Neurocomputing 406 (2020) 89–98

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep 3D Facial Landmark Localization on position maps

Jingchen Zhang[a], Kangkang Gao[a], Keren Fu[a,*], Peng Cheng[b]

[a] College of Computer Science, Sichuan University, Sichuan, China
[b] College of Aeronautics and Astronautics, Sichuan University, Chengdu, China

ARTICLE INFO

Article history:
Received 26 November 2019
Revised 4 March 2020
Accepted 2 April 2020
Available online 23 April 2020

Communicated by Dr. H. Yu

Keywords:
3D landmark detection
deep convolutional neural network
position map
UV map

ABSTRACT

3D facial landmark detection is a crucial step for many computer vision applications, such as 3D facial expression analysis, 3D face recognition, and 3D cephalometry. Pose variations, expression changes and self-occlusion yet make 3D facial landmark detection a very challenging task. In this paper, we propose a novel *3D Facial Landmark Localization Network* (*3DLLN*), which is robust to the above challenges. Different from existing methods, the proposed 3DLLN utilizes the position maps as an intermediate representation, from which 3DLLN detects 3D landmark coordinates. Further, we demonstrate the usage of a deep regression architecture to improve the accuracy and robustness of a large number of landmarks. The proposed scheme is evaluated on two public datasets FRGCv2 subset and BU-3DFE, and is shown to achieve state-of-the-art results.

1. Introduction

In the fields of biometrics, computer vision/graphics, and medical/surgery applications, finding key points has been a long-standing topic because of their rich semantic information. As the key points on faces, 3D facial landmarks are widely used in many face-related applications including recognition, expression analysis, semantic segmentation, rendering and relighting, and also cephalometry [1]. Therefore, an automatic and effective 3D landmark localization system is very valuable. There have been many literatures on 3D landmark detection/localization so far. Texture-based detection methods were reported to achieve decent results [2,3]. In our work, texture information is not mandatory since we aim at detecting landmarks from pure 3D shape information. This is a more general consideration for the cases where texture information acquired is inaccurate or not available. Here below we summarize three main challenges for 3D landmark detection under unconstrained conditions:

- *Expression variations*. Expression changes lead to point cloud deformation, which may make the landmarks scattered and hard to locate. Perakis et al. [4] propose a method to detect landmarks using Full Face Statistical Landmark Model (FLM) and a combination of Shape Index and Spin Image. Their method is designed to handle large expression variations. However, regarding to the accuracy of the eight landmarks on FRGCv2 dataset, this method achieves a mean error higher than 6.0mm for faces having extensive expressions, while the detection success rate is only 90.4%. Such results still have large room for further improvement.
- *Pose variations*. Pose variations may lead to the self-occlusion problem of 3D face data when viewing from an un-calibrated view, making some landmarks invisible. However, determining pose parameters without landmarks is difficult. To tackle this, Križaj J et al. [5] propose a method based on grid function and SIFT features. They use grid function to estimate the pose of a 3D face, and then with the estimated pose they fit the target landmarks by iterating on the positions of the initial landmarks. Note that this method is particularly designed for the cases where data of the invisible areas is missing. However for the cases where the invisible areas have complete data, we believe using pose correction is an essential step for subsequent landmark detection.
- *Detecting a large number of landmarks*. Some existing methods can only capture a few points of distinctive facial areas, such as nose tip, eye corners and mouse corners. This is because they resort to hand-craft features such curvature [6] and Spin Image [4]. Unfortunately, in many face-related applications such as deformation analysis, a lot of landmarks (e.g., 68 landmarks) are needed. Gilani et al. [7] solve this problem based on dense correspondence of 3D faces. Fan et al. [3] detect landmarks on a textured 3D model by conformal mapping. However on BU-3DFE dataset, their mean errors are higher than 4.3mm and standard deviations exceed 2.5mm, which are far from satisfactory.

* Corresponding author.
E-mail address: fkrsuper@scu.edu.cn (K. Fu).

https://doi.org/10.1016/j.neucom.2020.04.025
0925-2312/

Recently, deep convolutional neural networks have been deployed for 2D landmark detection [8–10] and shown to obtain impressive performance. Inspired by the work of [11–13] and to tackle the above-mentioned challenges, we make the following attempts for 3D landmark detection: (1) utilize pose normalization before detection so that the latter is facilitated; (2) find a better intermediate representation of 3D data for deep convolutional networks; (3) deploy deep convolutional network for locating a large number of 3D landmarks. As a result, we propose a novel *3D Facial Landmark Localization Network* (*3DLLN*), which is the first to introduce the UV[1] position map as an effective intermediate representation for 3D landmark discovery.

Motivation. A widely-used method to convert 3D shape information into 2D representation is using depth images, whose pixel values are relative depth from a sensor. Then, facial landmark detection [14] or even face recognition [15] can be done instead on such depth images. However, we notice two shortcomings of depth images when used for facial landmark detection. First, after conversion two points that are far away in 3D space can be spatially quite close on depth images. A small detection error in such area will inevitability lead to a high localization error on the 3D shape, e.g., the nose region that varies steeply. Second, pixel values on a depth image indicate only relative depth that may be insufficient to fully characterize a 3D shape. We introduce the UV position map [11] to remedy the above two issues. UV maps are widely used to represent 3D models [3,13]. They can fully represent 3D structures in a 2-dimensional space meanwhile avoiding self-occlusion. Position maps are recent advances [11] which go one step further upon UV maps. Position maps represent 3D coordinates as RGB information via an encoding scheme called *Projected Normalized Coordinate Code* (PNCC) [12]. By encoding, the coordinates of a 3D point are normalized into interval [0, 1] and can be represented as a color value on the UV map. The resultant UV map is called UV position map (as visualized in top-right of Fig. 3), which we believe is more informative and beneficial to 3D facial landmark detection.

In summary, the contributions of this paper are three-fold:

1. We propose a novel *3D Facial Landmark Localization Network*(*3DLLN*) that detects 3D landmarks from UV position maps. To the best of our knowledge, it is the first time that UV position maps are jointly used with deep convolutional neural network to locate a large number of 3D landmarks. Besides, unlike existing 2D landmark networks which output heatmaps, 3DLLN leverages a deep regression architecture to obtain landmarks' 3D coordinates directly. Also, unlike [11] which discovers landmarks jointly with the 3D shape by regressing an UV position map from an unconstrained 2D image, their landmarks have fixed known locations on the map (see Fig. 4 in [11]). By contrast we aim to "detect" landmarks "on" the UV position map.
2. We propose an effective pose calibration scheme for 3D facial data, which employs Shape Index and Principle Component Analysis (PCA). This ensures one to perform landmark detection regardless of pose changes.
3. The proposed 3DLLN is evaluated on two representative datasets FRGCv2 and BU-3DFE, which are widely-used for 3D point cloud research. 3DLLN is shown to surpass existing methods and attains state-of-the-art results.

The reminder of the paper is organized as follows. Section 2 describes related work on 3D facial landmark detection. Section 3 describes the proposed method in details. Experimental results, performance evaluation and comparisons are included in Section 4. Finally, conclusion is drawn in Section 5.

[1] A UV map means a 2-dimensional map projected from a 3D model's surface in UV space, where "U" and "V" denote the two axes of the map. This is because "X", "Y", and "Z" are already used to denote the axes in 3D space. Note that UV maps/mapping are common concepts in 3D modelling.

2. Related work

3D facial landmark detection has attracted the attention of researchers because of its importance in 3D research. In general, 3D facial landmark detection methods can be mainly divided into two categories: (a) Localization based on face geometry features; (b) Detection by point cloud distribution model.

In the first category, surface curvatures are firstly used as a 3D model geometric feature for landmark positioning [16–21]. For example, [17,18] use Gaussian and mean curvatures to detect landmarks of eyes, nose, lips, etc. Besides, by combining surface curvature classification and depth relief curves, peak and pit region of surfaces [18] are also detected to further obtain facial landmarks. Recently, [16] utilizes 12 descriptors coming from differential geometry to localize 13 soft-tissue landmarks from eye and nose areas based on a thresholding technique on Bosphorus and internal database. In general, this kind of methods is effective for detecting landmarks having distinct features, however the accuracy is very limited for the other landmarks left. In addition, they are not robustness for faces having extreme expression variations because the curvature will change greatly. As another representative work, Berretti et al. [22] propose a method based on SIFT features and depth map. They conduct evaluation for 9 landmarks on the BU-3DFE database. Yet even with fewer landmarks to detect, the accuracy seems low. Jahanbin et al. [23] use Gabor features computed at similarity maps for 3D landmark detection. This method relies on the pose of the 3D model, so any pose variations will affect the accuracy of landmark detection. Li et al. [6] propose using two curvature-based detectors that can repeatedly identify complementary locations with high local curvature, attempting to avoid the effect of pose on landmark detection by curvature. Their method is sensitive to noise and time consuming.

In recent years, some researchers have done further work on this basis, such as [3–5,24,25]. Among them, [5] proposes a landmark detection method for expression and pose variations by combining the SIFT feature with the designed Grid Function, and achieves a higher accuracy. However, large pose variations lead to large standard deviations and the robustness to expression variation is limited. Perakis et al. [4] propose the Full Face Statistical Landmark Model(FLM) and a combination of Shape Index and Spin Image to achieve higher accuracy on FRGCv2 for large yaw and expression variations. Fan at al. [3] attempt to solve the pose variations problem through projection, which maps the multi-pose 3D face model and corresponding texture to 2D space through Harmonic mapping, and try to accurately locate the feature points of the 3D face model in 2D space. However, this method relies on accurate 3D model texture information, and when the texture is inaccurate, accurate positions of landmarks cannot be obtained. Recently, Terada et al. [14] put forward the latest results, which use the deep convolution neural network to perform landmark detection on the depth map and return the detected landmarks to the 3D face model. Yet the results are not satisfactory. Subsequently, Paulsen et al. [26] use multi-view consensus convolutional neural networks to achieve outstanding results in 3D facial landmark detection tasks.

For the methods falling into the second category mentioned above, they try to achieve better results in 3D landmark detection research by model matching. Cootes et al. [27] propose *Active Shape Models* (*ASM*) for detecting objects. After that, Edward et al. [28] propose *Active Appearance Model*(*AAM*) to locate landmarks. Sun et al. [29] extend the *ASM* and *AAM* to 3D face modeling. Although these algorithms can locate a set of landmarks by fitting the active models, the correspondence between 2D and 3D infor-

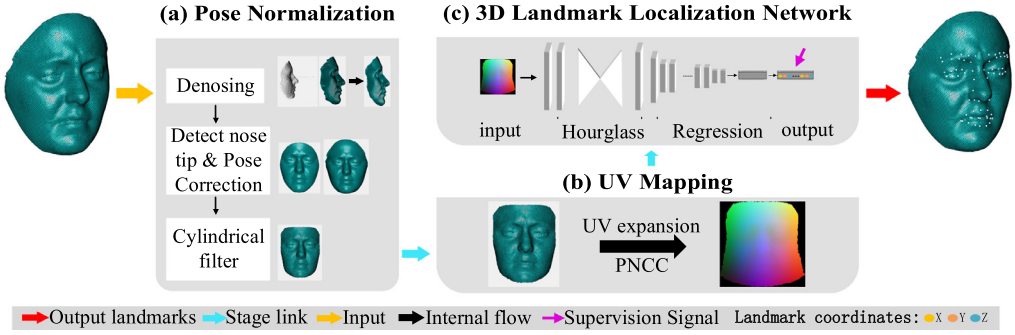

Fig. 1. The processing flow of our scheme. The flow arrows with different colors indicate different meanings as shown in the bottom.

mation is discarded. There are also works to combine the feature detection-based methods and AAM-based methods. Fenelli et al. [30] fit the 3D model to the depth and intensity image of the invisible face in the *AAM* frame. Later, Feng et al. [31] believe that this method cannot obtain accurate landmarks. In the same year, Gilani et al. [7] propose a shape-based 3D landmark detection algorithm for large number of 3D landmarks, which evolves level set curves with adaptive geometric speed functions to automatically extract effective seed points for dense correspondence.

Different from all the works mentioned above, our method utilizes the position maps as an intermediate representation. To the best of our knowledge, it is the first time that UV position map are jointly used with deep convolutional neural network to locate a large number of 3D landmarks.

3. Methodology

The processing flow of our method is given in Fig. 1. It consists of three key steps: (a) Pose normalization, which performs data pre-processing and pose correction; (b) UV position map generation, combining the expansion technology and pose-normalized 3D model to generate the UV position map; (c) 3D Landmark Localization Network (3DLLN) with the UV position map as input. The final output of 3DLLN is the vector of landmarks' 3D coordinates.

3.1. Pose normalization

Given a 3D facial data, we first conduct outlier filtering [32,33] so that holes and noises are removed. We fill holes by cubic interpolation [32]. As a result, outlier filtering benefits subsequent processing. To perform pose calibration, we first find the nose tip [19,32,33] as a key reference. The *Shape Index* measure [4,34] is employed in this task and we follow [4] to make a rough prediction of nose-tip area, and then the centroid of the area is used as the nose tip position. After the nose tip is obtained, the 3D facial data is cropped by a sphere centered on this location with a radius of 90mm (millimeter). Finally, the principle component analysis (PCA) is computed to find the three principle axes. Thanks to the intrinsic structure of 3D facial data and also the nose-tip reference, one can easily calibrate the 3D facial data to standard *X, Y, Z* axis-system with the frontal face pointing to the *Z*-axis. An example is shown in Fig. 2(a).

3.2. Position map generation

After pose normalization, we use two cylindrical filters centering on the nose tip to re-crop out a face area. One cylindrical filter is along horizontal direction (*Y*-axis) and has a radius of 90mm (Fig. 2(b)). The other cylindrical filter is along vertical direction (*X*-axis) and has a radius 80mm (Fig. 2(c)). Note that both radius values are determined empirically. The face area after cropping is then closer to a square size when viewing onto the $X-Y$ plane. This is beneficial to the UV position map generation. In addition, more face regions can be reserved than the spherical cropping used during the aforementioned PCA.

Unlike [11], we generate UV position map without the help of the Basel Face Model (BFM) [35], because aligning to the BFM model will change the structure and point number of input 3D facial data. In contrast, we use the cylinderical projection technique mentioned in [13] (the bottom flow in Fig. 3), which projects a 3D mesh to a 2D plane as below:

$$\mathcal{P} = (\mathcal{U}, \mathcal{V}) \tag{1}$$

In (1), $\mathcal{P} = (\mathcal{U}, \mathcal{V})$ denotes the projected 2D coordinates of a 3D point $V = (x, y, z)$ in UV space, where $\mathcal{U} = R \cdot atan2(\frac{x}{z})$, $\mathcal{V} = y$, and $R = \max(abs(x))$ is the radius of the cylinder. Note that here $V = (x, y, z)$ denotes the coordinate with the projection center C rather than the model centroid c as origin (as shown in Fig. 3 Cylinderical projection). We locate the projection center on the *Z*-axis and it is ε away from the model centroid. Fig. 3 bottom-right shows the effect of ε, where $\varepsilon = 0$ means using the model centroid as projection center. In practice, the choice of parameter ε has certain impact on the generated position maps, as exemplified in the bottom-right of Fig. 3, therefore on the landmark localization results. The influence of ε on detection performance is discussed by experimenting with different ε's values in Section 4.4, according to which we find $\varepsilon = 20$mm is a good choice for projection.

3.3. 3D Facial Landmark Localization Network

The UV position map inherits the advantages of both UV map and position map, and allows one to directly infer 3D location information. This section elaborates how we locate a large number of facial landmarks on UV position map. To the best of our knowledge, it is the first time that the UV position map are used for 3D landmark detection. Our deep network is called *3D Facial Landmark Localization Network* (Fig. 4), or 3DLLN for short.

The entire network architecture is shown in Fig. 4. In Stage 1, we adopt the hourglass structure proposed in [36] as our backbone (with input size 256 × 256) to predict a heatmap for each landmark from the input UV position map. We use a hierarchical, parallel and multi-scale structure in [37] throughout the entire Hourglass network. The position of the 2D landmark is considered as the position of the maximum value in the corresponding heatmap, which is shown by the red highlight of each heatmap in Fig. 4. In Stage 2, the input UV position map is concatenated with the obtained heatmaps from hourglass structure and then fed to a regression network (with a smaller input size 128 × 128 for efficiency). We base our regression architecture on the residual blocks described in [38]. The final output from the regression network is

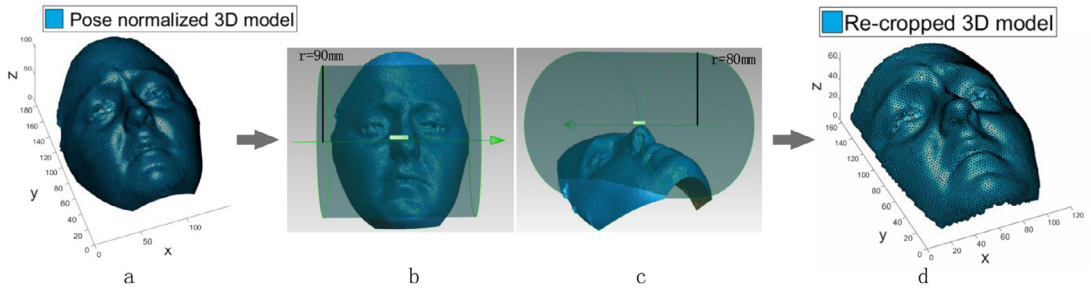

Fig. 2. Cylindrical filtering before position map generation. (a) 3D model after pose correction. (b) Cylindrical filter with a radius of 90 mm in the horizontal direction. (c) Cylindrical filter with a radius of 80 mm in the vertical direction. (d) Cropped 3D model.

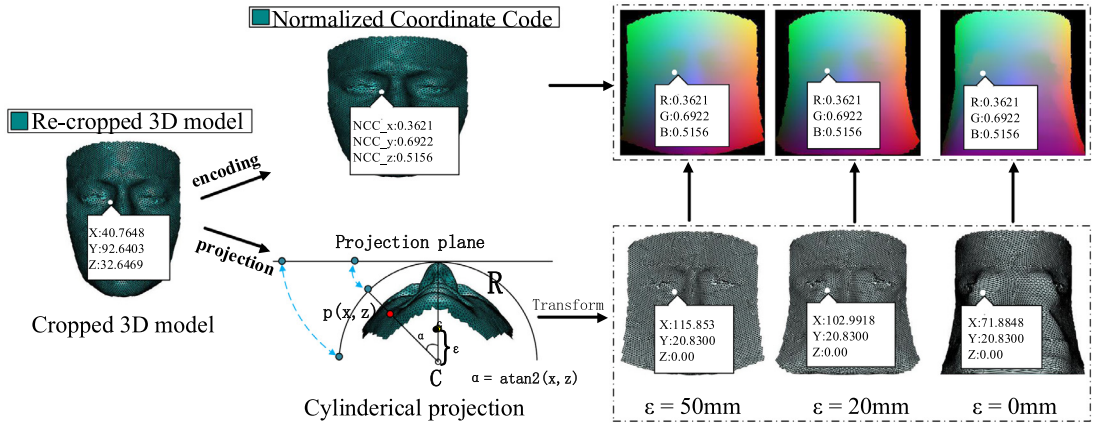

Fig. 3. Position map generation process. The cylinderical projection is employed to associate a 3D point with a 2D coordinates on the position map. The corresponding values on the map are the so-called *Projected Normalized Coordinate Code* (PNCC) [12].

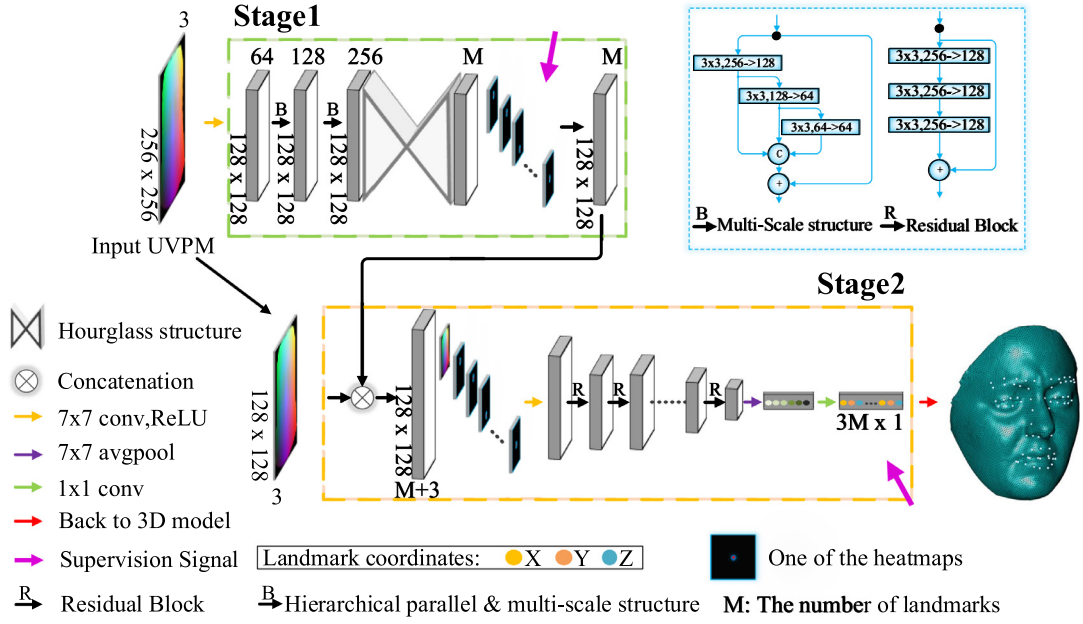

Fig. 4. The network architecture of 3DLLN. The input is the UV position map with 3 channels and the output is the 68 landmark coordinates.

a vector containing 3D coordinates of landmarks. It is worthy noting that since we use a UV position map which contains 3D coordinate information and combine it with the landmark location in the heatmaps, the regression of final coordinates is largely facilitated because the UV position map itself contains coordinate information. As comparison, we have tried to regress 3D coordinates by concatenating a 3-channel UV texture map with the heatmaps, but found the regression was likely to fail. Finally, we find the nearest neighbor on the original 3D point cloud of each landmark so that all found landmarks locate on the 3D mesh.

Rationale: Besides using the UV position map as an effective representation for 3D facial data, the regression network in Fig. 4 is

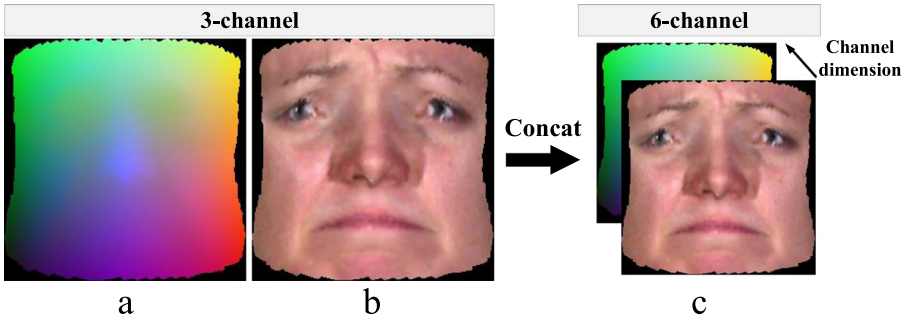

Fig. 5. (a) UV position map (abbreviated as UVPM) and (b) UV texture map (abbreviated as TexUV); (c) Merged UVPM and TexUV (abbreviated as TexUVPM).

employed for improving the localization accuracy of landmarks. Although when given the UV position map and the heatmaps obtained from Stage 1 (Fig. 4), one could infer the landmark coordinates by finding on UV maps the values whose locations corresponding to maximum values on those heatmaps, the found values have to be traced back onto the original 3D model and may still have errors. We use the regression network for two reasons: (1) Optimize such "trace back" operation and improve the accuracy (as will be shown in experiment Section 4.4). (2) Directly obtain final 3D coordinates in an "end-to-end" manner. Note that when concatenated with the UV position map, the heatmaps could be deemed as a kind of attention information telling which regions to focus on the position map.

During training, 3DLLN is jointly supervised by two Euclidean losses, namely the loss of heatmaps and the loss of landmark coordinates:

$$\mathcal{L} = \underbrace{\sum_{i=1}^{M} ||H_i - \hat{H}_i||^2}_{\text{heatmap loss}} + \underbrace{\lambda||L - \hat{L}||_2^2}_{\text{landmark loss}} \tag{2}$$

where $\mathcal{L}$ denotes the overall loss, H_i is the predicted heatmap and $\hat{H}_i$ is the ground truth heatmap with spatial size 128 × 128 associated with the ith landmark. M denotes the total number of landmarks. L is the final coordinate vector of the predicted landmarks and $\hat{L}$ is the corresponding ground truth. λ is the balancing weight between the two losses.

3.4. Utilization of texture information

The proposed method can be easily extended to deal with the case where the corresponding texture of a 3D model is available. In such a case, a UV texture map (abbreviated as TexUV shown in Fig. 5(b)) is generated in a similar process as that of the UV position map (abbreviated as UVPM shown in Fig. 5(a)) described in Section 3.2, and the correspondence relationship between them is maintained. We merge the UVPM and TexUV by channel-wise concatenation and then feed the resulting 6-channel map (abbreviated as TexUVPM shown in Fig. 5(c)) to the first stage of 3DLLN to generate heatmaps. Joint prediction with texture and 3D shape information is able to achieve more robust results due to cross-modal complementarity between texture and shape. Experimental comparisons are given in Section 4.4. Lastly, note that investigating more complex fusion strategies beyond simple concatenation will be an interesting future work.

4. Experiments and results

In this section, we validate the effectiveness of our approach on the three challenges mentioned in Section 1 and compare with the state-of-the-art methods.

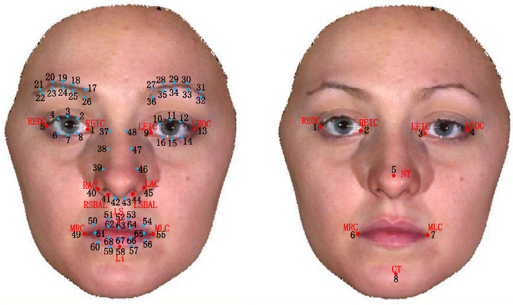

Fig. 6. (a) 68 landmarks on BU-3DFE database. (b) 8 landmarks on DB00F database.

4.1. Databases and evaluation criteria

We consider two widely-used databases BU-3DFE and a subset of FRGCv2 (namely DB00F). The BU-3DFE database contains 2500 3D facial models with 83 annotated facial landmarks. Different from FRGCv2, each 3D facial model from BU-3DFE has the associated 3D mesh, a cropped 3D face mesh, a pair of texture images from two view angles (about ± 45° yaw angle). Due to errors of some landmark data, we choose only 1486 models for training, 1000 models left for testing, and 68 facial landmarks for localization. An example of the 68 landmarks in this database is shown in Fig. 6(a). DB00F is a subset of FRGCv2, which is elaborately prepared by Perakis et al. [4]. DB00F has 300 training models and 975 testing models annotated with 8 landmarks shown in Fig. 6(b). These models are accompanied by minor pose variations and extreme expression variations, which are further divided into three subsets **"neutral", "mild", "extreme"** by [4].

Following previous works [3–5,14,26], the performance of landmark localization is generally evaluated through the three criteria below:

1. **Mean error** (*MeanErr*): The average Euclidean distance between the positions of the predicted landmarks and the ground truth landmarks.
2. **Success rate** (*SRate*): The ratio of successfully detected landmarks on a test set. A successful detection is considered as a landmark whose Euclidean distance from the ground truth is under a certain threshold. The threshold is set as 10mm, as suggested in [39].
3. **Standard deviation** (*Std*). The standard deviation of Euclidean errors of predicted landmarks.

For the three metrics above, lower *MeanErr, Std*, and higher *SRate* indicate better performance.

Table 1
Comparisons on landmark detection performance on FRGCv2 subset (DB00F) and BU-3DFE databases. "neutral","mild" and "extreme" are the subset of DB00F. The **bold** is the best in each row.

Methods		Perakis et al.	Križaj et al.	Gilani	Fan et al.	3DLLN	
		[4]	[5]	[7]	[3]	UVPM	TexUVPM
Texture		No	No	No	Yes	No	Yes
Test	DB00F	975	975	975	975	975	975
	BU-3DFE	0	0	2500	1100	1000	1000
neutral	*MeanErr*	4.52	3	3.38	2.59	2.78	**2.53**
	Std	1.51	1.6	**0.94**	–	1.79	1.62
	SRate	99.32%	99.60%	–	98.61%	99.56%	**99.89%**
mild	*MeanErr*	4.95	3.4	3.81	2.84	2.87	**2.57**
	Std	1.46	1.7	**1.28**	–	1.87	1.73
	SRate	99.72%	99.70%	–	97.95%	99.19%	**99.93%**
extreme	*MeanErr*	6.28	3.9	4.46	3.7	3.07	**2.93**
	Std	2.6	1.8	**1.77**	–	2.05	1.86
	SRate	90.40%	99.40%	–	92.73%	98.23%	**99.57%**
DB00F	*MeanErr*	5	3.2	3.72	2.89	2.87	**2.62**
	Std	1.85	1.7	**1.15**	–	1.81	1.71
	SRate	97.30%	99.60%	–	97.30%	99.58%	**99.85%**
BU-3DFE	*MeanErr*	–	–	5.85	4.66	2.66	**2.15**
	Std	–	–	4.26	2.5	1.89	**1.65**
	SRate	–	–	–	93.52%	99.54%	**99.89%**

4.2. Implementation details of 3DLLN

Architecture. The proposed network takes as input a 256 × 256 × 3 UVPM. It outputs predictions of the landmark heatmaps (output of Stage 1) and the coordinate vector (output of Stage 2). Different from [36], we only use one substructure of stacked Hourglass Networks with 5-layer symmetric convolution. The outputs are heatmaps with a size of 128 × 128 × 68 in Stage 1. Regarding the regression network, it is based on ResNet-50 backbone [38] with stacked UVPM and heatmaps as input, and with a modified fully-connected layer as output.

Training and testing. During training, data augmentation, such as flipping, rotation and noise, is randomly applied to input images. The network is trained for 190 epochs in total, including 80 epochs for Stage 1 and 50 epochs for Stage 2, respectively, and finally 60 epochs for the whole network with the joint loss as Eq. (2). The balancing weight λ in Eq. (2(is set as 1. Meanwhile, we adopt Stochastic Gradient Descent (SGD) optimization algorithm with an initial learning rate of 10^{-3} for Stage 1 and 10^{-5} for both Stage 2 and the final joint training. Specifically for Stage 1, the learning rate is reduced by a factor of 10% after 4 epochs and a factor of 1% after 37 epochs. For Stage 2 and also the final joint training, the learning rate is reduced by a factor of 10% after 20 epochs. Notably, the ground truth heatmaps supervising Stage 1 are generated by placing a 2D Gaussian centered around a ground truth landmark position. Moreover, because of the insufficient training data from DB00F, we train our 3DLLN based on BU-3DFE database for 68 landmarks, and then fine-tune it on DB00F for 8 landmarks. During testing, our model takes about 0.13s to process an image on an NVIDIA 1080Ti GPU.

4.3. Comparisons to state-of-the-art methods

We compare 3DLLN with two different input configurations, namely UVPM (Section 3.2) and TexUVPM (Section 3.4), to four representative state-of-the-art methods including: Perakis et al. [4], Križaj et al. [5], Gilani et al. [7], and Fan et al. [3]. The compared results of these four methods are the reported ones in their original papers.

Quantitative comparisons are shown in Table 1. By column comparisons, our 3DLLN with UVPM performs well on both DB00F (8 landmark detection) and BU-3DFE (68 landmark detection). Among the methods which only utilize 3D shape information, we achieve the best *MeanErr* and *SRate* on these two databases. Our results are even better than those of Fan's method [3] which incorporates texture. On the other hand, 3DLLN achieves the best results on BU-3DFE for a large number of 68 landmarks, and also it shows robustness (only 0.29mm *MeanErr* increased) when expressions vary from "nature" to "extreme". Regarding the effectiveness of incorporating texture, TexUVPM achieves better results than UVPM and is the best among all the competitors.

Following [3,4,7,18,26,40], detailed comparisons of individual *key* landmarks are shown in Table 2 (BU-3DFE) and Table 3 (DB00F). The landmarks considered are shown in Fig. 6 as red color dots. Here note that methods like [18,26,40] are not compared in the previous Table 1 because they did not consider all 8 or 68 landmarks in their experiments. For example, [26] only experiments with 11 landmarks on BU-3DFE.

By row comparisons in Table 2 and 3, we observe that 3DLLN with TexUVPM achieves the best result in average (see row "Mean"), followed by 3DLLN with UVPM. The comparisons between inner (rows "REIC" and "LEIC") and outer canthus (rows "REOC" and "LEOC") show a consistent trend for almost all methods that the landmarks of inner canthus are easier to detect. This is because: (1) The outer canthus is easily effected by expression

Table 2
Comparisons of 12 individual landmarks with the state-of-the-art methods on BU-3DFE databases. The abbreviations of these landmarks are shown in Fig. 6(a) in red color. The criteria used are *MeanErr* ± *Std* (defined in Section 4.1). The **bold** indicates the best in each row.

Method	Grewe and Zachow	Segundo et al.	Gilani et al.	Paulsen et al.	3DLLN	
	[40]	[18]	[7]	[26]	UVPM	TexUVPM
Images	2500	2500	2500	500	1000	1000
Texture	No	No	No	Yes	No	Yes
REIC	3.23 ± 1.86	–	3.29 ± 2.67	1.80 ± 0.89	1.75 ± 1.49	**1.42 ± 1.27**
REOC	3.22 ± 2.18	6.33 ± 5.04	4.35 ± 2.70	2.85 ± 1.50	2.58 ± 1.72	**2.11 ± 1.51**
LEIC	3.04 ± 1.75	6.33 ± 4.82	4.75 ± 2.64	1.89 ± 0.98	1.82 ± 1.46	**1.60 ± 1.37**
LEOC	2.95 ± 1.93	–	4.43 ± 2.74	2.59 ± 1.53	2.51 ± 1.79	**2.03 ± 1.49**
LS	–	–	3.20 ± 2.68	2.33 ± 1.31	2.05 ± 1.52	**1.48 ± 1.50**
LI	–	–	6.90 ± 5.31	2.50 ± 1.41	2.37 ± 1.81	**1.79 ± 1.73**
LAC	–	6.66 ± 3.36	4.30 ± 2.73	2.61 ± 1.41	2.51 ± 1.80	**2.28 ± 1.86**
RAC	–	6.49 ± 3.40	4.28 ± 2.71	2.96 ± 1.56	2.53 ± 1.77	**2.14 ± 1.57**
LSBAL	2.37 ± 1.37	–	4.86 ± 2.80	–	2.26 ± 1.77	**1.98 ± 1.72**
RSBAL	2.47 ± 1.29	–	3.57 ± 2.59	–	2.14 ± 1.71	**1.86 ± 1.63**
MLC	–	–	6.00 ± 3.94	2.18 ± 1.44	2.65 ± 1.76	**2.10 ± 1.58**
MRC	–	–	5.45 ± 3.12	2.42 ± 1.44	2.60 ± 1.80	**2.00 ± 1.56**
Mean	2.88 ± 1.73	6.45 ± 4.15	4.61 ± 4.05	2.41 ± 1.34	2.31 ± 1.70	**1.90 ± 1.56**

Table 3
Comparisons of 8 individual landmarks with the state-of-the-art methods on DB00F. The abbreviations of these landmarks are shown in Fig. 6(b) in red color. The criteria used are *MeanErr* ± *Std* (defined in Section 4.1). The **bold** indicates the best in each row.

Method	Passalis et al.	Perakis et al.	Gilani et al.	Fan et al.	3DLLN	
	[39]	[4]	[7]	[3]	UVPM	TexUVPM
Images	975	975	975	975	975	975
Texture	No	No	No	Yes	No	Yes
REIC	5.03 ± 1.66	4.15 ± 2.35	3.23 ± 2.29	**1.33 ± 1.47**	2.81 ± 1.81	2.39 ± 1.66
REOC	5.79 ± 3.45	5.58 ± 3.33	4.10 ± 3.04	**2.53 ± 1.62**	3.41 ± 2.13	3.13 ± 1.84
LEIC	5.48 ± 2.59	4.41 ± 2.49	2.76 ± 1.99	2.49 ± 1.67	2.63 ± 1.73	**2.30 ± 1.70**
LEOC	5.62 ± 3.47	5.83 ± 3.42	3.60 ± 2.74	**1.39 ± 1.84**	3.24 ± 1.93	3.04 ± 1.78
NT	4.91 ± 2.49	4.09 ± 2.41	2.67 ± 2.47	4.38 ± 2.90	1.87 ± 1.30	**1.41 ± 0.96**
CT	6.31 ± 4.43	4.92 ± 3.74	4.10 ± 3.04	3.85 ± 3.05	3.06 ± 1.85	**2.66 ± 1.51**
MLC	6.47 ± 4.26	5.42 ± 3.84	4.25 ± 3.12	3.06 ± 2.82	**2.90 ± 1.85**	3.20 ± 1.71
MRC	5.65 ± 4.34	5.56 ± 3.93	4.10 ± 3.04	4.07 ± 3.36	3.02 ± 1.85	**2.91 ± 1.64**
Mean	5.65 ± 3.34	5.0 ± 3.18	3.72 ± 1.15	2.89 ± 2.34	2.867 ± 1.806	**2.63 ± 1.6**

changes; (2) The inner canthus has more distinguishable features. Similarly, rows "LS", "LI", "MLC" and "MRC" show that landmarks around mouth lead to relatively high *MeanErr*, because they are easily affected by expression changes. With reference to rows "NT", "REIC" and "LEIC", we find that most methods can obtain good results for strong-robustness landmarks because of their expression-invariant geometric features. However, regarding the rest landmarks listed in Table 2, our method still achieves good performance at the locations whose geometric features are ambiguous or implicit on the 3D shape. Last but not least, by column comparisons, one can observe that 3DLLN with both UVPM and TexUVPM achieve consistent improvement on almost all landmarks.

Some visual examples on BU-3DFE for 68 landmark detection are shown in Fig. 7. The white dots mean the detected landmarks by 3DLLN with UVPM input, while the pink ones indicate the ground truth landmarks. The yellow dots indicate the coincidences of our results and ground truth.

4.4. Ablation studies

In this section, we validate three issues: (1) the impact of different input maps in order to validate the necessity of UV expansion and position coding; (2) the effectiveness of regression network; (3) the influence of different ε in cylindrical projection.

Regarding the first issue, in addition to TexUVPM, we compare UVPM with three other different 2D facial representations as follows: PM (Fig. 8 left), position map without expansion whose pixel values represent 3D point coordinates; UVDM (Fig. 8 right), expanded depth map whose pixel values are depth values (instead of 3D coordinates) along *Z*-axis; TexUV, described in Section 3.4.

Table 4
Comparisons of landmark detection results before and after regression network on BU-3DFE 1000 testing images. The **bold** is the best in each part.

3DLLN input maps	Texture	*MeanErr* ± *Std*(*mm*)(*SRate*)	
		Before	After
UVDM	No	3.16 ± 2.08(98.01%)	2.72 ± 1.93(99.48%)
PM	No	3.01 ± 2.03(98.83%)	**2.61 ± 1.90(99.52%)**
UVPM	No	**2.92 ± 1.98(98.94%)**	2.66 ± 1.89(99.54%)
TexUV	Yes	2.63 ± 1.9189(99.04%)	–
TexUVPM	Yes	**2.53 ± 1.9143(99.23%)**	**2.15 ± 1.65(99.89%)**

Regarding the second issue, we consider the performance before and after using regression network, which correspond to the results from Stage 1 and Stage 2 of 3DLLN, respectively. Regarding the third issue, we try different ε values every 10mm from 0mm to 50mm. All the above ablation studies are conducted on BU-3DFE, and the overall results are shown in Tables 4 and 6, from which interesting observations can be found as below:

UVPM *vs.* **PM**. By comparing 3DLLN (UVPM) and 3DLLN (PM) in column "Before" of Table 4, we can observe that before regression the UVPM achieves better performance than PM. This is because for PM simply using orthogonal projection to map the 3D face model onto X-Y plane makes some facial points locate very closer to each other and become nearly indistinguishable. By contrast, employing UV expansion is able to improve this problem. To validate this, we further analyze the nasal saddle landmarks which may have such a problem. Detailed results are shown in Table 5. The improvement of UVPM over PM can be clearly observed on

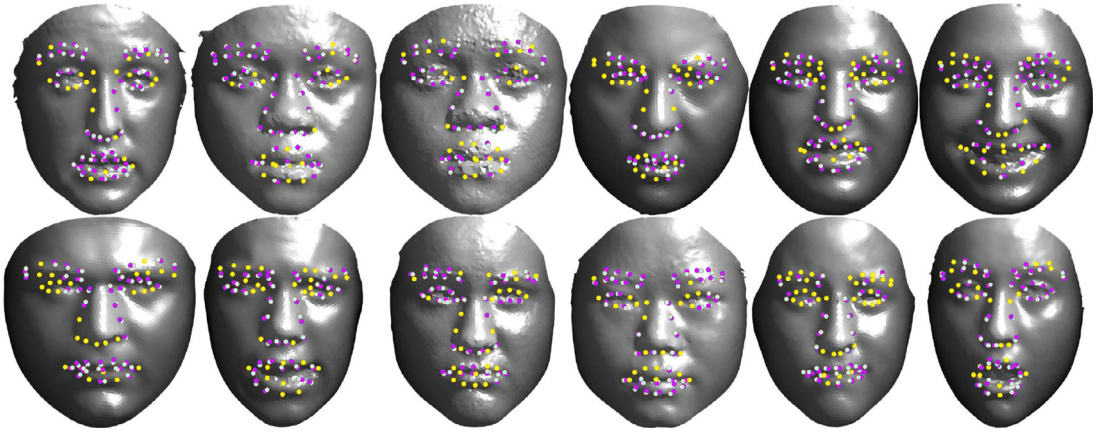

Fig. 7. Visual landmark detection examples of 3DLLN with UVPM input from BU-3DFE testing database.

Table 5
Comparisons between UVPM and PM on landmark localization in nasal saddle region on BU-3DFE database. We consider 8 landmarks, namely 37 – 40, 45 – 48 in Fig. 6(a).

Landmark Index		(37)	(38)	(39)	(40)	(45)	(46)	(47)	(48)	mean
PM	*MeanErr*	3.78	3.72	3.98	3.06	3.31	3.73	3.54	3.45	3.57
	Std	2.13	2.36	2.34	1.83	2.07	2.16	2.23	2.19	
	SRate	98.19%	96.38%	98.79%	99.59%	98.59%	97.79%	97.79%	**98.39%**	
UVPM	*MeanErr*	**2.99**	**3.38**	**3.60**	**2.50**	**2.90**	**3.39**	**3.46**	**3.20**	**3.17**
	Std	**1.96**	**2.11**	**2.07**	**1.74**	**1.91**	**2.03**	**2.06**	**2.10**	
	SRate	**98.59%**	**98.99%**	**98.99%**	**99.60%**	**99.80%**	**98.59%**	**99.60%**	98.18%	

Table 6
Quantitative landmark detection results from Stage 1 with different ε in cylindrical projection on BU-3DFE database.

Input	Criteria	ε value(mm)					
		0	10	20	30	40	50
UVPM	*MeanErr*	3.49	3.01	**2.92**	**2.92**	2.95	2.98
	Std	2.24	2.02	1.98	1.98	**1.97**	1.99
	SRate	93.77%	98.44%	98.94%	98.46%	**99.03%**	98.94%

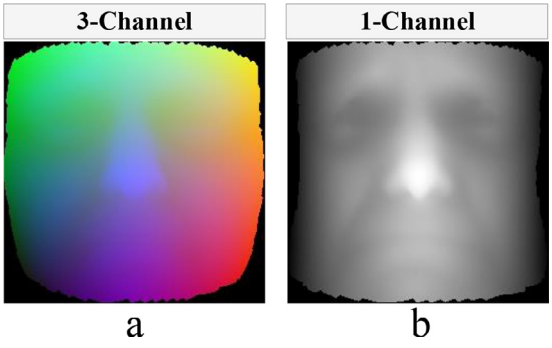

Fig. 8. (a) PM, position map without expansion, (b) UVDM, expanded depth map.

the 8 landmarks considered, demonstrating the effectiveness and necessity of UV expansion.

UVPM *vs.* **UVDM**. Comparing 3DLLN (UVPM) and 3DLLN (UVDM) in column "Before" of Table 4, the effectiveness of using position maps with PNCC scheme [11,12] for landmark detection is clearly validated. Comparing to commonly used depth maps [14,26], position maps encode more informative 3D shape information, leading to much better accuracy. We also show detailed results for all 68 landmarks in Fig. 9. The UVPM achieves consistent lower *MeanErr* and *Std* than UVDM for almost all of the landmarks.

UVPM *vs.* **TexUV** *vs.* **TexUVPM**. Comparing 3DLLN UVPM, TexUVPM, and TexUV in Table 4, one can observe that texture information is indeed more informative than shape information for landmark localization. Solely using TexUV achieves higher accuracy than UVPM. This actually makes sense because some landmarks such as those of eyebrows are distinguishable on texture map, but become no longer distinguishable on 3D shape. This implies that detecting landmarks on pure 3D shape is generally a more difficult task than on texture. Moreover, our experimental results show that the combination of both texture and shape, namely TexUVPM, yields the encouraging best performance, attributed to the benefit of cross-modality fusion.

Several visual examples from BU-3DFE for 68 landmark detection by 3DLLN with different facial maps are shown in Fig. 10. The white dots mean the detected landmarks, while the pink ones indicate the ground truth landmarks. The yellow dots indicate the coincidences of our results and ground truth. Generally, more yellow dots mean better performance. One can obsereve that 3DLLN with TexUVPM input performs better than other input maps.

Validation of regression network. For the proposed regression network in the Stage 2 of 3DLLN, by comparing columns "Before" and "After" in Table 4, we find that it leads to consistent improvement over all input types and on all metrics considered. One can see it becomes very essential to incorporate the regression network for further improvement. As mentioned before, the regression on TexUV is likely to fail due to lack of 3D coordinate coding information. Therefore, for Table 4 "Texture" row, the result "After" is not

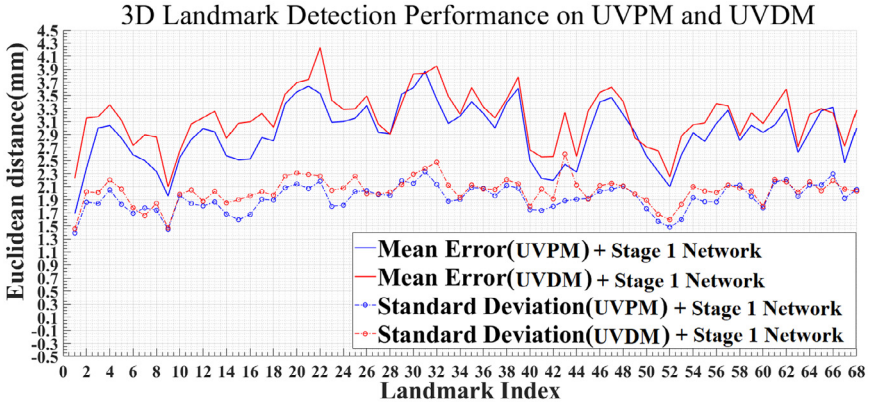

Fig. 9. Comparisons on all the 68 landmarks for UVPM and UVDM. Solid lines indicate *MeanErr* and dotted lines indicate *Std*.

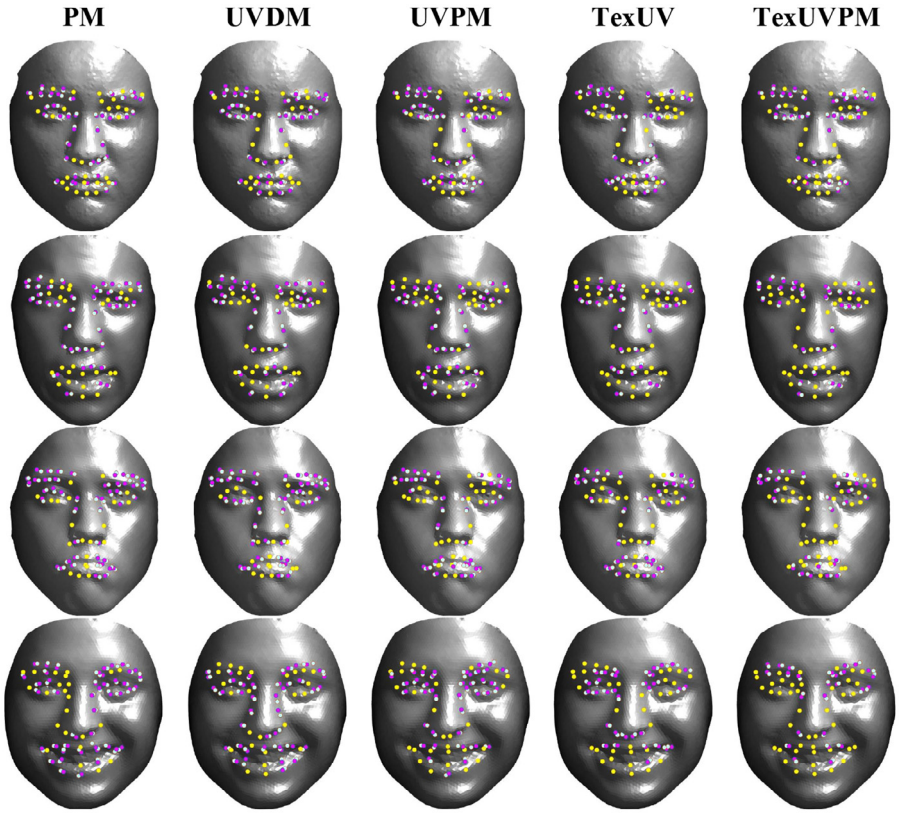

Fig. 10. Visual landmark detection comparison examples of 3DLLN with different facial maps (as shown in Fig. 5 and Fig. 8) inputs from BU-3DFE testing database including PM, UVDM, UVPM, TexUV, TexUVPM.

provided. We also find the regression could somewhat compensate the gap between UVPM and PM.

Influence of ε. Table 6 shows the results of using different ε. To remove the impact of Stage 2, only the results from Stage 1 are obtained. One can see that $\varepsilon = 20$mm achieves the best performance with the lowest *MeanErr*, decent *Std* and *SRate*. This indicates the fact ε can neither be too small nor too large, because it controls the expansion of facial areas. Basically, smaller ε tends to horizontally expand the central facial areas containing nose and eyes, meanwhile shrinking the cheek areas on the two sides. However, too small ε (e.g., $\varepsilon = 0$mm) may lead to too exaggerated deformation that in turn degenerates localization accuracy. Therefore, we choose $\varepsilon = 20$mm as the default setting in all experiments aforementioned.

5. Conclusion

We propose a novel method for 3D landmark detection using deep convolutional neural network based on UVPM. It has been evaluated using challenging 3D facial databases which contain 3D scans with extreme expression variations or a large number of landmarks to be detected. Our method achieves state-of-the-art accuracy with high success rate. Besides, we have performed ablation studies and achieved the following observations: (a) The position map retains richer 3D shape information than the depth map. Such information is very beneficial for 3D landmark detection. (b) The expanded position map performs favorably against the position map without expansion for landmarks near the nasal saddle region. (c) The regression architecture brings significant improvement on localization accuracy. (d) Combining texture and position maps is able to achieve even higher accuracy thanks to cross-modal fusion. In the future, we may investigate complex fusion strategies to fuse TexUV and UVPM beyond simple concatenation for more promising results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

CRediT authorship contribution statement

Jingchen Zhang: Conceptualization, Investigation, Software, Writing - original draft. **Kangkang Gao:** Data curation, Methodology, Software. **Keren Fu:** Project administration, Supervision, Writing - review & editing. **Peng Cheng:** Supervision, Writing - review & editing.

Acknowlgedgments

This work was supported by the National Science Foundation of China, No. 61703077, U1833128, the Fundamental Research Funds for the Central Universities, No. YJ201755, and the Sichuan Science and Technology Major Projects (2018GZDZX0029).

References

[1] G.R. Swennen, F.A. Schutyser, J.-E. Hausamen, Three-dimensional cephalometry: a color atlas and manual, Springer Science & Business Media, 2005.
[2] X. Zhao, E. Dellandrea, L. Chen, I.A. Kakadiaris, Accurate landmarking of three--dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41 (5) (2011) 1417–1428.

[3] X. Fan, Q. Jia, K. Huyan, X. Gu, Z. Luo, 3d facial landmark localization using texture regression via conformal mapping, Pattern Recognition Letters 83 (2016) 395–402.
[4] P. Perakis, G. Passalis, T. Theoharis, I.A. Kakadiaris, 3d facial landmark detection under large yaw and expression variations, IEEE transactions on pattern analysis and machine intelligence 35 (7) (2012) 1552–1564.
[5] J. Križaj, Ž. Emeršič, S. Dobrišek, P. Peer, V. Štruc, Localization of facial landmarks in depth images using gated multiple ridge descent, in: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), IEEE, 2018, pp. 1–8.
[6] H. Li, D. Huang, J.-M. Morvan, Y. Wang, L. Chen, Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors, International Journal of Computer Vision 113 (2) (2015) 128–142.
[7] S. Zulqarnain Gilani, F. Shafait, A. Mian, Shape-based automatic detection of a large number of 3d facial landmarks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4639–4648.
[8] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.
[9] W. Deng, Y. Fang, Z. Xu, J. Hu, Facial landmark localization by enhanced convolutional neural network, Neurocomputing 273 (2018) 222–229.
[10] X. Tang, F. Guo, J. Shen, T. Du, Facial landmark detection by semi-supervised deep learning, Neurocomputing 297 (2018) 22–32.
[11] Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou, Joint 3d face reconstruction and dense alignment with position map regression network, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 534–551.
[12] X. Zhu, X. Liu, Z. Lei, S.Z. Li, Face alignment in full pose range: A 3d total solution, IEEE transactions on pattern analysis and machine intelligence 41 (1) (2019) 78–92.
[13] J. Booth, S. Zafeiriou, Optimal uv spaces for facial morphable model construction, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 4672–4676.
[14] T. Terada, Y.-W. Chen, R. Kimura, 3d facial landmark detection using deep convolutional neural networks, in: 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), IEEE, 2018, pp. 390–393.
[15] Z. Hu, P. Gui, Z. Feng, Q. Zhao, K. Fu, F. Liu, Z. Liu, Boosting depth-based face recognition from a quality perspective, Sensors 19 (19) (2019) 4124.
[16] E. Vezzetti, F. Marcolin, S. Tornincasa, L. Ulrich, N. Dagnes, 3d geometry-based automatic landmark localization in presence of facial occlusions, Multimedia Tools and Applications 77 (11) (2018) 14177–14205.
[17] A. Colombo, C. Cusano, R. Schettini, 3d face detection using curvature analysis, Pattern recognition 39 (3) (2006) 444–455.
[18] M.P. Segundo, L. Silva, O.R.P. Bellon, C.C. Queirolo, Automatic face segmentation and facial landmark detection in range images, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 40 (5) (2010) 1319–1330.
[19] H. Dibeklioğlu, B. Gökberk, L. Akarun, Nasal region-based 3d face recognition under pose and expression variations, in: International Conference on Biometrics, Springer, 2009, pp. 309–318.
[20] M.P. Segundo, C. Queirolo, O.R. Bellon, L. Silva, Automatic 3d facial segmentation and landmark detection, in: 14th International Conference on Image Analysis and Processing (ICIAP 2007), IEEE, 2007, pp. 431–436.
[21] K.I. Chang, K.W. Bowyer, P.J. Flynn, Multiple nose region matching for 3d face recognition under varying facial expression, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1695–1700.
[22] S. Berretti, B.B. Amor, M. Daoudi, A. Del Bimbo, 3d facial expression recognition using sift descriptors of automatically detected keypoints, The Visual Computer 27 (11) (2011) 1021.
[23] S. Jahanbin, H. Choi, A.C. Bovik, Passive multimodal 2-d+ 3-d face recognition using gabor features and landmark distances, IEEE Transactions on Information Forensics and Security 6 (4) (2011) 1287–1304.
[24] C. Creusot, N. Pears, J. Austin, A machine-learning approach to keypoint detection and landmarking on 3d meshes, International journal of computer vision 102 (1-3) (2013) 146–179.
[25] N. Cihan Camgoz, V. Struc, B. Gokberk, L. Akarun, A. Alp Kindiroglu, Facial landmark localization in depth images using supervised ridge descent, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 136–141.
[26] R.R. Paulsen, K.A. Juhl, T.M. Haspang, T. Hansen, M. Ganz, G. Einarsson, Multi-view consensus cnn for 3d facial landmark placement, in: Asian Conference on Computer Vision, Springer, 2018, pp. 706–719.
[27] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Computer vision and image understanding 61 (1) (1995) 38–59.
[28] G.J. Edwards, C.J. Taylor, T.F. Cootes, Interpreting face images using active appearance models, in: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 1998, pp. 300–305.
[29] Y. Sun, X. Chen, M. Rosato, L. Yin, Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40 (3) (2010) 461–474.
[30] G. Fanelli, M. Dantone, L. Van Gool, Real time 3d face alignment with random forests-based active appearance models, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–8.
[31] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, X.-J. Wu, Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting, IEEE Transactions on Image Processing 24 (11) (2015) 3425–3440.
[32] A. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2d-3d hybrid approach to automatic face recognition, IEEE transactions on pattern analysis and machine intelligence 29 (11) (2007) 1927–1943.
[33] M. Emambakhsh, A.N. Evans, M. Smith, Using nasal curves matching for expression robust 3d nose recognition, in: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), IEEE, 2013, pp. 1–8.
[34] C.G. Harris, M. Stephens, et al., A combined corner and edge detector., in: Alvey vision conference, 15, Citeseer, 1988, pp. 10–5244.
[35] V. Blanz, T. Vetter, et al., A morphable model for the synthesis of 3d faces., in: Siggraph, 99, 1999, pp. 187–194.
[36] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, Springer, 2016, pp. 483–499.
[37] A. Bulat, G. Tzimiropoulos, Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3706–3714.
[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[39] G. Passalis, P. Perakis, T. Theoharis, I.A. Kakadiaris, Using facial symmetry to handle pose variations in real-world 3d face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (10) (2011) 1938–1951.
[40] C.M. Grewe, S. Zachow, Fully automated and highly accurate dense correspondence for facial surfaces, in: European Conference on Computer Vision, Springer, 2016, pp. 552–568.

Jingchen Zhang is currently working toward the Master degree in the College of Computer Science, Sichuan University, Chengdu, China. Her current research interests include learning-based 3D face landmark methods and 3D reconstruction.

Kangkang Gao is currently working toward the Master degree in the College of Computer Science, Sichuan University, Chengdu, China. His current research interests include learning-based 3D face landmark methods.

Keren Fu received the dual Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, and Chalmers University of Technology, Gothenburg, Sweden, under the joint supervision of Prof. Jie Yang and Prof. Irene Yu-Hua Gu. He is currently a research associate professor with College of Computer Science, Sichuan University, Chengdu, China. His current research interests include visual computing, saliency analysis, and machine learning.

Peng Cheng received Ph.D. degree from Sichuan University, Chengdu, China. Currently, he is an associate professor with College of Aeronautics and Astronautics in Sichuan University, Chengdu, China. His research interests include image registration, image fusion and computer vision.