Computer Vision and Image Understanding 250 (2025) 104227

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

2S-SGCN: A two-stage stratified graph convolutional network model for facial landmark detection on 3D data

Jacopo Burger, Giorgio Blandano, Giuseppe Maurizio Facchi [*], Raffaella Lanzarotti

PHuSe Lab, Department of Computer Science, University of Milan, Milan, Italy

ARTICLE INFO

Keywords:
3D facial landmark detection
Graph convolutional network
Heatmap regression
Stratified coarse-to-fine approach
Facescape and Headspace datasets
GCN lightweight model

ABSTRACT

Facial Landmark Detection (FLD) algorithms play a crucial role in numerous computer vision applications, particularly in tasks such as face recognition, head pose estimation, and facial expression analysis. While FLD on images has long been the focus, the emergence of 3D data has led to a surge of interest in FLD on it due to its potential applications in various fields, including medical research. However, automating FLD in this context presents significant challenges, such as selecting suitable network architectures, refining outputs for precise landmark localization and optimizing computational efficiency. In response, this paper presents a novel approach, the 2-Stage Stratified Graph Convolutional Network (`2S-SGCN`), which addresses these challenges comprehensively. The first stage aims to detect landmark regions using heatmap regression, which leverages both local and long-range dependencies through a stratified approach. In the second stage, 3D landmarks are precisely determined using a new post-processing technique, namely `MSE-over-mesh`. `2S-SGCN` ensures both efficiency and suitability for resource-constrained devices. Experimental results on 3D scans from the public Facescape and Headspace datasets, as well as on point clouds derived from FLAME meshes collected in the DAD-3DHeads dataset, demonstrate that the proposed method achieves state-of-the-art performance across various conditions. Source code is accessible at https://github.com/gfacchi-dev/CVIU-2S-SGCN.

1. Introduction

Facial Landmark Detection (FLD) algorithms (Wu and Ji, 2019) are designed to automatically detect the locations of key facial landmarks, such as the corners of the eyes, eyebrows and the tip of the nose, in images, 3D point clouds or videos. FLD approaches can be categorized in FLD on images and on 3D data.

FLD on images has drawn much attention for many years as a prerequisite in many computer vision applications. For example, it is adopted as the first step in a wide variety of tasks, including face recognition (Bodini et al., 2018), estimation of head pose (Yang et al., 2019), and evaluation of facial pain expression (Patania et al., 2022), up to cues related to its clinical status (Sforza et al., 2013). Recently, to obtain accurate facial landmark detection on in-the-wild images, 2.5D/3D FLD methods have been proposed (Zeng et al., 2023; Ferman et al., 2024). They still operate on 2D images, while requiring multi-view consistency or leveraging generative 3D visual models and neural rendering.

Instead, FLD on 3D data, the focus of our study, aims at detecting facial landmarks on point clouds. Indeed, with the evolution of 3D/4D capture technologies even with smartphone (Thurzo et al., 2022), the abundance of meshes and point clouds has increased (Cheng et al., 2018; Yang et al., 2020), fostering interest in this domain, which has found utility in diverse areas, including face recognition and biometrics (Kakadiaris et al., 2017; Zhou and Xiao, 2018), facial expression recognition (Sandbach et al., 2012), alignment and morphometric analysis (Blandano et al., 2024), animation and virtual reality (Choi et al., 2022), medical and dental applications (Hallgrímsson et al., 2020; Lee et al., 2022). In particular, it should be noted that in medical research, there is an extensive collection of studies investigating the use of facial anthropometry to diagnose genetic syndromes through craniofacial abnormalities, focusing on manually detected landmarks (Vu et al., 2022; Gibelli et al., 2020). However, these methods rely on operator skill, making them challenging and time-consuming. Thus, automating FLD on 3D data is crucial to improve reliability and efficiency, allowing rapid analysis of facial morphology through 3D scans.

Current methods for extracting landmarks from 3D data often rely on 2D images through projection techniques (Paulsen et al., 2018), conformal geometric mappings (Fan et al., 2016), or UV position maps (Zhang et al., 2020). Although these approaches allow for the use of well-established 2D image processing techniques, they have limitations. Specifically, conformal geometric mapping methods can robustly locate landmarks despite variations in facial expression and

[*] Corresponding author.
E-mail address: giuseppe.facchi@unimi.it (G.M. Facchi).

https://doi.org/10.1016/j.cviu.2024.104227
Received 31 May 2024; Received in revised form 17 October 2024; Accepted 6 November 2024
Available online 12 November 2024

rotation, but they rely on 2D landmark extraction techniques that are often appearance-based. This can be problematic when the appearance is missing or varies due to subject or acquisition conditions. Similarly, projection methods may inaccurately determine landmark positions because they do not fully account for the inherent 3D geometry of the different faces among the subjects.

Therefore, a continuous, compact, and easily accessible 3D representation is favored as input for 3D landmark localization models, enabling more direct and precise estimation. However, the pursuit of this approach presents significant challenges, mainly due to three key factors that must be addressed. First, select the appropriate network architecture to learn the 3D landmarks. Convolutional Neural Networks (CNNs) are inherently designed for regular grid structures like 2D images, and thus do not directly apply to 3D facial data, which is typically represented as unordered point clouds rather than regular grid structures. Graph Neural Network (GNN) models offer a promising alternative to process 3D data, but there has been a limited emphasis on optimizing predictive performance by defining a graph structure for the FLD task. A second challenge involves refining neural network outputs in post-processing to achieve maximal precision. In a FLD on images context, the soft-argmax method is often used to extract precise landmark coordinates from a 2D heatmap (Honari et al., 2018), enabling robust and accurate landmark localization. Unfortunately, applying this function to a set of 3D coordinates yields 3D target landmarks that may not lie on the facial surface. This is especially true in areas with significant curvatures, requiring a different approach. Lastly, it is important to minimize the computational time of the FLD method, as this is typically the first stage in a pipeline for further processing on 3D or 4D data and must avoid the creation of bottlenecks. Moreover, the suitability for deployment on low-resource or edge devices also requires the implementation of lightweight models. Effectively addressing these challenges is key to advancing FLD techniques on 3D data and their practical application in real-world scenarios.

In this study, we present a new method that addresses all the aforementioned challenges by detecting the position of 3D landmarks directly on the mesh surface, leveraging geometric information rather than relying on 2D image-based appearance. Specifically, we propose a 2-Stage Stratified Graph Convolutional Network, namely 2S-SGCN. In the first stage, a heatmap regression is proposed to detect the landmark regions, adopting a stratified Coarse-to-Fine GCN. Essentially, this method involves the construction, evolution and exchange of a fully connected coarse graph that fosters a global exchange of features, along with a locally connected fine graph designed to capture the local details inherent in the 3D facial structure. The advantage of integrating long-range dependencies is dual: it not only enhances performance, but also provides a lightweight solution that is suitable for implementation on resource-constrained devices. The efficiency of this step is further enhanced by working on resampled point clouds with a strictly lower number of points, which accelerates the process while still ensuring accurate localization of the landmark regions. In the second stage, starting from the landmark regions, the precise 3D landmarks are determined. To this aim, we introduce a novel method called MSE-over-mesh, which identifies, through a minimization process, the facial landmarks on the original mesh. Given the required precision, the computation of this stage is carried out over the original mesh points (not resampled) ensuring that the estimated landmarks are part of it, achieving high precision. Moreover, due to the localized nature of this process, the time consumption is minimal.

In summary, the main contributions of our work include:

- A novel Stratified GCN (SGCN) heatmap regressor which leverages both local and long-range dependencies. This leads to optimized utilization of computational capacity, making it well-suited for devices with limited hardware.
- A new 3D heatmap post-processing method for 3D landmarks regression (MSE-over-mesh), guaranteeing the landmarks are part of the mesh.

Our approach demonstrates state-of-the-art performance on publicly available 3D scan datasets, Facescape and Headspace, across different emotion and expression conditions, with additional evaluation on point clouds derived from FLAME meshes in the DAD-3DH dataset. Comprehensive ablation studies further reinforce the effectiveness of the stratified approach and the proposed refinement method.

2. Related works

In this section, we recall FLD methods on images, survey techniques for FLD on 3D data, and revisit deep learning architectures for point-cloud data.

2.1. FLD on images

Current facial landmark detection methods for images can be categorized into two types: 2D landmark detection, which focuses on localizing landmarks only on the visible portion of the face in the 2D image, and 2.5D/3D landmark detection, which identifies also landmarks that are not visible in the 2D image but are present on the underlying 3D facial structure eventually inferred.

In recent decades, the field of 2D FLD on images, has witnessed notable advances, with traditional FLD algorithms typically falling into three main categories: Holistic methods, Constrained Local Model (CLM) methods, and Regression-Based methods. Holistic methods rely on both the overall facial appearance and the broader facial shape patterns to detect landmarks. An example of this approach is the Active Appearance Model (AAM), pioneered by Taylor and Cootes (Edwards et al., 1998). AAM utilizes statistical techniques to fit facial images using a limited set of coefficients, thereby controlling variations in both facial appearance and shape. This method aims to minimize differences between a synthesized face and the target face. In contrast, CLM methods (Cristinacce et al., 2006) determine landmark positions by considering both global facial shape patterns and local appearance information surrounding each landmark. This approach is advantageous because it is better equipped to handle illumination variations and occlusion. Regression-Based methods (Tang et al., 2019) take a different approach by directly learning the mapping from image appearance to landmark locations. Unlike Holistic and CLM methods, they typically do not construct explicit global face-shape models. Instead, they may implicitly embed face shape constraints in their learning process. More recently, the advent of deep learning, particularly convolutional neural networks (CNNs), has significantly improved the precision of 2D FLD, marking a remarkable advancement in the field (Zhu and Ramanan, 2012; Burgos-Artizzu et al., 2013).

More recently, there has been a surge of research activity in this field, focusing on estimating 3D landmarks from images and videos. In this work, we broadly categorize these methods into three approaches: 2D-to −3D projection, 3D-aware techniques, and 3D Morphable Model (3DMM)-based methods. While 2D-to −3D projection methods directly regress 3D landmarks from 2D images (Bulat and Tzimiropoulos, 2017), 3D-aware techniques use volumetric representations to encode the 3D landmarks (Zhang et al., 2022), applying explicit multi-view image constraints to ensure consistent 3D landmark predictions across images (Zeng et al., 2023; Ferman et al., 2024). Lastly, 3DMM-based methods rely on 3D face models, such as BFM (Paysan et al., 2009) or FLAME (Piao et al., 2019), either to directly estimate model parameters (Wood et al., 2022) or as intermediate representations to refine 3D landmarks (Wu et al., 2021; Valle et al., 2019). Additionally, some methods project 3DMM vertices directly onto the image, as seen in 3DDFA-V2 (Guo et al., 2020b), or fit the FLAME model to the 3DMM and use the resulting vertices for projection, as demonstrated in DAD-3DNet (Martyniuk et al., 2022) and RingNet (Sanyal et al., 2019).

Although numerous efforts have been made to develop methods for extracting 3D landmarks from 2D images, the task remains fundamentally challenging due to the limited spatial information that 2D images can provide. Accurately determining the 3D positions of facial landmarks using only 2D data is an ill-posed problem, as it inherently lacks the depth and dimensionality required for precise localization. This approach does not fully capitalize on the potential of 3D data, which can significantly enhance accuracy and reliability. In fact, 3D data offers a detailed representation of facial surface geometry, capturing fine details and contours that 2D images cannot, leading to more accurate and robust landmark estimation.

2.2. FLD on 3D data

With the growing use of 3D acquisition devices, newer techniques have emerged that operate directly on 3D data. These techniques eliminate the need to infer the depth component from 2D images, leading to 3D facial landmark detection algorithms designed to identify the precise locations of facial landmarks in three-dimensional space. It is important to note that the method proposed in this paper, 2S-SGCN, falls into this category, using a 3D point cloud as input to fully exploit the capabilities of 3D data.

Conventional approaches have relied on geometric features, such as combining surface curvatures and depth relief curves for the location of landmarks (Segundo et al., 2010). However, the accuracy of these methods is often limited by their reliance on manually engineered features or custom 3D face models.

More recently, the “Multiview Consensus CNN for 3D Facial Landmark Placement” (MVLM) (Paulsen et al., 2018) has been proposed to leverage the robustness of 2D FLD methods. MVLM can be categorized into indirect 3D deep learning methods involving a multistep process: first, the 3D shape is projected onto a 2D plane from various angles. Then, a 2D CNN is employed to predict the landmark positions, forecasting a heatmap on these 2D images. Ultimately, it integrates the highest point of the heatmap across multiple views using a RANSAC procedure. Although this method yields superior accuracy compared to conventional techniques, it introduces numerical discrepancies due to the transition between 2D and 3D coordinates, as well as the integration of 2D landmarks from multiple viewpoints. In addition, it is time consuming.

The state-of-the-art performance are achieved by 3DFA-GCN (Wang et al., 2022) that employs a fully 3D approach to the FLD problem. It is based on the Position Adaptive Graph Convolution (PAConv) point cloud architecture (Xu et al., 2021), and incorporates a local surface unfolding and registration module to predict 3D landmarks from the heatmaps.

2.3. Deep learning architectures on 3D data

3D data are unstructured and inherently irregular, which prevents the application of standard CNNs designed for 2D image data. Here we briefly recall the point cloud network architectures (Guo et al., 2020a) that can be referred also to tackle the 3D landmark detection. They can be categorized in three main areas: Pointwise MLP networks, Convolutional-Based networks, and Graph-Based networks.

A pioneering work in Pointwise MLP networks is PointNet (Qi et al., 2017a). Specifically, it learns pointwise features independently with several MLP layers and extracts global features with a max-pooling layer. It achieves permutation invariance by summing up all representations and applying nonlinear transformations. As the core of the PointNet++ (Qi et al., 2017a) hierarchy, its set abstraction level is composed of three layers: the sampling layer, the grouping layer, and the PointNet based learning layer. By stacking several set abstraction levels, PointNet++ learns features from a local geometric structure and abstracts the local features layer by layer.

Regarding Convolution-Based networks, PointConv (Wu et al., 2019) performs 3D convolution by treating convolution kernels as nonlinear functions of local coordinates of points, comprising of weight functions (learned via MLP networks) and density functions (estimated through kernel density estimation). Similarly, KP-Conv (Thomas et al., 2019) constructs convolution kernels by combining predefined kernels with specific rules. These methods often exhibit high complexity, both in terms of memory usage and computational burden during learning.

Recent advancements in point-cloud learning have turned to Graph-Based Networks due to their ability to model neighboring information effectively for irregular data. With this in mind, DGCNN (Wang et al., 2019b) employs an EdgeConv module to extract local features from a dynamic graph, continually updating neighboring relationships at each feature layer. Another recently introduced module is PAConv (Xu et al., 2021), which has been integrated into widely-used architectures such as PointNet++ and DGCNN. PAConv dynamically constructs convolutional kernels based on positional information, assembling basic weight matrices stored in a Weight Bank. The coefficients for these matrices are self-adaptively learned from point positions through ScoreNet. This data-driven approach enables PAConv to build kernels with greater flexibility, enhancing its capability to handle irregular and unordered point cloud data efficiently and effectively.

3. Preliminaries

3.1. Notation

In our discussion, the set of vertices on the original mesh is denoted by P. We define the subset of P representing facial landmarks as L, where $|L| \ll |P|$. From a resampled version of P (cfr. Section 3.2) we construct a graph represented as a tuple $G = (V, E)$, where $e_{ij} = (v_i, v_j) \in E \subseteq V \times V$ represents an edge from the vertex v_i to the vertex v_j. The set of neighbors of a vertex v_i is denoted by $\mathcal{N}(v_i) = \{v_j \in V \mid (v_i, v_j) \in E\}$. The matrix of node features is denoted as $F \in \mathbb{R}^{|V| \times n}$, where each $f_i \in \mathbb{R}^n$ is the feature vector for vertex v_i.

3.2. Preprocessing

Consistency and effectiveness across datasets are ensured by implementing a pre-processing phase aimed at normalizing raw facial point clouds. First, to address meshes expressed in varying coordinate systems with different scales, rotations, and translations, a face standardization step is required. This involves centering data points to eliminate potential translations, rotating them to align their principal eigenvectors with reference ones, and scaling them so that all points fit within a volume where each side extends from −1 to 1. Standardization is applied to both the vertices in P and the landmarks in L. For simplicity, throughout the remainder of the paper, we will refer to standardized positions without explicitly stating it.

Furthermore, we conceived a resampling step aiming at normalizing the point-cloud cardinality, as well as the point-cloud sampling locations. Indeed, some devices employ a uniform sampling method across all surfaces, while others may sample more densely in areas of high curvature to better capture detailed surface information. To ensure uniformity, we resample all input meshes to have the same number of points $|V|$ uniformly distributed (see Fig. 1). To this aim, we adopted Poisson disk sampling as described in Yuksel (2015). The resampling step also serves the purpose of downsampling the mesh, significantly reducing the computational cost of the first stage of our method. In the remainder of the paper, we will denote the collection of resampled vertices as V (where $|L| \ll |V| \ll |P|$), which will serve as the basis for constructing the graphs in the next sections.

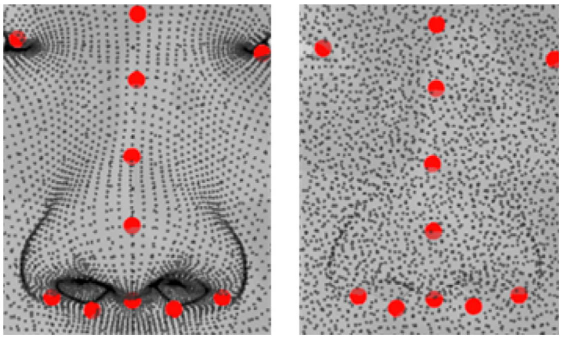

Fig. 1. (left) A detail of an original mesh from the Facescape dataset. **(right)** The corresponding mesh after uniform resampling. Red dots correspond to the landmarks in this area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

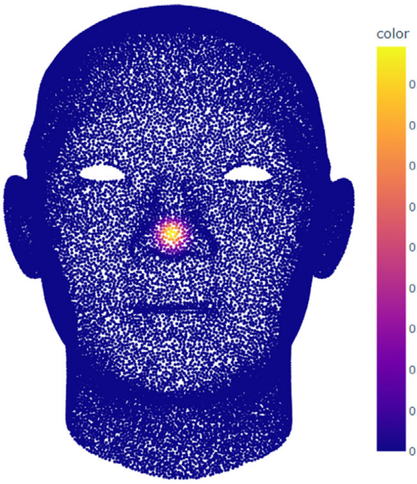

Fig. 2. An example of a 3D heatmap corresponding to the nose landmark. It is obtained applying Eq. (1) with $\sigma = 0.03$.

3.3. 3D heatmap generation

In order to train the heatmap regressor used in the first stage of the proposed method, it is essential to generate 3D heatmaps from landmark ground truths serving as soft labels (Wang et al., 2022).

Specifically, given the resampled vertices V of a 3D face point cloud, and the corresponding landmarks L, for each vertex $v \in V$, we compute the Euclidean distances to every landmark $l \in L$, producing a distance matrix Δ with dimensions $\mathbb{R}^{|V|\times|L|}$. This matrix is then transformed using a Gaussian function to encode the distances into a normalized probability matrix, represented as the 3D heatmap:

$$H = \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \in \mathbb{R}^{|V|\times|L|}, \tag{1}$$

where σ represents a hyperparameter that controls the spread of the heatmap. Essentially, the 3D heatmap indicates the probability of each landmark occurring at a specific location as shown in Fig. 2 for one example case.

3.4. GCN layer and loss function

A core component of our approach is the use of a GCN (Graph Convolutional Network) layer, as proposed in Kipf and Welling (2016). The iterative aggregation and update steps at layer l for node features f_i are defined by the following equation:

$$f_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(v_i)} \frac{1}{c_{ij}} W^{(l)} f_j^{(l)}\right). \tag{2}$$

In this expression, $W^{(l)}$ denotes the learnable weight matrix at layer l, c_{ij} is a normalization factor, which may be the product of degrees of the v_i and v_j among other potential normalization techniques, and $\sigma(\cdot)$ is the activation function applied element-wise.

Another essential component of the proposed solution is the use of a loss function specifically designed for unbalanced data. This is crucial because the generated 3D heatmaps feature only a few points within the Gaussian bell of each landmark, with many points falling outside of it, resulting in an unbalanced ground truth. Specifically, we adopt a modified version of the Adaptive Wing Loss (Wang et al., 2019a): let h and $\hat{h}$ represent the ground truth and estimated heatmap values for a point v, respectively. We calculate the point-wise loss as follows:

$$\ell(h,\hat{h}) = \begin{cases} (1+\beta h)\omega \ln|h-\hat{h}|^{\alpha-h} & \text{if } |h-\hat{h}| < \theta \\ (1+\beta h)M|h-\hat{h}| - N & \text{otherwise.} \end{cases} \tag{3}$$

In essence, the Adaptive Wing Loss dynamically adjusts its behavior based on the magnitude of prediction errors. For small errors, it behaves like a logarithmic curve, focusing on fine-tuning and detailed adjustments. For larger errors, which are common in cases of severe class imbalance or when the model prediction is far from the ground truth, the loss function transitions to a more linear behavior, ensuring that the model remains sensitive to these errors without being overwhelmed by them.

Our modification introduces a weight, $(1 + \beta h)$, proportional to the point-wise heatmap h, thus prioritizing errors near the landmark positions.

4. Proposed method

Estimating 3D landmarks presents unique challenges, which we address with a 2-Stage Stratified Graph Convolutional Network. The first stage focuses on identifying the landmark regions, while the second one focuses on the inferred areas to pinpoint the precise landmark positions. The implications of this two-stage approach are twofold: it ensures an efficient solution by allowing the determination of landmark regions on a rough, resampled mesh, while analyzing the original mesh just for the final point detection, which hence requires minimal computation. Additionally, high 3D landmark precision is ensured by the second stage working on the original mesh.

Furthermore, the first stage employs a stratified approach that combines a holistic yet sparser analysis (Coarse) with a more detailed local analysis (Fine). This method effectively addresses the imbalanced nature of the data (distinguishing between landmark and non-landmark points) and the lack of precise morphological features for certain landmarks, such as those on the face contour. A purely local analysis would struggle to tackle these challenges, while a purely global approach would be too rigid and coarse, failing to adapt to the subtle variations in facial features and expressions. In Fig. 4 we sketch our method comprising the two primary stages. The first is the *heatmap regression*, which produces $\hat{H} \in \mathbb{R}^{|V|\times|L|}$, that is the estimate of the 3D landmark heatmap H. The second is the *landmark estimation* stage, which takes $\hat{H}$ and the original mesh P as input, and predicts the final landmark positions $\hat{L}$, belonging to P.

4.1. Heatmap regression stage

Our model takes as input the resampled point cloud, treated as the set of nodes V for the graph G. For the local analysis, we connect each vertex in V with its k-nearest neighbors (G_F in Fig. 4). The choice of k impacts the method performance: higher k augments the receptive field for each GCN layer avoiding to get stuck in local minima, while introducing the risk of over-smoothing. In contrast, for the holistic analysis, our objective is to identify a small subset of vertices $C \subset V$, establishing a total connection among them to guarantee global communication (G_C in Fig. 4). Such set C should guarantee the total

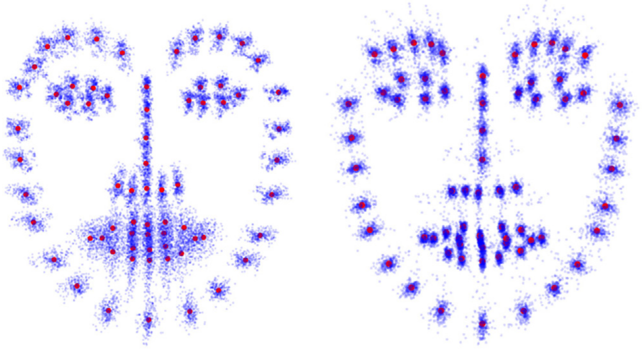

Fig. 3. 3D representation of landmark positions for 300 randomly selected subjects from Facescape Emotions (left) and Headspace datasets (right). The red dots constitute the average positions L^{mean} for each landmark type. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

coverage of the point cloud. One possible choice could be to perform a very sparse uniform resampling (e.g., with 100 vertices). Here we opt to a more effective solution that exploits the knowledge given by the labeled training set. Indeed, as shown in Fig. 3, all landmarks on the normalized training data are spread over specific areas of interest, even in the presence of multiple emotions (plot on the left). Computing the average for each landmark (red dots in Fig. 3) allows the construction of a prototype to refer to in the determination of the subset C for each graph at hand.

Specifically, we obtain for each dataset the prototype L^{mean} computing the mean position of each landmark across all subjects in the training set:

$$L^{mean} = \{l_i^{mean} = \frac{1}{|S|}\sum_{s\in S} l_i^s,\ i = 1...|L|\}, \tag{4}$$

where s represents a sample in the training set S.

Then, L^{mean} is aligned with the vertices V at hand, and the subset of vertices C is obtained by identifying the points in V closest to the positions in L^{mean} as:

$$C = \{c_i \mid c_i = \arg\min_{v\in V} d(v, l_i^{mean}), i = 1...|L|\}. \tag{5}$$

The resulting set C ensures coverage of the entire face, focusing specifically on areas likely to correspond to the landmarks to be identified. Given C, we can construct the Coarse graph $G_c = (V, E_c)$, where $E_c = \{(u, v) \mid (u, v) \in C \times C\}$ is the set of edges of G_c. For completeness, we specify that the Fine graph $G_f = (V, E_f)$, implements the local connection as $E_f = \{(u, v) \mid v \in \mathcal{N}(u), \forall u \in V\}$, where $\mathcal{N}$ resembles a k-nearest neighbors (k-NN) technique.

All vertices $v \in V$ are characterized by the feature vector $f(v) = [pos(v), normal(v)]$, where $pos(v)$ is the 3D coordinate set of v and $normal(v)$ is the normal of the triangle face from which v has been sampled.

To harness both coarse and fine-grained representations of 3D facial data, we devised two distinct branches within our model architecture. The main computational block is made up of a GCN layer (Eq. (2)), a normalization layer (LayerNorm), and a Rectified Linear Unit (ReLU) activation function. This is initially applied to G_c and G_f, computing the initial embeddings for each node in both the Coarse and Fine scenarios independently. The architecture is then designed to integrate the information from the two modalities using a linear layer, which furnishes all nodes with both local and global information. Following this integration, the previously introduced main computational block is applied iteratively to both branches D times, alternating the linear layer to fuse the local and global information. Here, to avoid over-smoothing, residual connections (Li et al., 2020) between consecutive layers are employed. Finally, a fully connected layer is adopted to produce the output $\hat{H} \in \mathbb{R}^{|V|\times|L|}$.

4.2. Landmark estimation stage

Given $\hat{H}$, a straightforward method for estimating a landmark l involves calculating the weighted average of the points with the highest predicted similarity values relative to the landmark itself.

Specifically, let $\text{top}_t(l)$ be the set of t points that have the highest values in $\hat{H}$ in relation to the landmark l. The weighted average is computed as:

$$m = \frac{1}{t}\sum_{v\in \text{top}_t(l)} \text{softmax}(\hat{H}(v, l)) \cdot pos(v), \tag{6}$$

where the softmax converts the heatmap values of the $\text{top}_t(l)$ vertices into a probability distribution, then used to weight the vertices positions. While this solution is adequate, it does not guarantee that the resulting point will be situated on the mesh manifold introducing an avoidable error.

Here, we introduce a new approach, namely MSE-over-mesh, that addresses the aforementioned problem, recasting the computation in the original mesh domain. Formally, for each landmark l, we first compute the mean m according to Eq. (6), and then identify the set B_m of vertices in the original mesh P that are located within a sphere centered on m. The radius of this sphere is set to include a broader locality:

$$\begin{aligned} B_m &= \{p \in P \mid d(p, m) \le \gamma r_m\}, \\ r_m &= \max\{d(m, v) \mid v \in \text{top}_t(l)\}, \end{aligned} \tag{7}$$

where $\gamma > 1$ and d is the Euclidean distance. The predicted landmark $\hat{l}$ is identified as the value p within B_m that, when used as the center of a Gaussian distribution (as specified in Eq. (1)), best aligns with the distribution represented by the estimated heatmap $\hat{H}(q, l)$ for the landmark l, throughout $q \in \text{top}_t(l)$. Mathematically, this is achieved by minimizing the sum of the squared differences between the Gaussian distribution centered at p and evaluated at q, and the estimated heatmap $\hat{H}(q, l)$ for the landmark l over the same points q:

$$\hat{l} = \arg\min_{p\in B_m} \frac{1}{t}\sum_{q\in \text{top}_t(l)} \left(\hat{H}(q, l) - \exp\left(-\frac{d(p, q)^2}{2\sigma^2}\right)\right)^2 \tag{8}$$

By selecting the point from the original mesh, we ensure that $\hat{l}$ is positioned accurately on the mesh itself, thus maintaining topological precision.

A representation of the whole landmark estimation stage is drawn in Fig. 5.

5. Experimental analysis

The experimental analysis has been conducted to evaluate various aspects of the proposed method, 2S-SGCN. First, a comprehensive application of 2S-SGCN to 3D face scans is performed. Specifically, we refer to two publicly available datasets, Facescape and Headspace, to evaluate the 2S-SGCN method alongside state-of-the-art (SOTA) methods acting on 3D scans, 3DFA-GCN and MVLM, while performing extensive ablation studies. Second, to demonstrate the effectiveness of 2S-SGCN even on large datasets, including subjects captured under diverse, uncontrolled conditions (pose, lighting, expression), we apply it to the FLAME meshes collected in the DAD-3DHeads dataset (Martyniuk et al., 2022). This allows for a comparison of our method with those that estimate 3D landmarks from images and videos, 3DDFA-V2, RingNet and DAD-3DNet, although the comparison is indirect.

5.1. Datasets

Facescape (FS) dataset (Yang et al., 2020). This dataset encompasses textured 3D faces from 847 subjects, acquired in 20 different conditions including four primary facial emotions (neutral, sadness, smile, and anger), while the other data capture facial elementary expressions (e.g. lip roll, brow raiser, mouth stretch). Each acquisition comprises approximately 25,000 vertices and 50,000 triangles. The dataset also includes 68 3D facial landmarks for each model (Fig. 6).

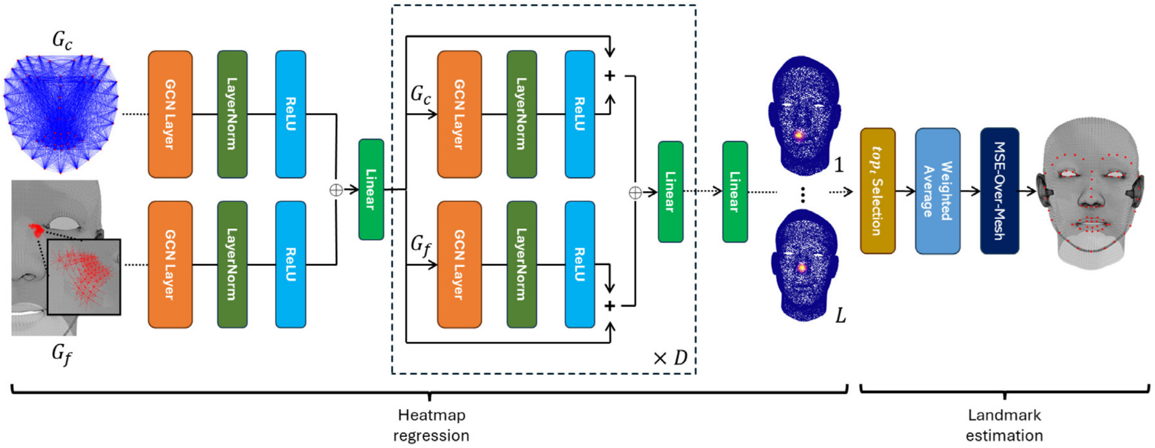

Fig. 4. Illustration of the proposed model architecture. Two stages are conceived: the Heatmap regression and the Landmark estimation. The input to the process is constituted by both a Coarse graph representation, G_c, where sparse nodes are fully connected, alongside a Fine representation, G_f, where each node is linked to its k-nn neighbors (for clarity, only a subset of the graph G_f is displayed). The symbol ⊕ signifies the concatenation operation. D refers to the model depth, while L indicates the number of landmarks.

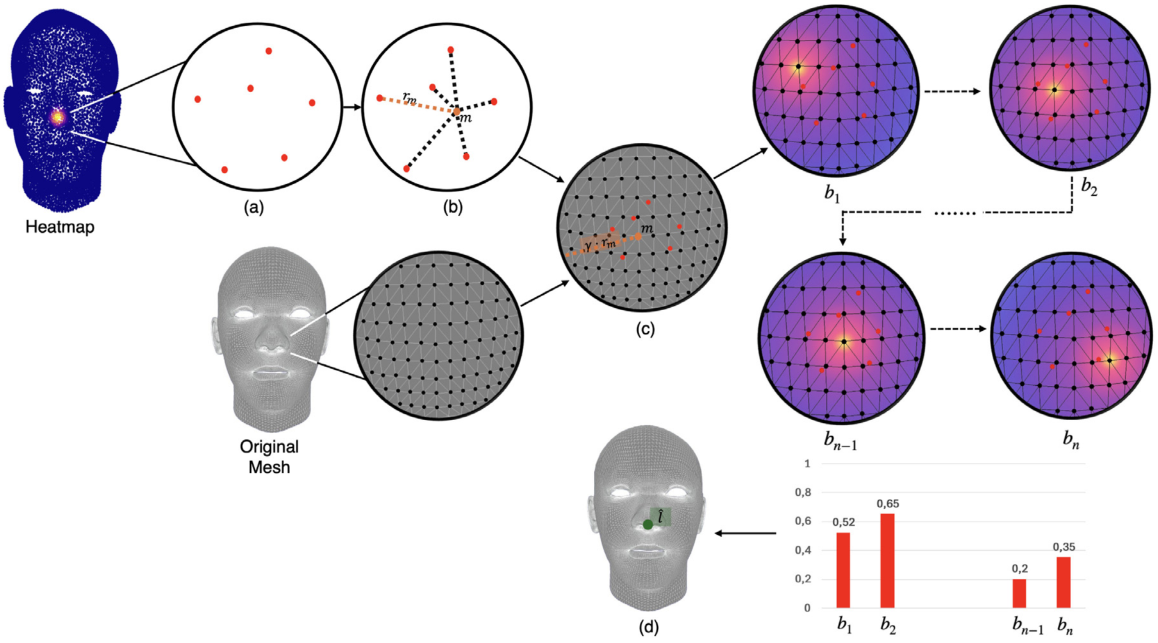

Fig. 5. Illustration of the refinement method. (a) The $\text{top}_t(l)$ points with the highest probability are identified on the learnt heatmap of the landmark l. (b) The weighted average, denoted as m, is calculated, and the distances from m to each of the $\text{top}_t(l)$ points are computed. (c) A sphere with radius $\gamma \cdot r_m$ is extrapolated on the original mesh centered at m. (d) For each $p_i \in B_m$, the error b_i is calculated as the average difference between the heatmap values of the $\text{top}_t(l)$ points, obtained by centering the heatmap at p_i, and their predicted probability values generated by the model. The predicted landmark on the original mesh is the one with minimum b_i relative error, as described in Eq. (8).

Headspace (HS) dataset (Dai et al., 2020). This dataset comprises 1519 3D subject acquisitions, including 25 non neutral expressions. The resolution is variable, but typically there are over 100,000 vertices and over 200,000 triangles. The dataset is provided together with 68 3D facial landmarks for 1223 models (Fig. 7).

DAD-3Dheads (DAD-3DH) dataset (Martyniuk et al., 2022). This dataset consists of 44,898 images captured in uncontrolled conditions, with 42,152 of these images linked to a corresponding FLAME mesh containing 5023 vertices. Of these, 3669 vertices represent the head, while the remainder correspond to the neck and eyeballs. Additionally, the dataset provides model-view and frustum projection matrices that map the 3D mesh from model space to 2D images, along with rich attribute data, including head poses, emotions, occlusions, gender, age, image quality, and lighting conditions. A sample from the dataset is shown in Fig. 12.

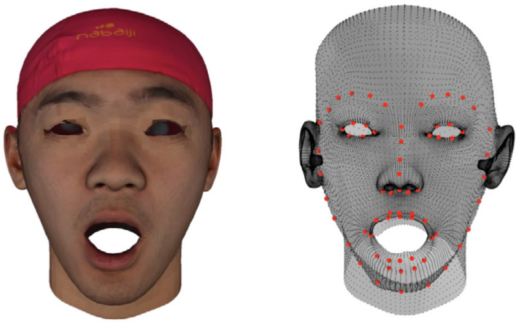

Fig. 6. **(left)** A textured mesh from Facescape (FS) dataset. **(right)** The mesh with landmarks.

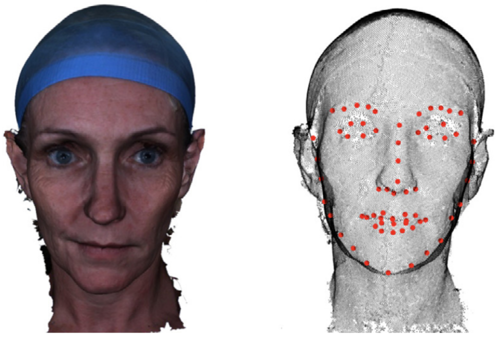

Fig. 7. **(left)** A textured mesh from Headspace (HS) dataset. **(right)** The mesh with landmarks.

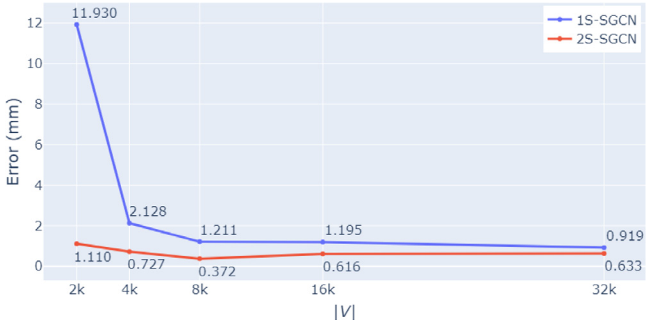

Fig. 8. Mean error (mm) obtained on the FS Neutral test set, varying the number of vertices $|V|$ of the input data.

5.2. Parameters setting and model configurations

The hyperparameter optimization and adjustment for the 2S-SGCN architecture were carried out with reference to the neutral subjects in the Facescape dataset, while also ensuring that they remain optimal for other trials considered, resulting in the following settings. The number of vertices of the graphs, $|V|$, has been set to 8000, according to the investigation reported in Fig. 8. This choice turns out to be a good trade-off between effectiveness and efficiency. Besides, the receptive field is defined by tuning the parameters k (the number of neighbors in G_f) and D (the number of layers in the central block of the architecture) together, finding an optimal balance with $k = 16$ and $D = 20$. The number of vertices t considered during stage 2 (Eq. (6) and subsequent) was determined by evaluating model performance while varying t within a range from 1 to 35. As shown in Fig. 9, when adopting the refinement MSE-over-mesh, t does not significantly affect performance, suggesting that the method is robust with respect to this parameter. We set it to 12 as it guarantees good performance and low computational costs. Finally, we set $\gamma = 5$ (Eq. (7)), to address the significant difference in point density between resampled and original meshes.

The heatmap regressor is configured to receive six input channels that represent initial node embeddings, which are then mapped to a 64-dimensional latent embedding. This dimensionality is kept constant until the final linear layer, which produces $|L|$ outputs. Concerning loss optimization parameters (Eq. (3)), we adhere to the settings recommended in the original paper (Wang et al., 2019a), specifically $\omega = 14, \theta = 0.5, \alpha = 2.1$, and we set $\beta = 50$.

The 3DFA-GCN architecture was trained using the original Adaptive-Wing loss, maintaining the parameter setting proposed in the original paper. We adhered to the code provided by the authors and applied their technique to resample point clouds using Farthest Point Sampling (FPS), as detailed in Qi et al. (2017b). We conducted numerous experiments to determine the ideal FPS setting, considering values such as 2048, 8192, and 32,768. Ultimately, we set the FPS to 8192, balancing computational efficiency with maintaining model accuracy, and allowing for direct comparison with 2S-SGCN.

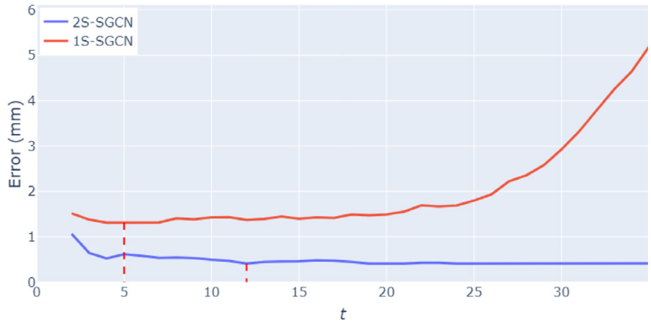

Fig. 9. Average landmark localization error produced varying the value of t in Eqs. (6)–(8). The plots refer to the method using the simple weighted average method m for landmark estimation, yielding the 1S-SGCN, or using the MSE-over-mesh method, resulting in the complete proposed method 2S-SGCN. The best accuracy for MSE-over-mesh is given by $t = 12$, while the best accuracy for m is given by $t = 6$.

All training procedures for both 3DFA-GCN and our method were performed over 500 epochs with the Adam optimizer (Kingma and Ba, 2014) and a batch size of 4.

The MVLM architecture was trained for 100 epochs with the default 3D processing configuration, including geometry and depth information while excluding the utilization of texture information during training.

5.3. Experiments on 3D scans

Experiments have been designed to take advantage of the characteristics of the available data. Specifically, the FS dataset is divided into three subsets: Neutral Emotion (1 per subject), Multi-Emotions (4 per subject including the neutral), and Multi-Expressions (16 per subject). In contrast, the HS dataset is treated as a whole, since only a few of the annotated data (specifically, 13) exhibit non-neutral expressions. Both the HS dataset and the first two FS subsets are used for training and testing (applying an 80% training and 20% testing split, identical for all models, to ensure a fair comparison), while the FS Multi-Expression set, which includes 2704 face models, is used in an inter-condition mode to test the models' ability to generalize. It is worth noting that although the number of landmarks is the same in the two datasets FS and HS, their positions differ, especially in their placement on the contours of the face and eyebrows (Fig. 3). This prevents them from being treated indistinguishably, making cross-dataset validation unfeasible.

Performance is assessed by calculating the Euclidean distance between the predicted landmarks and the actual ground truth. The experimental findings are shown in Table 1, which indicates the mean error for all landmarks across each test set evaluated. Furthermore, to provide an insight into the model behavior, we evaluated their performance in intra-condition mode averaging the errors with respect to specific facial regions by grouping semantically similar landmarks. Specifically we identified five groups: the facial contour, the eyebrows, the eyes, the nose and the mouth. In Fig. 10 the summarized performance are plotted. Concerning the experiments in inter-condition mode we report the performance obtained by each model on each expression (see Figs. 14 and 15).

Finally, the assessment of model sizes and prediction times is documented in Table 2. Although all experiments were carried out on an NVIDIA A100 GPU 80 GB, MVLM was trained on the same A100 GPU but underwent testing on an NVIDIA RTX 3060 GPU due to the method's reliance on 3D rendering of faces.

Table 1
FLD performance is reported as the mean and standard deviation of the Euclidean distance between predicted and true landmarks. The experiments reference the training and test sets used, with the size of each set indicated in brackets. Tests on FS Expressions represent inter-condition mode. The first three lines report our method and comparisons with the SOTA. The last three lines report the ablation studies: 1S-SGCN (the second-stage refinement is replaced by the simple weighted average (m)), 2S-GCN (stratification is not done in the first stage, referring only to the Fine graph), and 1S-GCN (both refinement and stratification are not done).

	Train FS Neutral (678)		Train FS Emotions (2712)		Train HS (978)
Method	Test FS Neutral (169)	Test FS Expressions (2704)	Test FS Emotions (676)	Test FS Expressions (2704)	Test HS (245)
MVLM	0.897 ± 0.177	2.134 ± 1.094	1.579 ± 0.782	1.965 ± 0.769	1.709 ± 0.568
3DFA-GCN	1.447 ± 0.149	2.556 ± 0.462	1.416 ± 0.145	2.226 ± 0.378	1.711 ± 0.477
2S-SGCN	**0.371** ± **0.227**	**1.932** ± **1.204**	**0.477** ± **0.255**	**1.825** ± **1.108**	**1.662** ± **0.727**
1S-SGCN	1.211 ± 0.313	2.490 ± 0.959	1.287 ± 0.301	2.322 ± 0.848	2.098 ± 0.625
2S-GCN	0.623 ± 0.267	2.352 ± 1.656	0.656 ± 0.376	1.875 ± 1.208	1.878 ± 0.766
1S-GCN	1.442 ± 0.282	2.834 ± 1.308	1.442 ± 0.304	2.354 ± 0.925	2.816 ± 0.701

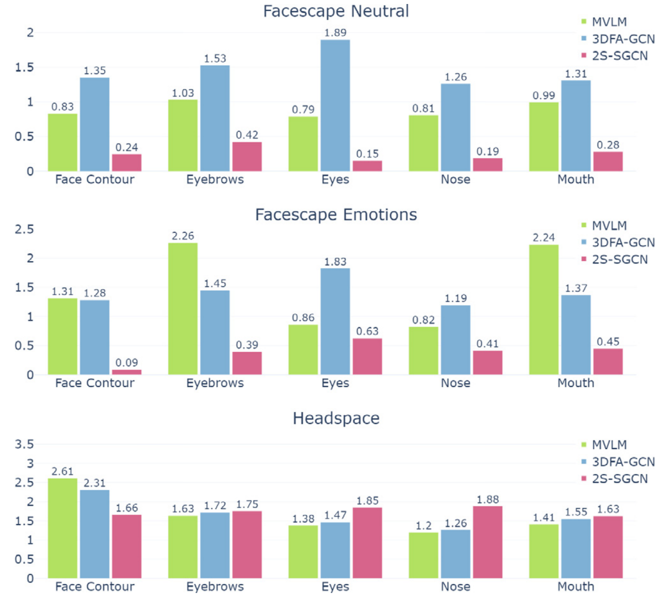

Fig. 10. FLD performance, expressed as the Euclidean distance (in millimeters) between predicted and true landmarks, for the three methods tested (MVLM, 3DFA-GCN, 2S-SGCN) on different landmark groups: Face Contour (17 landmarks), Eyebrows (10 landmarks), Eyes (12 landmarks), Nose (9 landmarks) and Mouth (20 landmarks).

Table 2
Comparison across model size, heatmap prediction time and landmark prediction time for the considered FLD methods.

	Model size	Heatmap prediction time	Landmark prediction time
MVLM[a]	283 MB	5.273 s	4.816 s
3DFA-GCN	4.1 MB	0.068 s	0.423 s
2S-SGCN	**1.5 MB**	**0.023 s**	**0.003 s**

[a] MVLM heatmap prediction time is obtained on NVIDIA RTX 3060 GPU.

5.4. Ablation studies

We conducted ablation studies to assess the impact of the key components of our proposed method, namely the stratified Coarse-to-Fine heatmap regression approach, and the contribution of the second stage MSE-over-mesh for final landmark estimation. Specifically, we setup three ablation studies: 1S-SGCN does not include the refinement stage MSE-over-mesh, using instead the simple weighted average (m) for landmark estimation, thus resulting in a one-stage method. The ablation 2S-GCN does not adopt the stratified approach in Stage 1, referring only to the Fine graph, resulting in a 2-Stage-GCN method. Finally, the ablation study 1S-GCN excludes both the Coarse graph in Stage 1 and the contribution of the MSE-over-mesh refinement method in Stage 2.

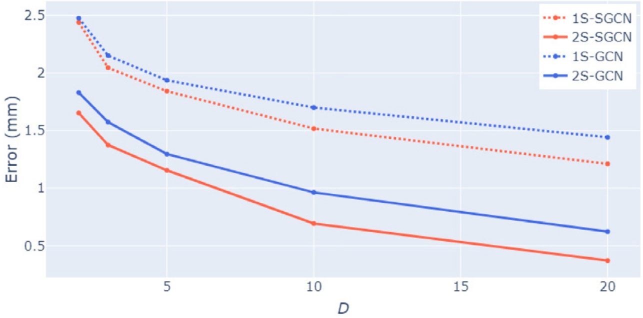

Fig. 11. Comparison of the mean error obtained employing either the stratified (Coarse + Fine) method (SGCN) or the Fine graph representations only (GCN) combined with either the weighted average (1S-) or the MSE-over-mesh methods (2S-). Evaluation is conducted in the 'Facescape Neutral' experimental setting.

The effectiveness and efficiency of the investigated models, along with the complete proposed method, have been evaluated by performing a series of experiments with varying model depth D. The results of these experiments, are presented in Fig. 11. Specifically, it can be inferred that omitting stratification requires a greater network depth to achieve the same performance: the plot shows that 2S-GCN with $D = 20$ has the same performance as 2S-SGCN with $D = 10$ (the same holds when comparing 1S-GCN and 1S-SGCN), indicating that 2S-GCN needs many more iterations of message-passing to reach the same level of information sharing obtained by 2S-SGCN with much less depth while including long-range dependencies. Moreover, it emerges clearly the importance of the second stage: by omitting it, the error significantly increases for any value of D.

A quantitative evaluation of the investigated models is provided in the second part of Table 1 (the last three lines), showing the results obtained for each experimental setting considered in the testing phase.

5.5. Experiments on FLAME meshes

This experimental session aims to evaluate the robustness and generalizability of the proposed method by testing it on meshes derived from images captured outside of a controlled laboratory environment, in real-world scenarios. For this purpose, we use the FLAME meshes provided in the DAD-3DH dataset, which encompass a wider range of facial expressions, ages, and acquisition conditions, covering a continuous spectrum.

The key difference between a 3D scan and a FLAME mesh is the regularity of the latter. This consistent structure allows us to precisely identify a subset of vertices, such as those consistently located at specific facial features, to serve as GT landmarks for our application. This regularity ensures accurate and repeatable landmark selection, which is not possible with raw 3D scans. In Fig. 12 (Right), the 122 GT landmarks are displayed. These landmarks have been selected to closely match those used in the Facescape and Headspace

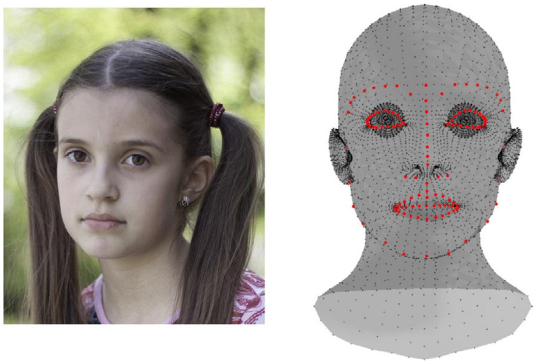

Fig. 12. **(Left)** A 2D image from the DAD-3DH dataset. **(Right)** The mesh generated from the 2D image using the FLAME model, with the selected subset of landmarks highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
NME of our method with SOTA 3D FLD methods on the DAD-3DH dataset across four different conditions: the Overall setting includes all data; the Pose setting focuses on atypical poses in the 2D images; the Expr setting includes instances with exaggerated facial expressions; and the Occl setting considers 2D images with significant facial occlusions. The column |**L**| indicates the number of landmarks detected. Notably, the comparison is indirect, as it refers to a different set of landmarks.

	|**L**|	Overall	Pose	Expr	Occl
3DDFA-V2	68	0.052	0.112	0.046	0.046
RingNet	68	0.128	0.393	0.073	0.186
DAD-3DNet	68	0.033	0.088	0.025	0.029
2S-SGCN	122	0.0073	0.0076	0.0078	0.0072

datasets, but with a denser distribution. Besides, to ensure a fair application of our model without leveraging knowledge of the vertex indexing, we resampled the FLAME models (considering the head portion only) as described in Section 3.2, generating generic point clouds with anonymized vertices. This process ultimately provides us with a large dataset of 42,152 labeled point clouds obtained under uncontrolled acquisition conditions.

For training, we split the dataset into two parts, with 80% used for the training set and 20% for the test set. Experiments demonstrate very high precision in 3D space, achieving a Normalized Mean Error (NME) of 0.0170 ± 0.0037. To provide an indicative measure of this error in a physical distance, we assume the human face is roughly 20 cm in size and scale the results accordingly, yielding a Mean Absolute Error (MAE) of 1.544 ± 0.342 mm.

Interestingly, in this experiment, we can compare our method with state-of-the-art 3D FLD methods on 2D data by using the transformation matrices provided by the DAD-3DH dataset, which map 3D points onto 2D images. These matrices allow us to project both the predicted and ground truth 3D landmarks onto the 2D image, enabling error computation in 2D and comparison with SOTA methods, as shown in Table 3. The results are also presented qualitatively in Fig. 13. While this comparison offers valuable insights, it remains an indirect comparison for two reasons. First, our division of the dataset into training and test sets may differ from those used by other methods. Second, the landmark sets are different: SOTA methods use 68 landmarks, whereas our method uses 122. This difference is due to the absence of the original 68 landmark indices in the dataset, prompting us to select 122 landmarks that closely resemble the original set but with a denser distribution. Nonetheless, given the large size of the dataset, we believe the observed performance is both relevant and consistent.

6. Discussion and conclusion

In this study, we introduced a novel approach, 2S-SGCN, to address challenges in Facial Landmark Detection on 3D data. 2S-SGCN is conceived as a two-stage process: the first stage employs a stratified Coarse-to-Fine Graph Convolutional Network (GCN) for heatmap regression, while the second stage focuses on refining these results to accurately locate points on the original mesh corresponding to the facial landmarks.

The advantage of using a stratified approach has been demonstrated by the ablation studies, where, in all experimental settings, the stratified approach significantly improves performance accuracy. Incorporating long-range dependencies through the Coarse graph not only enhances performance but also provides a lightweight solution by reducing errors with fewer layers. This makes it well-suited for implementation on resource-constrained devices, as shown in Table 2.

Even more effective is the contribution of the refinement stage MSE-over-mesh: it not only allows to improve significantly the performance (cfr. Fig. 11 and Table 1), but also makes the approach less sensitive to the parameter settings, as can be appreciated in Fig. 9 for the top rank definition, and in Fig. 8 for the choice of the graph cardinality. This last achievement is particularly important, as it renders the method agnostic to the size of the input graph, thereby making it especially reliable and adaptable across various scenarios.

Comparisons with the state-of-the-art demonstrate the effectiveness of 2S-SGCN. In particular, in the intra-condition experiments on the FS dataset, 2S-SGCN outperforms both MVLM and 3DFA-GCN. In the experiments involving the HS dataset, all methods achieve performances that differ by a negligible margin of less then 0.2 mm.

In the inter-condition experiments (Test FS Expressions), 2S-SGCN performs the best, although the advantage is reduced, especially compared to MVLM, which as expected behaves well on less controlled data being a non-geometric approach that imposes fewer constraints while requiring higher computational costs. In contrast, 3DFA-GCN shows a reduced capability to generalize. Finally, not surprisingly, the error of 2S-SGCN is slightly reduced when training on multi-emotions instead of solely on neutral expressions.

The investigation is enriched by the efficiency evaluation reported in Table 2, where both the model size and execution time attest to the effectiveness of both stages of 2S-SGCN.

Further insight into the behavior of the model is gained by analyzing the performance of different groups of landmarks in the intra-condition tests (Fig. 10). For the proposed method 2S-SGCN, it is observed that when faces exhibit emotions, a greater error is registered in correspondence to the eyes and the mouth, which are notably the parts of the face that undergo the most deformation during an expression.

Concerning the inter-condition tests, we can observe in Figs. 14 and 15 that 2S-SGCN generally performs better than the other methods, although it struggles with certain expressions such as 'Mouth stretch', 'Jaw left/right', and 'Lip funneler'. It should be noted that these expressions involve significant mouth deformations that were not present during training (especially when adopting the model trained on subjects with neutral expression only), making inference exceptionally challenging for all models.

Furthermore, the application of 2S-SGCN on meshes derived from single 2D images has allowed us to test our method on a large dataset acquired under uncontrolled conditions, further proving its robustness. Additionally, it has demonstrated that our method is agnostic to the original data source: as long as the input data can be mapped to a 3D mesh, our approach consistently performs well, regardless of how the 3D mesh is generated.

The encouraging results prompt further investigation. First, it would be valuable to evaluate the adoption of 2S-SGCN on genuine 3D data acquired in the wild. While such datasets are currently lacking, the growing availability of low-cost acquisition devices, such as depth cameras and smartphones, is expected to address this gap. These datasets would enable a comprehensive evaluation of our method in real-world settings, capturing a wider range of variations beyond those found in controlled environments or synthetic models, offering deeper insights

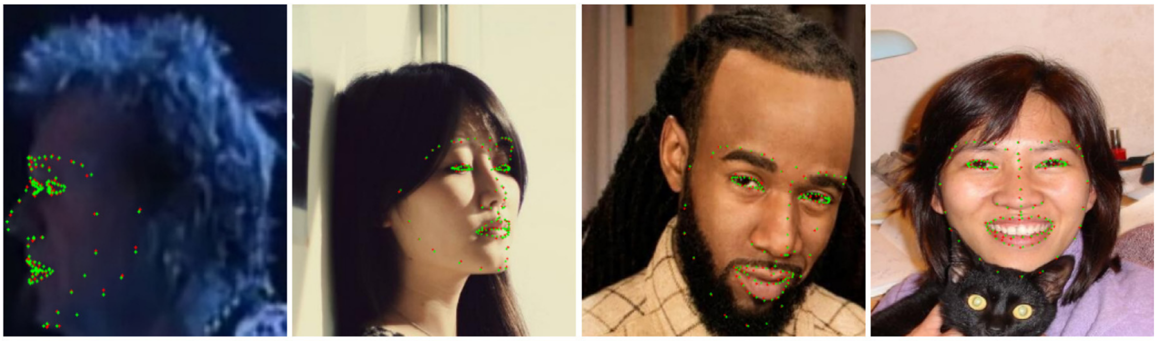

Fig. 13. Results on the DAD-3DH dataset. Green dots represent the projection of the ground truth landmarks, while red dots indicate the projection of the predicted landmarks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

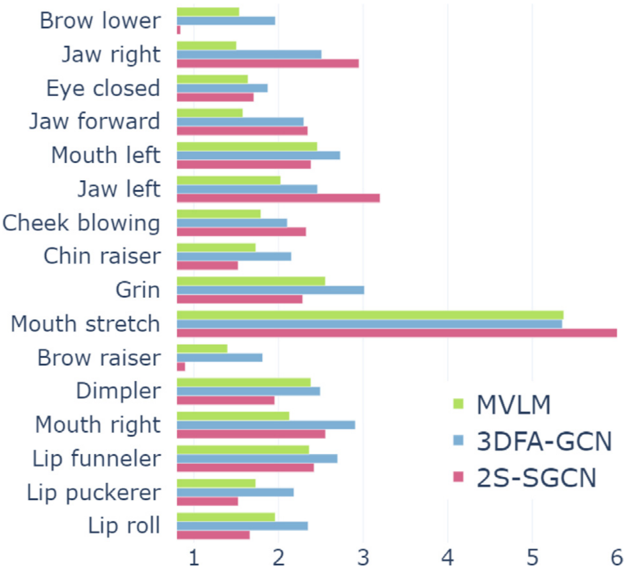

Fig. 14. Mean error (mm) for each expression in the FS Expressions subset (2704 face graphs), evaluated under inter-condition modality using a model trained on FS Neutral.

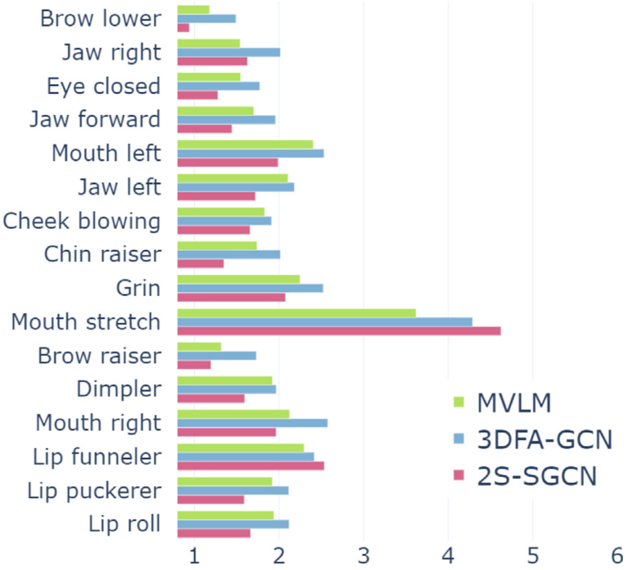

Fig. 15. Mean error (mm) for each expression in the FS Expressions subset (2704 face graphs), evaluated under inter-condition modality using a model trained on FS Emotions.

into the robustness and applicability of our approach in diverse, unconstrained conditions. Additionally, it would be highly interesting to assess the reliability of the identified points indirectly by utilizing them in downstream tasks. Beyond landmark localization, examining how our FLD impacts tasks such as facial expression recognition, emotion detection, or facial attribute classification could offer important insights into the effectiveness of the detected landmarks for more complex facial analysis. Furthermore, the generalization capabilities of the model open opportunities for applications involving rare or unique facial deformations (e.g., due to syndromes, pathologies, or accidents), where extensive datasets may not be readily available.

CRediT authorship contribution statement

Jacopo Burger: Writing – review & editing, Writing – original draft, Methodology, Software, Validation, Investigation, Conceptualization. **Giorgio Blandano:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Giuseppe Maurizio Facchi:** Writing – review & editing, Writing – original draft, Methodology, Software, Validation, Investigation, Conceptualization. **Raffaella Lanzarotti:** Writing – review & editing, Writing – original draft, Project administration, Resources, Conceptualization, Methodology, Supervision.

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.

Data availability

The authors do not have permission to share data.

References

Blandano, G., Facchi, G.M., Tartaglia, G.M., Burger, J., Pedersini, F., Dolci, C., Sforza, C., Cappella, A., 2024. Gender classification via graph convolutional networks on 3D facial models. In: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC 2024). ACM.

Bodini, M., D'Amelio, A., Grossi, G., Lanzarotti, R., Lin, J., 2018. Single sample face recognition by sparse recovery of deep-learned lda features. In: Advanced Concepts for Intelligent Vision Systems: 19th International Conference, ACIVS 2018, Poitiers, France, September 24–27, 2018, Proceedings. 19, Springer, pp. 297–308.

Bulat, A., 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: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030.

Burgos-Artizzu, X.P., Perona, P., Dollár, P., 2013. Robust face landmark estimation under occlusion. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1513–1520.

Cheng, S., Kotsia, I., Pantic, M., Zafeiriou, S., 2018. 4DFAB: A large scale 4D database for facial expression analysis and biometric applications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Choi, B., Eom, H., Mouscadet, B., Cullingford, S., Ma, K., Gassel, S., Kim, S., Moffat, A., Maier, M., Revelant, M., et al., 2022. Animatomy: An animator-centric, anatomically inspired system for 3d facial modeling, animation and transfer. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9.

Cristinacce, D., Cootes, T.F., et al., 2006. Feature detection and tracking with constrained local models.. In: Bmvc. vol. 1, (2), Citeseer, p. 3.

Dai, H., Pears, N., Smith, W., Duncan, C., 2020. Statistical modeling of craniofacial shape and texture. Int. J. Comput. Vis. 128 (2), 547–571.

Edwards, G.J., Taylor, C.J., Cootes, T.F., 1998. Interpreting face images using active appearance models. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, pp. 300–305.

Fan, X., Jia, Q., Huyan, K., Gu, X., Luo, Z., 2016. 3D facial landmark localization using texture regression via conformal mapping. Pattern Recognit. Lett. 83, 395–402.

Ferman, D., Garrido, P., Bharaj, G., 2024. FaceLift: Semi-supervised 3D facial landmark localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1781–1791.

Gibelli, D., Dolci, C., Cappella, A., Sforza, C., 2020. Reliability of optical devices for three-dimensional facial anatomy description: A systematic review and meta-analysis. Int. J. Oral Maxillofac. Surg. 49 (8), 1092–1106.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2020a. Deep learning for 3d point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 43 (12), 4338–4364.

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

Hallgrímsson, B., Aponte, J.D., Katz, D.C., Bannister, J.J., Riccardi, S.L., Mahasuwan, N., McInnes, B.L., Ferrara, T.M., Lipman, D.M., Neves, A.B., et al., 2020. Automated syndrome diagnosis by three-dimensional facial imaging. Genet. Med. 22 (10), 1682–1693.

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

Kakadiaris, I.A., Toderici, G., Evangelopoulos, G., Passalis, G., Chu, D., Zhao, X., Shah, S.K., Theoharis, T., 2017. 3D-2D face recognition with pose and illumination normalization. Comput. Vis. Image Underst. 154, 137–151.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lee, J.D., Nguyen, O., Lin, Y.-C., Luu, D., Kim, S., Amini, A., Lee, S.J., 2022. Facial scanners in dentistry: an overview. Prosthesis 4 (4), 664–678.

Li, G., Xiong, C., Thabet, A., Ghanem, B., 2020. Deepergcn: All you need to train deeper gcns. arXiv preprint arXiv:2006.07739.

Martyniuk, T., Kupyn, O., Kurlyak, Y., Krashenyi, I., Matas, J., Sharmanska, V., 2022. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20942–20952.

Patania, S., Boccignone, G., Buršić, S., D'Amelio, A., Lanzarotti, R., 2022. Deep graph neural network for video-based facial pain expression assessment. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. pp. 585–591.

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

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T., 2009. A 3D face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. Ieee, pp. 296–301.

Piao, J., Qian, C., Li, H., 2019. Semi-supervised monocular 3D face reconstruction with end-to-end shape-preserved domain transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9398–9407.

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

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. 30.

Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L., 2012. Static and dynamic 3D facial expression recognition: A comprehensive survey. Image Vis. Comput. 30 (10), 683–697.

Sanyal, S., Bolkart, T., Feng, H., Black, M.J., 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7763–7772.

Segundo, M.P.P., Silva, L., Bellon, O.R.P., Queirolo, C.C., 2010. Automatic face segmentation and facial landmark detection in range images. IEEE Trans. Syst. Man Cybern. B 40 (5), 1319–1330.

Sforza, C., de Menezes, M., Ferrario, V.F., 2013. Soft-and hard-tissue facial anthropometry in three dimensions: what's new. J. Anthropol. Sci. 91, 159–184.

Tang, Z., Peng, X., Li, K., Metaxas, D.N., 2019. Towards efficient u-nets: A coupled and quantized approach. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 2038–2050.

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

Thurzo, A., Strunga, M., Havlínová, R., Reháková, K., Urban, R., Surovková, J., Kurilová, V., 2022. Smartphone-based facial scanning as a viable tool for facially driven orthodontics? Sensors 22 (20), 7752.

Valle, R., Buenaposada, J.M., Valdés, A., Baumela, L., 2019. Face alignment using a 3D deeply-initialized ensemble of regression trees. Comput. Vis. Image Underst. 189, 102846.

Vu, N.H., Trieu, N.M., Anh Tuan, H.N., Khoa, T.D., Thinh, N.T., 2022. Facial anthropometric, landmark extraction, and nasal reconstruction technology. Appl. Sci. 12 (19), 9548.

Wang, X., Bo, L., Fuxin, L., 2019a. Adaptive wing loss for robust face alignment via heatmap regression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6971–6981.

Wang, Y., Cao, M., Fan, Z., Peng, S., 2022. Learning to detect 3D facial landmarks via heatmap regression with graph convolutional network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, (3), pp. 2595–2603.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019b. Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. 38 (5), 1–12.

Wood, E., Baltrušaitis, T., Hewitt, C., Johnson, M., Shen, J., Milosavljević, N., Wilde, D., Garbin, S., Sharp, T., Stojiljković, I., et al., 2022. 3D face reconstruction with dense landmarks. In: European Conference on Computer Vision. Springer, pp. 160–177.

Wu, Y., Ji, Q., 2019. Facial landmark detection: A literature survey. Int. J. Comput. Vis. 127 (2), 115–142.

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

Wu, C.-Y., Xu, Q., Neumann, U., 2021. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In: 2021 International Conference on 3D Vision (3DV). IEEE, pp. 453–463.

Xu, M., Ding, R., Zhao, H., Qi, X., 2021. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3173–3182.

Yang, T.-Y., Chen, Y.-T., Lin, Y.-Y., Chuang, Y.-Y., 2019. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1087–1096.

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

Yuksel, C., 2015. Sample elimination for generating poisson disk sample sets. In: Computer Graphics Forum. vol. 34, (2), Wiley Online Library, pp. 25–32.

Zeng, L., Chen, L., Bao, W., Li, Z., Xu, Y., Yuan, J., Kalantari, N.K., 2023. 3D-aware facial landmark detection via multi-view consistent training on synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12747–12758.

Zhang, H., Dai, T., Tai, Y.-W., Tang, C.-K., 2022. Flnerf: 3d facial landmarks estimation in neural radiance fields. arXiv preprint arXiv:2211.11202.

Zhang, J., Gao, K., Fu, K., Cheng, P., 2020. Deep 3D facial landmark localization on position maps. Neurocomputing 406, 89–98.

Zhou, S., Xiao, S., 2018. 3D face recognition: a survey. Human-centric Comput. Inf. Sci. 8 (1), 35.

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