DiffusionNet: Discretization Agnostic Learning on Surfaces

NICHOLAS SHARP, Carnegie Mellon University, University of Toronto
SOUHAIB ATTAIKI, LIX, École Polytechnique
KEENAN CRANE, Carnegie Mellon University
MAKS OVSJANIKOV, LIX, École Polytechnique

We introduce a new general-purpose approach to deep learning on 3D surfaces, based on the insight that a simple diffusion layer is highly effective for spatial communication. The resulting networks are automatically robust to changes in resolution and sampling of a surface—a basic property which is crucial for practical applications. Our networks can be discretized on various geometric representations such as triangle meshes or point clouds, and can even be trained on one representation then applied to another. We optimize the spatial support of diffusion as a continuous network parameter ranging from purely local to totally global, removing the burden of manually choosing neighborhood sizes. The only other ingredients in the method are a multi-layer perceptron applied independently at each point, and spatial gradient features to support directional filters. The resulting networks are simple, robust, and efficient. Here, we focus primarily on triangle mesh surfaces, and demonstrate state-of-the-art results for a variety of tasks including surface classification, segmentation, and non-rigid correspondence.

CCS Concepts: • Computing methodologies → Shape analysis.

Additional Key Words and Phrases: geometric deep learning, geometry processing, discrete differential geometry, partial differential equations

ACM Reference Format:

1 INTRODUCTION

Recently there has been significant interest in learning techniques for non-uniform geometric data, inspired by the tremendous success of convolutional neural networks (CNNs) in computer vision. A particularly challenging setting is extending the power of CNNs to learning directly on curved surfaces [Bronstein et al. 2017; Hanocka et al. 2019; Masci et al. 2015; Poulenard and Ovsjanikov 2018]. Unlike volumetric [Maturana and Scherer 2015] or point-based [Qi et al. 2017a] approaches, surface-based methods exploit the connectivity of the surface representation to improve performance, and furthermore can be robust in the presence of non-rigid deformations, making them a strong solution for many tasks such as deformable shape matching [Boscaini et al. 2016; Masci et al. 2015].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). © 2022 Copyright held by the owner/author(s).
https://doi.org/10.1145/3507905

Fig. 1. Surface learning methods must generalize to shapes represented differently from the training set to be useful in practice, yet many existing approaches depend strongly on mesh connectivity. Here, our DiffusionNet trained for human segmentation with limited variability seen during training automatically generalizes to widely varying mesh samplings (left), scales gracefully to resolutions ranging from a simplified model to a large raw scan (middle), and can even be evaluated directly on point clouds (right).

However, although the field has largely been focused on the benchmark accuracy of networks for such problems, at least two other major roadblocks remain for achieving the full potential of learning on surfaces. First, whereas real-world geometric data comes from a variety of sources, existing networks are strongly tied to a particular representation (e.g., triangulations or point clouds) or even discretization resolution. Hence, training cannot benefit from all available data. One popular strategy is to simply convert all data to a common representation (e.g., via point sampling), but this approach has well-known drawbacks (sampling a high-quality, detailed surface can alias thin features, lose informative details, etc.). Second, many existing mesh-based architectures do not scale well to high-resolution surface data. Though coarse inputs are sufficient for, e.g., classification tasks, they preclude potential future applications such as high-fidelity geometric analysis and synthesis.

A key technical difficulty in surface-based learning is defining appropriate notions of convolution and pooling—two main building blocks in traditional CNNs. Unfortunately, unlike the Euclidean case, there is no universal canonical notion of convolution on surfaces. Existing approaches have tried to address this challenge through a variety of solutions such as mapping to a canonical domain [Maron et al. 2017; Sinha et al. 2016], exploiting local parametrizations [Boscaini et al. 2016; Masci et al. 2015; Wiersma et al. 2020] or applying convolution on the edges of the mesh [Hanocka et al. 2019]. However, the
use of more advanced and delicate geometric operations, such as computing geodesics or parallel transport, has a significant impact on both the robustness and scalability of the resulting methods. Perhaps even more importantly, existing surface-based approaches are often too sensitive to the underlying mesh structure, and thus unable to generalize to significantly different sampling and triangulations between training and test sets. As a result, despite significant recent progress in geometric deep learning [Cao et al. 2020; Greengard 2020], current methods typically struggle to cope with the variability, complexity and scale encountered in the real-world surface and mesh-based settings.

In this work, we propose a method that exploits the surface representation, but is both scalable and robust in the presence of significant sampling changes (see Figure 1). Our main observation is that expensive and potentially brittle operations used in previous works [Masci et al. 2015; Poulenard and Ovsjanikov 2018; Wiersma et al. 2020] can be replaced with two basic geometric operations: a learned diffusion layer for information propagation and a spatial gradient for capturing anisotropy. Discretizing these operations with the principled techniques of discrete differential geometry [Crane et al. 2013; Meyer et al. 2003] then automatically endows the resulting networks with both robustness and scalability, while maintaining the simplicity of the learning framework.

Remarkably, we show that combining these basic geometric operations yields neural networks that are not only robust and scalable, but also achieve state-of-the-art results in a wide variety of applications, including deformable surface segmentation, classification, as well as unsupervised and supervised non-rigid shape matching. Perhaps even more fundamentally, our DiffusionNet offers a unified perspective across representations of surface geometry—in principle it can be applied to any geometric representation where one has a Laplacian and gradient operator. In this paper, for example, we show how the same architecture achieves accurate results for both meshes and point clouds, and even allows training on one and evaluating on the other.

Contributions. The main contributions of this work are:

• We show that a simple learned diffusion operation is sufficient to share spatial data in surface learning.
• We introduce spatial gradient features for learning local directional filters.
• Inspired by these insights, we present DiffusionNet, an architecture for learning on surfaces which has many advantages including robustness to discretization, and achieves state-of-the-art results on several benchmarks.

2 RELATED WORK

Applying deep learning techniques to 3D shapes is a rich and extensive area of research. Below we review the approaches most closely related to ours, and refer the interested readers to recent surveys, including [Bronstein et al. 2017; Cao et al. 2020; Xu et al. 2016].

View-based and volumetric methods. Most early geometric deep learning-based methods directly leveraged tools developed for 2D images, and thus mapped 3D shapes onto the plane either using multi-view renderings [Kalogerakis et al. 2017; Su et al. 2015; Wei et al. 2016] or more global, often parametrization-based techniques such as panoramas [Sifakis et al. 2017; Shi et al. 2015], geometry images [Sinha et al. 2016], or metric-preserving mappings [Ezuz et al. 2017], among many others.

Another direct approach to applying convolution to 3D shapes relies on volumetric voxel grid representations, which has led to a variety of methods, including [Maturana and Scherer 2015; Wu et al. 2015] and their efficient extensions [Klokov and Lempitsky 2017; Wang et al. 2017]. Such techniques can, however, be computationally expensive and difficult to apply to detailed deformable shapes.

2.1 Learning on Surfaces

Methods that learn on 3D surfaces directly typically fall into two major categories, based either on point cloud or triangle mesh representations.

Point-based methods. A successful set of methods for learning on 3D shapes represented as point clouds was pioneered by the PointNet [Qi et al. 2017a] and PointNet++ [Qi et al. 2017b] architectures, which have been extended in many recent works, including PointCNN [Li et al. 2018], DGCNN [Wang et al. 2019], PCNN [Atzmon et al. 2018] and KPConv [Thomas et al. 2019] to name a few (see also [Guo et al. 2020] for a recent survey). Moreover, recent efforts have also been made to incorporate invariance and equivariance of the networks with respect to various geometric transformations, e.g., [Deng et al. 2018; Hansen et al. 2018; Li et al. 2021; Poulenard et al. 2019; Zhang et al. 2019; Zhao et al. 2020]. The major advantages of point-based methods are their simplicity, flexibility, applicability in a wide range of settings and robustness in the presence of noise and outliers. However, first, their overall accuracy can often be lower than that of methods that explicitly use surface (e.g., mesh) connectivity when it is available. Second, though effective on static mechanical objects and scenes, point-based methods may not be well-suited for deformable (non-rigid) shape analysis, requiring extremely large training sets and significant data augmentation to achieve good results, e.g., for non-rigid shape matching applications [Donati et al. 2020; Groueix et al. 2018]. Globally supported point-based networks were recently considered in [Peng et al. 2020]; our method naturally allows global support via learned diffusion.

Fig. 2. Many recent mesh-based learning methods are applied only to dramatically simplified inputs (Section 5.6), while our method easily processes full-resolution models, preserving detail and facilitating adoption.
Surface and graph-based methods. To address the limitations of point-based approaches, several methods have been proposed that operate directly on mesh surfaces and thus can learn filters that are intrinsic and robust to complex non-rigid deformations. The earliest pioneering approaches in this direction generalize convolutions [Boscaini et al. 2016; Fey et al. 2018; Masci et al. 2015; Monti et al. 2017], typically using local surface parameterization via the logarithmic map. Unfortunately, local parameterizations are only defined up to rotation in the tangent plane, leading to several methods that address this issue through design of equivariant surface networks [Haan et al. 2021; He et al. 2020; Mitchel et al. 2021; Poulie and Ovsjanikov 2018; Wiersma et al. 2020; Yang et al. 2021]. Likewise, operating on vector-valued data in a local tangent space can expand the expressivity of the filter space [Beani et al. 2021; Mitchel et al. 2021; Wiersma et al. 2020]. Our method leverages learned gradient features (Section 3.4), which geometrically require only a local spatial gradient operation, and sidestep the challenge of equivariant filters by using only inner products, which are naturally invariant. These gradient features are built on local differential operators, which have also been exploited in other recent methods (e.g., [Eliaos and Treister 2020; Jiang et al. 2018]).

Surface mesh structure has also been used in a variety of graph-like approaches which specifically leverage mesh connectivity [Bodnar et al. 2021; Feng et al. 2019; Gong et al. 2019; Hajiji et al. 2020; Hanocka et al. 2019; Lim et al. 2018; Milano et al. 2020; Verma et al. 2018], the structure of discrete operators [Smirnov and Solomon 2021], or even random walks along edges [Lahav and Tal 2020], among others. While accurate, these methods can be costly on densely sampled shapes and often are not robust to significant changes in the mesh structure (Figure 3).

2.2 Spectral Methods

Our use of diffusion is also closely related to techniques that operate in the spectral domain and often exploit the link between convolution and operations in a derived (e.g., Fourier or Laplace-Beltrami) basis, including [Bruna et al. 2014; Levie et al. 2018; Sun et al. 2020]. Such methods have a long history in graph-based learning and are well-rooted in data analysis more broadly, including Laplacian eigenmaps [Belkin and Niyogi 2003], spectral clustering [Vallet and Lévy 2008] and diffusion maps [Coifman et al. 2005]. In geometry processing, spectral methods have been used for a range of tasks including multi-resolution representation [Levy 2006], segmentation [Rustamov 2007] and matching [Ovsjanikov et al. 2012], among others [Vallet and Lévy 2008; Zhang et al. 2010].

Unfortunately, Laplacian eigenfunctions depend on each shape and thus coefficients or learned filters from one shape are not trivially transferable to another. Levie et al. [2019] argue for transfer between discretizations of the same shape, but 3D geometric learning typically demands transfer between different shapes. Functional maps [Ovsjanikov et al. 2012] can be used to “translate” coefficients between shapes, and have been used, e.g., in [Yi et al. 2017] with spectral filter learning. Deep functional maps [Litany et al. 2017] propose to learn features in the primal domain, which are then projected onto the Laplace-Beltrami basis for functional map estimation. However, the features are still learned either with MLPs starting from pre-computed descriptors [Ginzburg and Raviv 2020; Halimi et al. 2019; Litany et al. 2017; Roufosse et al. 2019] or using point-based architectures [Donati et al. 2020].

Instead, we propose an approach that learns the parameters of a diffusion process which is directly transferable across shapes and, as we show below, can be used effectively in applications like non-rigid shape matching. We also stress that DiffusionNet is not spectral in nature and only uses spectral operations as an acceleration technique for evaluating diffusion efficiently.

We also note that our use of the Laplacian in defining the diffusion operator is related to methods based on polynomials of the Laplacian [Deferrard et al. 2016; Kostrikov et al. 2018], CayleyNets [Levie et al. 2018] and their recent application in shape matching using ACSCNNs [Li et al. 2020b]. However, we demonstrate that complex polynomial filters can be replaced with simple learned diffusion, and moreover that gradient features can inject orientation information into the network, improving performance and robustness. Similarly to our approach, diffusion for smooth communication has been explored on graphs [Klicpera et al. 2019; Xu et al. 2019], images [Liu et al. 2016], and point clouds [Hansen et al. 2018]. In contrast, our method directly learns a diffusion time per-feature (which significantly improves performance, Table 7), incorporates a learned gradient operation, and is applied directly to mesh surfaces.

Pooling. In surface learning it is nontrivial to define pooling, especially on meshes where it often amounts to mesh simplification [Hanocka et al. 2019]. Various recent operations have been proposed for point cloud [Hu et al. 2020; Lin et al. 2020], mesh [Milano et al. 2020; Zhou et al. 2020] or even graph pooling [Li et al. 2020a; Ma et al. 2020]. A key advantage of our approach is that it automatically supports global spatial support without any downsampling operation, simplifying implementation and improving learning.
3 METHOD

Our method consists of three main building blocks: multi-layer perceptrons (MLPs) applied at each point to model pointwise scalar functions of feature channels, a learned diffusion operation for propagating information across the domain, and local spatial gradient features to expand the network’s filter space beyond radially-symmetric filters. In this section, we describe these main numerical components, and then we assemble them into an effective architecture in Section 4. Our method is defined in a representation-agnostic manner; applying it to meshes or point clouds simply amounts to assembling the appropriate Laplacian and gradient matrices as we discuss below.

3.1 Pointwise Perceptrons

On a mesh or point cloud with $V$ vertices, we consider a collection of $D$ scalar features defined at each vertex. Our first basic building block is a pointwise function $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$, which is applied independently at every vertex to transform the features. We represent these pointwise functions as a standard multilayer perceptron (MLP) with shared weights across all vertices. Although these MLPs can fit arbitrary functions at each point, they do not capture the spatial structure of the surface, or allow any communication between vertices, so a richer structure is needed.

Past approaches for communication have ranged from global reductions to explicit geodesic convolutions—instead, we will demonstrate that a simple learned diffusion layer effectively propagates information, without the need for potentially costly or error-prone computations.

3.2 Learned Diffusion

In the continuous setting, diffusion of a scalar field $u$ on a domain is modeled by the heat equation

$$\frac{\partial}{\partial t} u_t = \Delta u_t,$$

where $\Delta$ is the Laplacian (or more formally: the Laplace-Beltrami operator). The action of diffusion can be represented via the heat operator $H_t$, which is applied to each node independently at each vertex to transform the features. We represent these pointwise functions as a standard multilayer perceptron (MLP) with shared weights across all vertices. Although these MLPs can fit arbitrary functions at each point, they do not capture the spatial structure of the surface, or allow any communication between vertices, so a richer structure is needed.

Past approaches for communication have ranged from global reductions to explicit geodesic convolutions—instead, we will demonstrate that a simple learned diffusion layer effectively propagates information, without the need for potentially costly or error-prone computations.

A note on generality. Remarkably, eschewing traditional representations of convolutions in favor of diffusion does not reduce the expressive power of our networks. This is supported by the following theoretical result (that we prove in Appendix A), which shows that radially-symmetric convolutions are contained in the function space defined by diffusion followed by a pointwise map:

**Lemma 1 (Inclusion of radially-symmetric convolutions).** For a signal $u : \mathbb{R}^2 \rightarrow \mathbb{R}$, let $U_r(p) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ denote the integral of $u$ along the r-sphere at $p$, and let $u_t(p) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ denote the signal value at $p$ after diffusion for time $t$. Then there exists a function $T$ which recovers $U_r(p)$ from $u_t(p)$:

$$U_r(p) = T[u_t(p)](r).$$

Thus convolution with a radial kernel $\alpha : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ is given by

$$(u * \alpha)(p) = \int_{\mathbb{R}^2} \alpha(|q - p|)u(q) \, dq = \int_0^\infty \alpha(r)T[u_t(p)](r) \, dr,$$

which is a pointwise operation at $p$ on the diffused values $u_t(p)$.
This fact is significant because it suggests that simple and robust diffusion can, without loss of generality, be used to replace complicated operations such as radial geodesic convolution.

Importantly, we will also extend our architecture beyond radially-symmetric filters by incorporating gradient features (Section 3.4).

3.3 Computing Diffusion

Many numerical schemes could potentially be used to evaluate the diffusion layer \( h_t(u) \), from direct solvers [Chen et al. 2008] to hierarchical schemes [Liu et al. 2021; Vaxman et al. 2010]. In particular, we seek schemes which are efficient as well as differentiable, to enable network training. Here we describe two simple methods considered in our experiments. The first scheme we consider is an implicit timestep, which is straightforward but requires solving large sparse linear systems, and the second is spectral expansion, which uses only efficient dense arithmetic at evaluation time but requires some modest precomputation. Both are easily implemented using common numerical libraries, and we observe that networks trained with either approach have similar accuracy. Efficiency is evaluated in Section 5.6; we generally recommend spectral acceleration.

3.3.1 Direct Implicit Timestep. Perhaps the simplest effective approach to simulate diffusion is a single implicit Euler timestep

\[
h_t(u) := (M + tL)^{-1}Mu
\]

which amounts to solving a (sparse) linear system for each diffusion operation. Using an implicit backward timestep rather than an explicit forward timestep is crucial as it makes the scheme stable, allows global support, and yields a reasonable approximation of diffusion after just one step. Solving linear systems (including derivative computation) is supported in modern learning software frameworks, allowing Equation 2 to implement a learnable diffusion block. However, this amounts to solving a distinct large linear system for each channel, and GPU-based computation may fall back to solving dense linear systems, which means that direct implicit timesteps may not scale well to large problems.

3.3.2 Spectral Acceleration. An alternate approach is to leverage a closed-form expression for diffusion in the basis of low-frequency Laplacian eigenfunctions [Vallet and Lévy 2008; Zhang et al. 2010]. Once the eigenbasis has been precomputed, diffusion can then be evaluated for any time \( t \) via elementwise exponentiation. Truncating diffusion to a low-frequency basis incurs some approximation error, but we find that this approximation has little effect on our method (Figure 13), perhaps because diffusion quickly damps high-frequency content regardless.

For weak Laplace matrix \( L \) and mass matrix \( M \), the eigenvectors \( \phi_i \in \mathbb{R}^V \) are solutions to:

\[
L\phi_i = \lambda_i M\phi_i,
\]

corresponding to the first \( k \) smallest-magnitude eigenvalues \( \lambda_1, \ldots, \lambda_k \).

We normalize them so that \( \phi_i^T M \phi_i = 1 \). These eigenvectors are easily precomputed for each shape of interest via standard numerical packages [Lehoucq et al. 1998]; the inset figure shows several example functions \( \phi_i \) for a surface of a human shape.

Let \( \Phi := [\phi_1] \in \mathbb{R}^{V \times k} \) be the stacked matrix of eigenvectors, which form an orthonormal basis with respect to \( M \). We can then project any scalar function \( u \) to obtain its coefficients \( \epsilon \) in the spectral basis via \( c \leftarrow \Phi \epsilon \), and recover values at vertices as \( u \leftarrow \Phi c \). Conveniently, diffusion for time \( t \) is easily expressed as an elementwise scaling of spectral coefficients according to \( c_i \leftarrow e^{-\lambda_i t} c_i \). The diffusion layer \( h_t(u) \) is then evaluated by projecting on to the spectral basis, evaluating pointwise diffusion, and projecting back

\[
h_t(u) := \Phi \left[ e^{-\lambda_1 t} \ldots e^{-\lambda_k t} \right] \odot (\Phi^TMu)
\]

where \( \odot \) denotes the Hadamard (elementwise) product. This operation is efficiently evaluated using dense linear algebra operations like elementwise exponentiation and matrix multiplication, and is easily differentiable with respect to \( u \) and \( t \).

Remarks. We emphasize that DiffusionNet can still learn high-frequency outputs despite the use of a low-frequency basis (e.g., in Section 5.4). Intuitively, diffusion is used for communication across points, for which a low-frequency approximation is typically sufficient, while MLPs and gradient features learn high-frequency features as needed for a task. Additionally, we note that DiffusionNet is not a spectral learning method—spectral coefficients are never used to represent filters or latent data, and thus no issues arise due to differing eigenbases on different shapes. Spectral acceleration is merely one possible numerical scheme to compute diffusion.

3.4 Spatial Gradient Features

Our learned diffusion layer enables propagation of information across different points on a shape, but it supports only radially-symmetric filters about a point. The last building block in our method enables a larger space of filters by computing additional features from the spatial gradients of signal values at vertices (Figure 6). Specifically, we construct features from the inner products between pairs of feature gradients at each vertex, after applying a learned scaling or rotation.

Evaluating gradients. We will express the spatial gradient of a scalar function on a surface as a 2D vector in the tangent space of each vertex. These gradients can be evaluated by a standard procedure, choosing a normal vector at each vertex (given as input or locally approximated), then projecting neighbors into the tangent plane—either 1-ring neighbors on a mesh, or \( m \)-nearest neighbors in a point cloud. The gradient is then computed in the tangent plane via least-squares approximation of the function values at neighboring points (see Mukherjee and Wu [2006] for analysis). These gradient operators at each vertex can be assembled into a sparse matrix \( G \in \mathbb{C}^{V \times V} \), which is applied to a vector \( u \) of real values at vertices to produce gradient tangent vectors at each vertex. This matrix does not depend on the features, and can be precomputed once for each shape. We use complex numbers as a convenient notation for tangent vectors, in an arbitrary reference basis in the tangent plane of each point (as in [Knöppel et al. 2013; Sharp et al. 2019]), etc.). If the normals are consistently oriented, then the imaginary axis is chosen to form a right-handed basis in 3D with respect to the outward surface normal.
Learned pairwise products. Equipped with per-vertex spatial gradients of each channel, we learn informative scalar features by evaluating an inner product between pairs of feature gradients at each vertex, after a learned linear transformation. Inner products are invariant to rotations of the coordinate system, so these features are invariant to the choice of tangent basis at vertices, as expected. Putting it all together, given a collection of $D$ scalar feature channels, for each channel $u$ we first construct its spatial gradient as $z_u \in \mathbb{C}^V$, a vector of local 2D gradients per-vertex

$$z_u := \mathbf{G}u$$  

then at each vertex $v$, we stack the local gradients of all channels to form $w_v \in \mathbb{C}^D$ and obtain real-valued features $g_v \in \mathbb{R}^D$ as

$$g_v := \tanh(\text{Re}(\mathbf{A}w_v))$$

where $A$ is a learned square $D \times D$ matrix, and taking the real part $\text{Re}$ of the complex conjugate $\overline{w_v}$ is again just a notational convenience for dot products between pairs of 2D vectors. This means the $i$th entry of the output at vertex $v$ is given by the dot product $g_v(i) = \tanh(\text{Re}\left\{\sum_{j=1}^{D} w_v(i) A_{ij} w_v(j)\right\})$, so that each inner product is scaled by a learned coefficient $A_{ij}$. The outer $\tanh(\cdot)$ nonlinearity is not fundamental, but we find that it stabilizes training.

The choice of $A$ as a complex or real matrix has a subtle relationship with the orientation of the underlying surface. Multiplying $w_v(j)$ by a complex scalar both rotates and scales local gradient vectors before taking inner products (recall that complex multiplication can be interpreted as a rotation and scaling in the complex plane). In contrast, real $A$ only allows scaling. However, the direction of rotation (clockwise or counter-clockwise) depends on the choice of the outward normal, and hence on orientation—so surfaces with consistently oriented normals gain a richer representation by learning a complex matrix, whereas surfaces without consistent orientation (e.g., raw point clouds) should restrict to real $A$.

In Figure 5, a small synthetic experiment (detailed in Appendix C) demonstrates how these learned rotations allow our method to disambiguate bilateral symmetry even in a purely intrinsic representation, a common challenge in non-rigid shape correspondence.

4 DIFFUSIONNET ARCHITECTURE

The previous section establishes three ingredients for learning on surfaces: an MLP applied independently at each vertex to represent pointwise functions, learned diffusion for spatial communication, and spatial gradient features to model directional filters. We combine these ingredients to construct DiffusionNet (Figure 7), composed of several DiffusionNet blocks. This simple network operates on a fixed channel width $D$ of scalar values throughout, with each DiffusionNet block diffusing the features, constructing spatial gradient features, and feeding the result to an MLP.

We include residual connections to stabilize training [He et al. 2016], as well as linear layers to convert to the expected input and output dimension. When appropriate, results at the edges or faces of a mesh can be computed by averaging network outputs from the incident vertices, e.g., to segment the faces of a mesh. Various activations can be appended to the end of the network based on the problem at hand, such as a softmax for segmentation, or a global mean followed by a softmax for classification; otherwise, this same architecture is used for all experiments.

Remarkably, we do not find it necessary to use any spatial convolutions or pooling hierarchies on surfaces—avoiding these potentially-complex operations helps keep DiffusionNet simple and robust.

Invariance. DiffusionNet is invariant to rigid motion of the underlying shapes as long as the input features remain unchanged, due to the intrinsic geometric nature of diffusion and spatial gradients. The overall invariance then depends on the choice of input features.

4.1 Input features

DiffusionNet takes a vector of scalar values per-vertex as input features. Here we consider two simple choices of features, others could easily be included when available. Most directly, we simply use the raw 3D coordinates of a shape as input; rotation augmentation can be used to promote rigid invariance when inputs are not consistently aligned. When rigid or even non-rigid invariance is desired, we instead use the Heat Kernel Signatures [Sun et al. 2009] as input; these signatures are trivially computed from the spectral basis in Section 3.3.2. Due to the intrinsic nature of our approach, with HKS as input, the networks are invariant to any orientation-preserving isometric deformation of the shape. Higher-order descriptors such as SHOT [Tombari et al. 2010] seem unnecessary, and may be unstable under remeshing [Donati et al. 2020].
5 EXPERIMENTS AND ANALYSIS

The same network architecture achieves state-of-the-art results across many tasks, and more importantly offers new and valuable capabilities. See the appendix for additional analyses.

Setup. We use the same basic 4-block DiffusionNet architecture and training procedure for all tasks, varying the network size from a small 32-width (30k parameter) to a large 256-width (1.8M parameter) DiffusionNet according to the scale of the problem. The shape of the first and last linear layers is adapted to the input and output dimension for the problem. MLPs use ReLU activations and optionally dropout after intermediate linear layers. We let “xyz” and “hks” denote networks with positions and heat kernel signatures as input, respectively. All inputs are centered and scaled to be contained in a unit sphere, and heat kernel signatures are sampled at 16 values logarithmically spaced on \([0.01, 1]\). We do not use any data augmentation, except random rotations in tasks where positions are used as features yet a rotation-invariant network is desired.

We fit DiffusionNet using the ADAM optimizer with an initial learning rate of 0.001 and a batch size of 1, training for 200 epochs and decaying the learning rate by a factor of 0.5 every 50 epochs. Cross-entropy loss is used for labelling problems. Spectral acceleration is used to evaluate diffusion except where noted, truncated to a \(k = 128\) eigenbasis. On point clouds, 30 nearest-neighbors are used to assemble matrices. Test accuracies are measured after the last epoch of training.

Implementation details. Precomputation to assemble matrices and compute the Laplacian eigenbasis for spectral acceleration is performed once as a preprocess on the CPU using SciPy [Lehoucq et al. 1998; Virtanen et al. 2020]. Networks are implemented in PyTorch [Paszke et al. 2019] and evaluated on a single GPU with standard backpropagation. Performance is discussed in Section 5.6: we find that DiffusionNet is very efficient and scalable compared to recent mesh-based learning methods. Code and reproducible experiments are available at github.com/nmwsharp/diffusion-net.

5.1 Classification

We first apply DiffusionNet to classify meshes in the SHREC-11 dataset [Lian et al. 2011], which has 30 categories of 20 shapes each. We demonstrate that DiffusionNet learns successfully even in the presence of limited data. As in the other cited results, we train on just 10 samples per class; our results are averaged over 10 trials of the experiment with random training splits. We fit a cross-entropy loss with a label smoothing factor of 0.2 (see discussion in [Goyal et al. 2021]). We use a 32-width DiffusionNet for hks features, and 64-width DiffusionNet for xyz features with rotation augmentation. DiffusionNet achieves the highest reported accuracy when applied directly on the original dataset models, or to the simplified models common in recent mesh-based learning work (Table 1).

Table 1. DiffusionNet achieves nearly-perfect accuracy classifying 30-class SHREC11 [Lian et al. 2011] while training on just 10 samples per class. Results marked by \(^\dagger\) are trained and tested on simplified models.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWCNN [Ezuz et al. 2017]</td>
<td>90.3%</td>
</tr>
<tr>
<td>MeshCNN(^\dagger) [Hanocka et al. 2019]</td>
<td>91.0%</td>
</tr>
<tr>
<td>HSN(^\dagger) [Wiersma et al. 2020]</td>
<td>96.1%</td>
</tr>
<tr>
<td>MeshWalker(^\dagger) [Lahav and Tal 2020]</td>
<td>97.1%</td>
</tr>
<tr>
<td>PD-MeshNet(^\dagger) [Milano et al. 2020]</td>
<td>99.1%</td>
</tr>
<tr>
<td>HodgeNet(^\dagger) [Smirnov and Solomon 2021]</td>
<td>94.7%</td>
</tr>
<tr>
<td>FC(^\dagger) [Mitchel et al. 2021]</td>
<td>99.2%</td>
</tr>
<tr>
<td>DiffusionNet - xyz(^\dagger)</td>
<td>99.4%</td>
</tr>
<tr>
<td>DiffusionNet - hks(^\dagger)</td>
<td>99.0%</td>
</tr>
<tr>
<td>DiffusionNet - hks</td>
<td>99.5%</td>
</tr>
<tr>
<td>DiffusionNet - hks</td>
<td>99.7%</td>
</tr>
</tbody>
</table>
Table 2. Accuracy of various mesh and point cloud schemes for RNA segmentation. DiffusionNet achieves state-of-the-art results, in part because it can be applied directly to the raw meshes. "xyz" and "hks" denote networks taking raw coordinates or heat kernel signatures as input, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PointNet++ [Qi et al. 2017a]</td>
<td>74.4%</td>
</tr>
<tr>
<td>PCNN [Atzmon et al. 2018]</td>
<td>78.0%</td>
</tr>
<tr>
<td>SPHNet [Poulenard et al. 2019]</td>
<td>80.1%</td>
</tr>
<tr>
<td>DiffusionNet - hks</td>
<td>84.0%</td>
</tr>
<tr>
<td>DiffusionNet - xyz</td>
<td>85.0%</td>
</tr>
<tr>
<td>SplineCNN [Fey et al. 2018]</td>
<td>53.6%</td>
</tr>
<tr>
<td>SurfaceNetworks [Kostrikov et al. 2018]</td>
<td>88.5%</td>
</tr>
<tr>
<td>DiffusionNet - hks</td>
<td>91.0%</td>
</tr>
<tr>
<td>DiffusionNet - xyz</td>
<td>91.5%</td>
</tr>
</tbody>
</table>

5.2 Segmentation

Molecular segmentation. We evaluate a 128-width DiffusionNet on the task of segmenting RNA molecules into functional components, using the dataset introduced by Poulenard et al. [2019]. This dataset consists of 640 RNA surface meshes of about 15k vertices each extracted from the Protein Data Bank [Berman et al. 2000], labelled at each vertex according to 259 atomic categories, with a random 80/20 train-test split. We learn these labels directly on the raw meshes, as well as on point clouds of 4096 uniformly-sampled points as in past work Poulenard et al. [2019]. For comparison, we cite point cloud results reported in [Poulenard et al. 2019], and additionally train methods related to ours, SplineCNN [Fey et al. 2018] and a Dirac Surface Network [Kostrikov et al. 2018] on meshes. We also attempted to train MeshCNN [Hancock et al. 2019] and HSN [Wiersma et al. 2020], but found the former prohibitively expensive, while the latter could not successfully preprocess the data. Our method achieves state-of-the-art accuracy on both the mesh and point cloud variants of the problem (Table 2, Figure 8). Learning directly on the mesh yields greater accuracy, perhaps because no information is lost when sampling a point cloud, and surface structure is preserved.

Human segmentation. We train a 128-width DiffusionNet with dropout to segment the human body parts on the composite dataset of [Maron et al. 2017], containing models from several other human shape datasets [Adobe 2016; Anguelov et al. 2005; Bogo et al. 2014; Giorgi et al. 2007; Vlasic et al. 2008]. Additionally, we cite a variety of reported results from other approaches on this task, as reported by the respective original authors and by Wiersma et al. [2020]. For clarity, we distinguish between variants of this task in past work which used simplified meshes and soft ground truth; more details in Appendix C. Our model is quite effective using both rotation-augmented raw coordinates or heat kernel signatures as input (Table 3).

Table 3. Human part segmentation on the dataset of Maron et al. [2017]. xyz and hks denote DiffusionNet with positions or heat kernel signatures as input, respectively. The † rows use simplified inputs with a soft ground truth at edges, and the ‡ rows use simplified inputs with hard ground truth at faces; details in Appendix C.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCNN [Masci et al. 2015]</td>
<td>86.4%</td>
</tr>
<tr>
<td>ACNN [Boscaini et al. 2016]</td>
<td>83.7%</td>
</tr>
<tr>
<td>Toric Cover [Maron et al. 2017]</td>
<td>88.0%</td>
</tr>
<tr>
<td>PointNet++ [Qi et al. 2017a]</td>
<td>90.8%</td>
</tr>
<tr>
<td>MDGCNN [Poulenard et al. 2018]</td>
<td>88.6%</td>
</tr>
<tr>
<td>DGCNN [Wang et al. 2019]</td>
<td>89.7%</td>
</tr>
<tr>
<td>SNGC [Haim et al. 2019]</td>
<td>91.0%</td>
</tr>
<tr>
<td>HSN [Wiersma et al. 2020]</td>
<td>91.1%</td>
</tr>
<tr>
<td>MeshWalker [Lahav and Tal 2020]</td>
<td>92.7%</td>
</tr>
<tr>
<td>CGConv [Yang et al. 2021]</td>
<td>89.9%</td>
</tr>
<tr>
<td>FC [Mitchel et al. 2021]</td>
<td>92.5%</td>
</tr>
<tr>
<td>DiffusionNet - xyz</td>
<td>90.6%</td>
</tr>
<tr>
<td>DiffusionNet - hks</td>
<td>91.7%</td>
</tr>
</tbody>
</table>

†

| MeshCNN [Hancock et al. 2019] | 92.3%    |
| MeshWalker [Lahav and Tal 2020] | 94.8%    |
| DiffusionNet - xyz       | 95.5%    |
| DiffusionNet - hks       | 95.5%    |

‡

| PD-MeshNet [Milano et al. 2020] | 85.6%    |
| HodgeNet [Smirnov and Solomon 2021] | 85.0%    |
| DiffusionNet - xyz       | 90.3%    |
| DiffusionNet - hks       | 90.8%    |
5.3 Functional Correspondence

Functional maps compute a correspondence between a pair of shapes by finding a linear transformation between spectral bases, aligning some set of input features [Ovsjanikov et al. 2012]. Recent work has shown that learned features can improve performance, e.g., Donati et al. [2020]; Litany et al. [2017]. Here we demonstrate that using DiffusionNet as a feature extractor outperforms other recent approaches, yielding to state-of-the-art correspondence results in both the supervised and weakly-supervised variants of the problem. We emphasize that the spectral representation in functional maps is unrelated to the spectral acceleration from Section 3.3.2, which is merely a scheme for evaluating diffusion; DiffusionNet itself does not learn in the spectral domain.

Our experiments follow the setup of Donati et al. [2020], training and evaluating on both SCAPE [Anguelov et al. 2005; Ren et al. 2018], and FAUST [Bogo et al. 2014] including training on one dataset and evaluating on the other. In the supervised setting we fit dataset-provided correspondences, and we generate rigid-invariant models by randomly rotating all inputs for training and testing. For the weakly-supervised setting, we use the dataset and losses advocated in [Sharma and Ovsjanikov 2020], where rigid alignment of the input is used as weak supervision, without known correspondences. In all cases we extract point-to-point maps between test shapes and evaluate them against ground truth dense correspondences; for simplicity we compare all methods without post-processing the maps, though we also report accuracies after postprocessing with ZoomOut for our method [Melzi et al. 2019]. In addition to citing results obtained using the KPConv feature extractor [Thomas et al. 2019] by Donati et al. [2020], we also train HSN [Wiersma et al. 2020], ACSCNN [Li et al. 2020b], and our own 128-width DiffusionNet with dropout. We also tried MeshCNN [Hanocka et al. 2019], but it proved to be prohibitively expensive at 14hr per epoch.

Table 4. Our approach yields state-of-the-art correspondences when used as a feature extractor for deep functional maps, both in the supervised (top, as in Donati et al. [2020]) and the weakly supervised setting (bottom, as in Sharma and Ovsjanikov [2020]). The dotted rows apply ZoomOut post-processing to the previous result [Melzi et al. 2019]. X on Y means train on X and test on Y. Values are mean geodesic error ×100 on unit-area shapes.

<table>
<thead>
<tr>
<th>Method / Dataset</th>
<th>FAUST</th>
<th>SCAPE</th>
<th>FonS</th>
<th>SonF</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPConv [Thomas et al. 2019]</td>
<td>3.1</td>
<td>4.4</td>
<td>11.0</td>
<td>6.0</td>
</tr>
<tr>
<td>KPConv - hks</td>
<td>2.9</td>
<td>3.3</td>
<td>10.6</td>
<td>5.5</td>
</tr>
<tr>
<td>HSN [Wiersma et al. 2020]</td>
<td>3.3</td>
<td>3.5</td>
<td>25.4</td>
<td>16.7</td>
</tr>
<tr>
<td>ACSCNN [Li et al. 2020b]</td>
<td>2.7</td>
<td>3.2</td>
<td>8.4</td>
<td>6.0</td>
</tr>
<tr>
<td>DiffusionNet - hks</td>
<td>2.7</td>
<td>3.0</td>
<td>3.8</td>
<td>3.0</td>
</tr>
<tr>
<td>DiffusionNet - xyz</td>
<td>2.7</td>
<td>3.0</td>
<td>3.3</td>
<td>3.0</td>
</tr>
<tr>
<td>+ ZoomOut</td>
<td>1.9</td>
<td>2.4</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>WSsupFMNet</td>
<td>3.3</td>
<td>7.3</td>
<td>11.7</td>
<td>6.2</td>
</tr>
<tr>
<td>WSsupFMNet + DiffusionNet - xyz</td>
<td>3.8</td>
<td>4.4</td>
<td>4.8</td>
<td>3.6</td>
</tr>
<tr>
<td>+ ZoomOut</td>
<td>1.9</td>
<td>2.6</td>
<td>2.7</td>
<td>1.9</td>
</tr>
</tbody>
</table>

As shown in Table 4, DiffusionNet yields state-of-the-art results for non-rigid shape correspondence in both the supervised and weakly-supervised settings, especially when transferring between datasets. One might wonder whether the improvement in this task truly stems from our DiffusionNet architecture, or from the use of HKS features. As shown in the same table, training KPConv with HKS as features shows DiffusionNet yields significant improvements regardless. Figure 9 visualizes the resulting correspondences on a challenging test pair, where only DiffusionNet achieves a high-quality correspondence when generalizing after training on a different dataset.
Table 5. DiffusionNet automatically retains highly accurate results under changes in meshing and sampling, while many other approaches overfit to mesh connectivity. Here we give correspondence errors on our remeshed FAUST dataset after training on template meshes, measured in mean geodesic distance ×100 after normalizing by the geodesic diameter.

<table>
<thead>
<tr>
<th>Method</th>
<th>orig</th>
<th>iso</th>
<th>dense</th>
<th>qes</th>
<th>cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACSCNN</td>
<td>0.05</td>
<td>35.29</td>
<td>19.09</td>
<td>41.15</td>
<td>N/A</td>
</tr>
<tr>
<td>SplineCNN</td>
<td>3.51</td>
<td>31.09</td>
<td>27.95</td>
<td>40.43</td>
<td>N/A</td>
</tr>
<tr>
<td>HSN</td>
<td>9.57</td>
<td>20.01</td>
<td>24.84</td>
<td>25.40</td>
<td>N/A</td>
</tr>
<tr>
<td>PointNet (vertices)</td>
<td>3.83</td>
<td>2.92</td>
<td>3.04</td>
<td>2.67</td>
<td>2.60</td>
</tr>
<tr>
<td>PointNet (sampled)</td>
<td>9.99</td>
<td>4.25</td>
<td>7.84</td>
<td>4.44</td>
<td>4.13</td>
</tr>
<tr>
<td>DGCNN (vertices)</td>
<td>2.44</td>
<td>14.58</td>
<td>15.72</td>
<td>27.00</td>
<td>32.13</td>
</tr>
<tr>
<td>DGCNN (sampled)</td>
<td>6.52</td>
<td>4.30</td>
<td>5.57</td>
<td>3.61</td>
<td>2.66</td>
</tr>
<tr>
<td>DiffusionNet</td>
<td>0.33</td>
<td>0.68</td>
<td>0.62</td>
<td>0.82</td>
<td>2.59</td>
</tr>
</tbody>
</table>

5.4 Discretization Agnostic Learning

A key benefit of DiffusionNet compared to many past approaches is that its outputs are robust to changes in the discretization of the input (e.g., different meshes of the same shape, or a mesh vs. a point cloud). This property is essential for practical applications where meshes for inference are likely to be tessellated differently from the training set, etc. Other invariants (e.g., rigid invariance) can be encouraged via data augmentation, but for discretization it is impractical to generate augmented inputs across all possible sampling patterns. Below, we demonstrate that DiffusionNet generalizes quite well across discretizations without any special regularization or augmentation, especially compared to recent mesh-based methods.

We study discretization invariance using a popular formulation of the shape correspondence task on the FAUST human dataset [Bogo et al. 2014], where each vertex of a mesh is to be labelled with the corresponding vertex on a template mesh. Importantly, these input meshes are already manually aligned templates, with exactly identical mesh connectivity. Past work has achieved near-perfect accuracy in this problem setup [e.g., Fey et al. 2018; Li et al. 2020b], however we suggest that these models primarily overfit to mesh graph structure. In contrast, DiffusionNet learns an accurate and general function of the shape itself, despite the synthetic setup.

To experimentally quantify this effect, we construct a version of the FAUST test set after remeshing with several strategies: orig is the original test mesh, iso is a uniform isotropic remeshing, dense refines the mesh in randomly sampled regions, qes first refines the meshes, then applies quadric error simplification [Garland and Heckbert 1997], and cloud is a point cloud with normals sampled from the surface (Figure 10). Unrelated remeshings have appeared in [Poulenard and Ovsjanikov 2018; Wiersma et al. 2020], but the procedure therein left large regions of the mesh unchanged. Ground truth for evaluation is defined via nearest-neighbor on the original test mesh. We train on the 3D coordinates of the 80 standard FAUST registered meshes, but evaluate on the remeshed set to mimic the practical scenario where the training set contains meshes tessellated via some particular common strategy, yet the fitted model must be applied to totally different meshes encountered in the wild. We also train several other methods—details in Appendix C.

Table 5 and Figure 11 show how other mesh-based approaches degrade rapidly under remeshing; only DiffusionNet yields accurate correspondences which are largely stable under remeshing and resampling. Some point-based methods avoid the dependence on connectivity, yet do not match the overall accuracy of the surface-based DiffusionNet.
5.5 Transfer Across Representations
Not only are the outputs of DiffusionNet consistent across remeshing and resampling of a shape, but furthermore the same network can be directly applied to different discrete representations. The only geometric data required for DiffusionNet are the Laplacian, mass, and spatial gradient matrices, which are easily constructed for many representations. Because all geometric operations in the network are defined in terms of these standard matrices, fitted network weights retain the same meaning across different representations. This enables us to train on one representation and evaluate on another as seen in Figure 1 and Section 5.4, cloud, without any special treatment or fine-tuning. In the future, DiffusionNet opens the door to heterogeneous training sets which intermingle mesh, point cloud, and other surface data from various sources.

5.6 Efficiency and Robustness

Runtime. DiffusionNet requires only standard linear algebra operations for training and inference, and is thus straightforward and efficient on modern hardware. As an example, DiffusionNet with spectral acceleration trains on the 14k-vertex RNA meshes (Section 5.2) in 38ms per input and requires 2.2GB of GPU memory. Preprocessing is performed on the CPU once for each input; for these RNA meshes preprocessing takes 5.4sec and generates 12MB of data each, composed mainly of the Laplacian eigenbasis $\Phi$ for spectral acceleration. Table 6 summarizes the runtime performance of DiffusionNet and several other recent methods on the human segmentation task (Section 5.2), including preprocessing, training with gradient computation, and inference. All timings are measured on a 24GB Titan RTX GPU and dual Xeon 5120 2.20GHz CPUs.

Table 6. Runtimes of DiffusionNet and other mesh-based methods across several different input mesh resolutions. Reported times are for one-time preprocessing (pre), a training evaluation with derivatives (train), and an inference evaluation (infer), each on a single input of the specified size. Entries marked by “—” were infeasibly expensive in time or memory usage. Unlike many recent mesh-based learning methods, DiffusionNet easily trains directly on medium-sized inputs, and even scales to very large meshes.

<table>
<thead>
<tr>
<th>Method</th>
<th>small</th>
<th>medium</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>752 vert</td>
<td>10k vert</td>
<td>184k vert</td>
</tr>
<tr>
<td><strong>DiffusionNet</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(spectral)</td>
<td>pre: 288ms</td>
<td>3.55sec</td>
<td>69.5sec</td>
</tr>
<tr>
<td></td>
<td>train: 19ms</td>
<td>25ms</td>
<td>379ms</td>
</tr>
<tr>
<td></td>
<td>infer: 7ms</td>
<td>10ms</td>
<td>154ms</td>
</tr>
<tr>
<td><strong>DiffusionNet</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(direct)</td>
<td>pre: 104ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>train: 329ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>infer: 81ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MeshCNN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Hanocka et al. 2019]</td>
<td>pre: 85ms</td>
<td>1.13sec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>train: 269ms</td>
<td>2.97sec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>infer: 194ms</td>
<td>2.71sec</td>
<td></td>
</tr>
<tr>
<td><strong>HSN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Wiersma et al. 2020]</td>
<td>pre: 905ms</td>
<td>162sec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>train: 188ms</td>
<td>1.08sec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>infer: 68ms</td>
<td>389ms</td>
<td></td>
</tr>
<tr>
<td><strong>HodgeNet</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Smirnov et al. 2021]</td>
<td>pre: n/a</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>train: 752ms</td>
<td>7.61sec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>infer: 645ms</td>
<td>6.87sec</td>
<td></td>
</tr>
</tbody>
</table>

Scaling. Most significantly, DiffusionNet’s efficiency enables direct learning on common mesh data without dramatic simplification, in contrast to other recent mesh-based schemes. As an example, the segmentation meshes from Maron et al. [2017] have up to 13k vertices, yet recent approaches simplify/downsampling to roughly 1k vertices for training, as shown in Figure 2 [Hanocka et al. 2019; Lahav and Tal 2020; Mitchel et al. 2021; Wiersma et al. 2020]. In contrast, our networks easily run at full resolution on this and other datasets, paving the way for adoption in practice and improving accuracy due to preserved details (e.g., in Table 2). We even demonstrate DiffusionNet on a large, 184k vertex raw scan mesh from FAUST—again no special treatment is needed (Table 6, Figure 1).

Robustness. DiffusionNet is also very robust to poor-quality input data; diffusion is a stable smoothing operation, and our method does not require any complex geometry processing operations such as geodesic distance [Masci et al. 2015], edge collapse [Hanocka et al. 2019], parallel transport [Wiersma et al. 2020], or managing pooling hierarchies with upsampling/downsampling. Even the gradient matrix $\mathcal{G}$ is the result of a stable least-squares fit. If desired, techniques like the intrinsic Delaunay Laplacian on meshes can be used to further increase robustness [Bobenko and Springborn 2007; Sharp and Crane 2020], though we do not find it necessary in our experiments. We demonstrate in Figure 1 that DiffusionNet can be applied directly to a low-quality, nonmanifold raw scan mesh without any issues.

6 CONCLUSION

We present a new approach for learning on surfaces that is built by using learned diffusion as the main network component, with spatial gradient features to inject directional information. Our method is very efficient to train and evaluate, is robust to changes in sampling, and even generalizes across representations, in addition to achieving state-of-the-art results on a range of tasks.

Limitations. DiffusionNet is designed to leverage the geometric structure of a surface; consequently it is not automatically robust to topological errors or outliers. In fact, diffusion does not allow any communication at all between distinct components of a surface, leading to nonsensical outputs in the presence of spurious disconnected components (see inset). Subsequent work might mitigate the limitation by combining diffusion with other notions of communication, such as global pooling (à la [Qi et al. 2017a]) or edge convolutions over latent nearest-neighbors [Wang et al. 2019].
Our networks are intentionally agnostic to local discretization, and thus may not be suited for tasks where one learns some property of the local discrete structure, such as denoising or mesh modification. Finally, although our method discourages overfitting to mesh sampling (Section 5.4), it cannot guarantee to totally eliminate it, and we still observe a small drop in performance when transferring between representations—further investigation will seek to close this gap entirely.

Future work: DiffusionNet can be applied to any surface representation for which a Laplacian matrix and spatial gradients can be constructed. This opens the door to directly learning—and even transferring pretrained networks—on a wide variety of surface representations, from occupancy grids [Caissard et al. 2019] to subdivision surfaces [De Goes et al. 2016]. More broadly, DiffusionNet need not be restricted to explicit surfaces, and could easily be adapted to other geometric domains like volumetric meshes, curve networks, implicit level sets, depth maps, or images. We believe that grounding geometric deep learning in the mathematically and computationally well-established diffusion operation will offer benefits across surface learning and beyond.

ACKNOWLEDGEMENTS
The authors thank Adrien Poulenard and Ruben Wiersma for assistance configuring experiments, and Mark Gillespie for suggesting the mirrored training for Figure 5. Parts of this work were generously supported by the Fields Institute for Research in Mathematical Sciences, the Vector Institute for AI, an NSF Graduate Research Fellowship, ERC Starting Grant No. 758800 (EXPROTEA) the ANR (C3SHAPES No. ANR-18-CE36-0001-01) and the German Research Foundation (DFG) under the priority program 1736 (DataSpace). We also thank Adobe. 2016. Adobe Mixamo 3D characters. www.mixamo.com.

REFERENCES


Richard B Lehoucq, Danny C Sorensen, and Chao Yang. 1998. ARPACK users' guide: Solution of Large-Scale Eigenvalue Problems with Arnoldi methods. SIAM.


Zheng Ma, Junyu Xuan, Yu Guang Wang, Ming Li, and Pietro Liò. 2020. Path Integral Coupling for Deep Learning on Surfaces. in CVPR.


A\textsc{AN ARGUMENT FOR GENERALITY}

In Section 3.2, we propose diffusion at various learned timescales followed by a learned pointwise function as the essential components of our method. Although this formulation clearly offers nonlocal support to the pointwise functions, it is not immediately clear how general the resulting function space is. In particular, it is significant to show that this function space includes at least radially symmetric convolutions, a basic building block which has appeared widely in past work. The treatment of radially symmetric convolutions arises because points on surfaces do not generally have canonical tangent coordinates, though it should be noted that recent work has since focused on expanding beyond symmetric filters, and our own method includes gradient features for precisely this purpose. Lemma 1 states that, at least in the flat, continuous setting, this function space is sufficiently general to represent radially symmetric convolutions.

Here, we give a full version of this argument and some discussion. Consider a scalar field $u : \mathbb{R}^2 \to \mathbb{R}$ in the plane. Let $U_i(r) : \mathbb{R}_{>0} \to \mathbb{R}$ denote the integral of the field $u$ along the sphere with radius $r$ centered at $p$, i.e., $U_i(r) = \int_{\partial B_i(r)} u(\mathbf{p}) d\mathbf{p}$. Recall that $U_i(r) : \mathbb{R}_{>0} \to \mathbb{R}$ denotes the value of $u$ at $p$ after diffusion for time $t$. We are interested in $U_i(r)$, because it will enable the evaluation of radially-symmetric convolutions against $u$. The crux of our argument is to show that $U_i(r)$ can be recovered from $u_i(r)$, which will be formalized by showing the existence of a function transform $T : (\mathbb{R}_{>0} \to \mathbb{R}) \to (\mathbb{R}_{>0} \to \mathbb{R})$ such that

$$U_i(r) = T[u_i(r)](r).$$

The heat kernel solution for $u_i(r)$ is given by

$$u_i(t) = \int_{\mathbb{R}^2} u(q) \frac{1}{4\pi t} e^{-\frac{|q-p|^2}{4t}} dq = \int_0^\infty U_i(r) \frac{1}{4\pi r} e^{-\frac{r^2}{4t}} dr,$$

where the second equality moves to a radial integral, recalling that $U_i(r)$ is defined as the integral of $u$ along the sphere of radius $r$ at $p$. Calculation verifies that this integral has the form of a Laplace transform of $U_i(r)$

$$u_i(r) = \int_0^\infty \frac{1}{4\pi t} \mathcal{L} \left[ \frac{1}{2\sqrt{\pi U_i(r)}} \right] \left( \frac{1}{4t} \right).$$
Table 7. An ablation study, evaluated on the human segmentation task. Omitting any of the components of our method leads to a significant drop in performance. Manually fixing a non-optimal diffusion time also impairs performance—our learned procedure automatically optimizes a diffusion time for each channel.

<table>
<thead>
<tr>
<th>Ablation</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>no diffusion</td>
<td>31.4 %</td>
</tr>
<tr>
<td>fixed-time diffusion ( t = 0.1 )</td>
<td>89.1 %</td>
</tr>
<tr>
<td>fixed-time diffusion ( t = 0.5 )</td>
<td>81.6 %</td>
</tr>
<tr>
<td>no gradient features</td>
<td>84.1 %</td>
</tr>
<tr>
<td>unlearned gradient features</td>
<td>85.6 %</td>
</tr>
<tr>
<td>(full method)</td>
<td>90.6 %</td>
</tr>
</tbody>
</table>

The Laplace transform is injective [Lerch 1903], which allows us to consider the inverse transform

\[ U_r(p) = T[u_r(p)](r). \]

And in fact, \( T \) will have the form of an inverse Laplace transform, up to reparameterization by \( \frac{1}{t} \) and constant coefficients.

Now that we have established the existence of \( T \), it is straightforward to evaluate a radially-symmetric convolution via a pointwise map applied to diffused values. Convolution against any radially symmetric kernel \( a(r) : \mathbb{R} \rightarrow \mathbb{R} \) is given by

\[
(u * a)(p) = \int_{\mathbb{R}^2} a(|p - q|)u(q)dq \\
= \int_0^\infty a(r)U_r(p)dr \\
= \int_0^\infty a(r)T[u_r(p)](r)dr.
\]

In this sense, the function space defined by diffusion followed by a pointwise map contains the space of radially symmetric convolutions, completing our argument.

Extending this treatment from \( \mathbb{R}^2 \) to curved manifolds would require a deeper analysis, though the same essential properties hold for diffusion on surfaces. Furthermore, we treat only the continuous setting above, rather than the discrete setting where pointwise maps are approximated via finite-dimensional MLPs, and diffusion is evaluated at a collection of times \( t \). More generally, it would be valuable to extend this analysis to formalize the stability properties of diffusion, à la [Kostrikov et al. 2018; Perlmutter et al. 2020]. Nonetheless, we consider this argument to be important evidence that diffusion followed by pointwise functions is an expressive function space, supported by the strong results of our method in practical experiments.

B ANALYSIS

Ablation. To validate the components of our approach, we consider a simple ablation study on the full-resolution human segmentation task from Section 5.2, using rotation-augmented raw coordinates as input. The variant no diffusion omits the diffusion layer from each DiffusionNet block, fixed-time diffusion manually specifies a diffusion time, no gradient features omits the gradient features, and unlearned gradient features includes gradient features but omits the learned transformation of gradient vectors \( A \). We observe a noticeable drop in accuracy when omitting any of the components of the method (Table 7). Manually specifying shared, non-optimal diffusion times \( (t = 0.1, t = 0.5) \) yields a network with significantly worse accuracy compared to our learned approach. A key advantage of our learned diffusion is that this time is automatically tuned by the optimization process, individually for each feature channel.

Spectral basis size. When evaluating diffusion with spectral acceleration (Section 3.3.2), increasing the size \( k \) of the spectral basis more accurately resolves diffusion at the cost of increased computation. In Figure 13 we vary \( k \) for the FAUST vertex-labeling correspondence task as in Table 5, measuring accuracy on the original test set. We find performance degrades significantly with fewer than 64 eigenvectors on this problem, while larger bases offer negligible benefit—our experiments use \( k = 128 \) eigenvectors as a safe default.

C EXPERIMENT DETAILS

Here we provide additional methodology details for experiments.

Orientation. Figure 5 shows the results of a simple artificial experiment in which we segment the left vs. right side of human models from the FAUST dataset [Bogo et al. 2014] using a purely intrinsic 32-width DiffusionNet with HKS as input. On the original dataset, asymmetric biases—such as a template mesh with asymmetric connectivity—make it unintentionally easy to distinguish left from right. We cancel the effect of these biases by augmenting the dataset with a copy of each mesh that has been mirrored across the left-right axis (preserving orientation by inverting triangles). With a complex-valued \( A \), our network is able to easily distinguish left from right with 99.9% accuracy, despite both a purely intrinsic architecture and intrinsic input features. Restricting to real-valued \( A \) removes the effect; the network is unable to disambiguate the symmetry, with a totally random 50.0% accuracy.

Human segmentation. All results are given in Table 3. Past work has used different variations of this dataset, both in terms of the input data and evaluation criteria. The original dataset presented by Maron et al. [2017] contains moderately large meshes of up to 12K vertices, with segmentations labeled per-face, and accuracy is reported as the fraction of faces in the entire test set which were classified correctly. The experiments from Wiersma et al. [2020]...

![Fig. 13. The effect of varying the size of the truncated basis for spectral diffusion evaluation, measured via error on the FAUST vertex-labeling correspondence from Table 5. orig. We use 128 eigenvectors in all experiments.](image-url)
deviate slightly: they remap the ground truth to vertices, and train and test on a subsampling of the vertex set; nonetheless we group these results with the original dataset for the sake of simplicity as they are very similar.

MeshCNN [Hanocka et al. 2019] generated a simplified version of the dataset where the meshes have < 1k vertices, and segmentations have been remapped to edges. Additionally, when reporting test evaluation, a soft ground truth is used allowing for multiple correct segmentation results for edges at the boundary between two regions. For comparison we also apply DiffusionNet to this variant of the task, denoted by † in Table 3—we directly generate a prediction per-edge by averaging per-vertex outputs to edges before applying the final softmax, and evaluate test results against the same soft ground truth. Finally, PD-MeshNet [Milano et al. 2020] generated per-face labels for the MeshCNN simplified models and trained and tested on these without any soft ground-truth—we denote this variant by ‡ and again evaluate DiffusionNet with per-face predictions.

Across all variants, DiffusionNet achieves highly accurate performance. Unlike many of these methods, DiffusionNet can easily be trained directly on the original meshes without any special treatment. Even methods which evaluate on full-resolution models may be scalable only due to special pre- and post-processing schemes, which add complexity to adoption in practice—for instance, Mesh-Walker [Lahav and Tal 2020] trains on simplified meshes then applies an upsampling and smoothing scheme to handle full resolution data.

Discretization agnostic learning. To investigate robustness to discretization on our remeshed FAUST dataset, we train several recent mesh-based and point-based surface learning methods, in addition to our own 256-width DiffusionNet with dropout. For mesh-based methods, we also train SplineCNN [Fey et al. 2018], ACSCNN [Li et al. 2020b], HSN [Wiersma et al. 2020]; we also tried MeshCNN [Hanocka et al. 2019], but found it prohibitively expensive. For point-based methods, we train PointNet [Qi et al. 2017a] and DGCNN [Wang et al. 2019], and consider using both the vertex set as a point set, as well as sampling a random point cloud on the surface, predicting there, then projecting the results back to vertices according to nearest-neighbors. For equivalent comparison, all models are trained with only vertex positions as input (or the constant function, for ACSCNN and SplineCNN), and we augment during training with random rotations about the vertical axis to encourage rotation-invariance. Wherever possible, we mimic the training configuration of the original work or make a best-effort to find suitable parameters for this task. We note that some models perform slightly worse than previously reported results, presumably due to the use of simpler input features or learning in a rotation-invariant setting rather than aligned.

In general, only DiffusionNet learns accurate correspondences which are robust to remeshing and resampling. In particular, ACSCNN still produces nearly perfect results on the original template meshes even in the rotation-invariant setting, but yields essentially random noise after any remeshing. Perhaps unsurprisingly, point-based methods are less prone to overfitting the mesh connectivity (though the DGCNN on vertices still manages to do so), but are still notably less accurate than mesh-based techniques. Figure 11 gives full geodesic error plots corresponding to Table 5.