Three-dimensional Facial Landmark Detection in 3D Photos

Master's Thesis in Data Science

LUCA CAROTENUTO
s1047465

February 4, 2022

Daily Affiliated Supervisor and Second Reader:
dr. Guido de Jong*

Internal Supervisor:
prof. dr. Tom Heskes**

*
3D Lab, Radboudumc, Nijmegen, The Netherlands
Deparment of Oral and Maxillofacial Surgery, Radboudumc, Nijmegen, the Netherlands

**
Department of Data Science, Radboud University, Nijmegen, The Netherlands

Radboudumc

Radboud University

Abstract

Three-dimensional (3D) landmarks are used in various fields within medicine. Oral and maxillofacial surgery involves reconstructive operations on the head, face and jaw as well as facial cosmetic surgery. Landmarks are being during the planning, follow-up and diagnostics of surgical interventions. However, placing 3D landmarks manually can be tedious and inconsistent. AI-assisted landmark detection can help to automatize this process by making use of recent developments in the field of 3D deep learning. This work leverages DiffusionNet for feature extraction of 3D triangle meshes and point clouds. DiffusionNet is a representation independent network structure based on heat diffusion. This work presents a point-wise regression method that predicts heatmaps around landmark points. First results already show promising localization accuracy. As DiffusionNet suffers from time-consuming preprocessing, CUDA-acceleration is added to enable real-time inference scenarios. A direct coordinate regression shows lower localization accuracy and only performs better than point-wise regression for high error thresholds.

1 Introduction

Three-dimensional (3D) landmarks find application in various fields within medicine, such as cephalometry, the study and measurement of the head. Jaw surgery, also known as orthognatic surgery, deals with correcting irregularities of the jaw bones. Orthognatic surgery and orthodontics often aims at creating facial balance or harmony [1]. Landmarks help the surgeon during the diagnosis, planning and documentation of surgical interventions. Moreover, they are used to retrospectively judge whether the interventions have been successful.

Conventionally, landmarks are placed by the surgeon manually. This process is tedious and introduces human error. It can suffer from a high variability caused by the way the same surgeon places landmarks due to human error (intraobserver variability) and from how different surgeons place landmarks due to different landmarking habits (interobserver variability). AI-assisted landmarking can automate the landmarking procedure by making use of recent advances in the field of deep learning.

In machine learning, a distinction is made between object localization and object detection. The former only locates the presence of an object, whereas the latter also assigns a class label to the object. In this work, we tackle a landmark detection problem to allow for a distinction between landmark types.

Different modalities can be used for landmarking, such as 3D photos, textured 3D photos, bony-tissue CT-scans and soft-tissue CT-scans. 3D photographs only capture soft tissue. Nonetheless, [1] showed a high reproducibility for most soft tissue landmarks, suggesting that no hard tissue data, i.e. bony structure, is needed to perform accurate soft tissue analysis. Acquiring hard tissue data requires radiation-based capturing devices such as X-ray or CT whereas soft tissue images can be recorded by 3D stereophotogrammetry (see Figure 1). Stereophotogrammetry is a three-dimensional registration method for quantifying facial morphology and detecting changes in facial morphology during growth and development [2]. The images come from the Headspace dataset [3], a public dataset that comprises 3D images of the human head for 1519 subjects. The majority of the images come with landmark annotations. However, the annotations from the Headspace dataset have been localized in an automatic manner. Specifically, they were determined by the Zhu-Ramanan mixture-of-trees algorithm [4] applied on the texture of each mesh. Dai et al. project the 2D points to 3D using the texture coordinates which adds inaccuracies to the resulting landmarks. Thus, we manually annotate 3D landmarks for around 400 3D photos to ensure accurate testing scores and to improve training by using more accurate annotations.

Deep learning is becoming an increasingly powerful tool for data processing in computer vision tasks. Especially 2D computer vision tasks can be solved with high accuracy. Convolutional neural networks (CNNs) have delivered excellent results in computer vision tasks such as classification, object detection and segmentation. However, 3D deep learning faces several challenges. 3D data can be represented in different formats, including depth images, multi-view images, volumetric grids, point clouds or meshes. Which data representation should be used depends heavily on the application and on the data acquisition device. Point clouds and meshes do not suffer from discretization or projection loss and are therefore the preferred method for surface-based learning [5]. Point clouds and meshes are intrinsically non-euclidean data representations. Due to the irregular distribution of the points in space, conventional techniques such as convolution are not directly transferable to 3D. Moreover, there is a lack of big data sets. There exist several big data sets for 2D facial landmarking, such as the Annotated Facial Landmarks in the Wild (AFLW) [6] collection that comprises 25,993 faces. However, even with bigger 3D data sets, training would remain difficult due to high computation costs and memory footprints.Despite these challenges, in recent years, the field of geometrical deep learning comes up with increasingly powerful techniques to tackle surface learning problems. In this work, we leverage DiffusionNet [7], a discretization agnostic network by Nicholas Sharp et. al, to extract features for the prediction of heatmaps around landmarks.

The paper starts with related works of facial landmarking and important methods for geometrical deep learning in Section 2. Then, the data preparation, pipeline and networks are described in Section 3. Section 4 presents quantitative results on the Headspace dataset and quantitative results on data from Radboudumc. In Section 5, limitations and future works are discussed. Section 6 finally concludes the paper.

2 Related Work

2.1 Facial landmarking

There is extensive reasearch about 2D facial landmarking. Many non-medical tasks such as person identification, expression transfer or emotion recognition require automatic landmarking as a necessary step [9]. Existing methods for facial landmark detection can be classified into two categories: generative and discriminative. Generative methods model the facial shape as a probabilistic distribution. This category includes part-based generative models such as ASM and holistic generative models such as AAM, that capture variations in the shape or texture by Principal Component Analysis (PCA), or Gauss- Newton Deformable Part Models (GN-DPM) [10]. Discriminative models take a different approach and directly look for relevant features which can be used to localize the landmarks given the input. Discriminative methods include Cascaded Regression models, but also neural networks. With the emergence of Convolutional Neural Networks (CNNs),

Table 1: **Landmarks that are considered in this work.** Definitions from [8] and [1]

Landmark	Abbreviation	Definition
Pogonion	pg	The most anterior midpoint of the chin
Nasion	n	The Point in the midline of both the nasal root and nasofrontal suture
Pronasale	prn	The most anterior midpoint of the nasal tip
Subnasale	sn	The midpoint on the nasolabial soft tissue contour between the columella crest and the upper lip
Alar curvature (right)	ac-r	The point located at the facial insertion of the alar base (right)
Alar curvature (left)	ac-l	The point located at the facial insertion of the alar base (left)
Exocanthion (right)	ex-r	The soft tissue point located at the outer commissure of the right eye fissure
Endocanthion (right)	en-r	The soft tissue point located at the inner commissure of the right eye fissure
Endocanthion (left)	en-l	The soft tissue point located at the inner commissure of the left eye fissure
Exocanthion (left)	ex-l	The soft tissue point located at the outer commissure of the left eye fissure
Cheilion (right)	ch-r	The point located at the right labial commissure
Cheilion (left)	ch-l	The point located at the left labial commissure

many traditional methods have been outperformed by neural networks. Most research on facial landmarking focuses on 2D. In this paper, we focus on 3D facial landmarking with deep neural networks. As there is little research about deep learning for the case of 3D facial landmark detection, we focus in this chapter on more general works on 3D deep learning.

2.2 3D Deep Learning

In the past years, point cloud understanding is receiving increasing attention from the research community, as practical applications such as autonomous driving and robotics emerge. Such applications require more information than flat images can provide to obtain a better sense of the environment. The 3D data is captured by cameras with depth sensor such as lidar or RGB-D cameras.

Most authors that develop novel network architectures only report their results for the more common 3D tasks classification, segmentation, object detection or shape correspondence. Since results for keypoint detection are less common, the performance results of the networks can only be regarded as a rough reference for their potential feature extraction.

Unlike for the euclidean case, there is no universal concept for convolutions in 3D. Different types of approaches have been developed to address this problem. unorderedness and irregularity in the data. Promising approaches are PointNet [11] that use point clouds and consider the permutation invariance of points in the input. Variations of PointNet are PointNet++ [12] that manage to improve classification and segmentation performance by modelling local regions through sampling and grouping or PointCNN [13] that take into account the correlation between points in the local regions. An approach that operates directly on triangular mesh data is MeshCNN [14], which applies convolutions on edges and the four edges of their incident triangles, and pooling is applied via an edge collapse operation.

general work: [15]

3 Methods

3.1 Data set

Figure 1: **Stereophotogrammetry.** The patient is being photographed from 5 different angles around the head. Subsequently, the mesh is created by combining the different views into a single 3D mesh. Photo from Headspace dataset[3].

The models are trained predominantly on the Headspace dataset [3], a set of 3D images of the human head that is available for university-based non-commercial research. The collection consists of 1519 subjects each wearing tight fitting latex caps. This is done to avoid holes in the mesh on the scalp of the patient. The photos are captured by a 5-camera setup around the head of the person (Fig. 1). The images have a high quality, consistent illumination, and are pose normalized. 1200 samples include annotations, but as they were automatically generated, the quality of the labels differs for each landmark type and for each sample. Nonetheless, due to the sheer number of annotated samples, the Headspace data is used as training data for the first network that extracts rough landmark locations. Manual inspection shows that the Zhu-Ramanan mixture of trees algorithm in combination with the subsequent 2D to 3D projection make repeating errors such as consistently placing the landmarks a little too low. In total, the headspace data set comes with 68 landmarks per image. However, many of them are non-surgical and ill-defined and can be discarded for the purpose of finding landmarks to assist oral and maxillofacial surgery. Out of the 68 landmarks we keep 12 medically relevant, anatomical landmarks (see table 2). For properly evaluating the results and for training the refinement network, we manually annotate those 12 landmark positions for around 350 Headspace samples. Although not placed by a specialist, these labels are considerably more accurate, as they are human-annotated as opposed to machine-annotated. Furthermore, we use 3D photos from Radboudumc to further validate our method. The in-house meshes have a lower quality and the illumination varies greatly. The 3D photos are not pose normalized and differ and differ substantially to the head positions in the Headspace data set. As the patients do not wear any latex-caps, holes and artefacts in the region of the hair is common. Also, most 3D photos are captured by a 2-camera-setup, namely one from the front-left and one from the front-right. This means that meshes reconstructed from such photos have big holes in the back of the cranium as they only capture the frontal view.

3.2 Pre-processing

The 3D meshes are stored as 'wavefront object' files (.obj file). This file format contains information for vertices, edges, faces, normal vectors and texture. Vertices are points in the Cartesian coordinate system defined by x, y and z. Meshes also contain surface data in the form of edges and faces that define the interconnectivity between vertices. Normals [16]. To simplify the problem, our network processes point clouds instead of meshes. The meshes of the Headspace dataset are already pose normalized.

The 68 landmarks in the Headspace data are given by a reference to the vertex index in the mesh. The manually annotated landmarks in 3DMedX[1][17] are saved as coordinates in Comma-separated values files. As the meshes are simplified before being fed into the network,the original landmark coordinates do not point to a vertex in the downsampled mesh anymore and have to be re-calculated. The corresponding landmark point for the downsampled mesh is re-calculated by picking the vertex with the smallest distance to the original coordinate. Currently, this is done in a naive way by iterating over each vertex in the mesh. This step can be sped up significantly by applying a more sophisticated algorithm. The ground truth not only consists of point landmarks as it is difficult to train neural networks on such sparse positive cases. Instead, we create point clusters (heatmaps) around the landmark point to create regions that the network can learn more easily. The point closest to the landmark has the highest activation (1.0), points in the 3mm neighborhood are assigned an activation of 0.75, in the 4.5mm neighborhood 0.5 and in the 6mm neighborhood 0.25. An alternative, perhaps more balanced solution would be to create a gaussian heatmap. The heatmap approach increases the proportion of points with an activation higher than zero and improves the class imbalance problem.

To build a landmark detector that can discriminate between landmark types and to allow for overlapping activation clusters (meaning one point can be part of the neighborhood of multiple landmarks), each landmark cluster is stored in its own channel..

However, storing the activation for each point is very memory-consuming. To compress the label files, only the vertex indices and activation for points with an activation higher than zero are explicitly stored. All other points are assumed to have an activation equal to zero. This reduces the necessary information to store from $\textit{total vertices} \times \textit{channels}$ for the sparse matrix representation to $\textit{verices in landmark regions} \times \textit{2} \times \textit{channels}$ where the factor 2 arises as the compressed representation not only needs to save the activation but also additional vertex index information.

3.3 DiffusionNet

We define a point set $X = \{X_i \in \mathbb{R}^F, \quad i = 1, 2, ..., N\}$ as the input of our model, where N defines the number of points in the point cloud, F the dimension, and x_i is the 3D coordinate of each point in the Cartesian reference system. Note, that even in a 3-dimensional reference system, F is not restricted to 3 as we can use other point-based features such as the color or the normal vector with respect to the surface.

Graph-based method in spatial-domain combined with a point-wise MLP. Spectral methods are used for accelerating for the computation of the diffusion operation. Most machine learning algorithms require a fixed input size. DiffusionNet is able to deal with a flexible input size, making sampling or simplification for the purpose of standardizing the number of vertices unnecessary.

1. 3DMedX is a software from Radboudumc that allows 3D reconstruction using DICOM files from (CB)CT-scans or MRI scans and offers tools for the evaluation of orthognathic surgery. It also supports the creation of custom workflows e.g. for registering landmarks.

3.4 Shape Variants

3D shapes can come in different variants that the network should be invariant to, such as different orientations or different discretization. Different camera setups or different pre-processing can lead to very different orientations of the head in the space. The network should give the same result regardless of how the head is rotated. The perhaps most straightforward approach is to perform data augmentation. While it can work to make the network more robust to the presented augmented variants, data augmentation does not scale well as it is not feasible to sample all variations. Additionally, including slightly varied samples in the training quickly increases training times. The preferred approach to deal with shape variants is to design a network that is inherently invariant to rotations. There is still ongoing research on how to most efficiently design such invariant networks. One way is to use input features such as the Heat Kernel Signature (HKS) that are invariant to isometric deformations, thus also to different poses. DiffusionNet deals with the problem by... adds robustness but not true invariance.

4 Results

5 Discussion

5.1 Limitations

Requires pre-computed operations. Processes point clouds instead of meshes. Not universally applicable: subjects should be able to be landmarked independently of variations in pose, expression, illumination, background, occlusion, and image quality.

6 Conclusion

Further work: focal loss?

Table 2: **Landmarks that are considered in this work.** Definitions from [8] and [1]

Landmark	Error in mm	
Pogonion	pg	$\pm$
Nasion	n	
Pronasale	prn	
Subnasale	sn	
Alar curvature (right)	ac-r	
Alar curvature (left)	ac-l	
Exocanthion (right)	ex-r	
Endocanthion (right)	en-r	
Endocanthion (left)	en-l	
Exocanthion (left)	ex-l	
Cheilion (right)	ch-r	
Cheilion (left)	ch-l	

References

[1] J. M. Plooij et al. "Evaluation of reproducibility and reliability of 3D soft tissue analysis using 3D stereophotogrammetry". In: *International Journal of Oral and Maxillofacial Surgery* 38 (3 2009). ISSN: 09015027. DOI: 10.1016/j.ijom.2008.12.009.

[2] F. Ras et al. "Quantification of facial morphology using stereophotogrammetry - Demonstration of a new concept". In: *Journal of Dentistry* 24 (5 1996). ISSN: 03005712. DOI: 10.1016/0300-5712(95)00081-X.

[3] Hang Dai et al. "Statistical Modeling of Craniofacial Shape and Texture". In: *International Journal of Computer Vision* 128.2 (Nov. 2019), pp. 547–571. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01260-7. URL: https://doi.org/10.1007/s11263-019-01260-7.

[4] Xiangxin Zhu and Deva Ramanan. "Face detection, pose estimation, and landmark localization in the wild". In: 2012. DOI: 10.1109/CVPR.2012.6248014.

[5] Yulan Guo et al. *Deep Learning for 3D Point Clouds: A Survey*. 2021. DOI: 10.1109/TPAMI.2020.3005434.

[6] Martin Köstinger et al. "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization". In: 2011. DOI: 10.1109/ICCVW.2011.6130513.

[7] Nicholas Sharp et al. *DiffusionNet: Discretization Agnostic Learning on Surfaces*. 2022. arXiv: 2012.00888 [cs.CV].

[8] Avinash S. Bidra et al. "The relationship of facial anatomic landmarks with midlines of the face and mouth". In: *Journal of Prosthetic Dentistry* 102 (2 2009). ISSN: 00223913. DOI: 10.1016/S0022-3913(09)60117-7.

[9] Romuald Perrot, Pascal Bourdon, and David Helbert. "Implementing Cascade of Regression-based Face Landmarking: an in-Depth Overview". In: *Image and Vision Computing* 102 (Oct. 2020), p. 103976. DOI: 10.1016/j.imavis.2020.103976. URL: https://hal.archives-ouvertes.fr/hal-02884592.

[10] Yongzhe Yan et al. "A survey of deep facial landmark detection". In: *RFIAP*. Paris, France, June 2018. URL: https://hal.archives-ouvertes.fr/hal-02892002.

[11] Charles R. Qi et al. "PointNet: Deep learning on point sets for 3D classification and segmentation". In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. 2017. ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.16. arXiv: 1612.00593.

[12] Charles R. Qi et al. "PointNet++: Deep hierarchical feature learning on point sets in a metric space". In: *Advances in Neural Information Processing Systems*. 2017. arXiv: 1706.02413.

[13] Yangyan Li et al. "PointCNN: Convolution on X-transformed points". In: *Advances in Neural Information Processing Systems*. 2018.

[14] Rana Hanocka et al. "MeshCNN". In: *ACM Transactions on Graphics* (2019). ISSN: 0730-0301. DOI: 10.1145/3306346.3322959.

[15] Saifullahi Aminu Bello et al. *Review: Deep learning on 3D point clouds*. 2020. DOI: 10.3390/rs12111729. arXiv: 2001.06280.

[16] Guido de Jong. *Lecture notes: 3D Mesh Processing and Analysis, 3D Computer vision for Medical Applications*. Oct. 2021.

[17] 3D Lab Radboudumc Nijmegen. *3DMedX® (v1.2.23.0), The all-in-one solution for 3D research and the home of the OrthoGnathicAnalyser*. URL: https://www.3dmedx.nl (visited on 02/04/2022).

A Explanation of important concepts

Rigid deformation Deformation describes the transformation of an object from an initial to some final geometry. A rigid deformation, in contrast to non-rigid deformation, does not change the position and orientation of the object relative to the internal reference frame. Rotation around an axis is an example of a rigid operation that changes the configuration of the points relative to the external, but not to the internal reference frame. Translation is another rigid transformation, in the sense that it only affects the external reference frame, as the points within the object are all moved along parallel paths to the axis.

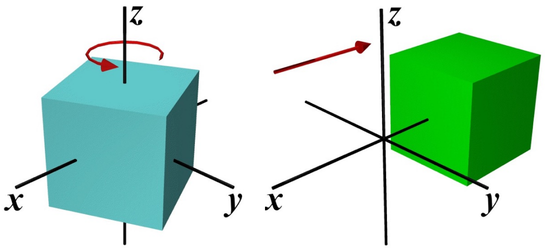

Figure 2: **Rigid transformations.** Rotation (left), translation (right). Images from [1].

Non-rigid deformation Non-rigid deformations can affect points within the object relative to both the internal and external reference frame. Distortion is an example of a non-rigid operation that changes the spacing of points within the object and consequently changes the overall shape of the object. Dilation or scaling is another non-rigid operation that changes the volume of the object, but differently to distortion, retains the same shape for the object.

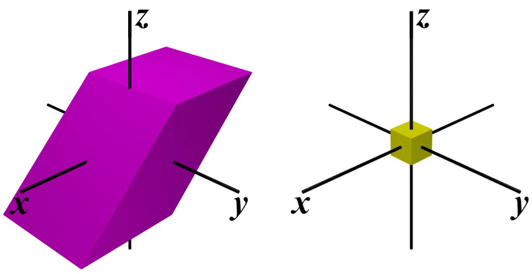

Figure 3: **Non-rigid transformations.** Distortion (left), dilation (right). Images from [1].

Isometric deformation/Isometry

Intrinsic/extrinsic deep learning Extrinsic deep learning methods treat the geometric data as Euclidean data. Voxel-based representation methods discretize the 3D data by defining a voxel as the smallest unit in the 3D space. Then, the object can be divided into a 3D grid. A typical network architecture that falls into this category is 3DSN. However, most voxel-based methods suffer from a large memory consumption and long training times due to the added third dimension. Also, some information is lost due to the discretization. Other deep learning methods are based on multi-view representations, such as the multiview-based CNN MVCNN [2]. Similarly to how the human eye manages to perceive depth, the methods combine views from multiple angles and process them into a single 3D image. Multi-view CNNs render the images from many different views, apply a CNN to each of the 2D image and perform a view pooling operation to combine the features. However, multi-view based methods suffer from different illumination, object occlusion and information loss during the reconstruction of the objects from different views.

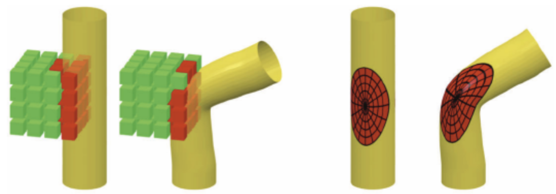

Figure 4: **Extrinsic (left) versus intrinsic (right) deep learning methods.** Image from [3].

Shape descriptor A shape descriptor characterizes the local geometriy of the surface. Examples for shape descriptors are the Gaussian curvature $K(x) = \kappa_1(x)\kappa_2(x)$ and the mean curvature $H(x) = \kappa_1(x)+\kappa_2(x)$. Good shape descriptors are robust to noise in the triangulation and against small deformations. They should also be invariant under rigid transformation and other isometries. [4]

Heat Kernel Signature The Heat Kernel Signature (HKS) is a popular shape descriptor that is derived from the Laplacian.

For a fixed time t, it is defined as

$$HKS(x) = k_t(x,x) = \sum_i e^{-\lambda_i t}\phi_i(x)^2 \tag{1}$$

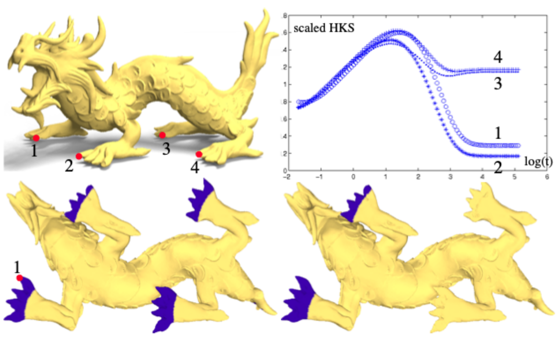

Figure 5: **Heat Kernel Signature.** At four different points on a triangulated surface. For small values of t ($t < 1$), the heat kernel signature $k_t(x, x)$ is almost the same. This is, because the local geometry, i.e. the tips of the dragon's feet is approximately equivalent. As t increases ($t > 1$), more global information is considered and the heat kernel signatures diverge, whereas the two points at the front feet (1, 2) and the two points at the back feet (3, 4) still capture more common surface information about the shape compared to one at the front and one at the back. Image from [5].

The Wave Kernel Signature (WKS) is a shape descriptor similar to HKS but is based on the Schrödinger wave equation. HKS and WKS both have the advantage of isometry-invariance and being easy to compute.

B Background for choice of the network

The project started with exploring different networks that can tackle the problem of 3D landmark detection. This phase also lead to insights regarding networks that do not work well for the problem at hand. PointNet is one of the earliest and simpler model architectures that operates on point clouds was a straightforward choice We tried the Pytorch implementation of the extension of PointNet, called PointNet++. The extensio in MeshCNN and Pointnet; many network architectures don't scale well

References

[1] San Diego State University. , *Visualizing Strain*. URL: http://www.sci.sdsu.edu/visualstructure/vss/htm_hlp/def_d.htm (visited on 02/04/2022).

[2] Hang Su et al. "Multi-view Convolutional Neural Networks for 3D Shape Recognition". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 945–953. DOI: 10.1109/ICCV.2015.114.

[3] Wenming Cao et al. "A Comprehensive Survey on Geometric Deep Learning". In: *IEEE Access* 8 (2020). ISSN: 21693536. DOI: 10.1109/ACCESS.2020.2975067.

[4] Stanford University Justin Solomon. *CS 468 Lecture 16:, Isometry Invariance and Spectral Techniques*. URL: https://graphics.stanford.edu/courses/cs468-13-spring/assets/lecture16-gawlik.pdf (visited on 02/04/2022).

[5] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. "A concise and provably informative multi-scale signature based on heat diffusion". In: vol. 28. 2009.