b）在 DR-GAN 的基础上进一步抽取了 3DMM 的系数作为特征，从而更好地保持了人脸结构信息。Cao 等人（2018）首先将形变场作用于原图得到正面化人脸，再以此为中间结果做进一步调整。在 IJB-A 的实验中，本文对比了不同类型的方法，包括特征解耦方法（Crosswhite 等人，2017；Yang 等人，2017；Zhao 等人，2017）、人脸增广方法（Zhu 等人，2016；Masi 等人，2017；Chang 等人，2017）和人脸正面化方法（Luan 等人，2017；Yin 等人，2017b；Zhao 等人，2018b；Cao 等人，2018）。值得一提的是，2019 年以后出现的方法多为通用人脸识别方法，极少针对大姿态人脸识别这一特定问题专门研究，本文与 ArcFace（Deng 等人，2019）采用 ResNetSE50 网络结构（网络能力与本文方法网络差不多）的版本[2]进行比较。

通过表 3 可以看到，在 MultiPIE 数据集上，本文方法得到了比之前方法更好的结果，尤其是在 75°到 90°的大姿态人脸下。值得一提的是，同样是利用深度网络自动学习形变场的方法，Hu 等人（2017）由于结构简单，在表 3 所采用的更复杂的数据集上并不能收敛，且在较小分辨率 32×32 的 MultiPIE 上

[1] 训练数据列表见 https://github.com/whobefore/MF-Net/tree/master/Data/300W-LP

[2] https://github.com/TreB1eN/InsightFace_Pytorch

的平均性能比 CFMF-Net 低 0.4%。

表3 MultiPIE数据集上的识别率

Table 3 Face Recognition Accuracy (mAC) on MultiPIE dataset

方法	±90	±75	±60	±45	±30	±15
Luan 等，2017	-	-	83.20	86.20	90.10	94.00
Huang 等，2017	64.64	77.43	87.72	95.38	98.06	98.68
Yin 等，2017a	76.96	87.83	92.07	90.34	98.01	99.19
Zhao 等，2018a	86.73	95.21	98.37	98.81	99.48	99.64
Cao 等，2018	92.32	96.40	99.14	99.88	**99.98**	99.99
CFMF-Net（本文方法）	**94.00**	**97.76**	**99.36**	**99.94**	**99.98**	**100.0**

表4 LFW数据集上的人脸验证准确率ACC和AUC

Table 4 Face Verification Accuracy (ACC) and Area Under Curve (AUC) on LFW dataset

方法	ACC	AUC
Zhu 等人（2015）	96.25	99.39
Yin 等人（2017b）	96.42	99.45
Cao 等人（2018）	**99.41**	**99.92**
CFMF-Net（本文方法）	99.02	**99.92**

LFW（Huang等人，2014）和CFP（Sengupta 等人，2016）上的实验结果如表4和5所示。可以看出，本文方法在正面人脸居多的测试中与当前最好的方法性能相当，包括采用更大训练集的Deng 等人（2019）方法。而在正脸-侧脸的识别上则取得了更好的性能（如表5所示）。这验证了CFMF-Net的正面化对侧面人脸的识别起到了重要的作用。从图4中也可以看到，在LFW数据集上，本方法得到了保持原始信息的正面化人脸。

表5 CFP数据集上的人脸验证准确率ACC

Table 5 Face Verification Accuracy on CFP dataset

方法	ACC（正脸-正脸）	ACC（正脸-侧脸）	ACC（平均）
Luan 等人（2017）	97.84	93.41	95.63
Yin 等人（2017b）	97.79	94.39	96.09
Zhao 等人（2018b）	99.44	93.10	96.27
Deng 等人（2019）	**99.62**	95.04	**97.33**
CFMF-Net（本文方法）	99.34	**95.17**	97.26

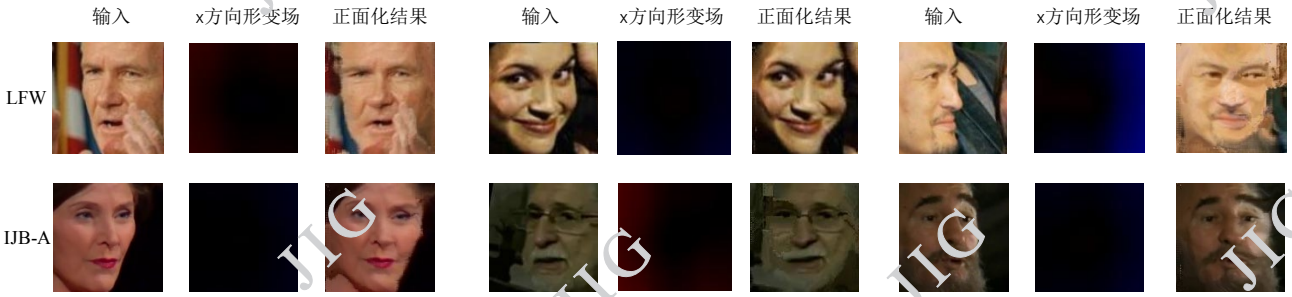


图 6 CFMF-Net 在 LFW 和 IJB-A 上的正面化结果示例

Fig.6 Exemplars of frontalization results on LFW and IJB-A of CFMF-Net

IJB-A 的实验结果如表 6 所示。在人脸正面化类方法中，本文方法与当前最好的方法效果相当。

在表 6 所有方法中，Masi 等人（2017），Luan 等人（2017）以及 Yin 等人（2017b）与 CFMF-Net1 具有相似的训练集和识别网络，这里将它们单独进行对比。可以看到，CFMF-Net1 取得了更好的识别效果。这是因为 Masi 等人（2017）是基于 3D 模型的规则进行正面化的方法，生成的正面人脸不够逼真，而 2D 回归生成方法（Luan 等人，2017；Yin 等人，2017b）没有充分地保留原图中的有效信息。本文方法以最大化真实正面人脸与生成正面人脸的相似度为目标，学习了原图与正面化图像的形变场，通过重组原图像素点得到正面化的图像，保证了生成图像的所有像素都来自于原图，因此本文方法结合了 3D 和 2D 方法的优势，既保持了原始身份信息，又保证了生成图像足够逼真，其正面化结果如图 6 所示。与结合了 GAN 与密集形变场的方法（Zhao 等人，2018b；Cao 等人，2018）相比，本文方法 CFMF-Net2 仅仅通过简单的形变场回归来正面化人脸，得到了和这些复杂方法持平的效果。

值得一提的是，可以看到，当前数据集的人脸图像主要的变化在 yaw 方向，即本文中的 x 方向，一种自然的想法是是否能够通过加强 x 方向形变场的训练权重来提升性能。然而实际上这种做法对性能几乎没有影响，因为 CFMF-Net 可以自动学习到形变场的主要变化在 x 方向。此外给 x 方向形变场

更多训练权重可能对可扩展性有影响，因为现实中的人脸图像还会存在其它方向上的姿态变化。

3.3 消融实验

为了分析 CFMF-Net 每个模块对人脸正面化和识别的影响，本文还进行了一系列消融实验。在 300W-LP 数据集上消融实验的可视化结果如图 7 所示。可以看到，通过 TPS 可以得到一个基本的人脸正面化结果，如图 7(b)所示。直接利用粗粒度形变场得到的人脸正面化图像，由于自遮挡问题，依然存在一定程度的失真，如图 7(c)所示。而借助细粒度形变场，可以得到逼真的正面化人脸图像，如图 7(d)所示。这验证了 CFMF-Net 各部分对正面化的作用。

从识别结果的角度来看，CFMF-Net 的每一部分对人脸识别的准确率也都有重要的作用。以在 IJB-A 上的实验为例，从表 7 可以看出，相比于不进行人脸正面化直接用 Fast AlexNet 进行人脸识别，使用粗粒度形变场进行正面化能一定程度地提升人脸识别的准确率。而使用细粒度形变场进行人脸正面化，则能进一步提升识别的准确率。

表6 IJB-A数据集上验证和识别结果

Table 6 Face Verification and Identification on IJB-A dataset

方法		人脸验证		人脸识别		训练集	方法
		FAR=0.01	FAR=0.001	Top-1	Top-5	图像转化/识别	
特征解耦	Crosswhite 等人（2017）	93.9	83.6	92.8	97.7	-/VGGFace	VGGNet16
	Yang 等人（2017）	94.1	88.1	95.8	98.0	-/3M ims of 50K ids	GoogLeNet-BN
	Zhao 等人（2017）	97.6	93.0	97.1	98.9	300W-LP/MS-Celeb	ResNeXT50+GoogLeNet-BN
人脸增广	Zhu 等人（2016）	89.0	82.8	90.3	92.8	300W-LP/MS-Celeb	-
	Masi 等人（2017）	88.8	75.0	92.6	96.6	-/Webface	VGGNet
	Chang 等人（2017）	90.1	85.2	91.4	93.0	-/MS-Celeb	ResNet101
人脸正面化	Luan 等人（2017）	77.4	53.9	85.5	95.7	MultiPIE/ Webface	CASIA-Net
	Yin 等人（2017b）	85.2	66.3	90.2	95.4	300W-LP/ Webface	CASIA-Net
	Zhao 等人（2018b）	93.3	87.5	94.4	-	MultiPIE/ -	LightCNN-29
	Cao 等人（2018）	95.2	**89.7**	**96.1**	97.9	CelebA-HQ/-	LightCNN-29
	CFMF-Net1（本文方法）	90.7	77.3	94.7	98.1	300W-LP/ Webface	Fast AlexNet
	CFMF-Net2（本文方法）	**95.3**	86.4	95.4	**98.4**	300W-LP/ Webface	LightCNN-29

(a)原始输入

(c)粗粒度形变场（CFMF-Net w/o $\boldsymbol{F^r}, \boldsymbol{F^d}$）转化结果

(b)TPS 转化结果

(d)CFMF-Net 最终转化结果

图 7 CFMF-Net 在 300W-LP 上消融实验的结果

Fig. 7 Ablation study of frontalization on 300W-LP

((a)input；(b)TPS；(c)CFMF-Net w/o $\boldsymbol{F^g}, \boldsymbol{F^d}$；(d)CFMF-Net)

表7 CFMF-Net在IJB-A上的消融实验

Table 7 Ablation study of CFMF-Net on IJB-A

方法	人脸验证		人脸识别	
	FAR=0.01	FAR=0.001	Top-1	Top-5
w/o CFMF-Net	85.2	60.0	90.9	97.1
CFMF-Net w/o F^g, F^d	88.8	70.7	92.1	97.2
CFMF-Net	**90.7**	**77.3**	**94.7**	**98.1**

表8 IJB-A上不同姿态子集的TOP-1识别率

Table 8 Top-1 recognition accuracy in our self-defined pose-subdivision test protocol on IJB-A

方法	[0, ±30)	[±30, ±60)	[±60, ±90)
w/o CFMF-Net	96.2	89.1	73.9
Deng 等人（2019）	97.6	95.2	85.2
CFMF-Net	**98.8**	**97.1**	**87.6**

为了进一步验证 CFMF-Net 对大姿态人脸的效果，本文将 IJB-A 测试集按照姿态大小划分为三组：[0, ±30)、 [±30, ±60)、 [±60, ±90) [3]。测试协议仍然与 IJB-A 人脸识别测试相同，只是每组实验再细划分为三组不同姿态的实验，即[0, ±30)的子集作为 gallery，[0, ±30)、 [±30, ±60)、 [±60, ±90)作为 probe 分别进行人脸识别测试。在每组数据上，首先用 CFMF-Net 进行人脸正面化，再用 Fast AlexNet 进行人脸识别，以测试识别准确率，并将其与直接使用 Fast AlexNet 进行识别的准确率相比较，结果如表 8 所示。通过表 8 可以看到，本文方法相比通用人脸识别方法 Deng 等人（2019），在能力相当的网络结构下取得了更好的结果，说明了现在仍然存在对姿态特殊处理的必要。另外在大姿态[±60, ±90)的测试集上，正面化后图像的识别率得到了显著提升，这进一步验证了本文方法对大姿态人脸识别的有效性。

[3] 按姿态划分后的 IJB-A 文件列表见 https://github.com/whobefore/MF-Net/tree/master/Data/IJBA

4 结 论

针对大姿态人脸识别问题，本文提出了一种基于由粗到细形变场回归的人脸正面化的方法 CFMF-Net。在实验结果中，尤其是大姿态的人脸识别实验中，本文方法表现出了比相关方法更好或持平的效果，表明该方法可以有效结合 2D 和 3D 人脸正面化方法的优点，既充分保留了原始图像中的信息，又保证了生成的正面图像足够逼真。与通用人脸识别方法的对比结果也说明了固然可以通过数据集的丰富和损失函数的设计显著提升直接进行人脸识别方法的性能，目前对人脸姿态的处理仍然存在其必要性。然而在本方法中，虽然通过由粗到细的学习方式提升了密集形变场回归的鲁棒性，但这样的算法仍然有很高的自由度，压缩形变场的冗余信息是一种更好的解决方式。未来的工作一方面希望对密集形变场进行结构可保持的稀疏化，一方面希望能够进一步设计出识别性能驱动的自动人脸或人脸特征对齐方法，发掘出最佳人脸对齐角度，并应用到更复杂场景的人脸识别中。

参考文献(References)

Asthana A, Marks T K, Jones M J, Tieu K H, and Rohith M V. 2012. Fully automatic pose-invariant face recognition via 3d pose normalization//IEEE International Conference on Computer Vision: 937-944 [DOI: 10.1109/ICCV.2011.6126336]

Bookstein L F. 2002. Principal warps: Thin-plate splines and the decomposition of deformations//IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(6): 567-585.

Cao J, Hu Y, Zhang H, He R, and Sun Z. 2018. Learning a high fidelity pose invariant model for high-resolution face frontalization//Advances in Neural Information Processing Systems: 2867-2877

Chang F J, Tran A T, Hassner T, Masi I, Nevatia R, Medioni G. 2017. Faceposenet: Making a case for landmark-free face alignment//IEEE International Conference on Computer Vision Workshop: 1599-1608 [DOI: 10.1109/ICCVW.2017.188]

Crosswhite N, Byrne J, Stauffer C, Parkhi O, Cao Q, Zisserman A. 2017. Template adaptation for face verification and identification//IEEE International Conference on Automatic Face

and Gesture Recognition: 1-8 [DOI: 10.1109/FG.2017.11]

Deng J, Guo J, Xue N, Zafeiriou S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition// IEEE Conference on Computer Vision and Pattern Recognition [DOI: 10.1109/CVPR.2019.00482]

Ding C, Xu C, and Tao D. 2015. Multi-task pose-invariant face recognition//IEEE Transactions on Image Processing, 24(3): 980-993 [DOI: 10.1109/TIP.2015.2390959]

Goodfellow I, Pougetabadie J, Mirza M, Xu B, Wardefarley D, Ozair S, Courville A, and Bengio Y. 2014. Generative adversarial nets//Advances in Neural Information Processing Systems: 2672-2680

Hu L, Kan M, Shan S, Song X and Chen X. 2017. LDF-Net: Learning a Displacement Field Network for Face Recognition across Pose. //IEEE International Conference on Automatic Face & Gesture Recognition: 9-16 [DOI: 10.1109/FG.2017.12]

Huang G B and Learned-Miller E. 2014. Labeled faces in the wild: Labeled Faces in the Wild: Updates and New Reporting Procedures. University of Massachusetts, Amherst, Technical Report UM-CS-2014-003

Huang R, Zhang S, Li T, and He R. 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis//IEEE International Conference on Computer Vision: 2439-2448 [DOI: 10.1109/ICCV.2017.267]

Kan M, Shan S, Chang H, and Chen X. 2014. Stacked progressive auto-encoders (spae) for face recognition across poses//IEEE Conference on Computer Vision and Pattern Recognition: 1883-1890 [DOI: 10.1109/CVPR.2014.243]

Kan M, Shan S, Zhang H, Lao S, and Chen X. 2016. Multi-view discriminant analysis//IEEE Transactions on Pattern Analysis and Machine Intelligence, 38 (1): 188-194 [DOI: 10.1109/TPAMI.2015.2435740]

Klare B F, Klein B, Taborsky E, and Blanton A. 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a//IEEE Conference on Computer Vision and Pattern Recognition: 1931-1939 [DOI: 10.1109/CVPR.2015.7298803]

Li A, Shan S, Chen X, and Gao W. 2009. Maximizing intra-individual correlations for face recognition across pose differences//IEEE Conference on Computer Vision and Pattern Recognition: 605-611 [DOI: 10.1109/CVPR.2009.5206659]

Li H, Hua G, Lin Z, Brandt J, and Yang J. 2013. Probabilistic elastic matching for pose variant face verification//IEEE Conference on Computer Vision and Pattern Recognition: 3499-3506 [DOI: 10.1109/CVPR.2013.449]

Li S, Liu X, Chai X, Zhang H, Lao S, and Shan S. 2012. Morphable displacement field based image matching for face recognition across pose//European Conference on Computer Vision: 102-115 [DOI: 10.1007/978-3-642-33718-5_8]

Luan T, Yin X, and Liu X. 2017. Disentangled representation learning gan for pose-invariant face recognition//IEEE Conference on Computer Vision and Pattern Recognition: 1283-1292 [DOI: 10.1109/CVPR.2017.141]

Luan X, Geng H, Liu L, Li W, Zhao Y, and Ren M. 2020. Geometry structure preserving based gan for multi-pose face frontalization and recognition//IEEE Access, 8: 104676-104687[DOI: 10.1109/ACCESS.2020.2996637]

Masi I, Hassner T, Tran A T, Medioni G. 2017. Rapid synthesis of massive face sets for improved face recognition//IEEE International Conference on Automatic Face and Gesture Recognition: 604-611. [DOI: 10.1109/34.24792]

Peng X, Yu X, Sohn K, Metaxas D N and Chandraker M. 2017. Reconstruction-Based Disentanglement for Pose-invariant Face Recognition//IEEE Conference on Computer Vision and Pattern Recognition: 1623-1632 [DOI: 10.1109/ICCV.2017.180]

Prabhu U, Heo J, and Savvides M. 2011. Unconstrained pose-invariant face recognition using 3d generic elastic models//IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(10): 1952-1961[DOI: 10.1109/TPAMI.2011.123]

Rong C, Zhang X, and Lin Y. 2020. Feature-improving generative adversarial network for face frontalization//IEEE Access, 8: 68842-68851[DOI: 10.1109/ACCESS.2020.2986079]

Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, and Pantic M. 2016. 300 faces in-the-wild challenge: database and results//Image and Vision Computing, 47(2016): 3-18 [DOI: 10.1016/j.imavis.2016.01.002]

Sengupta S, Chen J-C, Castillo C, Patel V M, Chellappa R, and Jacobs D W. 2016. Frontal to profile face verification in the wild//Winter Conference on Applications of Computer Vision (WACV): 1-9 [DOI: 10.1109/WACV.2016.7477558]

Sharma A, Jacobs D W. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch//IEEE Conference on Computer Vision and Pattern Recognition: 593-600 [DOI: 10.1109/CVPR.2011.5995350]

Sharma A, Kumar A, Daume H., and Jacobs D W. 2012. Generalized multiview analysis: A discriminative latent space//IEEE Conference on Computer Vision and Pattern Recognition: 2160-2167 [DOI: 10.1109/CVPR.2012.6247923]

Sim T, Baker S, and Bsat M. 2003. The cmu pose, illumination, and expression database//IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12): 1615-1618 [DOI: 10.1109/TPAMI.2003.1251154]

Wu X, He R, Sun Z, and Tan T. 2018. A light cnn for deep face representation with noisy labels//IEEE Transactions on Information Forensics and Security, 13(11): 2884-2896 [DOI: 10.1109/TIFS.2018.2833032]

Yang J, Ren P, Zhang D, Chen D, Wen F, Li H, Hua G. 2017. Neural aggregation network for video face recognition//IEEE Conference on Computer Vision and Pattern Recognition: 5216-5225 [DOI: 10.1109/CVPR.2017.554]

Yi D, Lei Z, Liao S, and Li S Z. 2014. Learning face representation from scratch. [EB/OL][2014-11] https://arxiv.org/pdf/1411.7923.pdf

Yin X, Xiang Y, Sohn K, Liu X, and Chandraker M. 2017. Towards large-pose face frontalization in the wild//IEEE International Conference on Computer Vision: 4010-4019 [DOI: 10.1109/ICCV.2017.430]

Yin X, and Liu X. 2017. Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing, 27(2): 964-975 [DOI: 10.1109/TIP.2017.2765830]

Zhang S, Miao Q, Huang M, Zhu X, Chen Y, Lei Z, and Wang J. 2019. Pose-weighted gan for photorealistic face frontalization//IEEE International Conference on Image Processing: 2384-2388[DOI: 10.1109/ICIP.2019.8803362]

Zhang Y, Shao M, Wong E K, and Fu Y. 2013. Random faces guided sparse many-to-one encoder for pose-invariant face recognition//IEEE International Conference on Computer Vision. 2416-2423 [DOI: 10.1109/ICCV.2013.300]

Zhao J, Xiong L, Jayashree K, Pranata S, Shen S, Feng J. 2017. Dual-agent gans for photorealistic and identity preserving profile face synthesis//Advances on Neural Information Processing Systems: 66-76

Zhao J, Xiong L, Cheng Y, Cheng Y, Li J, Zhou L, Xu Y, Karlekar J, Pranata S, Shen S, Xing J, Yan S, and Feng J. 2018. 3d-aided deep pose-invariant face recognition//International Joint Conference on Artificial Intelligence: 1184-1190 [DOI: 10.24963/ijcai.2018/165]

Zhao J, Cheng Y, Xu Y, Xiong L, Li J, Zhao F, Jayashree K, Pranata S, Shen S, Xing J, Yan S, and Feng J. 2018. Towards pose invariant face recognition in the wild//IEEE Conference on Computer vision and Pattern Recognition: 2207-2216 [DOI: 10.1109/CVPR.2018.00235]

Zhu X, Lei Z, Yan J, Dong Y, and Li S Z. 2015. High-fidelity pose and expression normalization for face recognition in the wild//IEEE Conference on Computer Vision and Pattern Recognition: 787-796 [DOI: 10.1109/CVPR.2015.7298679]

Zhu X, Lei Z, Liu X, Shi H, Li S Z. 2016. Face alignment across large poses: A 3d solution//IEEE Conference on Computer Vision and Pattern Recognition: 146-155 [DOI: 10.1109/CVPR.2016.23]

Zhu Z, Luo P, Wang X, and Tang X. 2013. Deep learning identity-preserving face space//IEEE International Conference on Computer Vision: 113-120 [DOI: 10.1109/ICCV.2013.21]

作者简介

胡蓝青，1992 年生，女，博士研究生，主要研究方向为计算机视觉、人脸识别与迁移学习。
E-mail：lanqing.hu@vipl.ict.ac.cn

山世光，通信作者，男，研究员，主要研究方向为计算机视觉、模式识别和机器学习。
E-mail：sgshan@ict.ac.cn

阚美娜，女，副研究员，主要研究方向为计算机视觉、模式识别、迁移学习与弱少监督学习。
E-mail：kanmeina@ict.ac.cn

陈熙霖，男，研究员，主要研究领域为计算机视觉、模式识别、多媒体技术以及多模式人机接口。
E-mail：xlchen@ict.ac.cn