PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

Learning to detect 3D facial landmarks via heatmap regression with Graph Convolutional Network

Yuan Wang[1,4], **Min Cao**[2], **Zhenfeng Fan**[3,4*], **Silong Peng**[1,4]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Computer Science and Technology, Soochow University, 215006 Suzhou, China.
[3]Institute of Computing Technology, Chinese Academy of Sciences
[4]University of Chinese Academy of Sciences
{wangyuan2020, silong.peng}@ia.ac.cn, mcao@suda.edu.cn, fanzhenfeng@ict.ac.cn

Abstract

3D facial landmark detection is extensively used in many research fields such as face registration, facial shape analysis, and face recognition. Most existing methods involve traditional features and 3D face models for the detection of landmarks, and their performances are limited by the hand-crafted intermediate process. In this paper, we propose a novel 3D facial landmark detection method, which directly locates the coordinates of landmarks from 3D point cloud with a well-customized graph convolutional network. The graph convolutional network learns geometric features adaptively for 3D facial landmark detection with the assistance of constructed 3D heatmaps, which are Gaussian functions of distances to each landmark on a 3D face. On this basis, we further develop a local surface unfolding and registration module to predict 3D landmarks from the heatmaps. The proposed method forms the first baseline of deep point cloud learning method for 3D facial landmark detection. We demonstrate experimentally that the proposed method exceeds the existing approaches by a clear margin on BU-3DFE and FRGC datasets for landmark localization accuracy and stability, and also achieves high-precision results on a recent large-scale dataset.

Introduction

Facial landmark detection aims to localize feature points on 2D images or videos with anatomical significance, such as the nose tip, the eye corner, and the pupils. It is also referred as 2D face alignment, which is fundamental to various face-related applications such as face recognition (Taigman et al. 2014), expression analysis (Yang, Ciftci, and Yin 2018), face animation (Cao et al. 2013), and 3D face reconstruction (Blanz and Vetter 1999; Liu et al. 2018a). 2D face alignment has experienced rapid development over the past few decades, which can be traced back to the traditional hand-crafted feature (Baker and Matthews 2004; Zhu and Ramanan 2012; Cao et al. 2014) and active-shape-based regression methods (Cootes, Edwards, and Taylor 2001; Xiong and De la Torre 2013; Zhu et al. 2015). The recent progress in convolutional neural networks (CNNs) has pushed the precision of 2D face alignment to a new milestone.

The CNNs-based studies can be mainly divided into coordinate-based methods (Miao et al. 2018; Liu et al. 2018b; Deng et al. 2020) and heatmap-based methods (Bulat and Tzimiropoulos 2017; Wang, Bo, and Fuxin 2019; Wang et al. 2020). The coordinate-based methods regress the coordinates of landmarks directly with CNNs. However, regressing coordinates directly from 2D images is a highly nonlinear process due to the discrete locations of landmarks, which reduces the generalization ability of the methods. The heatmap-based methods predict the probability instead of the location of each landmark on the image plane, and allow the network to be fully convolutional. Since the heatmap-based methods make full use of the continuity of heatmaps and capture the correlation between landmarks effectively, they achieve better robustness and performance over the coordinate-based methods in practice.

2D face alignment has achieved significant progress in the literature but still encounters some bottlenecks for complex scenarios such as extreme illumination and head pose. Meanwhile, 3D face alignment (Fan et al. 2016; Zhang et al. 2020) that locates the feature points on 3D images recently raises increasing attention by computer vision communities due to its robustness to illumination and pose variations. The current development of commercial 3D sensors and data acquisition technologies seeks for urgent real-life applications. Like the 2D case, 3D face alignment is also fundamental to many downstream 3D face applications, such as 3D face recognition (Soltanpour, Boufama, and Wu 2017), 3D face animation (Zollhöfer et al. 2018), and statistical 3D face modeling (Galteri et al. 2019). 3D facial data include additional geometry information compared with 2D images, which is more tolerant of makeup, expression, and age. In this paper, we denote 3D face alignment as landmark detection on 3D facial data, which differs from the current literature using a 3D face model to align 2D face (Gou et al. 2016; Feng et al. 2018a; Guo et al. 2020).

3D face alignment is a challenging task for two main reasons. Firstly, there is no feasible and effective network for the learning of 3D landmarks. The well-known CNNs are intrinsically designed on regular grid data such as the 2D image. However, 3D facial data such as point clouds are composed of unordered vertices, which hinders the direct application of the state-of-the-art CNN backbone on 3D face alignment. Secondly, there is a lack of effective post-processing strategies to convert deep features to 3D landmark coordinates. Regressing coordinates directly from 3D

*Corresponding author.
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

point cloud is even more difficult than the 2D case, since it contradicts with the translation-invariant property of convolutional networks (Liu et al. 2018c).

In this work, for the first obstacle, we employ Graph Convolution Networks (GCNs) as backbones to process the 3D facial point cloud under the natural advantages of GCNs (Kipf and Welling 2017) on modeling irregular format of data. We can learn the geometric features adaptively from the 3D points in this way. Correspondingly, we incorporate 3D heatmaps for ease of training GCN for 3D face alignment. Given the second difficulty, we develop a post-processing method based on local surface unfolding and registration to output 3D coordinates of landmarks. The main contributions of our work are summarized as follows:

- We construct a novel framework to learn the landmark coordinates from 3D point clouds by means of 3D heatmaps. To the best of our knowledge, it is the first deep point cloud learning framework for 3D face alignment.
- We propose a novel 3D heatmap post-processing method for 3D face alignment with local surface unfolding and registration, which effectively generalizes 2D heatmap post-processing methods for 2D face alignment.
- Our proposed method achieves the state-of-the-art performance on public and representative 3D face datasets, and suppresses the existing methods by a clear margin.

Related Work

Our proposed 3D face alignment method is closely related to deep point cloud learning, heatmap in 2D face alignment, and 3D face alignment.

Deep point cloud learning

Due to the irregular structure of point cloud data, the existing deep point cloud learning works design various schemes to imitate the convolution and pooling operations in 2D CNNs. PointNet (Qi et al. 2017a) as a pioneer of deep point cloud learning, and its effective variant (Qi et al. 2017b) use multi-layer perception (MLP) and global aggregation to replace 2D convolution and pooling. In a departure from PointNet, PointConv (Wu, Qi, and Fuxin 2019) considers the convolution kernels as nonlinear functions of 3D local coordinates composed of weight and inverse density coefficients. KPConv (Thomas et al. 2019) generates convolution kernels by combining some pre-defined kernels with specific rules. However, these methods commonly have high complexity (memory and computation burden) for learning.

The recent state-of-the-art works commonly use GCN for point cloud learning because of its advantage in modeling the neighboring information for irregular data. For example, ECC (Simonovsky and Komodakis 2017) proposes a dynamic graph edge convolution method for point cloud learning. DGCNN (Wang et al. 2019b) employs an EdgeConv module for local feature extraction on a dynamic graph for point cloud through renewing the neighboring relationship at each feature layer. Liu *et al.* (Liu et al. 2019) propose a dynamic aggregation module (DPAM) to simplify point agglomeration (sampling, grouping, and pooling). Recently, Xu *et al.* (Xu et al. 2021) propose a position adaptive graph convolution module (PAConv), which constructs the convolution kernels dynamically from position information.

The current GCNs work well for several tasks, such as 3D shape analysis, and object detection. However, applying GCN to 3D face alignment has not been studied thoroughly yet. In this work, we generalize the GCN backbone to the 3D face alignment task with constructed 3D heatmaps.

Heatmap in 2D face alignment

The current state-of-the-art deep-learning-based methods (Tang et al. 2019; Huang et al. 2020) commonly employ heatmap regression strategies for 2D face alignment. The heatmap is constructed with a Gaussian function with a small variance (commonly 2-3 pixels) of the distance to each landmark. Some early methods (Newell, Yang, and Deng 2016; Bulat and Tzimiropoulos 2017) employ the $\mathcal{L}_1$, $\mathcal{L}_2$, or smooth $\mathcal{L}_1$ loss function for training. More recently, Wang *et al.* (Wang, Bo, and Fuxin 2019) propose an adaptive Wing (AWing) loss, which is suitable for heatmap regression by adaptively adjusting the importance of foreground and background pixels in training. For heatmap post-processing methods, the coordinates of landmarks are estimated by either the *argmax* method (Wu et al. 2018) or the *soft-argmax* method (Honari et al. 2018).

In our proposed 3D face alignment framework, we migrate the Gaussian heatmap and Awing loss function from 2D to 3D. We also propose an effective post-processing method with local surface unfolding and registration to accurately estimate the coordinates of the 3D landmarks.

3D face alignment

Most works on 3D face alignment belong to traditional methods, which involve geometric features or 3D face models for landmark detection. For example, Segundo *et al.* (Segundo et al. 2010) combine surface curvatures and depth relief curves for landmark detection. Perakis *et al.* (Perakis et al. 2012) propose a facial analytical model to extract candidate landmarks with shape index and spin image, which improves robustness for faces with large expression and pose. Gilani *et al.* (Gilani, Shafait, and Mian 2015) propose a shape-based algorithm for dense facial landmark detection, which evolves level set curves with adaptive geometric functions to extract seed points for dense correspondence. Fan *et al.* (Fan et al. 2016) propose a novel method by mapping a 3D face model and corresponding texture to a 2D image plane. Križaj *et al.* (Križaj et al. 2018) propose a landmark detection algorithm by combining SIFT feature and grid function. These traditional methods are commonly effective for detecting landmarks with distinctive features under frontal views. However, their accuracy is commonly limited by hand-crafted features or tailored 3D face models.

Some recent works (Paulsen et al. 2018; Zhang et al. 2020) project 3D shape to 2D plane under multiple directions and employ 2D CNN to regress the landmarks. The accuracy of these methods commonly outperform the traditional ones. However, the projection between 2D and 3D points and the ensemble of 2D landmarks in multiple directions may introduce intrinsic numeric errors. Therefore, we

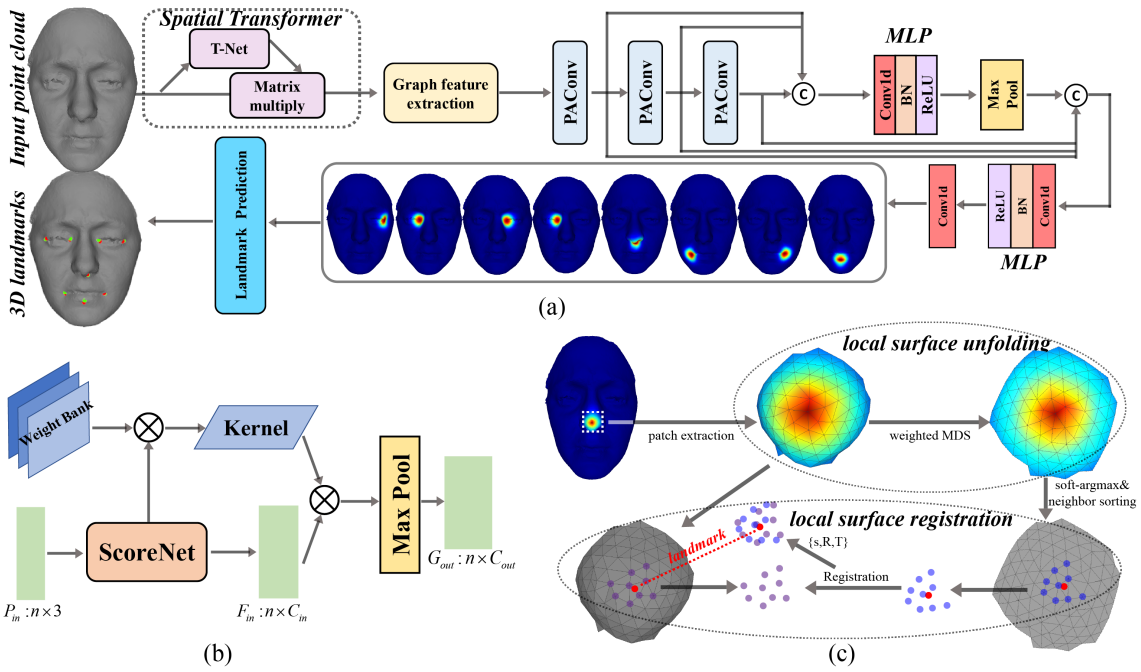


Figure 1: (a) The overall framework of our model; (b) The PAConv module; (c) Landmark prediction module.

consider them as indirect 2D deep learning methods, and their performance is still limited.

Different from the current works for 3D face alignment, we propose a novel framework to learn geometric features directly with GCN. Our method takes 3D heatmaps as intermediate representations and boosts the accuracy notably.

Method

In this section, we first review the general pipelines for deep point cloud learning, including the convolution-aggregation operation and the resampling of point cloud. Then, we describe the 3D heatmap construction, architecture of GCN, and training loss function tailored for 3D face alignment. Finally, we illustrate a post-processing algorithm to obtain landmark coordinate from the 3D heatmaps.

Point cloud deep learning with GCN

Given a point cloud $P = \{p_i \mid i = 1, ..., n\} \in \mathbb{R}^{n \times 3}$, we denote $F = \{f_i \mid i = 1, ..., n\} \in \mathbb{R}^{n \times C_{in}}$ and $G = \{g_i \mid i = 1, ..., n\} \in \mathbb{R}^{n \times C_{out}}$ as the input and output features of a graph convolution layer, respectively. The convolution-aggregation on a graph can be formulated as

$$g_i = \pi\left(\{K\left(p_i, p_j\right) \cdot f_j \mid j \in \mathcal{N}_i\}\right), \tag{1}$$

where $\pi(.)$ is the aggregation function and $K(.)$ outputs the convolutional weights according to the neighboring relationship between the central node p_i and its neighboring nodes $p_j (j \in \mathcal{N}_i)$. The weights for a GCN are calculated dynamically according to the relative position between the central node and its $\mathbf{k}$ neighboring nodes in some state-of-the-art networks (Simonovsky and Komodakis 2017; Wang et al. 2019b). A cascade of several (commonly 3 to 5) convolution-aggression layers, which are not as deeper than the CNNs because of over-smoothing (Alon and Yahav 2020), formulates the trunk of the whole GCN.

The GCN commonly involves resampled point cloud as the input. In this work, we use the farthest point sampling (**FPS**) method (Eldar et al. 1997) to resample the point cloud to a fixed number of n points. FPS guarantees uniform spacing of each point as a standard method for processing the input of GCN. The output for GCN is usually a probability vector for different classes in the classification task and a tensor with 3D binary features for the segmentation task. In a departure from that, the output of our GCN model for 3D face alignment is a tensor composed of multiple 3D heatmaps for landmarks.

3D heatmap construction

The construction of a 3D heatmap, which is a generalization of the 2D heatmap in the recent state-of-the-art 2D face alignment works (Feng et al. 2018b; Wang, Bo, and Fuxin 2019; Kumar et al. 2020), is described as follows.

Let n be the number of points in each resampled point cloud, and let l be the number of landmarks for 3D face alignment. On each resampled facial point cloud, we calculate the Euclidean distance between each point and each specific landmark to obtain a distance matrix $D \in \mathbb{R}^{n \times l}$. Similar to the heatmap construction pipeline in the recent works for 2D face alignment, we use the Gaussian function to encode the distance matrix to obtain a normalized distance matrix as the 3D heatmap

$$H = \exp\left(-\frac{D^2}{2\sigma^2}\right), \tag{2}$$

where σ is a hyper-parameter for the width of the heatmap. The 3D heatmap reaches the climax at each landmark location and decreases with the distance to each landmark. In fact, the 3D heatmap represents the probability of each landmark in each location. It can also be considered as a *"soft segmentation label"*, which differs from the binary segmentation label in the general segmentation task (Qi et al. 2017a; Wang et al. 2019a,b). The soft segmentation enables the network to learn the local geometric features on a face from the neighboring points around each landmark robustly.

Network architecture

We employ the PAConv (Xu et al. 2021) as one of the basic building blocks for feature extraction of landmarks. PAConv (Fig. 1(b)) is plug-and-play and position-correlated adaptive convolution module. The key element of PAConv, as an improvement of DGCNN (Wang et al. 2019b) and ScoreNet (Xu et al. 2021), is a weight bank module composed of some weighting matrices, which enables the network to learn a coefficient vector from point positions for self-attention adaptively. In this way, the PAConv can learn the local structure information hidden in the positions of neighboring points, which is directly associated with our task for 3D face alignment. In addition, We include a spatial transformer net (Qi et al. 2017a) in our model to adaptively learn an affine transformation applied to the input point clouds for better generalization. Figure 1(a) shows the overall network architecture.

Loss function

We take the adaptive wing loss (AWing) (Wang, Bo, and Fuxin 2019) as the loss function in the training stage, which is an effective variant of Wing loss (Feng et al. 2018b) and is particularly suitable for heatmap regression for face alignment. It alleviates the small gradient problem with a piecewise analytical function and adjusts the importance of foreground and background pixels in the training process. The AWing loss is formulated as

$$\mathcal{L}_{AWing}(h,\hat{h}) = \begin{cases} \omega \ln\left(1+\left|\frac{h-\hat{h}}{\varepsilon}\right|^{\alpha-h}\right), |h-\hat{h}| < \theta \\ A|y-\hat{h}| - \Omega, |h-\hat{h}| \geq \theta \end{cases}, \tag{3}$$

where h and $\hat{h}$ are the ground truth heatmap and the predicted heatmap, respectively. α, ϵ, θ, ω, A, and Ω are some hyper-parameters that control the shape of the loss function and satisfy

$$A = \omega(\alpha-h)\left(\frac{\theta}{\varepsilon}\right)^{\alpha-h-1} / \left(\left(1+\frac{\theta}{\varepsilon}\right)^{\alpha-h}\right)/\varepsilon \tag{4}$$

and

$$\Omega = \theta A - \omega l \ln\left(1+(\theta/\varepsilon)^{\alpha-h}\right). \tag{5}$$

In this work, we use the default setting of these parameters as in the work (Wang, Bo, and Fuxin 2019).

Coordinate prediction from 3D heatmaps

In the current heatmap-based 2D face alignment method, a common way to obtain the landmark position from 2D heatmap is *soft-argmax* method, which is an effective variant of *argmax* method. The *soft-argmax* method reduces the sensitivity to noisy predictions of heatmaps by an ensemble of multiple locations, which differs from the *argmax* method by a single location of maximum heatmap value. In the 2D case, the *soft-argmax* operation is formulated as:

$$\begin{aligned} S_{max}(H) &= \sum_{x,y} softmax\left(\beta H_{x,y}\right)\cdot(x,y) \\ &= \sum_{x,y} \frac{e^{\beta H_{x,y}}}{\sum_{x,y} e^{\beta H_{x,y}}}\cdot(x,y), \end{aligned} \tag{6}$$

where $H_{x,y}$ is the predicted heatmap (probability) at location (x,y), and β is an annealing parameter. In discrete case for 3D heatmap, the *soft-argmax* operation can be formulated as:

$$\begin{aligned} S_{max}(H) &= \sum_{i\in Q} softmax\left(\alpha H_i\right)\cdot p_i \\ &= \sum_{i\in Q} \frac{e^{\beta H_i}}{\sum_{i\in Q} e^{\beta H_i}}\cdot p_i, \end{aligned} \tag{7}$$

where i $(i = 1,2,...,r)$ are indices for the discrete point subset $Q \subset P$ with maximum heatmap values.

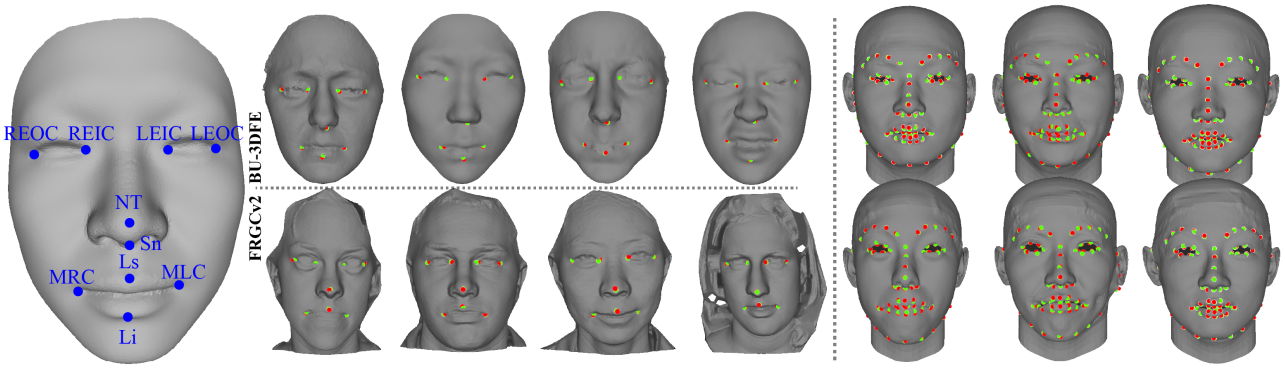
Unfortunately, the *soft-argmax* in Eq.7 is not suitable for 3D landmark prediction directly. The reason is two-fold: 1) Contrary to the 2D image grid, point indices do not corresponded to the locations of 3D point cloud; 2) The resulting 3D location as an ensemble of some 3D coordinates does not necessarily lie on the facial surface, especially in areas with large curvatures. To deal with these problems, we propose an effective landmark prediction method from 3D heatmaps based on local surface unfolding and registration.

Algorithm 1: Landmark Prediction from 3D Heatmaps

Input: 3D facial point cloud $P \in \mathbf{R}^{n\times 3}$; 3D heatmap $H \in \mathbf{R}^{n\times l}$; the number of neighbors r for local unfolding.
Output: 3D landmark coordinates $Z \in \mathbf{R}^{l\times 3}$.

1: **for** $i = 1 : l$ **do**
2: Select a local patch Q_i of r points with maximum values on each 3D heatmap;
3: Compute the distance matrix $E \in \mathbf{R}^{r\times r}$ for Q_i;
4: Compute the weighted distance matrix $E_w \in \mathbf{R}^{r\times r}$ with the corresponding heatmap value by Eq. 8;
5: Apply MDS to E_w to acquire $\tilde{Q}_i$ as a dimension-degraded version of Q_i;
6: Compute the centroid u of $\tilde{Q}_i$ by Eq. 7;
7: Select a few nearest points (U) around u in $\tilde{Q}_i$, add zeros to the third dimension, and register them to the corresponding points in Q_i by Eq. 9;
8: Set $Z_i = u$ after registration.
9: **return** Z

Local surface unfolding. We employ the multi-dimensional scaling (**MDS**) method (Cox and Cox 2008) for the local surface unfolding. MDS is a classical dimension reduction method and minimizes the pairwise distances



(a) BU-3DFE&FRGCv2 (b) FaceScape

Figure 2: Some qualitative results by our method on (a) the BU-3DFE and FRGCv2 datasets, and (b) the FaceScape dataset. The red and green dots represent the ground truth landmarks and the detected landmarks by our method, respectively. The anatomical meanings of some landmarks for the BU-3DFE and FRGCv2 datasets are shown in the first column.

of all sampling data in a least-square sense. It involves pairwise distance matrix construction, double centering, and eigenvalue decomposition to obtain a dimension-degraded version of the original data finally. We denote the distance matrix as $E_{r\times r}$ by selecting $\mathbf{r}$ points (which constitute a local patch $Q \subset P$) with maximum heatmap values, where each element is the pairwise distance of two points in Q. In this work, we make some modifications on $E_{r\times r}$ on the classical MDS method by multiplying a weight for heatmaps, as

$$(E_w)_{r\times r} = H_{r\times r} \odot E_{r\times r}. \tag{8}$$

The purpose of weighting is to aggravate the locations with large heatmaps (probabilities).

Local surface registration. After the unfolding of a local patch is completed, we apply *soft-argmax* in Eq. 7 to obtain the centroid u. Then, we select a few nearest neighboring points around u, which constitute a set U, and register them to the corresponding 3D points in Q by a scaled rigid transform as

$$\{\mathrm{s}, \mathrm{R}, \mathrm{T}\} = \underset{s\in\mathbb{R}, R\in SO(3), T\in\mathbb{R}^3}{\operatorname{argmin}} \sum_{j\in U} \|sRU_j + T - \mathrm{Q}_j\|_2^2, \tag{9}$$

where s, R, and T denote the scaling factor, the rotation matrix, and the translation, respectively. Eq. 9 can be solved by singular value decomposition efficiently.

The detailed process is summarized in Algorithm 1 referring to Fig. 1(c). In this way, we can predict the landmarks from 3D heatmaps on facial surfaces.

Experiments

In this part, we carry out the experiments on several publicly available datasets, including BU-3DFE (Yin et al. 2006), FRGCv2 (Phillips et al. 2005), and FaceScape (Yang et al. 2020) to demonstrate the effectiveness of the proposed method.

Datasets and Implementation details

BU-3DFE includes $2,500$ 3D facial samples from 100 subjects. Each subject is composed of 6 different expressions in 4 growing levels in addition to a neutral face. This dataset is constructed by depth scans from 2 different directions. The resolution is around $10{,}000$ vertices per face. There are 83 annotated landmarks for each face, and we select 8 of them for comparison with the existing works. **FRGCv2** dataset includes $4,007$ 3D facial samples from 466 subjects with natural expressions. Similar to BU-3DFE, we select 8 provided landmarks by the work (Creusot, Pears, and Austin 2013) compared to the existing literatures. **FaceScape** dataset is a recently published large-scale 3D dataset and consists of $18,760$ 3D faces from 938 subjects with 20 different expressions. The landmark locations are hidden in the topological uniform models provided by this dataset. We keep all 68 landmarks for evaluation in the experiment.

Evaluation criteria

We use the following evaluation metrics for 3D face alignment referring to some previous works (Fan et al. 2016; Zhang et al. 2020).

Mean error (**ME**) evaluates the average Euclidean distance between the predicted and ground truth landmarks.

Success rate (**SR**) represents the proportion of successfully detected landmarks for which the ME is within a fixed threshold. We set the threshold to $10mm$ following some prior works for a fair comparison.

Standard deviation (**Std**) of the ME across all testing samples qualifies the robustness of a method.

Evaluation on BU-3DFE and FRGC dataset

We compare our method with several representative prior works for 3D facial landmark detection. Most of these works belong to the traditional methods. As an exception, Zhang *et. al* (Zhang et al. 2020) use 2D CNN to learn depth features on projection planes along multiple directions.

First, we summarize the quantitative results in terms of ME, SR, and Std for the BU-3DFE and FRGC dataset in Table 1 and Table 2, respectively. The results also include the performance in terms of ME and Std for each specific landmark, the anatomical meaning of which is marked in the

Table 1: Comparisons of 8 individual landmarks in terms of *ME* $\pm$ *Std* and *SR* on the BU-3DFE datasets. The **bold** indicates the best in each row.

	(Segundo et al. 2010)	(Gilani, Shafait, and Mian 2015)	(Grewe and Zachow 2016)	(Fan et al. 2016)	(Paulsen et al. 2018)	(Zhang et al. 2020)	Ours
REIC	6.33 ± 5.04	3.29 ± 2.67	3.23 ± 1.86	–	1.80 ± 0.89	1.75 ± 1.49	$\mathbf{1.58 \pm 1.33}$
REOC	–	4.35 ± 2.70	3.22 ± 2.18	–	2.85 ± 1.50	2.58 ± 1.72	$\mathbf{2.06 \pm 1.54}$
LEIC	6.33 ± 4.82	4.75 ± 2.64	3.04 ± 1.75	–	1.89 ± 0.98	1.82 ± 1.46	$\mathbf{1.67 \pm 1.37}$
LEOC	–	4.43 ± 2.74	2.95 ± 1.93	–	2.59 ± 1.53	2.51 ± 1.79	$\mathbf{2.24 \pm 1.58}$
Sn	–	3.90 ± 3.26	1.97 ± 1.06	–	2.52 ± 1.69	–	$\mathbf{1.82 \pm 1.44}$
MRC	–	5.45 ± 3.12	–	–	2.42 ± 1.44	2.60 ± 1.80	$\mathbf{2.05 \pm 1.51}$
MLC	–	6.00 ± 3.94	–	–	2.18 ± 1.44	2.65 ± 1.76	$\mathbf{2.00 \pm 1.58}$
Li	–	6.90 ± 5.31	–	–	2.50 ± 1.41	2.37 ± 1.81	$\mathbf{2.33 \pm 1.63}$
Mean	6.33 ± 4.93	4.88 ± 3.30	2.88 ± 1.76	4.66 ± 2.50	2.34 ± 1.36	2.32 ± 1.69	$\mathbf{1.97 \pm 1.50}$
SR	–	–	–	93.52%	–	99.54%	**100.0%**

Table 2: Comparisons of 8 individual landmarks in terms of *ME* $\pm$ *Std* and *SR* on the FRGCv2 datasets. The **bold** indicates the best in each row.

	(Segundo et al. 2010)	(Perakis et al. 2012)	(Gilani, Shafait, and Mian 2015)	(Fan et al. 2016)	(Križaj et al. 2018)	(Zhang et al. 2020)	Ours
REIC	3.35 ± 2.33	4.15 ± 2.35	2.73 ± 2.14	$\mathbf{1.33 \pm 1.47}$	3.1 ± 3.8	2.81 ± 1.81	2.70 ± 1.69
REOC	–	5.58 ± 3.33	3.74 ± 2.79	2.53 ± 1.62	3.6 ± 4.1	3.41 ± 2.13	$\mathbf{3.20 \pm 1.90}$
LEIC	3.69 ± 2.26	4.41 ± 2.49	3.12 ± 2.09	2.49 ± 1.67	3.1 ± 3.8	2.63 ± 1.73	$\mathbf{2.51 \pm 1.65}$
LEOC	–	5.83 ± 3.42	4.50 ± 2.97	$\mathbf{1.39 \pm 1.84}$	3.6 ± 4.1	3.24 ± 1.93	3.15 ± 1.89
NT	2.73 ± 1.39	4.09 ± 2.41	2.68 ± 1.48	4.38 ± 2.90	3.6 ± 5.7	1.87 ± 1.30	$\mathbf{1.30 \pm 1.14}$
MRC	–	5.56 ± 3.93	4.38 ± 2.08	3.85 ± 3.05	3.4 ± 3.4	3.02 ± 1.85	$\mathbf{2.66 \pm 1.65}$
MLC	–	5.42 ± 3.84	5.31 ± 2.05	4.07 ± 3.36	3.4 ± 3.4	2.90 ± 1.85	$\mathbf{2.76 \pm 1.68}$
Ls	–	–	3.31 ± 2.65	–	–	–	$\mathbf{2.06 \pm 1.50}$
Mean	3.25 ± 1.99	5.00 ± 3.11	3.72 ± 2.28	2.86 ± 2.27	3.40 ± 4.04	2.84 ± 1.80	$\mathbf{2.54 \pm 1.64}$
SR	–	97.85%	–	97.30%	99.60%	99.58%	**100.0%**

Table 3: Comparisons of ME, Std and SR with different point cloud learning methods on the BU-3DFE dataset.

Dataset	BU-3DFE		
Metric / Model	ME	Std	SR
PointNet++ (Qi et al. 2017b)	2.59	1.95	100.0%
DGCNN (Wang et al. 2019b)	2.08	1.55	100.0%
GACNet (Wang et al. 2019a)	2.37	1.85	100.0%
DeepGCN (Li et al. 2019)	2.62	1.57	100.0%
Bow Pool (Zhang 2021)	2.22	1.92	100.0%
Ours	**1.97**	**1.50**	**100.0%**

leftmost column of Fig. 2(a). We can see that our method achieves considerable improvement for all the evaluation metrics on all datasets. Specifically, our proposed method achieves $\mathbf{15.1}\%$ and $\mathbf{10.6}\%$ improvements in terms of the average ME on the BU-3DFE dataset and FRGCv2 dataset, respectively. Among all the landmarks, the corners of the mouth ("MRC" and "MLC"), the lower lip (Li), and the upper lip (Ls) are challenging and easily affected by facial expressions as indicated by some prior works (Perakis et al. 2012; Zhang et al. 2020). Nevertheless, our proposed method achieves significant improvement over the existing works on these landmarks, demonstrating its effectiveness.

Then, we show some qualitative results in Fig. 2(a). The detected landmarks are almost indistinguishable from the ground truth. Since we only use 3D shapes for landmark detection, the visual differences to the ground truth (commonly labeled with both shapes and textures) are considered reasonable. We owe the successful detection of landmarks to effective learning of the geometric features of 3D face.

Evaluation on FaceScape

FaceScape is a recently published dataset on which we are the first to report the landmark detection results. We conduct experiments on all 68 landmarks in order to obtain a whole benchmark for future research. The ME and Std scores by the proposed method reach $\mathbf{1.60}$ and $\mathbf{1.18}$, respectively, with better quality for facial point cloud and sufficient training data in this dataset. Some qualitative results in the publishable list for this dataset are shown in Fig. 2(b).

Ablation Study

In this part, we evaluate the impact of two key components of our proposed method: the validity of the customized GCN network for feature (heatmap) extraction and the post-processing method for landmark prediction from heatmaps.

Table 4: Comparisons between *soft-argmax*, *argmax*, and the proposed post-processing method for 3D landmark detection.

Dataset	BU-3DFE			FRGCv2		
Metric / Method	ME	Std	SR	ME	Std	SR
soft-argmax	2.61	5.87	97.86%	3.42	9.96	96.08%
argmax	2.26	1.61	100.0%	2.84	2.26	99.92%
Ours w/o weight	1.99	1.50	100.0%	2.59	1.77	99.92%
Ours w/ weight	**1.97**	**1.50**	**100.0%**	**2.54**	**1.64**	**100.0%**

First, we compare the GCN structure in the proposed method with some state-of-the-art networks on BU-3DFE in Table 3. We replace the main feature extraction block with the other networks, while modifying the last feature layer to be compatible with the output heatmaps. The results show that the customized network in this work significantly outperforms the baseline networks of DGCNN, GAC-Net, PointNet, DeepGCN, and Bow Pool, demonstrating the PAConv and spatial transformer modules are effective for learning adaptive geometric features for 3D landmark detection. In addition, we find that very deep GCN networks, such as the DeepGCN, do not necessarily benefit the landmark detection task in this work.

Then, we compare the performance of the proposed post-processing method with the common *argmax* and *soft-argmax* methods as shown in Table 4. We set the hyper-parameter $\beta = 1000$ in Eq. 6 and Eq. 7 for the *soft-argmax* operation. Table 4 shows the comparative results, where the proposed post-processing method gains the best performance. Compared with the *argmax* method, we attribute the improvement to less sensitivity to noise by an ensemble of multiple coordinates. Compared with the *soft-argmax* method, we use a local surface unfolding and registration method to ensure the predicted landmark lying on the facial surface. In addition, the weighting strategy also gains some improvement. It demonstrates the effectiveness of the proposed post-processing method for the prediction of 3D facial landmarks.

Hyper-parameter settings

In this part, we evaluate the influence of two hyper-parameters for the proposed method. 1) We use k nearest neighbors for the dynamic construction of the adjacent matrix for GCN. 2) The post-processing of 3D heatmap involves r vertices with maximum heatmap values.

Number of neighbors. Table 5 shows the result of using different k on the detection results. In some prior works for GCN, k should not be too large or too small. In this work, we observe that our model achieves the best performance in terms of ME when setting k to 30.

Number of regression points. We observe that larger r reduces the effectiveness of MDS for local surface unfolding, while smaller r hinders the robustness of the *soft-argmax* method. Table 6 shows the impact of different r in terms of ME. We set $r = 10$ in our experiments.

Table 5: Results in terms of ME for different number of neighboring points in the graph construction.

ME / Data / k	BU-3DFE	FRGCv2
20	3.54	3.62
25	2.13	2.76
30	**1.97**	**2.54**
35	2.03	2.64

Table 6: Results in terms of ME for different number of regression points for local surface unfolding.

ME / Data / r	BU-3DFE	FRGCv2
15	1.98	2.59
12	2.13	2.60
10	**1.97**	**2.54**
8	2.00	2.61

Discussion and Conclusion

In this work, we propose a novel 3D face alignment method[1], which localizes some feature points given the input point cloud of a 3D face. The proposed method is motivated by the recent progress in deep point cloud learning and heatmap-based 2D face alignment. The key element of our proposed method is an advanced GCN structure for adaptive heatmap regression and a compatible post-processing method to predict landmarks from regressed heatmaps. Extensive experiments on some representative 3D face datasets demonstrate the effectiveness of the proposed method.

A limitation is that the Farthest Point Sampling method as a pre-processing step is not efficient for real-time performance. The accuracy is also limited by the number of sampling points. In the future, we will study some adaptive sampling methods for both lifted accuracy and efficiency.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant 2021YFF0602101, National Science Foundation of China under Grant NSFC 62106250 and Liaoning Collaboration Innovation Center For CSLE.

[1]Code at https://github.com/wangyuan123ac/3DFA-GCN

References

Alon, U.; and Yahav, E. 2020. On the Bottleneck of Graph Neural Networks and its Practical Implications. In *International Conference on Learning Representations*.

Baker, S.; and Matthews, I. 2004. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3): 221–255.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 187–194.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision*, 1021–1030.

Cao, C.; Weng, Y.; Lin, S.; and Zhou, K. 2013. 3D shape regression for real-time facial animation. *ACM Transactions on Graphics*, 32(4): 1–10.

Cao, X.; Wei, Y.; Wen, F.; and Sun, J. 2014. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2): 177–190.

Cootes, T. F.; Edwards, G. J.; and Taylor, C. J. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6): 681–685.

Cox, M. A.; and Cox, T. F. 2008. Multidimensional scaling. In *Handbook of data visualization*, 315–347. Springer.

Creusot, C.; Pears, N.; and Austin, J. 2013. A machine-learning approach to keypoint detection and landmarking on 3D meshes. *International Journal of Computer Vision*, 102(1-3): 146–179.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5203–5212.

Eldar, Y.; Lindenbaum, M.; Porat, M.; and Zeevi, Y. Y. 1997. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9): 1305–1315.

Fan, X.; Jia, Q.; Huyan, K.; Gu, X.; and Luo, Z. 2016. 3D facial landmark localization using texture regression via conformal mapping. *Pattern Recognition Letters*, 83: 395–402.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018a. Joint 3d face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision*, 534–551.

Feng, Z.-H.; Kittler, J.; Awais, M.; Huber, P.; and Wu, X.-J. 2018b. Wing loss for robust facial landmark localisation with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2235–2245.

Galteri, L.; Ferrari, C.; Lisanti, G.; Berretti, S.; and Del Bimbo, A. 2019. Deep 3D morphable model refinement via progressive growing of conditional Generative Adversarial Networks. *Computer Vision and Image Understanding*, 185: 31–42.

Gilani, S. Z.; Shafait, F.; and Mian, A. S. 2015. Shape-based automatic detection of a large number of 3D facial landmarks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4639–4648.

Gou, C.; Wu, Y.; Wang, F.-Y.; and Ji, Q. 2016. Shape augmented regression for 3D face alignment. In *European Conference on Computer Vision*, 604–615.

Grewe, C. M.; and Zachow, S. 2016. Fully automated and highly accurate dense correspondence for facial surfaces. In *European Conference on Computer Vision*, 552–568.

Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; and Li, S. Z. 2020. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, 152–168.

Honari, S.; Molchanov, P.; Tyree, S.; Vincent, P.; Pal, C.; and Kautz, J. 2018. Improving landmark localization with semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1546–1555.

Huang, X.; Deng, W.; Shen, H.; Zhang, X.; and Ye, J. 2020. PropagationNet: Propagate Points to Curve to Learn Structure Information. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7265–7274.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Križaj, J.; Emeršič, Ž.; Dobrišek, S.; Peer, P.; and Štruc, V. 2018. Localization of facial landmarks in depth images using gated multiple ridge descent. In *IEEE International Work Conference on Bioinspired Intelligence*, 1–8.

Kumar, A.; Marks, T. K.; Mou, W.; Wang, Y.; Jones, M.; Cherian, A.; Koike-Akino, T.; Liu, X.; and Feng, C. 2020. LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8236–8246.

Li, G.; Muller, M.; Thabet, A.; and Ghanem, B. 2019. Deepgcns: Can gcns go as deep as cnns? In *IEEE/CVF International Conference on Computer Vision*, 9267–9276.

Liu, F.; Zhao, Q.; Liu, X.; and Zeng, D. 2018a. Joint face alignment and 3D face reconstruction with application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3): 664–678.

Liu, H.; Lu, J.; Guo, M.; Wu, S.; and Zhou, J. 2018b. Learning reasoning-decision networks for robust face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3): 679–693.

Liu, J.; Ni, B.; Li, C.; Yang, J.; and Tian, Q. 2019. Dynamic points agglomeration for hierarchical point sets learning. In *IEEE/CVF International Conference on Computer Vision*, 7546–7555.

Liu, R.; Lehman, J.; Molino, P.; Such, F. P.; Frank, E.; Sergeev, A.; and Yosinski, J. 2018c. An intriguing failing of convolutional neural networks and the CoordConv solution. In *International Conference on Neural Information Processing Systems*, 9628–9639.

Miao, X.; Zhen, X.; Liu, X.; Deng, C.; Athitsos, V.; and Huang, H. 2018. Direct shape regression networks for end-to-end face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5040–5049.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 483–499.

Paulsen, R. R.; Juhl, K. A.; Haspang, T. M.; Hansen, T.; Ganz, M.; and Einarsson, G. 2018. Multi-view consensus CNN for 3D facial landmark placement. In *Asian Conference on Computer Vision*, 706–719.

Perakis, P.; Passalis, G.; Theoharis, T.; and Kakadiaris, I. A. 2012. 3D facial landmark detection under large yaw and expression variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1552–1564.

Phillips, P. J.; Flynn, P. J.; Scruggs, T.; Bowyer, K. W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; and Worek, W. 2005. Overview of the face recognition grand challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, 947–954.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Advances in Neural Information Processing Systems*, 30.

Segundo, M. P. P.; Silva, L.; Bellon, O. R. P.; and Queirolo, C. C. 2010. Automatic face segmentation and facial landmark detection in range images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(5): 1319–1330.

Simonovsky, M.; and Komodakis, N. 2017. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *IEEE conference on Computer Vision and Pattern Recognition*, 3693–3702.

Soltanpour, S.; Boufama, B.; and Wu, Q. J. 2017. A survey of local feature methods for 3D face recognition. *Pattern Recognition*, 72: 391–406.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708.

Tang, Z.; Peng, X.; Li, K.; and Metaxas, D. N. 2019. Towards efficient u-nets: A coupled and quantized approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8): 2038–2050.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *IEEE/CVF International Conference on Computer Vision*, 6411–6420.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; and Shan, J. 2019a. Graph attention convolution for point cloud semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10296–10305.

Wang, X.; Bo, L.; and Fuxin, L. 2019. Adaptive wing loss for robust face alignment via heatmap regression. In *IEEE/CVF International Conference on Computer Vision*, 6971–6981.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019b. Dynamic graph cnn for learning on point clouds. *ACM Transactions On Graphics*, 38(5): 1–12.

Wu, W.; Qi, Z.; and Fuxin, L. 2019. Pointconv: Deep convolutional networks on 3d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9621–9630.

Wu, W.; Qian, C.; Yang, S.; Wang, Q.; Cai, Y.; and Zhou, Q. 2018. Look at boundary: A boundary-aware face alignment algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2129–2138.

Xiong, X.; and De la Torre, F. 2013. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 532–539.

Xu, M.; Ding, R.; Zhao, H.; and Qi, X. 2021. PAConv: Position Adaptive Convolution with Dynamic Kernel Assembling on Point Clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3173–3182.

Yang, H.; Ciftci, U.; and Yin, L. 2018. Facial expression recognition by de-expression residue learning. In *IEEE Conference on Computer Vision and Rattern Recognition*, 2168–2177.

Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 601–610.

Yin, L.; Wei, X.; Sun, Y.; Wang, J.; and Rosato, M. J. 2006. A 3D facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, 211–216.

Zhang, J.; Gao, K.; Fu, K.; and Cheng, P. 2020. Deep 3D Facial Landmark Localization on position maps. *Neurocomputing*, 406: 89–98.

Zhang, Z. 2021. BoW Pooling: A Plug-and-Play Unit for Feature Aggregation of Point Clouds. In *AAAI Conference on Artificial Intelligence*, volume 35, 3403–3411.

Zhu, S.; Li, C.; Change Loy, C.; and Tang, X. 2015. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4998–5006.

Zhu, X.; and Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2879–2886.

Zollhöfer, M.; Thies, J.; Garrido, P.; Bradley, D.; Beeler, T.; Pérez, P.; Stamminger, M.; Nießner, M.; and Theobalt, C. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, 523–550.