DOI: 10.1111/cgf.14599
Eurographics Symposium on Geometry Processing 2022
M. Campen and M. Spagnuolo
(Guest Editors)

Volume 41 (2022), Number 5

Localized Shape Modelling with Global Coherence: An Inverse Spectral Approach

M. Pegoraro[1] and S. Melzi[1,2] and U. Castellani[3] and R. Marin[†1] and E. Rodolà[†1]

[1]Sapienza University of Rome, Italy
[2]Bicocca University of Milan, Italy
[3]University of Verona, Italy

Abstract

Many natural shapes have most of their characterizing features concentrated over a few regions in space. For example, humans and animals have distinctive head shapes, while inorganic objects like chairs and airplanes are made of well-localized functional parts with specific geometric features. Often, these features are strongly correlated – a modification of facial traits in a quadruped should induce changes to the body structure. However, in shape modelling applications, these types of edits are among the hardest ones; they require high precision, but also a global awareness of the entire shape. Even in the deep learning era, obtaining manipulable representations that satisfy such requirements is an open problem posing significant constraints. In this work, we address this problem by defining a data-driven model upon a family of linear operators (variants of the mesh Laplacian), whose spectra capture global and local geometric properties of the shape at hand. Modifications to these spectra are translated to semantically valid deformations of the corresponding surface. By explicitly decoupling the global from the local surface features, our pipeline allows to perform local edits while simultaneously maintaining a global stylistic coherence. We empirically demonstrate how our learning-based model generalizes to shape representations not seen at training time, and we systematically analyze different choices of local operators over diverse shape categories.

CCS Concepts

• ***Computing methodologies*** → ***Shape analysis;*** *Shape representations;*

1. Introduction

Generative modeling of 3D shapes is a fascinating problem that unifies geometry and statistics at their finest. The interest is both theoretical and practical; we aim to better understand the rules lying "under the surface" and, as a direct consequence, to solve real-world problems like automatic content generation, 3D shape reconstruction, and body tracking [CRXZ20, EST*20], among many others. This problem is as compelling as it is challenging, and the research in the field has a long history. For several decades, many linear statistical approaches have been proposed [RDP99, LMR*15, RTB17], until deep learning methods took the stage, unleashing powerful non-linear methods [CRXZ20]. However, even if the results are getting better every day, many underlying properties remain mysterious, hardly interpretable and controllable.

Inspired by recent work in spectral geometry processing, with this paper we propose a new shape modelling paradigm that addresses the following question: How should the global geometry of a shape be changed, to make it coherent with user-defined local edits that modify its semantics?

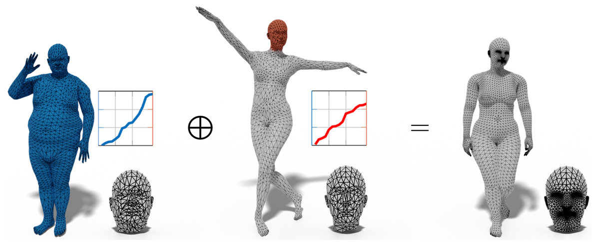


Figure 1: *An example of semantic control provided by our method. By combining a global spectrum (in blue) with a local spectrum (in red) of two shapes with different discretizations, we generate a new shape (on the right) possessing the global characteristics of the blue shape and the local features of the red one.*

Instead of attempting to rigorously define a geometrical notion of 'coherence', which is ultimately subjective, we propose to learn this concept from examples, namely, by finding latent correlations between locally and globally supported geometrical features on a collection of representative shapes. Technically, we achieve this by learning a sum operation between the spectra of standard and

† Equal contribution

localized Laplacians, followed by an inverse learnable solver that recovers a 3D embedding from the combined spectrum. Operating with spectra endows our pipeline with invariance to near-isometries by construction, as well as robustness to mesh discretization, and makes it applicable to any shape representation that admits the definition of a Laplace operator (e.g., point clouds).

We position our work within a recent line of approaches that emphasize the practical potential of using spectra as a rich, albeit compact representation of the shape geometry [CPR*19, RTO*19, MRC*20, RPC*21, MRC*21]. Differently from these approaches, which focus on applications such as shape correspondence, region detection, style transfer and adversarial attacks, here we regard the spectra as *manipulable* representations. Further, instead of operating with the standard Laplacian eigenvalues, we demonstrate for the first time how the combination of multiple spectra from different operators can lead to a practical benefit in shape modelling applications.

While working with multiple spectra at once makes the shape recovery step more challenging (since the network must interpret them all in one shot), we show that mixing spectra also improves the reconstruction quality, as well as generalizing more easily to unseen shapes. In the example of Fig. 1, we combine the Laplacian eigenvalues of the blue shape with the eigenvalues of a Laplace operator localized on the red region (head) of the shape in the middle, generating a new shape that globally reflects the features of the former (height and robustness), but which is coherent with the semantics of the latter (physiological gender). As shown in the example, our method can deal with shapes that do not share the same connectivity or pose. As a final point, we show that this statistical correlation between local spectra and geometry is so strong that it holds even on unorganized point clouds, i.e., with noisy spectra and without known correspondence across the training data.

To summarize, our contribution is threefold:

1. We address the task of enforcing global semantic consistency of 3D geometry, when this undergoes local user edits. We do this in a shape-from-spectrum setup, by proposing a generative model from *multiple* spectra. The combination of global and local information is a novelty of our work, showing that this representation is capable of providing not only better reconstruction, but also new application possibilities;
2. We propose a decoder-only architecture that directly connects the spectrum to the 3D geometry. We show that this simple approach outperforms previous more sophisticated methods, it is more interpretable, and provides new insights on inverse spectral geometric problems;
3. We propose a new dataset designed for analyzing inverse spectral methods, together with new error measures, establishing a sound protocol for evaluating the performance on this task.

Code and data are available online[†].

[†] https://github.com/Marco-Peg/Localized-Shape-Modelling-with-Global-Coherence

2. Related work

Generative models. From a technical perspective, our method can be classified as a generative model due to its ability to generate new shapes by sampling a learned parametric space. In the realm of 3D shapes, existing generative models differ depending on the final application, and on the chosen representation for the 3D geometry. Popular representations include voxels [WZX*16], triangle meshes [RBSB18, TTZ*20, GFK*18], implicit functions [BSTPM20, CYAE*20], and point clouds [QSMG17, ADMG18]. While each representation requires a specific architecture, shapes that undergo non-rigid deformations, and in particular the class of human bodies, have received increasing attention in the recent literature [XBZ*20, BSTPM20, JZCZ20, JSS18].

While most of these works focus on reconstruction quality, only a few have investigated ways to inject semantics in the generation process, in a controllable way. [AATJD19] proposed an autoencoder with a disentangled latent space, enabling a separate control of intrinsic and extrinsic deformations; [CNH*20] showed that plausibility of the generated shapes can be improved by promoting metric preservation in the loss function; [ZBPM20] proposed an unsupervised technique to disentangle shape and pose in the latent space representation; [CKF*21] encoded geometric details as a style property that conditions the refinement of a low-resolution coarse voxel shape through a generative adversarial network (GAN). Other related works addressed the generation of rigid composite-objects [LHW*19, GGC*20, WZX*20, MGY*19, YCC*20] exploiting hierarchical neural network architectures or probabilistic mixture models to manipulate shape parts [ADMG18, HHGCO20, LZZ*21].

Shape from spectrum. Recently, the Laplacian eigenvalues have been used as a compact representation to recover and manipulate 3D geometry. According to a physical interpretation, the eigenfunctions of the Laplace operator on a surface relate to the evolution of waves over it, and the associated eigenvalues are the frequencies of such waves. These are determined uniquely by the intrinsic geometry of the shape, and are fully invariant to isometric deformations. However, the inverse problem (i.e., determining the intrinsic geometry from a set of Laplacian eigenvalues) has been an open question for a long time [Kac66], with the negative result of Gordon and colleagues [GWW92] posing a theoretical tombstone to the problem.

The vision community has recently rediscovered interest in this problem from a practical perspective, with [CPR*19, RTO*19] showing that this inverse problem can be solved through a complex optimization. The recent work [MRC*20], and its extension [MRC*21], replace the costly optimization with a data-driven framework, where a latent encoding is connected with the Laplacian spectrum via trainable maps. At test time, the network can instantaneously recover a shape from its spectrum. While we consider these works the closest to ours by their data-driven nature, the authors of [MRC*20, MRC*21] limit their analysis to the standard Laplacian, without investigating localized operators such as those studied in [MRCB18, CSBK20]. Importantly, the methods of [MRC*20, MRC*21] address the generation problem by designing a network that is hard to interpret, without providing the user with a way to exert control on the desired output.

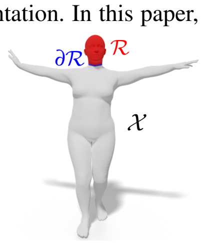
Our method. With this paper, we propose a generative model for 3D shapes that makes full use of the spectrum as an informative, compact, and manipulable representation. Our method is straightforward as it relies upon a simple decoder-only network, and considers a combination of different spectra. The only loss we use is a standard reconstruction loss, without any ad-hoc regularizer. This way, our network is encouraged to discover by itself the hidden mechanism that links a (combined) spectrum to the corresponding 3D shape. Hence, the statistical relations within an object class emerge, providing in turn a better control of the generative process.

3. Background and notation

Smooth setting. In this setting, a 3D shape is modelled as a compact and connected Riemannian surface (2-dimensional manifold) $\mathcal{X}$ embedded in $\mathbb{R}^3$. Each surface $\mathcal{X}$ is equipped with a Laplace-Beltrami operator (LBO) $\Delta_{\mathcal{X}}$, which generalizes the classical Laplacian operator to non-Euclidean domains. From now on, we will refer to this operator as Laplacian to streamline our notation. The Laplacian $\Delta_{\mathcal{X}}$ is a positive semi-definite operator. From its eigendecomposition we obtain the eigenvalues $\{\lambda_\ell\}$ with $\lambda_\ell \in \mathbb{R}, 0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_\infty$, and associated orthogonal eigenfunctions $\{\phi_\ell\}$. When $\mathcal{X}$ is a manifold with boundary $\partial\mathcal{X}$, we consider homogeneous Dirichlet boundary conditions as done in [RTO*19]:

$$\phi_\ell(x) = 0, \ \forall x \in \partial\mathcal{X}\,. \tag{1}$$

The ordered discrete sequence of Laplacian eigenvalues of $\mathcal{X}$ is usually referred to as the *spectrum* of $\mathcal{X}$. When $\mathcal{X}$ is a 1D Euclidean domain, the eigenvalues coincide with the squares of frequencies of Fourier basis functions. Following this analogy, we only consider a truncated spectrum corresponding to the k eigenvalues with smallest absolute values as a band-limited representation. In this paper, we test values of k in the range from 10 to 80.



In this work, we also consider connected submanifolds $\mathcal{R} \subset \mathcal{X}$ with boundary $\partial\mathcal{R}$ (see inset figure for an illustration). The region $\mathcal{R}$ inherits the metric from the complete manifold $\mathcal{X}$, and is similarly equipped with a Laplacian operator $\Delta_{\mathcal{R}}$.

Discrete setting. We discretize smooth surfaces, submanifolds, and their Laplacians in two alternative ways: (i) *Triangle meshes*: $\mathcal{M} = (X, F)$ with n vertices X and m triangular faces F. Submanifolds are constructed as subsets of vertices and faces, with their local connectivity preserved. We adopt the cotangent formula to discretize the Laplacian [PP93]. (ii) *Unorganized point clouds*: $\mathcal{P} = (X)$. Submanifolds are simply subsets of the vertices. We compute the Laplacian using the implementation of [SC20][‡]. In both representations, the matrix $X \in \mathbb{R}^{n\times 3}$ encodes vertex coordinates.

4. Method

This section outlines the proposed method, providing details about its implementation and theoretical insights. A visualization of our pipeline is given in Fig. 2.

‡ https://github.com/nmwsharp/robust-laplacians-py

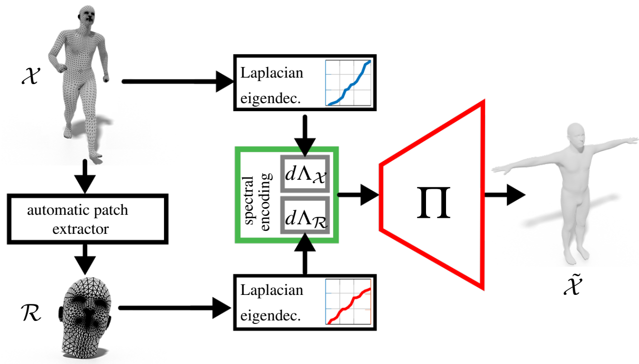


Figure 2: *The proposed model. A neural network Π is trained to reconstruct 3D shapes from spectral encodings, obtained as the simple concatenation of a global and a local spectral encoding. Note the lack of a learnable block between shapes and spectral encoding, which differentiates our model from a classical autoencoder architecture. Therefore, Π directly maps spectral encodings to 3D coordinates, without ever seeing the input region.*

Overall intuition. From a technical perspective, the main idea behind our method is to learn a map that directly translates eigenvalues to 3D coordinates. However, the eigenvalues are a mixture of a *global* spectrum (computed from the classical Laplacian) and a *local* spectrum (computed from a Laplace-like operator localized on a region); crucially, the learnable map does *not* know the region used for computing the local spectrum. This has two important consequences:

- Since the map is required to reconstruct the entire shape as accurately as possible, it must learn on its own that the local spectrum encodes geometric details of some surface region, while the global spectrum encodes the overall geometry.
- By seeing global and local spectra jointly, the network learns how the two interact; namely, it learns to associate changes in the local spectrum with changes in the global one.

In other words, the network learns (i) correlations between eigenvalues and geometric features, and (ii) correlations between global and local spectra.

Both properties have a practical impact. Property (i) allows us to use the eigenvalues directly as a parametric encoding of 3D shapes, without the need to learn a new representation as done with the autoencoder paradigm [MRC*20]. For this reason, we can adopt a simple decoder-only architecture. Further, eigenvalues are interpretable due to their analogy with Fourier analysis, follow a natural ordering, are easy to compute, and are robust to discretization. Property (ii), on the other hand, allows to recover statistical correlations between local details and global shapes. This enables new paradigms for shape modeling, as shown in Figure 1 and several other examples in Section 5.

4.1. Proposed model

Given an input shape $\mathcal{X}$, our pipeline involves the computation of a local spectrum on a given region $\mathcal{R} \subset \mathcal{X}$. Here we detail the selection criteria for the region, and the choice of a localized operator.

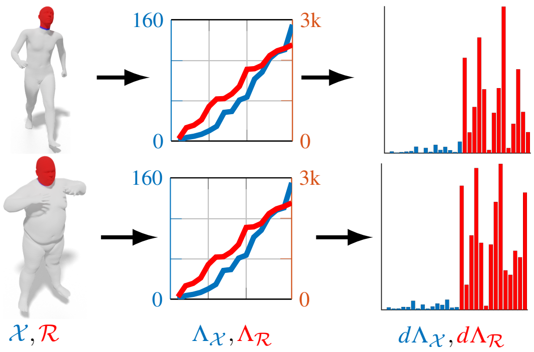


Figure 3: *Two examples of the computation of the proposed spectral encoding.* Left*: Starting shapes and regions (in red);* Middle*: Global and local spectra, in blue and red respectively;* Right*: Bar plots representing the differences of Eq.* (4)*, which compose our encoding.*

Local region. The region is selected to be informative for the final task, and the choice should be coherent across all shapes in the training set. For example, if the application expects the user to modify facial features, then the human head should be included as a region of interest in the training data. Well-established segmentation approaches may be used to select $\mathcal{R}$ automatically, e.g. for man-made objects such as airplanes [KAMC, SACO20, DMB*17, KO19] or for organic shapes such as humans (e.g., see the head extractor proposed in [MMRC20]).

Localized operator. Once a region $\mathcal{R} \subset \mathcal{X}$ has been identified, we compute a localized operator over it and, in turn, its truncated spectrum $\Lambda_{\mathcal{R}}$. Perhaps the most natural choice is to disconnect the region from the rest of the shape, and compute the standard Laplacian on the resulting surface with boundary conditions; we refer to this choice as **PAT**. Other possibilities include the definition of a Hamiltonian operator with a sharp potential [CSBK20] (**HAM**), and the localized manifold harmonics [MRCB18], which yields a Hamiltonian-like operator whose eigenfunctions are orthogonal to the Laplacian eigenbasis (**LMH**). In Section 6 we compare these choices experimentally.

Spectral encoding. Given the (truncated) global spectrum $\Lambda_{\mathcal{X}} \in \mathbb{R}^k$, sorted non-decreasingly, we first compute the differences:

$$d\lambda_\ell = \lambda_\ell - \lambda_{\ell-1}, \; \forall \ell \in 2, \dots, k, \tag{2}$$

where $d\lambda_\ell \geq 0$, $\forall \ell$, and store them in a vector:

$$d\Lambda_{\mathcal{X}} = [d\lambda_2, \dots, d\lambda_k] \in \mathbb{R}^{k-1}. \tag{3}$$

We do the same for the local spectrum $\Lambda_{\mathcal{R}} \in \mathbb{R}^h$, obtaining $d\Lambda_{\mathcal{R}} \in \mathbb{R}^{h-1}$. Finally, we simply concatenate the two vectors $d\Lambda_{\mathcal{X}}$ and $d\Lambda_{\mathcal{R}}$ to generate our *spectral encoding*:

$$d\Lambda = [d\Lambda_{\mathcal{X}} \, ; d\Lambda_{\mathcal{R}}] \in \mathbb{R}^{(k-1)+(h-1)}. \tag{4}$$

An illustration of this computation is depicted in Fig. 3, and is represented in green in the middle of Fig. 2. This encoding exploits the natural hierarchy carried by each set of eigenvalues. We observed that using differences between subsequent eigenvalues, in place of their absolute values, has a regularizing effect that helps training more effectively. Taking differences does not disrupt the geometric information encoded in the spectra, and can be effectively used by the network to recover the original eigenvalues if needed.

Map training. We aim to learn the map Π, which receives as input a spectral encoding $d\Lambda$, and outputs a 3D shape that corresponds to that encoding. Given a collection of training shapes $\{\mathcal{X}_i\}$ from a given class, we compute for each of them the spectral encoding $\{d\Lambda_i\}$ following the process described above. We implement the map Π as a fully-connected decoder, and train it to minimize a standard reconstruction loss:

$$Loss = \sum_i \|\Pi(d\Lambda_i) - X_i\|_F^2 \,. \tag{5}$$

When we deal with point clouds, we replace the Frobenius norm with the Chamfer distance defined in [ADMG18]. We remark that while local regions are involved in the computation of the spectral encoding, we are not using any specific loss to guide the reconstruction of the corresponding local geometry.

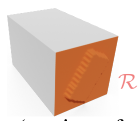
Modeling at test time. Once the model is trained, one can feed the network a previously unseen spectral encoding $d\Lambda$. The resulting 3D shape exhibits the geometric details encoded in $d\Lambda$, but with the discretization of the training set. One can also compose global and local encodings from different shapes, interpolate the encodings, or perform other operations as shown in the next Section.

Network architecture. The proposed network is composed of 4 fully connected layers. We refer to the supplementary materials for further details.

5. Results

5.1. Datasets

CUBE.



This is a synthetic dataset comprising 1000 cube meshes with 7350 vertices each. Each cube shows one of 125 different patterns on its front face, and has a depth picked at random from 8 possible values (see inset for an example). These two factors of variation are uncorrelated and provide a controlled setup for our tests. As region $\mathcal{R}$ on each of these samples, we select the front face since this is where we apply the local variations.

SURREAL. To challenge our model on more realistic data, we collect 2337 human shapes from SURREAL [VRM*17]. As $\mathcal{R}$ on these human bodies, we choose the head for three reasons: 1) it encodes the identity characteristics; 2) it is unique in the body and hard to confuse with other body parts; 3) in contrast with cubes, where pattern and depth are uncorrelated, head and body tend to correlate. We are interested in verifying how this impacts our learning process.

SMAL. We test our model with another example of a realistic dataset: SMAL [ZKJB17], a dataset of 3D mesh animals generated by a morphable model learned by scanning toy figurines. As a dataset, we choose 4872 animals belonging to five different classes:

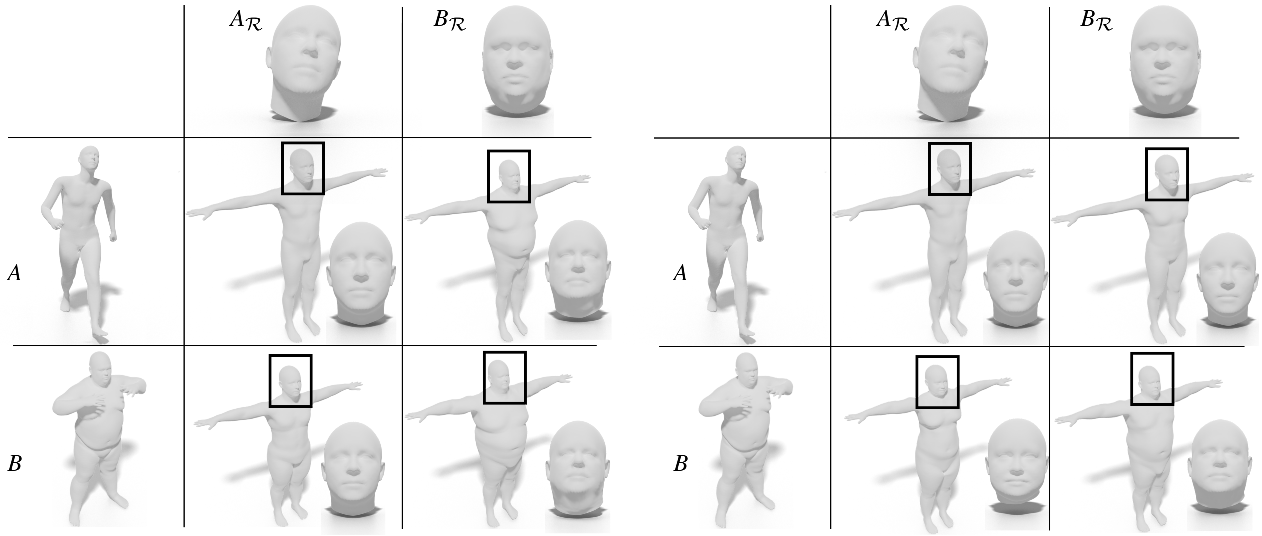


Figure 4: *On the left, semantic swap experiment with PAT15+15. On the right, the same using LBO30.*

tiger, wolf, cow, zebra, hippo. Each mesh has 3889 vertices and represents the animal in the rest pose. As $\mathcal{R}$ we consider the head for the same reasons as above. Like SURREAL, SMAL allows us to test the model in a realistic setting where each local variation may correlate with the rest of the body. Moreover, SMAL is composed of multiple kinds of animals with distinctive and different features. The higher diversity allows us to discriminate better the different contribution of the local and global spectra during reconstruction.

AIRPLANES. The above datasets have a common discretization. To test if our model can discover the underlying relation between spectrum and geometry even in more general cases, we selected 448 airplanes from ShapeNet [YSS*17] and sampled 500 points from each of them, producing unordered point clouds without known correspondence. As $\mathcal{R}$ we selected the tail segment.

We report complete details about the datasets we used in the supplementary materials. If not differently stated, the shapes we adopt in all our experiments and figures have never been seen during training, and belong to the test set or a completely different dataset.

5.2. Shape modeling

Here and throughout the manuscript, we adopt the notation PAT$_\mathcal{R}$15+15 to denote the spectral encoding composed of 15 eigenvalues of the standard (global) Laplacian and 15 eigenvalues of the local PAT operator defined by the region $\mathcal{R}$. Other choices of the operator, local region, and dimensions of the spectral encoding follow the same notation. When the region we are considering is clear and unique, we remove the subscript $\mathcal{R}$, and we only write PAT15+15. As the main baseline, we consider LBO30, which corresponds to the encoding provided by the first 30 eigenvalues of the global Laplacian.

Semantic swap. To better present the impact of our method in shape modeling applications, we provide qualitative examples in different contexts. We start by showing how our method allows to recombine the encoding of different objects to generate novel shapes with natural coherence.

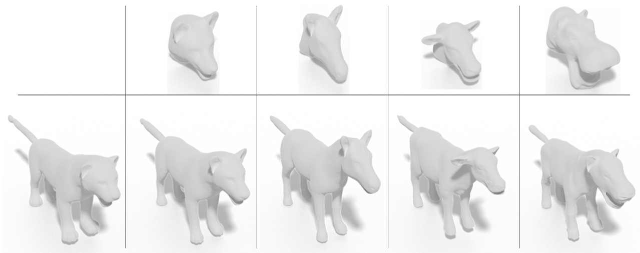


Figure 5: *Shapes obtained by combining the global spectrum of a tiger with the local spectra of the heads of different animals (one per column). Observe the change in body shapes induced by the different heads; in the last column, the body does not change much, but the head has an intermediate shape between tiger and hippo. The operator used in these tests is PAT$15+15$.*

In Figure 4 we show an example on humans, where we consider two shapes, namely A and B, from different datasets([VRM*17] and [PMRMB15], respectively), and fix for both the local region on the head, respectively denoted as $A_\mathcal{R}$ and $B_\mathcal{R}$. On the left we use the map recovered for the PAT15+15 input, whereas on the right we use the map obtained from the standard LBO30 input. On the main diagonal of each grid, we report the reconstruction results from the original encoding of the two shapes. We notice that both approaches return reliable results, generalizing to datasets unseen at training time (such as B).

The off-diagonal entries show the results produced by mixing the spectra of the two inputs. In the top right, we concatenated the first 15 values of the encoding of A with the last 15 of the encoding of B. Both PAT15+15 and LBO30 start with the first 15 eigenvalues

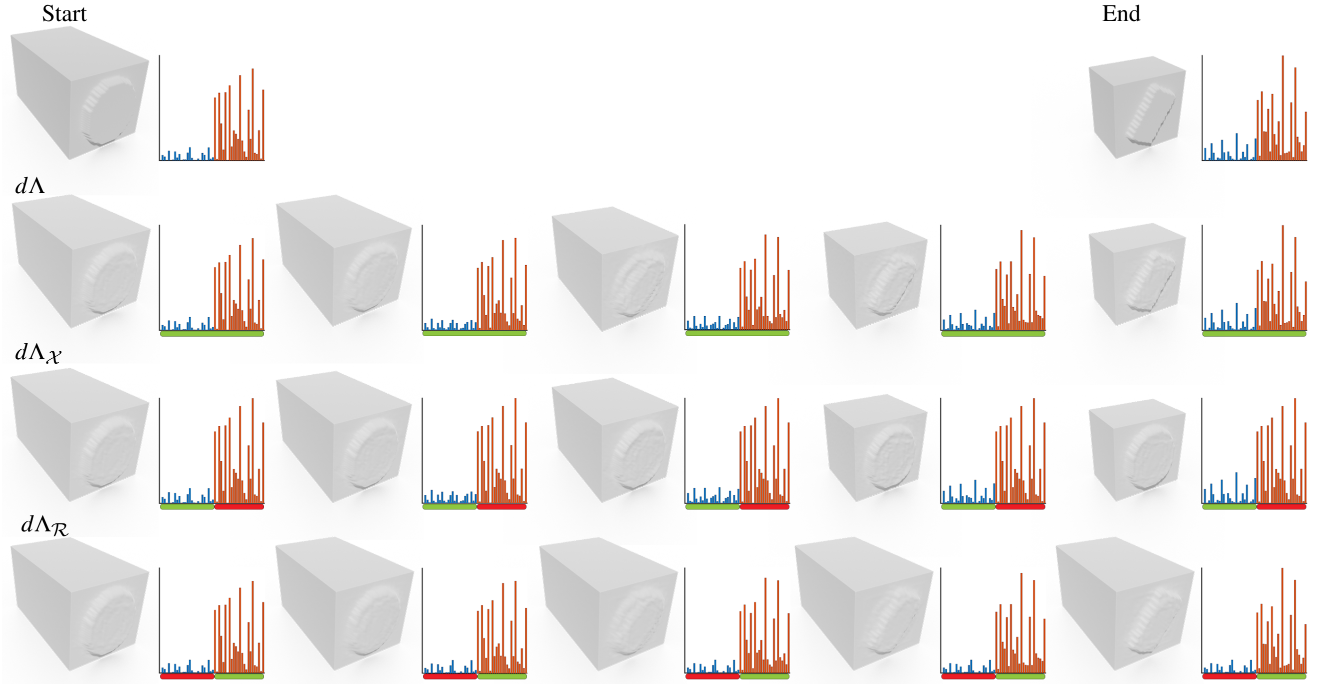


Figure 6: *Interpolation results for a pair of cubes: the left one is taken from the test set, while the rightmost from the train set. We plot the spectral encoding associated with each shape as a bar plot; we highlight in green the part of the encoding that we are interpolating in each row, and in red the values that we keep fixed.* First row*: Two input cubes.* Second row*: Interpolation of the entire encoding.* Third row*: Interpolation of the global part only (blue in the bar plots); observe how the pattern on the front face does not change, while the volume of the entire cube is correctly interpolated.* Last row*: Interpolation of the local part of the encoding (red in the bar plots), inducing a change in the front pattern only.*

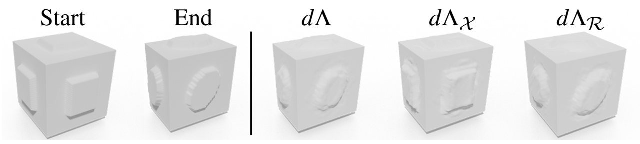


Figure 7: *Final results of the interpolation from Start to End. On the right side we show the interpolation of the entire encoding ($d\Lambda$), of the global part only ($d\Lambda_{\mathcal{X}}$) and of the local part only ($d\Lambda_{\mathcal{R}}$).*

of the LBO from A while for PAT15+15 the second part is given by the first 15 eigenvalues of the localized operator on the region $B_{\mathcal{R}}$ region, while for LBO30 the eigenvalues from 16 to 30 of the LBO of B. In the bottom left, the same with the inverse role of B and A. We notice that PAT15+15 succeeded to produce meaningful modifications but keeping an overall coherence. Injecting into the network the spectra of A and the one of $B_{\mathcal{R}}$ produces a man with similar height and proportions of A, but more robust (while not as robust as B). Using the head of $A_{\mathcal{R}}$ on the body of B has the effect to obtain similar proportions to B (producing a shorter person), but respecting the semantics suggested by $A_{\mathcal{R}}$, which does not suggest a robust human. On the contrary, the grid on the right shows that working with the LBO30 representation does not provide the same level of control.

A clearer example is depicted in Figure 5. We take the global spectrum of a tiger and combine it with the local spectrum (one per column) of all the others animals in the dataset. As before, when we change the local spectrum, the identity of the shape changes. The most peculiar instance is the combination between the global spectrum of the tiger and the local spectrum of the hippo (last column). The resulting mesh has ears similar to the tiger, but has a snout different from both animals. This intermediate result may be due to the dominance of the global features of the tiger that prevent the snout to puff up like in the hippo. The rest of the body remains more similar to the tiger.

Shape space exploration. Our encoding also allows to explore the space of shapes. We report what we believe is a significant example from the CUBE dataset. This dataset has, by construction, two factors of variation: the depth of the cube and the pattern on the front face. These two features have no statistical correlation and are thus fully disentangled by purpose.

In Figure 6 we report the results obtained by interpolating the spectra of two different cubes ('Start' and 'End' respectively). We interpolate the entire encoding (global+local, second row), only the global part (third row) or only the local part (last row). These tests suggest that the learned map has learned to disentangle between the factors of variations in this dataset, which would not be possible by using only a global spectrum. It would also be hard to obtain with a standard autoencoder architecture, unless a disentanglement technique is explicitly implemented. To see how our method can

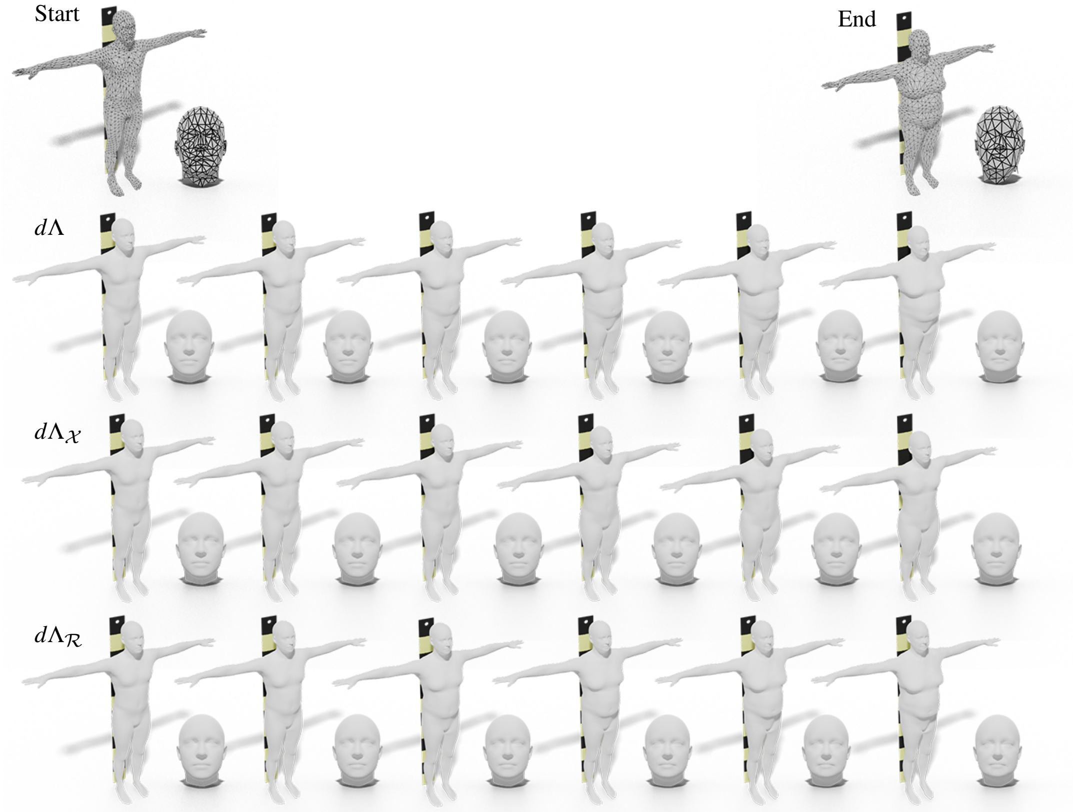


Figure 8: *Different interpolations between two inputs from human shapes (first row): Global+Local ($d\Lambda$-second row), only Global($d\Lambda_{\mathcal{X}}$-third row) and only Local($d\Lambda_{\mathcal{R}}$-fourth row). We add a ruler behind each shape to emphasize the height variation.*

effectively learn correlations between local regions and the whole shapes, we train our network on a different setting where cubes have the same pattern applied to all the faces and the local region $\mathcal{R}$ is a single face as before. On the right of Figure 7, we show the final results of the interpolations from the 'Start' to the 'End' cube (depicted on the left). The interpolation of the global encoding changes only the length of the cube without modifying the pattern on the faces, while the local interpolation changes all the patterns coherently but leaves the cube's depth unchanged. We report the full experiment in the supplementary materials. These two experiments confirm the ability of our encoding to discover the correlation between local details and global variation if present (Figure 7) and to preserve their independence if it is the case (Figure 6). A variation in the local encoding produces more or less localized changes according to the level of dependency between the selected region $\mathcal{R}$ and the rest of the shape. In any case, the generated deformation, even if local, can induce a wider variation to maintain global consistency.

A similar example is depicted in Figure 8. In this example, the two shapes have different discretizations, emphasizing that our method is agnostic to them. We observe that the method discovered a relation between the local (head) region and the global region, as we expected for human data. Interpolating the global spectra change keeps the identity suggested by the head while changing the body proportion (the ruler behind each shape eases the comparison of the heights). Interpolating the head requires keeping the body dimension while changing the subject identity; it is worth noting that the final head is coherent with the target shape.

Finally, our method is efficient at inference time, enabling real-time shapes explorations. We attach a video of interactive navigation via an intuitive interface.

Remark. At training time, our network does *not* know the association between local spectrum and region, but it just sees 30 values without knowing where they originate from. Therefore, it effectively learns what values are responsible for local changes and what are those responsible for the global changes in geometry.

Different discretizations and point clouds. In previous experiments, we require all the shapes in the training set to be in full point-to-point correspondence, which is of great help for the network to discover patterns in the data. To investigate if the relationship between the input eigenvalues and the 3D points is strong enough to arise even without providing point correspondences, we remove the mesh connectivity and consider a noisy scenario of unordered point clouds of airplanes. Differently from [MRC*20], which requires redesigning the encoder to handle unordered point clouds, our model does not require any modification.

We considered the tail segment as the local region. The tails of

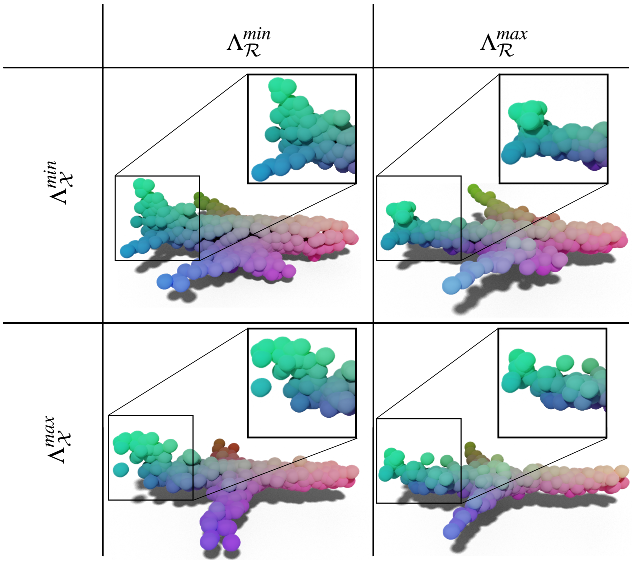


Figure 9: *Minimum and maximum spectra mixing on point clouds (more details in the text).*

airplanes are interesting because their size and shape vary and identify specific categories (such as jet and passenger transport aircraft, among others).

In Fig. 9, we propose a spectra swap similar to Fig. 4. For all the point clouds in the AIRPLANES dataset, we consider the PAT15+15 encoding. For each of the 30 dimensions, we compute the minimum and the maximum among all the 448 shapes. Then we divide the sequence of 30 maxima and the sequence of the 30 minima into global (the first 15 values) and local information (the last 15). We refer to these four output vectors respectively as $\Lambda_{\mathcal{X}}^{\mathbf{max}}$, $\Lambda_{\mathcal{R}}^{\mathbf{max}}$ from the maximum values, and $\Lambda_{\mathcal{X}}^{\mathbf{min}}$, $\Lambda_{\mathcal{R}}^{\mathbf{min}}$ for the minimum ones. We denote the global values with $\mathcal{X}$ and the local ones with $\mathcal{R}$. We then compose 4 new spectra as the possible different combinations of the local and global parts. On the main diagonal, the generated airplanes are a large aircraft associated with the frequencies' minimum values (top left) and a thinner and longer jet for the maximum values (bottom right). As a first observation, this behavior is coherent with our expectations in terms of spectrum-geometry association. The two airplanes exhibit different empennage of the tails: the large aircraft has a conventional tail, while the jet has a T-tile. The kind of airplane determines the shape of the tail, and the coherence between the two would be critical for some applications. What we observe in this case is that the global spectrum $\Lambda_{\mathcal{X}}$ (rows) captures the type of airplane and determines the tail shape. Instead, by editing the local spectrum $\Lambda_{\mathcal{R}}$ (columns), we obtain a variation in tail dimensions, with slight modifications to airplane structure to adapt to this change, but without changing the class.

We consider this a fascinating result because the network is trained *without* a point-to-point correspondence across the training shapes. The spectrum statistics are informative enough to relate spectrum and geometry through the unsupervised Chamfer distance. We report further examples of point clouds in the supplementary materials.

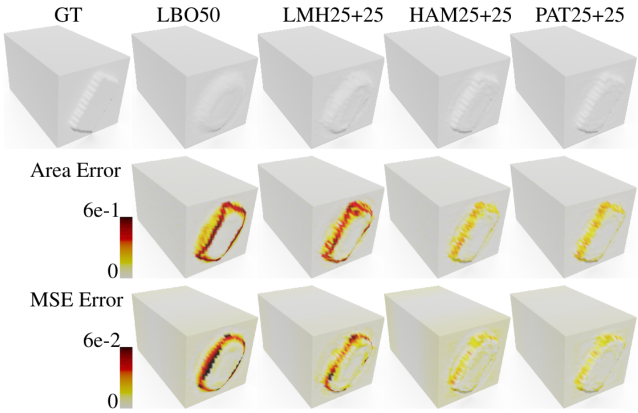


Figure 10: *Qualitative results on the CUBE with extrinsic and intrinsic measures. With MSE, a wrong scale causes an error that accumulates on the cube extremities, while the Area error highlights the wrong pattern extrusion. Errors are color-coded, growing from white to dark red.*

6. Evaluation and analysis

Here we report our analysis of different elements of the method, providing justification of our choices and useful insights.

6.1. Different local operators

In Table 1, we report the performance of our method under different choices of local operators. We consider three different error measures:

1. MSE: mean squared error between the 3D coordinates of our reconstruction and the ground-truth, measured on the entire shape;
2. MSE-$\mathcal{R}$: mean squared error defined as above, measured only on the region $\mathcal{R}$;
3. Area: average difference between the area elements of each vertex of our reconstruction, and the corresponding area elements of the ground-truth shape; this measure quantifies the intrinsic metric distortion caused by the reconstruction module.

The values of MSE and MSE-$\mathcal{R}$ in the table are all multiplied by 10e6, while the Area values are multiplied by 10e3. We notice that PAT maintains the best performance in general, across different dataset and measures. We suppose that this is due to the following reason: PAT is the only operator that is fully localized on the region, because it treats the region as a disconnected component. The other two, HAM and LMH, have both global support and are computed by localizing the LBO by means of a scalar potential, which is known to cause leaking outside of the region [RTO*19]. Moreover, LMH has an additional term of orthogonality that enforces its dependency on the LBO, thus potentially increasing redundancy and, in turn, reducing the amount of information that can be used for a faithful reconstruction.

6.2. Dimension of the embedding

Since previous works do not highlight the relation between the input dimension and the obtained reconstruction, we analyzed the

	CUBE			SURREAL			SMAL		
Method	MSE	MSE-$\mathcal{R}$	Area	MSE	MSE-$\mathcal{R}$	Area	MSE	MSE-$\mathcal{R}$	Area
LBO30	11	62.5	1.96	1.7	2.12	0.82	1.39	1.48	1.90
PAT15+15	**5.66**	**27.1**	2.31	**0.71**	**0.5**	**0.46**	**1.08**	**1.07**	1.63
HAM15+15	6.61	32.87	**1.86**	0.86	0.57	0.51	1.11	1.13	**1.54**
LMH15+15	13.64	61.25	3.54	1.5	0.98	0.71	1.78	1.9	1.93

Table 1: *In all settings we considered* 30 *dimensional encodings; for LBO we used the first* 30 *eigenvalues, for PAT, HAM, and LMH we considered the first* 15 *eigenvalues from LBO and the first* 15 *from the local operators.*

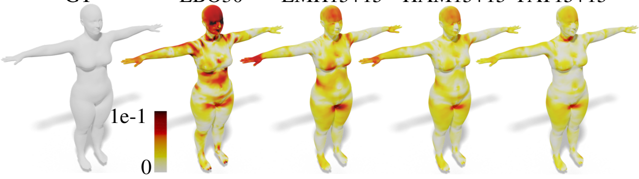


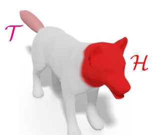
Figure 11: *Qualitative result on the SURREAL. The Area error is shown on the reconstructed surfaces, encoded by color growing from white to dark red.*

	CUBE		
Method	MSE	MSE-$\mathcal{R}$	Area
LBO30	11	62.5	1.96
LBO50	10.7	62	1.80
LBO80	11.1	63.5	1.73
PAT15+15	5.66	27.1	2.31
PAT25+25	**3.59**	**19.1**	**1.33**
HAM25+25	4.07	20.1	**1.33**
LMH25+25	14.7	69.2	1.96

Table 2: *Reconstruction error on CUBE at varying size of input*

method scalability for the different operators. In Table 2 we report the results with LBO at varying dimensions. We see that the instability of higher frequencies dramatically impacts the scalability, preventing further improvement. Instead, considering local operators, both PAT and HAM show improvement when the encoding grows in size. In the supplementary, we report further results on different ratios between global and local information, showing that, in general, an even splitting between the two provides the best results.

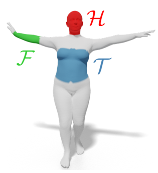
6.3. Number and kind of local regions



In this section, we investigate the importance of the choice of the selected region. Previously, we consider a unique region $\mathcal{R}$ for each class, claiming that it characterizes the objects. In particular, we stated that this selection lets the network correlate between local geometric patterns and global features of the shape, and that this relation respects some semantics. In the SMAL dataset, we tested two different regions: head $\mathcal{H}$ and tail $\mathcal{T}$, respectively depicted in red and pink in the inset figure, and as PAT$_{\mathcal{H}}$15+15 and PAT$_{\mathcal{T}}$15+15 in Table 3. With respect to LBO30, with PAT$_{\mathcal{H}}$15+15 the error decreases, while PAT$_{\mathcal{T}}$15+15 worsens significantly. We justify this behavior with the absence of enough high-frequency details on the tail to produce an informative spectrum for the entire shape.

	SURREAL		SMAL	
Method	MSE	Area	MSE	Area
LBO30	1.7	0.82	1.39	1.90
PAT$_{\mathcal{H}}$15+15	**0.71**	**0.46**	**1.08**	**1.63**
PAT$_{\mathcal{F}}$15+15	4.33	1.3		
PAT$_{\mathcal{T}}$15+15	1.71	0.94	3.93	3.05
PAT$_{\mathcal{H}+\mathcal{T}}$ 10+10+10	0.89	0.56	1.96	2.37
PAT$_{\mathcal{H}+\mathcal{T}}$ 15+15+15	*0.68*	0.48	1.82	2.24

Table 3: *Reconstruction error on SURREAL and SMAL test sets considering various regions*



In the same spirit of the previous experiment, we validate this hypothesis on the SURREAL dataset, selecting another region with few features, i.e., in which the details are less characterizing. We choose the right forearm as local region $\mathcal{F}$ on the human body (highlighted in green in the inset figure), and we trained the model PAT$15+15$. We refer to this method as PAT$_{\mathcal{F}}15+15$, while for the method with the head as region $\mathcal{H}$ we use the notation PAT$_{\mathcal{H}}15+15$. In the third row of Table 3, we report the errors for this choice. Similar to the tail region in SMAL, the errors are higher than PAT$_{\mathcal{H}}15+15$ and LBO30. We also test PAT$15+15$ with the torso as region $\mathcal{T}$, (highlighted in blue in the inset figure) and refer to it as PAT$_{\mathcal{T}}15+15$. The torso region comprises both the chest (that can significantly change between men and women) and the abdomen (that is larger or thinner in different subjects). $\mathcal{T}$ can be comparable to the tail region but with more continuous and significant variations. In all cases, on SURREAL, *PAT*$15+15$ performs significantly better. These experiments emphasize the importance of the input representation and that some local regions contain more information on the whole shape than others. In particular the torso region, besides having important features, has a central position in the shape. Thus its improvement affects the MSE of all the other body parts. On the contrary, the tail region on the SMAL dataset has indeed similar important features but it is a peripheral region and its improvement does not flow on other regions. The elbow instead is a region with poor features that do not add enough information to the global spectrum.

Multi-region. Until now, we experimented only using a single local spectrum as input. To underline how our method can be used with multiple regions, we perform additional experiments on SURREAL and SMAL without modifying the parameters of the decoder. In SMAL, we have already tested the single efficacy of the head and tail. In the next experiment, we test their efficacy when combined together. We train the PAT version using as input the local spectra computed both on the head and tail surface. We indicate the results with PAT$_{\mathcal{H}+\mathcal{T}}k+h'+h''$ where k is the number of LBO eigenvalues, h' the number of local eigenvalues for the first region (head) and h'' the number of local eigenvalues for the second region (tail for SMAL and torso for SURREAL). In particular, we consider the cases $10+10+10$ and $15+15+15$. Notice that in this second case the input encoding is larger then other methods. In the last two rows of Table 3 we show the instrinsic and extrinsic errors. The addition of multiple regions does not improve the performance of PAT$_{\mathcal{H}}$15+15, but neither worsens them as in PAT$_{\mathcal{T}}$15+15.

Our conclusions are that including other regions can be done, giving more freedom during the generation. At the same time, there is a trade-off between the control and the reconstruction quality, since increasing the number of regions limits the amount of encoding assigned to each part (PAT$_{\mathcal{T}}$10+10+10). Discovering the patterns across multiple regions is more challenging; it requires a design in the encoding partition (i.e., assigning to each region eigenvalues proportional to its semantic significance, as we report in supplementary materials) or a deeper network to properly exploit the encoding information. Even when we increased the number of total eigenvalues in input (PAT$_{\mathcal{H}+\mathcal{T}}$ 15+15+15) the performance did not always improve. We attribute this slight drop to the fact that, for a fair comparison, we trained all the models with the same number of epochs, while PAT$_{\mathcal{H}+\mathcal{T}}$ 15+15+15 may have needed more time to optimize the larger input information effectively.

6.4. Decoder-only vs Autoencoder

In Table 4, we compare our architecture (32M parameters) against the one proposed in [MRC*20] (9M parameters) also by empowering its decoder and, coherently, its encoder (namely [MRC*20]$_{big}$ in the table, 90M parameters). We trained [MRC*20] using the absolute value of eigenvalues, as proposed in the original paper. Results show not only that our decoder approach is better than the full architecture even in the LBO setting, but that [MRC*20] is not equally capable of combining local information with its latent space.

7. Conclusions

This paper presented a novel approach for generating and modeling 3D shapes from a canonical and ubiquitous spectral representation. We consider this theoretically exciting task helpful for shape manipulation, especially in combining semantic characteristics of local and global parts. We highlighted several properties of local spectral operators and their relation with the standard Laplacian in

	SURREAL		
Method	MSE	MSE-$\mathcal{R}$	Area
LBO30	1.7	2.12	8.19
PAT15+15	**0.71**	**0.5**	**4.58**
[MRC*20] PAT15+15	3.1	3.89	19.81
[MRC*20]$_{big}$ PAT15+15	2.51	1.8	16.9
[MRC*20]$_{big}$ LBO30	2.51	2.26	15.74

Table 4: *In all settings we considered* 30 *dimensional encodings; for LBO we used the first* 30 *eigenvalues, for PAT, HAM, and LMH we considered the first* 15 *eigenvalues from LBO and the first* 15 *from the local operators computed on the head region.*

this encoding. For the first time, we performed a shape from spectrum pipeline from a mix of spectra of different operators. Furthermore, we have also shown the close relationship between spectra and the geometry from which they come, even in a noisy and unsupervised scenario. Dealing with multi-region localization would be a stimulating future direction, considering parts with different semantics and proportions. While a complete analysis is beyond this paper's scope, our preliminary evidence suggests this is a promising field for further exploration.

The main theoretical limitation of our study is not considering the spectra of extrinsic operators, like the Dirac operator [LJC17]. Extrinsic operators could be successfully injected into our pipeline, providing a mixture of intrinsic and extrinsic information. Moreover, we do not consider inter-class experiments since the spectrum may be ambiguous among different classes. We believe that our work may elicit discussion in the community on these topics. The main applicative limitations arise from the limitation of the intrinsic spectral representations in the presence of shapes with different topology, significant noise, or outliers. The research on these aspects is quite active in the community, and our method lends itself well to methodological progress.

Acknowledgements

Parts of this work were supported by the ERC Starting Grant No. 802554 (SPECGEO), the SAPIENZA BE-FOR-ERC 2020 Grant (NONLINFMAPS) and the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of the Sapienza University of Rome and the University of Verona.

References

[AATJD19] Aumentado-Armstrong T., Tsogkas S., Jepson A., Dickinson S.: Geometric disentanglement for generative latent shape models. In *International Conference on Computer Vision (ICCV)* (2019). 2

[ADMG18] Achlioptas P., Diamanti O., Mitliagkas I., Guibas L.: Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning* (2018), pp. 40–49. 2, 4

[BSTPM20] Bhatnagar B. L., Sminchisescu C., Theobalt C., PONS-MOLL G.: Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)* (2020). 2

[CKF*21] CHEN Z., KIM V. G., FISHER M., AIGERMAN N., ZHANG H., CHAUDHURI S.: Decor-gan: 3d shape detailization by conditional refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 15740–15749. 2

[CNH*20] COSMO L., NORELLI A., HALIMI O., KIMMEL R., RODOLÀ E.: LIMP: Learning Latent Shape Representations with Metric Preservation Priors. In *European Conference on Computer Vision (ECCV)* (2020). 2

[CPR*19] COSMO L., PANINE M., RAMPINI A., OVSJANIKOV M., BRONSTEIN M. M., RODOLÀ E.: Isospectralization, or how to hear shape, style, and correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 7529–7538. 2

[CRXZ20] CHAUDHURI S., RITCHIE D., XU K., ZHANG H.: Learning generative models of 3d structures. *Computer Graphics Forum (CGF) 39*, 2 (2020), 643–666. 1

[CSBK20] CHOUKROUN Y., SHTERN A., BRONSTEIN A., KIMMEL R.: Hamiltonian operator for spectral shape analysis. *IEEE Transactions on Visualization and Computer Graphics 26*, 2 (2020), 1320–1331. 2, 4

[CYAE*20] CAI R., YANG G., AVERBUCH-ELOR H., HAO Z., BELONGIE S., SNAVELY N., HARIHARAN B.: Learning gradient fields for shape generation. In *European Conference on Computer Vision(ECCV)* (2020). 2

[DMB*17] DENITTO M., MELZI S., BICEGO M., CASTELLANI U., FARINELLI A., FIGUEIREDO M. A. T., KLEIMAN Y., OVSJANIKOV M.: Region-based correspondence between 3d shapes via spatially smooth biclustering. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 4270–4279. 4

[EST*20] EGGER B., SMITH W. A., TEWARI A., WUHRER S., ZOLLHOEFER M., BEELER T., BERNARD F., BOLKART T., KORTYLEWSKI A., ROMDHANI S., ET AL.: 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG) 39*, 5 (2020), 1–38. 1

[GFK*18] GROUEIX T., FISHER M., KIM V. G., RUSSELL B., AUBRY M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2018). 2

[GGC*20] GADELHA M., GORI G., CEYLAN D., MECH R., CARR N., BOUBEKEUR T., WANG R., MAJI S.: Learning generative models of shape handles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). 2

[GWW92] GORDON C., WEBB D. L., WOLPERT S.: One cannot hear the shape of a drum. *Bulletin of the American Mathematical Society 27*, 1 (1992), 134–138. 2

[HHGCO20] HERTZ A., HANOCKA R., GIRYES R., COHEN-OR D.: Pointgmm: A neural gmm network for point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12054–12063. 2

[JSS18] JOO H., SIMON T., SHEIKH Y.: Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8320–8329. 2

[JZCZ20] JIANG B., ZHANG J., CAI J., ZHENG J.: Disentangled human body embedding based on deep hierarchical neural network. 2

[Kac66] KAC M.: Can one hear the shape of a drum? *The american mathematical monthly 73*, 4P2 (1966), 1–23. 2

[KAMC] KALOGERAKIS E., AVERKIOU M., MAJI S., CHAUDHURI S.: 3d shape segmentation with projective convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6630–6639. 4

[KO19] KLEIMAN Y., OVSJANIKOV M.: Robust structure-based shape correspondence. *Comput. Graph. Forum 38*, 1 (2019), 7–20. 4

[LHW*19] LIU X., HAN Z., WEN X., LIU Y.-S., ZWICKER M.: L2g auto-encoder: Understanding point clouds by local-to-global reconstruction with hierarchical self-attention. In *Proceedings of the 27th ACM International Conference on Multimedia* (2019). 2

[LJC17] LIU H.-T. D., JACOBSON A., CRANE K.: A dirac operator for extrinsic shape analysis. *Computer Graphics Forum 36*, 5 (2017), 139–149. 10

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graph. 34*, 6 (2015), 248:1–248:16. 1

[LZZ*21] LUO Z., ZHOU J., ZHU H., DU D., HAN X., FU H.: Simpmodeling: Sketching implicit field to guide mesh modeling for 3d animalmorphic head design. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (2021), pp. 854–863. 2

[MGY*19] MO K., GUERRERO P., YI L., SU H., WONKA P., MITRA N., GUIBAS L.: Structurenet: Hierarchical graph networks for 3d shape generation. *ACM Transactions on Graphics (TOG), Siggraph Asia 2019 38*, 6 (2019), Article 242. 2

[MMRC20] MARIN R., MELZI S., RODOLÀ E., CASTELLANI U.: Farm: Functional automatic registration method for 3d human bodies. *Computer Graphics Forum 39*, 1 (2020), 160–173. 4

[MRC*20] MARIN R., RAMPINI A., CASTELLANI U., RODOLÀ E., OVSJANIKOV M., MELZI S.: Instant recovery of shape from spectrum via latent space connections. In *International Conference on 3D Vision (3DV)* (2020). 2, 3, 7, 10

[MRC*21] MARIN R., RAMPINI A., CASTELLANI U., RODOLÀ E., OVSJANIKOV M., MELZI S.: Spectral shape recovery and analysis via data-driven connections. *International Journal of Computer Vision* (2021), 1573–1405. 2

[MRCB18] MELZI S., RODOLÀ E., CASTELLANI U., BRONSTEIN M. M.: Localized manifold harmonics for spectral shape analysis. *Computer Graphics Forum 37*, 6 (2018), 20–34. 2, 4

[PMRMB15] PONS-MOLL G., ROMERO J., MAHMOOD N., BLACK M. J.: Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH) 34*, 4 (Aug. 2015), 120:1–120:14. 5

[PP93] PINKALL U., POLTHIER K.: Computing discrete minimal surfaces and their conjugates. *Experimental mathematics 2*, 1 (1993), 15–36. 3

[QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 652–660. 2

[RBSB18] RANJAN A., BOLKART T., SANYAL S., BLACK M. J.: Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)* (2018). 2

[RDP99] ROBINETTE K. M., DAANEN H., PAQUET E.: The caesar project: a 3-d surface anthropometry survey. In *Proc. Second International Conference on 3-D Digital Imaging and Modeling* (Washington, DC, USA, oct 1999), IEEE, pp. 380–386. 1

[RPC*21] RAMPINI A., PESTARINI F., COSMO L., MELZI S., RODOLÀ E.: Universal spectral adversarial attacks for deformable shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2021), pp. 3216–3226. 2

[RTB17] ROMERO J., TZIONAS D., BLACK M. J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6 (Nov. 2017). 1

[RTO*19] RAMPINI A., TALLINI I., OVSJANIKOV M., BRONSTEIN A. M., RODOLÀ E.: Correspondence-free region localization for partial shape similarity via hamiltonian spectrum alignment. In *International Conference on 3D Vision (3DV)* (2019). 2, 3, 8

[SACO20] SHARP N., ATTAIKI S., CRANE K., OVSJANIKOV M.: Diffusion is all you need for learning on surfaces, 2020. arXiv:2012.00888. 4

[SC20] SHARP N., CRANE K.: A Laplacian for Nonmanifold Triangle Meshes. *Computer Graphics Forum (SGP) 39*, 5 (2020). 3

[TTZ*20] TRETSCHK E., TEWARI A., ZOLLHÖFER M., GOLYANIK V., THEOBALT C.: DEMEA: Deep Mesh Autoencoders for Non-Rigidly Deforming Objects. *European Conference on Computer Vision (ECCV)* (2020). 2

[VRM*17] VAROL G., ROMERO J., MARTIN X., MAHMOOD N., BLACK M. J., LAPTEV I., SCHMID C.: Learning from synthetic humans. In *CVPR* (2017). 4, 5

[WZX*16] WU J., ZHANG C., XUE T., FREEMAN W. T., TENENBAUM J. B.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems* (2016), pp. 82–90. 2

[WZX*20] WU R., ZHUANG Y., XU K., ZHANG H., CHEN B.: Pq-net: A generative part seq2seq network for 3d shapes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 2

[XBZ*20] XU H., BAZAVAN E. G., ZANFIR A., FREEMAN W. T., SUKTHANKAR R., SMINCHISESCU C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6184–6193. 2

[YCC*20] YIN K., CHEN Z., CHAUDHURI S., FISHER M., KIM V. G., ZHANG H.: Coalesce: Component assembly by learning to synthesize connections. In *2020 International Conference on 3D Vision (3DV)* (2020), IEEE, pp. 61–70. 2

[YSS*17] YI L., SHAO L., SAVVA M., HUANG H., ZHOU Y., WANG Q., GRAHAM B., ENGELCKE M., KLOKOV R., LEMPITSKY V., ET AL.: Large-scale 3d shape reconstruction and segmentation from shapenet core55. *arXiv preprint arXiv:1710.06104* (2017). 5

[ZBPM20] ZHOU K., BHATNAGAR B. L., PONS-MOLL G.: Unsupervised shape and pose disentanglement for 3d meshes. In *European Conference on Computer Vision (ECCV)* (August 2020). 2

[ZKJB17] ZUFFI S., KANAZAWA A., JACOBS D., BLACK M. J.: 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (July 2017). 4