Cross-atlas Convolution for Parameterization Invariant Learning on Textured Mesh Surface

Shiwei Li[1] Zixin Luo[1] Mingmin Zhen[1] Yao Yao[1]*
Tianwei Shen[1] Tian Fang[2]† Long Quan[1]
{slibc|zluoag|mzhen|yyaoag|tshenaa|quan}@cse.ust.hk fangtian@altizure.com
[1]The Hong Kong University of Science and Technology
[2]Shenzhen Zhuke Innovation Technology (Altizure)

Abstract

We present a convolutional network architecture for direct feature learning on mesh surfaces through their atlases of texture maps. The texture map encodes the parameterization from 3D to 2D domain, rendering not only RGB values but also rasterized geometric features if necessary. Since the parameterization of texture map is not pre-determined, and depends on the surface topologies, we therefore introduce a novel cross-atlas convolution to recover the original mesh geodesic neighborhood, so as to achieve the invariance property to arbitrary parameterization. The proposed module is integrated into classification and segmentation architectures, which takes the input texture map of a mesh, and infers the output predictions. Our method not only shows competitive performances on classification and segmentation public benchmarks, but also paves the way for the broad mesh surfaces learning.

1. Introduction

The 3D mesh is one of the most popular representation for 3D shape, which consists of an array of vertices and an array of face indices indicating the surface geometry. It could be augmented with texture coordinates and texture maps to render the color appearance of the mesh.

Feature learning on textured meshes is challenging: on the geometry side, the arrays of vertices and faces are permutable. The mesh geometry by its design, could be very adaptive: a similar shape can be meshed in very different patterns, with irregular vertex densities and inconsistent triangle qualities. On the texture side, the parameterization of texture map could be arbitrary and still renders the same appearance, as long as the texture coordinate is in accordance with the content in texture map.

Existing methods that can directly or indirectly operate on textured meshes include 1) multi-view projection [37, 11, 10], which renders the RGB(D) images from all perspectives of the mesh. The multi-view images can be consumed via 2D convolutional neural networks (CNNs) and the global feature is aggregated by view pooling. However, this is only feasible for small objects [43] where occlusion is not significant. When it comes to larger scenes where occlusion and view selection are non-trivial, their performance would degrade . 2) Volumetric models can be easily obtained from meshes, which could be consumed by 3D CNNs [43, 5, 25]. As mentioned in [6], the volumetric representation is memory-consuming and loses rich valuable image details, thus hindering the performance. 3) Point-based learning is a recent breakthrough [28, 30]. The colored point cloud can be sampled from the mesh. As the point cloud learning employs multi-layer perceptrons rather than convolutions, it takes significantly more parameters and thus the maximum number of points can be learned is far less than image pixels under the same configuration. 4) Geometric deep learning [3] techniques can directly process on meshes via spectral analyses [4, 8, 16, 13, 46] or geodesic convolutions [24, 2, 26]. While their methods focus more on learning the intrinsic geometries for the dynamic correspondence task [1]. It's unclear how they can combine the texture or image for semantic feature learning.

In fact, the textured mesh is a self-contained combination of 2D texture and 3D shape. Some recent works are aware of this importance and perform joint learning on 2D images and 3D shapes [29, 6, 36], achieving improved performances. However, processing on multi-view images, rather than a single texture map, is redundant as the same 3D area is processed by multiple times. Besides, such bundled data-structure of 3D shape, images and camera poses is difficult to acquire. With the popularity of 3D sensing techniques, it would be easier and more common to obtain the standalone textured mesh model.

To this end, we propose to perform direct learning on the textured mesh via its texture map. The texture map can in-

[*] Intern at Shenzhen Zhuke Innovation Technology (Altizure)
[†] Corresponding author

clude not only the color information as it usually does, but also arbitrary geometric features as long as they are rasterized in the map. Each pixel in texture map becomes a generic feature vector, encoding both color and geometric information. More importantly, the texture map is already in 2D domain, enjoying numerous benefits: 1) it can be learned via the standard CNNs, which can leverage the efficient designs from rich previous researches. 2) The texture map rasterization is analogous to point sampling, and thus invariant to the irregular meshing. 3) The hierarchy of a 2D map is clearly defined as the image pyramid, making multi-scale learning (*e.g.*, for global feature extraction) easily achievable. 4) The texture map inherits the geodesic neighborhood of meshes, whereas other (volumetric or point-based) representations discard the geodesic neighborhood information.

The practice of learning 3D meshes via 2D domain is seen in previous methods [35, 23], but their performances suffer from two crucial problems. The first one is distortion: unlike the generic texture map parameterization that segments the mesh to multiple atlases and pack them tightly in the texture map [17, 48], their parameterization performs only one cut and unfolds the mesh to a complete 2D squared map. This inevitably introduces the distortion, which could be unpredictably large [35]. The second problem is their networks are not invariant to parameterization, which is further determined by the cut on the mesh surface. To get around, the author suggests to try multiple cuts on testing stage [23], and select the most responsive one.

In this paper, we address two above problems. First, for the distortion problem, we do not unfold the 3D surface onto a complete 2D map. Instead, we segment the mesh to multiple atlases, project them to 2D domain and pack them tightly inside the texture map. By doing so, each atlas finds its best projection to minimize the distortion. Second, regarding the parameterization, the atlas composition in texture map is unpredictable, depending on the packing algorithm. Each atlas is isolated, and thus the neighborhood is taken apart when crossing the texture seams on mesh surface. To tackle these issues, we introduce the *cross-atlas convolution* – we redirect the pixel positions at atlas boundaries to the actual neighborhood on mesh surface, which is done with a precomputed offset map. We have integrated the cross-atlas convolution into the classification and segmentation networks, and verified the effectiveness on public benchmarks. Overall, our contributions are threefold:

1. We unlock a novel approach to mesh learning directly through their atlases of texture maps.
2. Our method addresses the distortion and the variance of parameterization problems in previous related methods [35, 23]. We also impose no restriction to the input mesh whereas they requires genus-0 meshes.
3. Our designed model is flexible and introduces no extra parameters: it can be trained on natural images, and tested on texture maps using our cross-atlas modules (see also Section 5.3).

2. Related works

Deep learning on non-Euclidean geometric 3D data is an active and ongoing research topic. The mesh is one of the most commonly used representations in 3D vision and graphics, yet the irregularity of mesh makes it challenging to learn. We survey existing approaches in three categories.

Converting to regular structures The most straightforward approach is to convert the irregular mesh to regular data structure suitable for CNN processing. This can be done by projecting the mesh to multi-view 2D images, and then applying 2D CNN and view pooling to aggregate the global feature [37, 11, 10]. These methods show great performance on small object classification such as ModelNet [43], but may degrade when it comes to self-occluded objects or large scenes where view selection is non-trivial. Another branch of methods convert the mesh to volumetric domains, and extract deep features from volumes by 3D convolutions [43, 5, 25]. The voxelization may introduce discretion errors and is highly memory-consuming. Using Octrees [31, 40] can alleviate the resolution problem to some extent. Recently, some approaches combine both 2D views and 3D volumes and achieve even better results [29, 6].

Point cloud approaches Point cloud can be easily obtained by sampling on meshes. The irregular point cloud data can be learned by PointNet [28], and its extended hierarchical version [30]. They use combinations of multi-layer perceptions and pooling operations to achieve permutation invariance on the point set. Their insights also inspire several following works on point cloud learning in terms of improving the scalability and enhancing the local information of point cloud structure [20, 36, 14, 41]. In these methods, the neighborhood of a point is found by the radius-search or K nearest neighbors, which does not keep the original geodesic neighborhood on meshes. This could be a potential problem when the geodesic distance and Euclidean distance vary significantly in the mesh model.

Geometric deep learning on meshes The mesh can be consumed by geometric deep learning techniques [3] for the non-rigid shape correspondence task. 1) If the mesh is seen as a graph, several works have proposed to apply the spectral analysis on the eigen decomposition of the Laplacian of mesh graph [4, 8] to establish dense correspondences between deformable shapes. A general limitation is the cross-domain generalization issue, which is later addressed by spectral transformers [46] to some extent. Dirac operator is an alternative to Laplacian operator which yields better stability in some scenarios [16]. 2) If seen as a manifold surface, the mesh can be applied with geodesic convolutions [24, 2, 26] on local patches, enjoying better generalization across domains than spectral methods. These techniques show great performance on non-rigid shape correspondence task but the receptive field of a polar filter [24] is very small and thus it is unclear how to extract high-level features. Besides of using polar filter, a recent work [27] proposes to apply standard convolution on tangential projections of local patches and construct the hierarchy via mesh simplification. 3) The third class of techniques apply global parameterization to the mesh and flatten it to 2D images [35, 23]. Our work belongs to this class, while the other two works are the GeometryImage [35] and the "flat-torus" method [23]. These methods enjoy common benefits derived from CNNs, but both of them unfold the mesh to a complete squared map, which induces considerable distortions. Besides, their methods require genus-0 input, otherwise they would crudely fill all topological holes. Regarding the parameterization, the one in GeometryImage [35] is not seamless, while the network in "flat-torus" method [23] is not invariant to the parameterization, depending on the cut of three chosen points on the mesh. Our work addresses the distortion and parameterization-dependent issues and presents a practical approach to mesh learning.

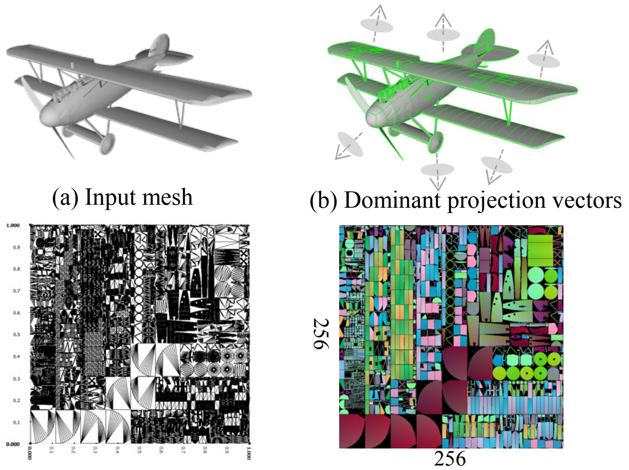
3. Mesh learning via texture maps

This section illustrates the detailed procedure of mesh learning in the texture map space. In Section 3.1, we describe how the input texture map is generated from a pure mesh or a textured mesh. In Section 3.2, we introduce the cross-atlas module to recover the connectivities of separated atlases. In Section 3.3, we present network architectures for classification and semantic segmentation tasks.

3.1. Generating texture maps

In the preprocessing step, the input is a triangular mesh $\mathcal{M} = \{V, T\}$, where $V = \{v_i\}$ and $T = \{t_i\}$ correspond to the vertices and triangles respectively. If it's a polygon mesh we simply triangulate the faces. The mesh can be with or without textures. The output includes a texture map $(H \times W \times C)$ and an offset map $(H \times W \times 2k^2)$ for the network input, where k is the kernel size for convolution.

Mesh without textures: if a pure mesh is given (*e.g.*, Fig. 1(a) in ModelNet [43]), we create the UV parameterization by segmenting the mesh to multiple atlases (Fig. 1(c)), and then rasterize the geometric features to a $H \times W \times C$ texture map Fig. 1(d). Our goal is similar to previous parameterization algorithms [17, 48] aiming at how to find the cuts for minimized atlas distortions. Specifically, we first find a minimum set of dominant projection directions $\mathbf{P} = \{P_i \in \mathbb{R}^3\}$ such that the angle between each triangle normal $n(t_i)$ and its best projection vector P_{t_i} is less than a threshold, *i.e.*, $\angle(\overrightarrow{n(t_i)}, \overrightarrow{P_{t_i}}) < \tau_{angle}$. By default $\tau_{angle} = 40°$. When the angle $\angle(\overrightarrow{n(t_i)}, \overrightarrow{P_{t_i}})$ approaches $0°$, the projection is exactly tangential and has minimized local distortion [27]. After that, we cluster the projected triangles to atlases via connected components, and pack them into one squared map using bin-packing [15]. With this UV parameterization, we can rasterize the geometric feature of the mesh to a $H \times W$ texture map $\mathbf{T}$. Note that the pixel in texture map $\mathbf{T}(x, y) = [f_1, f_2, ..., f_c]^T$ is a C-dimensional feature vector, instead of RGB values of the standard definition of "texture maps". The choice of geometric feature could be intrinsic features (*e.g.*, curvatures, heat kernel signatures) or extrinsic properties (*e.g.*, spatial coordinates, normals), or even concatenating multiple of them, depending on specific tasks. For invalid regions we simply fill zeros. The output texture map is of size $H \times W \times C$, where $H \times W$ is the spatial resolution and C the feature channel.


(a) Input mesh

(b) Dominant projection vectors

(c) UV parameterization by bin-packing

(d) Rasterized geometric features as texture map

Figure 1: For a pure mesh (a), we compute its dominant projections (b), and pack atlases in one UV map (c). The geometric feature (*e.g.*, vertex coordinate) is rasterized to a texture map (d).

Mesh with textures: if the mesh comes with textures, we still generate our own parameterization using above algorithm. The RGB color in original texture maps is an additional feature that can be concatenated to the geometric feature vector $\mathbf{T}(x, y)$. We do not use the original parameterization as we need to ensure two criteria. 1) The parameterization should be *area-preserving*, which means the areas of mesh triangles and their projected areas in 2D maps are proportional. It ensures the receptive field of 2D convolution over the texture map corresponds to equal geodesic area over the mesh surface (also mentioned in [35]). 2) It should be *rotation-aligned*: the rotation of atlas in the original texture map could be arbitrary, making the later convolution suffer from rotation ambiguities. In our generated texture map, the rotation of each atlas is aligned with the negative Z-axis (usually the gravity direction), such that the visual content of atlases is upright.

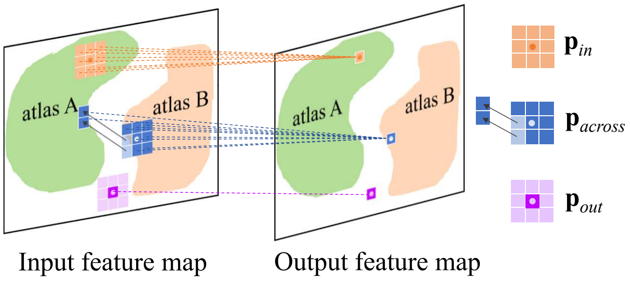


Figure 2: Illustration of three situations in cross-atlas convolution: $\mathbf{p}_{in}$ (standard convolution); $\mathbf{p}_{across}$ (cross-atlas convolution with offsets); $\mathbf{p}_{out}$(invalid pixel always = 0).

Offset maps: Unlike natural images, atlases in texture map are discontinuous and locate unpredictably. To bridge the atlases and apply convolution across them, we also generate the offset map of size ($H \times W \times 2k^2$, k is the kernel size), which encodes the neighborhood information between atlas boundaries. We will describe how to create it in the next section together with the cross-atlas convolution.

3.2. Cross-atlas convolution via kernel offsets

The standard 2D convolution computes the output feature map $\mathbf{F}_o$ via the weighted sum of a k size regular patch over every pixel $\mathbf{p} = (x, y)$ at the input feature map $\mathbf{F}_i$:

$$\mathbf{F}_o(\mathbf{p}) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{F}_i(\mathbf{p} + \mathbf{p}_n) \cdot g(\mathbf{p}_n), \quad (1)$$

where $g(\cdot)$ is the kernel weight and $\mathcal{R} = \{\mathbf{p}_n\} = \{(x, y) : -\frac{k-1}{2} \leq x, y \leq \frac{k-1}{2}\}$ enumerates neighboring locations of the center pixel and indicates the receptive field.

This equation holds when the $k \times k$ local neighborhood is straightforward in natural images. For texture maps where atlases are isolated, two neighboring surface points on the mesh can lie at two separated atlases. To recover the original mesh geodesic neighborhood, the 2D convolution should be able to apply across atlases.

To this end, when rasterizing the texture maps (Section 3.1), we encode the atlas connectivity information by generating the corresponding offset map $\mathcal{R}_{\mathit{offset}}$ with the same spatial resolution $H \times W$ and channel length $2k^2$, where each pixel $\mathbf{p}$ in the offset map has a $k \times k$ patch indicating the offsets of x-axis and y-axis:

$$\begin{aligned} \mathcal{R}_{\mathit{offset}}(\mathbf{p}) &= \{(\triangle x, \triangle y)\} & (2) \\ &= \{\triangle \mathbf{p}_n\}, \quad |\mathcal{R}_{\mathit{offset}}(\mathbf{p})| = k^2. & (3) \end{aligned}$$

The offset values $\mathcal{R}_{\mathit{offset}}(\mathbf{p}) = \{\triangle \mathbf{p}_n\}$ are augmented to the standard neighboring locations $\mathcal{R} = \{\mathbf{p}_n\}$ and redirect the pixel to another location which corresponds to the actual mesh geodesic neighborhood, *i.e.*, $\mathcal{R} + \mathcal{R}_{\mathit{offset}}(\mathbf{p}) = \{\mathbf{p} + \mathbf{p}_n + \triangle \mathbf{p}_n\}$ are the geodesic neighboring locations of $\mathbf{p}$. The original equation in Eq. 1 becomes

$$\mathbf{F}_o(\mathbf{p}) = \sum_{\substack{\mathbf{p}_n \in \mathcal{R} \\ \triangle \mathbf{p}_n \in \mathcal{R}_{\mathit{offset}}(\mathbf{p})}} \mathbf{F}_i(\mathbf{p} + \mathbf{p}_n + \triangle \mathbf{p}_n) \cdot g(\mathbf{p}_n). \quad (4)$$

The offset value can be fractional, and we use bilinear interpolation to sample the pixel. As illustrated in Fig. 2, we classify the texture map into three regions, denoted by pixels $\mathbf{p}_{in}, \mathbf{p}_{out}, \mathbf{p}_{across}$.

$\mathbf{p}_{in}$: When the convolution is applied over the inner-atlas region, the standard pixel neighborhood is corresponding to the mesh geodesic neighborhood, and thus no offset should be added: $\mathcal{R}_{\mathit{offset}}(\mathbf{p}_{in}) = \{(0, 0)\}$.

$\mathbf{p}_{out}$: When applying over the out-of-atlas region, this pixel value is invalid and should be kept isolated in order not to contaminate other pixels. We use offset values inverse to $\mathcal{R}$, *i.e.*, $\mathcal{R}_{\mathit{offset}}(\mathbf{p}_{out}) = -\mathcal{R}$ and $\mathbf{p}_{out} + \mathbf{p}_n + \triangle \mathbf{p}_n = \mathbf{p}_{out}$. This ensures $\mathbf{F}_o(\mathbf{p}_{out}) = \mathbf{F}_i(\mathbf{p}_{out}) = 0$ throughout the network.

$\mathbf{p}_{across}$: When the standard $\mathcal{R}$ is just across the border of an atlas, we add the precomputed offset values to its standard locations, so $\mathbf{p}_{across} + \mathbf{p}_n + \triangle \mathbf{p}_n$ should just locate at the true mesh geodesic neighborhood.

Therefore, only pixels $\mathbf{p}_{across}$ need to compute their offset values. The pixels $\mathbf{p}_{across}$ are determined by the kernel size: if we place a k-size filter kernel within the atlas at $\mathbf{p}$ and there are some pixels of this kernel locate out of the atlas, then $\mathbf{p} = \mathbf{p}_{across}$. For a 3×3 kernel as an example, we compute the offsets for 1-ring atlas boundary pixels.

To compute offset values for a center pixel $\mathbf{p}$ regarding its neighbor pixel $\mathbf{p} + \mathbf{p}_n$ lying out of atlas, we rasterize the vertex coordinates with resolution $H \times W$ to a map, where we can instantly query from the pixel $\mathbf{p}$ to the 3D point $\mathbf{X}$ on mesh surface. Then we use the Fast Marching [38] to search the geodesic neighbor point of $\mathbf{X}$ along $\overrightarrow{\mathbf{p}_n}$ direction, yielding the point $\mathbf{X}'$. The $\mathbf{X}'$ finds its corresponding texture coordinates $\mathbf{p}'$ on the texture map. Finally $\triangle \mathbf{p} = \mathbf{p}' - \mathbf{p} - \mathbf{p}_n$ is the offset value.

Note that unlike standard convolution on natural images can add paddings to image boundaries, the cross-atlas convolution has no “image boundaries” – if the mesh surface is water-tight, every pixel in texture map can find its neighborhood. If it is an open mesh and the pixel exactly corresponds to the mesh boundary that finds no neighborhood, it simply fills zero value (analogous to the zero padding effect).

Deconvolution In semantic segmentation tasks where the feature map is finally up-sampled to the input resolution, deconvolution (or transposed convolution) is a popular operation [21]. Essentially, deconvolution can be decomposed into 1) scattering pixels evenly from the sparse feature map to the dense feature map, and 2) applying convolution over this map. We can simply replace the convolution with the cross-atlas version when up-sampling texture maps.

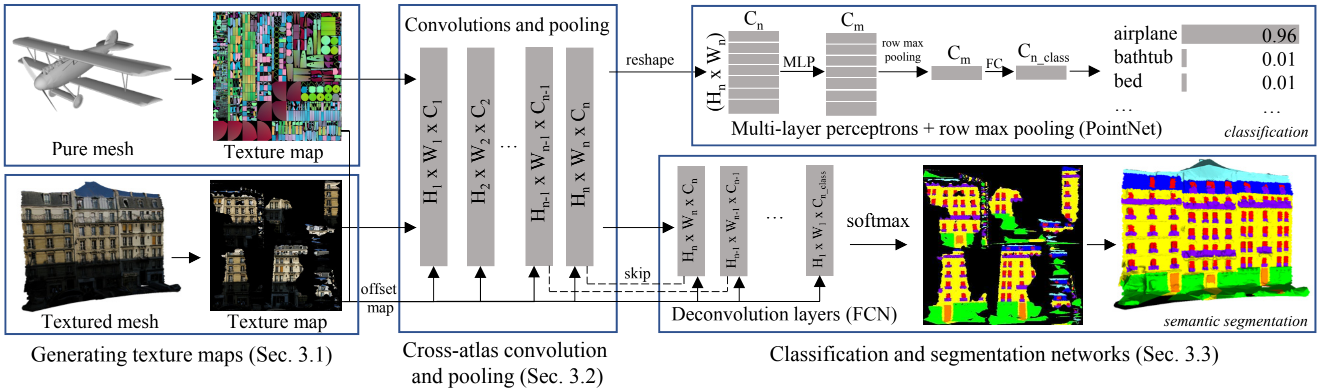

Pure mesh
Texture map
Textured mesh
Texture map
Generating texture maps (Sec. 3.1)
Convolutions and pooling
$H_1 \times W_1 \times C_1$
$H_2 \times W_2 \times C_2$
…
$H_{n-1} \times W_{n-1} \times C_{n-1}$
$H_n \times W_n \times C_n$
offset map
Cross-atlas convolution and pooling (Sec. 3.2)
reshape
$(H_n \times W_n)$
C_n
MLP
C_m
row max pooling
C_m
FC
C_{n_class}
airplane 0.96
bathtub 0.01
bed 0.01
… …
Multi-layer perceptrons + row max pooling (PointNet)
classification
skip
$H_n \times W_n \times C_n$
$H_{n-1} \times W_{n-1} \times C_{n-1}$
…
$H_1 \times W_1 \times C_{n\_class}$
softmax
Deconvolution layers (FCN)
semantic segmentation
Classification and segmentation networks (Sec. 3.3)

Figure 3: The classification and segmentation network architectures of our method. We use 4∼19 convolutional/pooling layers, 5∼7 MLP layers, 2∼3 FC layers and 4∼8 deconvolution layers. The concrete number of layers varies in specific tasks.

Pooling The standard pooling operation is often used to reduce dimensions of feature maps:

$$\mathbf{F}_o(\mathbf{p}_o) = pool_{k\times k}(\mathbf{F}_i(\mathbf{p}), \mathbf{F}_i(\mathbf{p}+\mathbf{p}_n), ...), \forall \mathbf{p}_n \in \mathcal{R} \quad (5)$$

where $\mathbf{p}_o$ is the location at the lower spatial resolution output feature map. The $pool_{k\times k}$ can be *max*, *average* or *sum* operation.

In cross-atlas pooling, the behavior is similar as cross-atlas convolution by replacing the standard image pixel neighborhood with the mesh geodesic neighborhood:

$$\begin{aligned}\mathbf{F}_o(\mathbf{p}_o) =& pool_{k\times k}(\mathbf{F}_i(\mathbf{p}), \mathbf{F}_i(\mathbf{p}+\mathbf{p}_n+\triangle\mathbf{p}_n), ...), \\ & \forall \mathbf{p}_n \in \mathcal{R}, \triangle\mathbf{p}_n \in \mathcal{R}_{offset}(\mathbf{p}). \qquad (6)\end{aligned}$$

Hierarchies For a specific $H \times W$ dimension feature map and the kernel size k, the corresponding offset map is unique. In the CNN pipeline, the spatial dimension of the feature map may change by in-network upsampling and downsampling. Therefore, the corresponding offset maps for all possible feature map dimensions need to be computed beforehand. Generally, we compute the hierarchy of offset maps by rasterizing a pyramid of resolutions, where each lower level is half in width and height of the upper level. The offset map with larger kernel size can be reused in the operation with smaller kernel. *e.g.*, we can take the central 3×3 from 5×5 offset locations. Note that in lower spatial dimensions, some small atlases might disappear as the rendering area is smaller than one pixel, but their feature information does not lose – it is absorbed by pixels in other atlases via cross-atlas convolution.

There are some similarities between our design and the deformable convolution [7] as they both leverage offset values to the convolution. However, our offset map is precomputed in order to recover the geodesic neighborhood, while offset values in [7] are trainable for a more flexible receptive field in objection detection problem.

3.3. Network architecture

We have integrated the cross-atlas convolution into classification and semantic segmentation architectures. The comprehensive architecture is illustrated in Fig. 3.

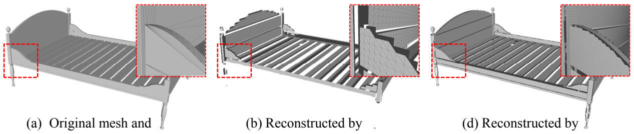
Classification A generic classification network extracts the high-level feature from the input. The common practice is to apply convolutions or pooling to reduce the spatial dimension and increase the channel dimension, and then flatten to a 1D global feature vector, followed by fully connected (FC) layers and softmax, yielding the class label.

To apply classification on the (textured) mesh, we first convert it to the $H_1 \times W_1 \times C_1$ input feature map (Section 3.1) and its corresponding offset map. Then, we replace the standard convolution and pooling with our cross-atlas version. After bypassing n layers of cross-atlas convolutions, it obtains a feature map with dimensions $H_n \times W_n \times C_n$. At this moment, we cannot simply flatten it to a 1D feature vector like the standard network does, because the spatial locations of pixels in this feature map are permutable when atlases are packed in a different way (*e.g.*, swap the atlas A and B in Fig. 2). Inspired by PointNet [28], we regard each valid pixel in the $H_n \times W_n \times C_n$ feature map is a permutable "point". We reshape it to $(H_n \times W_n) \times C_n \times 1$, *i.e.*, each C_n-channel pixel is expanded to one row. Then we apply multi-layer perceptrons (MLPs) and row-wise max pooling to obtain a m-channel 1D feature vector. Finally fully connected layers and softmax are applied and outputs the n_{class}-channel 1D feature vector indicating the probabilities of classes. We use ReLU as our activation function. The specific numbers of layers and the feature map dimension vary according to the task complexity.

Segmentation The segmentation is a pixel-wise classification task. Its former part is similar to classification which yields the $H_n \times W_n \times C_n$ feature map via several layers of convolution and pooling. Unlike classification which extracts the global feature in the later modules, it up-samples the feature map to an original resolution annotation map $H_1 \times W_1 \times C_{n_{class}}$. Here, we leverage the deconvolution in FCN [21] as our up-sampling layers, and replace their convolution with our cross-atlas version. We add three skip connections between convolution and deconvolution layers.

4. Understanding the cross-atlas convolution

In this section, we discuss three important properties of the proposed method. 1) It is robust to irregular meshes. 2) It is invariant to texture parameterization. 3) The receptive field follows the mesh geodesic distance.



(a) Original mesh and the meshing structure

(b) Reconstructed by the 256 x 256 texture map

(d) Reconstructed by the 1024 x 1024 texture map

Figure 4: The original mesh (a) and reconstructed meshes from texture maps (b)(c). Our method is invariant to irregular mesh input: applying convolution over the texture map is analogous to applying over the vertices of reconstructed regular meshes.

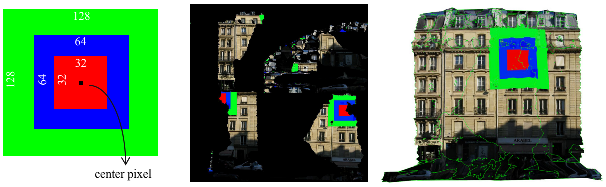
Robust to irregular meshes By design, a 3D shape can be meshed with very different vertex connectivities, but it is undesirable that different meshing would cause inconsistent semantic predictions. In our approach, the texture map rasterization in the first step can be deemed as "remeshing" – if we connect every 4-neighborhood of vertices corresponding to texture map pixels, it reconstructs a regular mesh as shown in Fig. 4. Therefore, applying convolution over the texture map is analogous to applying over the vertices of these regular meshes. It is straightforward that using a higher resolution of texture map retains more geometric details (*e.g.*, bed legs in Fig. 4(c)(d)).

Invariant to parameterization When generating the texture map (Section 3.1), the parameterization is unpredictable. Essentially, it has two uncertainties: 1) *how atlases are cut* and 2) *how atlases are packed*. Here we explain why our method is invariant to these two uncertainties.

Atlas cutting: recall that in Section 3.1, we set an angle threshold τ_{angle} to tradeoff the amount of atlases and their distortions – setting τ_{angle} too small may lead to many fragmentary atlases, and vice versa. Despite how they are cut, their neighborhood information is encoded in the offset map, informing the network to convolve across discontinuous atlases. With a proper value of τ_{angle}, we can assume the distortion is relatively small. Although such a small distortion of atlas still has an unpredictable variance, it is analogous to the practice of augmenting training data where a random mild transform is applied to training samples, which helps prevent from over-fitting.

Atlas packing: the only uncertainty in atlas packing is their spatial locations (*i.e.*, translations), as we have already aligned the rotation and preserve the scale of atlas area. Our network architectures are invariant to the translation of atlas. Imagine that we exchange locations between atlas A and B in Fig. 2: regarding the segmentation task, the network is fully convolutional, which is translation equivalent. Therefore, altering locations of atlas shall yield consistent predictions if the correct offset map is given. As for classification task, the global feature is extracted by PointNet [28], which is by design invariant to set permutation. It extracts identical features regardless of the pixel spatial location variance in the last convolutional feature map ($H_n \times W_n \times C_n$).

Note that the rotation alignment of Z-axis only disambiguates X- and Y- axises rotational varieties. The in-plane rotation perpendicular to Z-axis is still ambiguous, which is a common problem in many previous works [25, 35, 28]. Likewise, we alleviate it by augmenting the training data with randomly rotations along Z-axis. Some related works use polar convolution [24, 2, 26] or angular pooling [27, 42] to achieve fully rotational invariance, as their task is more sensitive to geometries. We do not use them in our current framework and leave it as a potential improvement.



(a) The illustration of three range receptive fields

(b) The receptive field on 512 x 512 texture map

(c) The corresponding receptive field on the textured mesh

Figure 5: Illustration of the receptive field. If we dilate the receptive field from a center pixel on the texture map and redirect when reaching to atlas boundaries, the field would be separated in severel atlases (b). Its corresponding field on textured mesh is approximately the geodesic field.

Receptive fields Although the convolution is applied over 2D texture maps, its receptive field follows the mesh geodesic distance. This is done by using the offset map to redirect pixel locations. Fig. 5 illustrates this behavior: we color-code the receptive fields of 32 × 32, 64 × 64, 128 × 128 of a center pixel in red, blue and green respectively. As we dilate a pixel to a block-wise field in the texture map (redirecting to the offset location when reaching to atlas boundaries), the field is actually separated in several atlases (Fig. 5(b)). On the contrary, its corresponding textured mesh shows the block-wise geodesic receptive field over its surface (Fig. 5(c)). Note that the receptive field on the mesh is not strictly following the geodesic distance due to the distortion in atlas projections, but it is a good approximation given the distortion is constrained by τ_{angle} in Section 3.1.

5. Experiments

We implement the texture map generation (Section 3.1) in C++, and the network (Section 3.3) using Tensorflow. Our method is evaluated on MeshMNIST [16], ModelNet [43] and Ruemonge2014 [32].

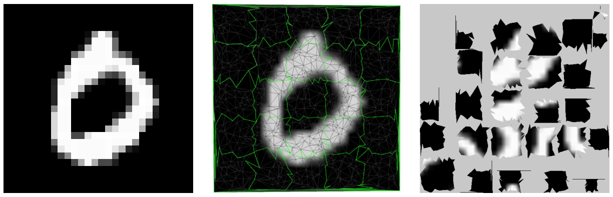
The network for MeshMNIST and ModelNet classification follows the architecture in Section 3.3. We use 4(conv.)+5(MLP)+3(FC) layers for the MeshMNIST task, and 8(conv.)+5(MLP)+3(FC) layers for ModelNet task. Both are optimized with 100 epochs by AdamOptimizer with learning rate = 0.001. The batch sizes are 100 and 5 for two task respectively.

The network for Ruemonge2014 segmentation is modified based on the fully convolutional networks (FCN) [21] with VGG19 [34]. This network exactly corresponds to our generic segmentation architecture in Fig. 3: the VGG part corresponds to the cross-atlas convolution. We replace all 3×3 convolutions and 2×2 pooling in VGG and the deconvolution in FCN with our cross-atlas version. We train this network with 100 epochs with AdamOptimizer (learning rate = 1e-4) and 10 batch size.

5.1. MeshMNIST

The original MNIST dataset contains handwritten digit images at 28×28 resolution (Fig. 6(a)). The MeshMNIST [16] converts the digit image to a triangulated mesh by mapping the intensity to the height-field of the mesh. Although MeshMNIST is first used in [16] for the evaluation of their generative model, we conduct a classification experiment using the meshes.

We inversely map the height-field back to intensity (Fig. 6(b)), segment the mesh and pack atlases in the texture map (Fig. 6(c)). The texture map and its corresponding offset map are fed into our network to train a classifier.



(a) The original MNIST (b) The MeshMNIST (c) The texture map of mesh

Figure 6: The MNIST data sample (a) is texture mapped to a mesh (b), and its texture map is not visually recognizable as in (c).

Conv. layers	FC layers	MNIST Acc.	MeshMNIST Acc.
Standard conv.	Standard FC	99.2%	30.8%
Standard conv	MLP+MaxPool+FC	96.8%	88.6%
Cross-atlas conv.	MLP+MaxPool+FC	96.8%	**96.5%**

Table 1: The testing accuracy of different combinations. MNIST – normal digit images, MeshMNIST – texture maps.

Ablation study To better validate the effectiveness of each component, we start with the standard LeNet5 on original MNIST dataset, and then replace with our components one by one. Table 1 shows the comparisons of using differnet modules and dataset.

The standard LeNet5 consists of 4 conv. layers and 3 FC layers, achieving 99.2% on the original MNIST dataset, and 30.8% if directly applied on texture maps. If we replace the standard FC layers with the MLP+max pooling+FC modules, the accuracy drops by 2.4% on standard image, and increases to 88.6% on texture maps, due to the loss of spatial information and meanwhile achieving translational invariance. If we further integrates the cross-atlas convolution and pooling in the network, the accuracy on texture maps increases to 96.5%.

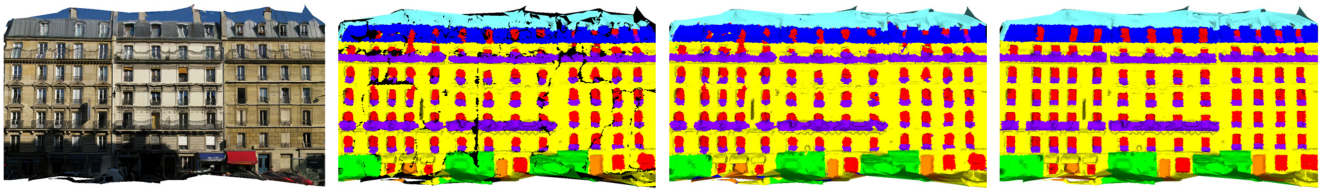
5.2. ModelNet classification

We evaluate our approach for 3D shape classification task on the two versions of the large scale Princeton ModelNet dataset [43]: ModelNet40 and ModelNet10, which consist of 40 and 10 classes respectively. We follow the same experiment setting as in [43]. The vertex coordinates are rasterized to the texture map at 256×256 resolution for the network input. As our texture bin-packing algorithm is randomized, we generate the input texture map by running multiple times to augment the training data, as well as randomly rotating the model along Z-axis.

Method	Input	ModelNet40 accuracy	ModelNet10 accuracy
MVCNN [37]	image	90.1%	-
RotationNet [11]	image	97.37%	98.46%
VoxNet [25]	volume	83%	92%
MVCNN+MultiRes [29]	img.+vol.	91.4%	-
PointNet [28]	point	89.2%	-
PointNet++ [30]	point	91.9%	-
SHR [12]	mesh	68.2%	79.9%
GeometryImage[35]	mesh	83.9%	88.4%
Ours	mesh	87.5%	91.2%

Table 2: The overall classification accuracies of the multi-view image, volume, point and mesh representations.

Table 2 shows the classification accuracy in testing. We have listed representative state-of-the-art methods of using different geometry representations, namely multi-view images, volumes, points and meshes. Our method achieves better results than the other two mesh-based methods [12, 35], while it is overall not as competitive as multi-view images or point-based methods. We perceive the CAD mesh has a signification problem that hinders the performance of mesh-based methods: some structure in the mesh model should have been topologically connected, but in fact they are just overlaid together. This makes the geodesic receptive field erroneous. On the contrary, multi-view image projection or point-based method can avert the problem.



(a) Input textured mesh (b) Our result (before fusion) (c) Our result (after fusion) (d) Ground truth

Figure 7: The qualitative comparison of semantic segmentation results on the Ruemonge2014 dataset [32].

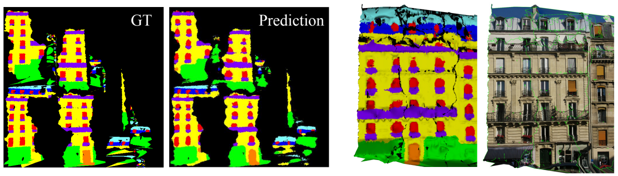
5.3. Ruemonge2014 segmentation

We evaluate the semantic segmentation performance on the Ruemonge2014 dataset [32], which consists of 428 street images, their camera poses and reconstructed meshes by multi-view stereo. The images and mesh come with ground truth semantic labels of seven classes, namely the *window*, *wall*, *balcony*, *door*, *roof*, *sky* and *shop*. The dataset is separated into training samples and testing samples. To evaluate, the class-averaged intersection over union (IoU) are per-triangle label accuracy are used.

To generate the textured mesh, we run the multi-view texturing algorithm [39] using the given images and camera poses. Then, we segment the training mesh part into 100 overlapped mesh segments, each has a 512 × 512 texture map. Here, the RGB + height values are used in the texture map channel.

Method	**triangle accuracy**	**class avg. IoU**
Riemenschneider [32]	-	41.92%
Gadde [9]	-	63.70%
Ours (texture)	65.90%	61.58%
Ours (texture+fusion)	74.27%	67.15%
Ours (image+fusion)	71.81%	64.84%
Ours (texure+image+fusion)	**75.09%**	**67.91%**

Table 3: The evaluation of mesh-based semantic segmentation on the Ruemonge2014 dataset [32]



(a) Inconsistent atlas boundaries between GT and prediction (b) Erroneous labels at texture seams and small atlases

Figure 8: The problem of semantic segmentation on texture maps: the atlas boundaries are not alway consistent with ground truth (a), leading to erroneous labels at texture seams (b).

Fig. 8 shows the testing result on one sample. We notice that the predicted annotation map does not have exactly the same atlas boundaries as the ground truth, and some small atlases are even filtered out (Fig. 8(a)). This problem is analogous to "over-rounded" artifacts in segmentation, whereas for texture maps the atlas may slightly dilate or erode. This issue leads to erroneous labels at texture seams when mapping the annotation map to the mesh (Fig. 8(b)).

Fusion in testing stage Inspired by the multi-scale testing trick used in common semantic segmentation methods, we conduct the testing on 50 individual and overlapped parts, and fuse them afterwards: each triangle label is finally determined by the label of majority pixels. This improves the result significantly, from 61.58% to 67.15% IoU as shown in Table 3. Overall, our method surpass the second place in the Ruemonge2014 challenge benchmark [32] by a large margin. Fig. 7 shows the qualitative results.

Train on images, test on textured meshes With cross-atlas convolution, the network can be trained and tested on texture maps with corresponding offset maps. One may wonder if the network can also be trained on natural images and tested on texture maps – we can regard the natural image as a single-atlas "texture map" with zero offsets. To validate, we take street view images (augmented with the height value) and the ground truth given in the dataset for training. The testing accuracy turns out decent, achieving 71.81% triangle label accuracy and 64.84% IoU. If we combine images and texture maps for training, the result is by 0.76% marginally better than only training on texture maps. This shows our method is flexible in terms of the training data – it can be trained on natural images and tested on texture maps of meshes as long as they have the same content.

6. Conclusion

We have proposed a parameterization invariant approach to textured mesh learning. The key to this method is the cross-atlas operations which recover the mesh geodesic neighborhood although it applies on 2D domain. An interesting future research would be adding the 3D object proposal on textured meshes and bridge the gap between semantic understanding and large-scale 3D reconstruction [22, 33, 47, 49, 44, 45, 19, 18], to potentially enable a fully automatic pipeline of reconstruction and semantic understanding in large-scale 3D scenes.

Acknowledgment. This work is supported by Hong Kong RGC GRF 16203518, T22-603/15N, ITC PSKL12EG02. We thank the support of Google Cloud Platform.

References

[1] F. Bogo, J. Romero, M. Loper, and M. J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ, USA, June 2014. IEEE. 1

[2] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 3189–3197, 2016. 1, 3, 6

[3] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 1, 2

[4] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013. 1, 2

[5] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, volume 2, page 10, 2017. 1, 2

[6] A. Dai and M. Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *CoRR, abs/1703.06211*, 1(2):3, 2017. 5

[8] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016. 1, 2

[9] R. Gadde, V. Jampani, R. Marlet, and P. V. Gehler. Efficient 2d and 3d facade segmentation using auto-context. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1273–1280, 2018. 8

[10] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri. 3d shape segmentation with projective convolutional networks. In *Proc. CVPR*, volume 1, page 8, 2017. 1, 2

[11] A. Kanezaki, Y. Matsushita, and Y. Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5010–5019, 2018. 1, 2, 7

[12] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003. 7

[13] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1

[14] R. Klokov and V. Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 863–872. IEEE, 2017. 2

[15] R. E. Korf. A new algorithm for optimal bin packing. In *AAAI/IAAI*, pages 731–736, 2002. 3

[16] I. Kostrikov, Z. Jiang, D. Panozzo, D. Zorin, and B. Joan. Surface networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 2018. 1, 2, 7

[17] B. Lévy, S. Petitjean, N. Ray, and J. Maillot. Least squares conformal maps for automatic texture atlas generation. In *ACM transactions on graphics (TOG)*, volume 21, pages 362–371. ACM, 2002. 2, 3

[18] S. Li, S. Y. Siu, T. Fang, and L. Quan. Efficient multiview surface refinement with adaptive resolution control. In *European Conference on Computer Vision*, pages 349–364. Springer, 2016. 8

[19] S. Li, Y. Yao, T. Fang, and L. Quan. Reconstructing thin structures of manifold surfaces by integrating spatial curves. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2896, 2018. 8

[20] Y. Li, R. Bu, M. Sun, and B. Chen. Pointcnn. *arXiv preprint arXiv:1801.07791*, 2018. 2

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4, 6, 7

[22] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 168–183, 2018. 8

[23] H. Maron, M. Galun, N. Aigerman, M. Trope, N. Dym, E. Yumer, V. G. Kim, and Y. Lipman. Convolutional neural networks on surfaces via seamless toric covers. *ACM Trans. Graph*, 36(4):71, 2017. 2, 3

[24] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015. 1, 3, 6

[25] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 922–928. IEEE, 2015. 1, 2, 6, 7

[26] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proc. CVPR*, volume 1, page 3, 2017. 1, 3, 6

[27] H. Pan, S. Liu, Y. Liu, and X. Tong. Convolutional neural networks on 3d surfaces using parallel frames. *arXiv preprint arXiv:1808.04952*, 2018. 3, 6

[28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017. 1, 2, 5, 6, 7

[29] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 1, 2, 7

[30] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017. 1, 2, 7

[31] G. Riegler, A. O. Ulusoy, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 3, 2017. 2

[32] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool. Learning where to classify in multi-view semantic segmentation. In *European Conference on Computer Vision*, pages 516–532. Springer, 2014. 7, 8

[33] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph-based consistent matching for structure-from-motion. In *European Conference on Computer Vision*, pages 139–155. Springer, 2016. 8

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[35] A. Sinha, J. Bai, and K. Ramani. Deep learning 3d shape surfaces using geometry images. In *European Conference on Computer Vision*, pages 223–240. Springer, 2016. 2, 3, 6, 7

[36] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz. SPLATNet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2530–2539, 2018. 1, 2

[37] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 1, 2, 7

[38] V. Surazhsky, T. Surazhsky, D. Kirsanov, S. J. Gortler, and H. Hoppe. Fast exact and approximate geodesics on meshes. In *ACM transactions on graphics (TOG)*, volume 24, pages 553–560. Acm, 2005. 4

[39] M. Waechter, N. Moehrle, and M. Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *European Conference on Computer Vision*, pages 836–850. Springer, 2014. 8

[40] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):72, 2017. 2

[41] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *arXiv preprint arXiv:1801.07829*, 2018. 2

[42] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 6

[43] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1, 2, 3, 7

[44] Y. Yao, S. Li, S. Zhu, H. Deng, T. Fang, and L. Quan. Relative camera refinement for accurate dense reconstruction. In *2017 International Conference on 3D Vision (3DV)*, pages 185–194. IEEE, 2017. 8

[45] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 8

[46] L. Yi, H. Su, X. Guo, and L. J. Guibas. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In *CVPR*, pages 6584–6592, 2017. 1, 2

[47] R. Zhang, S. Zhu, T. Fang, and L. Quan. Distributed very large scale bundle adjustment by global camera consensus. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 29–38, 2017. 8

[48] K. Zhou, J. Synder, B. Guo, and H.-Y. Shum. Iso-charts: stretch-driven mesh parameterization using spectral analysis. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 45–54. ACM, 2004. 2, 3

[49] S. Zhu, R. Zhang, L. Zhou, T. Shen, T. Fang, P. Tan, and L. Quan. Very large-scale global sfm by distributed motion averaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2018. 8