ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447519

SC-MAD: MIXTURES OF HIGHER-ORDER NETWORKS FOR DATA AUGMENTATION

Madeline Navarro and Santiago Segarra

Electrical and Computer Engineering, Rice University, USA

ABSTRACT

The myriad complex systems with multiway interactions motivate the extension of graph-based pairwise connections to higher-order relations. In particular, the simplicial complex has inspired generalizations of graph neural networks (GNNs) to simplicial complex-based models. Learning on such systems requires large amounts of data, which can be expensive or impossible to obtain. We propose *data augmentation of simplicial complexes through both linear and nonlinear mixup mechanisms* that return mixtures of existing labeled samples. In addition to traditional pairwise mixup, we present a convex clustering mixup approach for a data-driven relationship among several simplicial complexes. We theoretically demonstrate that the resultant synthetic simplicial complexes interpolate among existing data with respect to homomorphism densities. Our method is demonstrated on both synthetic and real-world datasets for simplicial complex classification.

***Index Terms*—** Simplicial complex, complexon, data augmentation, mixup, convex clustering

1. INTRODUCTION

Simplicial complexes unlock useful topological tools for data science [1–5] and practical applications [6, 7] due to their ability to model higher-order interactions. Simplicial complex-based learning has received much attention lately, with the classical graph-based architectures naturally being extended to higher-order networks [8–11]. However, graph datasets suffer from limited data due to the complexity of obtaining labeled samples, a problem which is exacerbated for higher-order simplicial complex data.

Data augmentation enables generating synthetic labeled samples from existing data, where the new samples embody characteristics that promote desirable model behavior. This procedure is not affected by any machine learning model restrictions as we merely add to the samples present in the dataset, affecting neither model capacity nor the original data [12, 13]. Mixup serves as an efficient data augmentation method that generates new labeled data as mixtures of existing samples [13], and its benefits enjoy copious empirical and theoretical validation [14, 15].

While graph mixup is still nascent, it has exploded in popularity due to the myriad interesting approaches for interpolating such discrete complex objects [16–18]. However, data augmentation for higher-order networks is extremely limited [19], and to the best of our knowledge mixup for higher-order networks has never been considered. Indeed, even in the case of graphs, interpolation of these non-Euclidean objects is nontrivial due to their irregular structure. This difficulty extends further for simplicial complexes as we must obtain mixtures accounting not only for interconnected entities but also for information shared across dimensions. Even data augmentation methods amenable to discrete graph objects struggle as higher dimensions are considered [19, 20]. We are thus prompted to turn to the attractive approach of performing mixup in a continuous latent embedding space. The choice and design of this embedding space allow us to control which characteristics are preserved during the mixup process.

Defining limits of discrete objects enables useful operations for moving within a space of objects as if they are continuous. The invention of the graphon, the limit object of a convergent sequence of dense graphs, provides a compact continuous space in which the graphs are dense [21]. Graphons allow us to perform tasks on graph data typically restricted to continuous objects, such as barycenter obtention and interpolation for mixup [16, 17, 22]. We can extend this benefit to higher-order networks through the complexon [23], an analogous limit object for simplicial complexes.

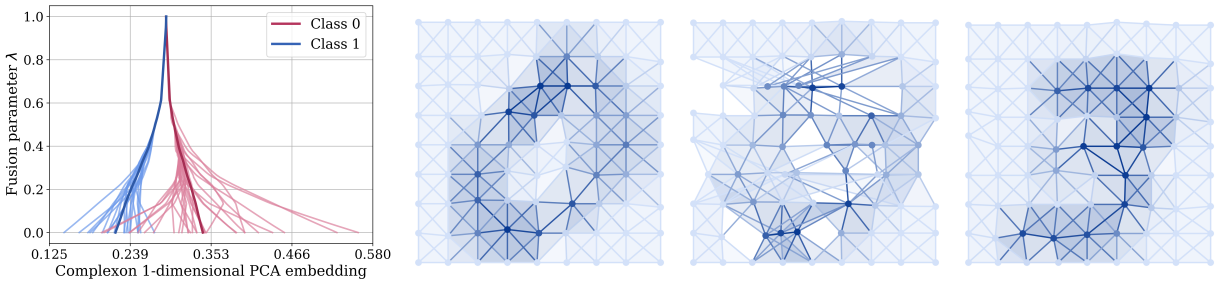
In this work, we present an inaugural method for Simplicial Complex Mixup for Augmenting Data (SC-MAD). Similarly to existing graph mixup methods [16–18], we consider a continuous embedding space for the practical implementation of simplicial complex mixup. We use the space of complexons, as its being the closure of the space of simplicial complexes means that we can directly compare objects in the original and embedding spaces. Furthermore, we theoretically show that any continuous interpolant that our method obtains preserves useful structural characteristics [21, 23]. In addition to traditional pairwise linear mixup [13], we apply convex clustering for mixup [24–26], where new samples describe the mixture of several simplicial complexes [17].

2. PRELIMINARIES

Simplicial complexes. A simplicial complex K is a finite collection of finite sets of elements, or simplices, that are closed under restriction, that is, for every subset $\sigma \in K$, all strict subsets $\sigma' \subset \sigma$ must also be in K [27]. We let $K^{(d)} \subseteq K$ denote the subset of K containing simplices in K with cardinality $d+1$, which are said to have dimension d. The dimension of a simplicial complex K is d, where d is the largest dimension for which $K^{(d)}$ is not empty. We may view the subset $K^{(0)}$ as the nodes in K, $K^{(1)}$ as edges, $K^{(2)}$ as triangles, and so on. We further define the degree of node i at dimension d of K as $D_i^{(d)} := |\{\sigma \in K^{(d)} : i \in \sigma\}|$.

For a pair of simplicial complexes F and K, the homomorphism density of F in K is $t(F, K) = \hom(F, K)/|F^{(0)}|^{|K^{(0)}|}$, where $\hom(F, K)$ denotes the number of homomorphisms from F to K [23]. Intuitively, $t(F, K)$ represents the number of occurrences of F in K while preserving simplices.

This work was partially supported by the NSF under award CCF-2008555. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-17-S-0002. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Army or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. The authors thank T. Mitchell Roddenberry for enriching discussions around the convergence of complexons. Emails: nav@rice.edu, segarra@rice.edu



(a) (b) (c) (d)

Fig. 1: Linear and convex clustering mixup of complexons. (a) Clusterpath from complexons estimated from two sets of Vietoris-Rips complexes, where each complex is formed from i.i.d. points sampled from one of two shapes in $\mathbb{R}^2$, a circle and a figure eight. As λ increases, complexes generated from the same shape fuse together before both shapes coalesce, but we maintain knowledge of the overall spread of data, as Class 1 fuses before Class 0. (b) Superpixel simplicial complex for MNIST digit 0. (c) Superpixel simplicial complex sampled from the complexon $W_{\text{new}} = 0.5\hat{W}_0 + 0.5\hat{W}_3$, where $\hat{W}_0$ and $\hat{W}_3$ denote complexons estimated from the MNIST digits 0 and 3. (d) Superpixel simplicial complex for MNIST digit 3.

Mixup for graph data augmentation. Mixup has enjoyed well-deserved popularity as an intuitive and efficient data augmentation method [13]. The classical mixup method obtains new samples as convex combinations of pairs of samples from different classes. Variants have been proposed for several domains and applications, including graph mixup [16, 17], interpolation in an embedding space [28], and nonlinear implementations [17].

Despite the rapid development of mixup for graphs, it remains difficult due to their non-Euclidean nature, so mixup in a continuous embedding space remains a popular approach [16–18]. However, projection from the graph domain onto a lower-dimensional space may lose critical semantic information, and the potential for information loss is even greater for higher-order networks, for which there are far fewer data augmentation methods [19].

Limit objects for networks. The increasing presence of large graphs, such as the internet, motivates the concept of graph limits. The graphon was thus introduced as the limit of a convergent sequence of dense graphs [21]. Simplicial complex sequences have recently received an analogous limit object known as the complexon [23]. Formally, a complexon W is a measurable function

$$W : \bigsqcup_{d\geq 1} [0,1]^{d+1} \to [0,1]$$

that is symmetric in its coordinates at every dimension. We represent the complexon at dimension d as $W^{(d)}$, and we may view a graphon as a complexon of dimension 1, $W^{(1)}$. Similarly to graphons [21], complexons not only represent limit objects but also can be used as a generative model to sample new simplicial complexes [23].

We may also define the homomorphism density of the simplicial complex F in the complexon W as

$$t(F,W) = \int_{[0,1]^{F^{(0)}}} \prod_{\sigma\in F} W(\zeta_\sigma) d\zeta,$$

where ζ_σ corresponds to indexing $(\zeta_1,\ldots,\zeta_{|F^{(0)}|}) \in [0,1]^{F^{(0)}}$ by $\sigma \subseteq F^{(0)}$ [23].

3. METHODOLOGY

Given a dataset of labeled simplicial complexes $\mathcal{D} = \{(K_i, y_i)\}_{i=1}^T$, we aim to generate a synthetic dataset $\mathcal{D}' = \{(K_i', y_i')\}_{i=1}^{T'}$ such that a classifier f trained on the augmented dataset $\mathcal{D} \cup \mathcal{D}'$ achieves a higher accuracy predicting the labels of unseen samples compared to training solely on the original dataset $\mathcal{D}$. We present SC-MAD for simplicial complex data augmentation following a three-step procedure: (1) We embed the existing simplicial complexes onto a continuous space, which we select as the space of complexons [23], (2) we perform mixup either via the efficient pairwise linear mixup [13] or the more informative convex clustering mixup [17], and (3) we sample complexon mixtures from the interpolants obtained from mixup and generate new simplicial complexes from those mixtures.

We now discuss the intuition behind the complexon as the embedding space. Step (1) of SC-MAD is common for mixup methods, where samples are interpolated in an embedding space [16, 17, 28]. The choice of embedding space is adaptable to a user's desired preserved characteristics when obtaining mixtures, and the complexon is a natural choice for the continuous treatment of simplicial complexes. First, as a Euclidean object, it enjoys amenability to interpolation for mixup. Second, the complexon can be used as a random simplicial complex model, representing a family of simplicial complexes [23]. For complex objects such as ours, a stochastic inversion is desirable for generating many views of simplicial complexes from the same complexon mixture. Third, invertible embeddings permit learning in the original space, mitigating information loss from lower-dimensional projections.

3.1. SC-MAD steps

We elaborate on each step of SC-MAD in the sequel. Of primary importance is how to convert simplicial complexes into complexons.

Step (1) Complexon estimation. We perform complexon estimation for each labeled simplicial complex $\{(K_i, y_i)\}_{i=1}^T$ to obtain a set of complexon embeddings $\{(\hat{W}_i, y_i)\}_{i=1}^T$. The task of estimating a graphon from a single graph is well studied, for which there are several computationally efficient and effective methods [29–31]. We adapt sorting-and-smoothing (SAS) for *graphon* estimation [29] to *complexon* estimation, where SAS consists of (1) sorting nodes by degree and (2) estimating edge probability by computing network histograms. Inspired by this, we obtain node orderings at every dimension and jointly apply them to sort nodes with more information than if we were to only sort by the number of edges as with graphons.

We first sort nodes in a given d-dimensional simplicial complex

K with N nodes by computing the following sum

$$D_i = \sum_{c=1}^{d} \tau^c D_i^{(c)} \tag{1}$$

for every node $i \in \{1, 2, \ldots, N\}$, where $\tau \in (0, 1)$ and $D_i^{(c)}$ is the degree of node i at dimension c as in Section 2. Reordering the nodes in K by the degree sum in (1) gives the sorted simplicial complex K_ϕ. We obtain a piecewise constant complexon $\hat{W}$ as a simplicial complex histogram, whose values at dimension c measure the frequencies of c-simplices of K_ϕ in histogram bins [29]. More specifically, for any $\zeta^{(c)} = (\zeta_1, \ldots, \zeta_{c+1}) \in [0, 1]^{c+1}$, we obtain

$$\hat{W}_\circ^{(c)}\left(\zeta^{(c)}\right) = \frac{1}{h^{c+1}} \sum_{j_1=1}^{h} \cdots \sum_{j_{c+1}=1}^{h} \mathbb{I}\left\{(q(\zeta_1)h + j_1, \ldots, q(\zeta_{c+1})h + j_{c+1}) \in K_\phi\right\}, \tag{2}$$

where $h > 0$ denotes the number of nodes in each bin and we let $q(\zeta) = \max\{\lceil \zeta \lfloor N/h \rfloor \rceil - 1, 0\}$. The estimate $\hat{W}_\circ$ approximates the *faceted* complexon [23]. Hence, we obtain the final complexon estimate $\hat{W}$ by computing

$$\hat{W}^{(c)}(\zeta^{(c)}) = \hat{W}_\circ^{(c)}(\zeta^{(c)}) \left(\prod_{\zeta \subset \zeta^{(c)}} \hat{W}_\circ^{(|\zeta|-1)}(\zeta) \right)^{-1}, \tag{3}$$

for every $\zeta^{(c)} = (\zeta_1, \ldots, \zeta_{c+1}) \in [0, 1]^{c+1}$. The complexon estimation in (2) and (3) generalizes the popular SAS graphon estimation while accounting for interactions across dimensions for higher-order objects.

Step (2) Complexon mixup. Once the labeled complexon estimates $\{(\hat{W}_i, y_i)\}_{i=1}^T$ are obtained, we can then apply linear or convex clustering mixup. For pairwise linear mixup, we select a pair of complexons $\hat{W}_i$ and $\hat{W}_j$ such that $y_i \neq y_j$ and interpolate as

$$W_{\text{new}} = (1 - \lambda)\hat{W}_i + \lambda \hat{W}_j, \tag{4}$$

where $\lambda \in [0, 1]$. For convex clustering mixup, we solve the following optimization problem [24, 25]

$$\hat{U}(\lambda) = \underset{U}{\text{argmin}} \sum_{i=1}^{T} \rho_{\text{fid}}(U_i, \hat{W}_i) + \frac{\lambda}{1 - \lambda} \sum_{i<j} w_{ij} \rho_{\text{fus}}(U_i, U_j), \tag{5}$$

where $\lambda \in [0, 1)$ is the tunable mixup parameter, $w_{ij} \geq 0$ is the weight determining the level of fusion between $\hat{W}_i$ and $\hat{W}_j$, and the functions ρ_{fid} and ρ_{fus} respectively quantify fidelity and fusion [17]. We choose the following convex functions

$$\rho_{\text{fid}}(W_1, W_2) = \textstyle\sum_{c=0}^{d} \int_{[0,1]^{c+1}} (W_1^{(c)}(\zeta) - W_2^{(c)}(\zeta))^2 d\zeta,$$
$$\rho_{\text{fus}}(W_1, W_2) = \textstyle\sum_{c=0}^{d} \int_{[0,1]^{c+1}} |W_1^{(c)}(\zeta) - W_2^{(c)}(\zeta)| d\zeta.$$

The clusterpath $\hat{U}(\lambda) = \{\hat{U}_i(\lambda)\}_{i=1}^T$ returns complexon mixtures at each $\lambda \in [0, 1]$, with $\hat{U}(1) = \{\frac{1}{T}\sum_j \hat{W}_j\}_{i=1}^T$ by definition. When $\hat{U}_i(\lambda) = \hat{U}_j(\lambda)$, we say that the value is the mixture of $\hat{W}_i$ and $\hat{W}_j$, where $\hat{W}_i$ and $\hat{W}_j$ are fused. The mixup parameter λ determines how similar to the original complexons $\hat{W}$ the mixtures should be. When $\lambda = 0$, $\hat{U}(0) = \{\hat{W}_i\}_{i=1}^T$ returns the original complexons, and as λ increases, complexons begin to fuse into clusters. We encourage the clusterpath $\hat{U}(\lambda)$ to identify class differences for downstream classification by letting $w_{ij} = 1$ when $y_i = y_j$ and $w_{ij} = \epsilon$ otherwise for some $\epsilon > 0$. For further implementation details, we refer the reader to [17]. Once we obtain the clusterpath $\hat{U}(\lambda)$ from (5), we select complexon mixtures $W_{\text{new}} = \hat{U}_i(\lambda)$ by choosing $\lambda \in [0, 1]$ and $i \in \{1, 2, \ldots, T\}$. A visualization of the clusterpath $\hat{U}(\lambda)$ for two sets of Vietoris-Rips complexes is shown in Fig. 1a.

Step (3) Simplicial complex sampling. As with graphons, there is an analogous process for sampling simplicial complexes from complexons [23]. Given a set of nodes $K_{\text{new}}^{(0)}$, we sample edges from the complexon W_{new} as

$$\zeta_i \sim \text{Unif}([0, 1]) \qquad \forall\, i \in K_{\text{new}}^{(0)},$$
$$\mathbb{P}\left[(i, j) \in K_{\text{new}}^{(1)}\right] = W_{\text{new}}^{(1)}(\zeta_i, \zeta_j) \qquad \forall\, (i, j) \in K_{\text{new}}^{(0)} \times K_{\text{new}}^{(0)},$$

identical to that of graphons. Beyond edges, to retain closure under restriction, we must preclude simplices whose proper subsets are not all already present in the sampled simplicial complex. At dimension $d > 1$, we add a d-simplex σ to $K_{\text{new}}^{(d)}$ with probability

$$\mathbb{P}\left[\sigma \in K_{\text{new}}^{(d)}\right] = W_{\text{new}}^{(d)}(\zeta_\sigma) \prod_{\sigma' \subset \sigma} \mathbb{I}\left\{\sigma' \in K_{\text{new}}\right\},$$

where $W_{\text{new}}^{(d)}(\zeta_\sigma)$ represents the probability of $\sigma \in K_{\text{new}}$ conditioned on the existence of all its proper subsets in K_{new}. Once a desired dimension is reached, the result is a simplicial complex K_{new} satisfying closure under restriction. Further details are provided in [23]. We can then sample any number of new simplicial complexes from one complexon W_{new}, generating multiple views from the same model whose structural characteristics are preserved.

3.2. Class structure in complexon mixtures

Mixup aims to generate new samples with characteristics from multiple classes. We theoretically show that the complexon mixtures W_{new} from linear mixup (4) or convex clustering mixup (5) contain a mixture of class-dependent structural characteristics from multiple simplicial complexes. In particular, we assume that for each class y, there is a finite set of discriminative simplicial complexes $\mathcal{F}_y$ such that for every labeled simplicial complex (K, y), there exists at least one $F \in \mathcal{F}_y$ that is a subcomplex of K [16], that is, there is a homomorphism from F to K. We present the following result on the structural similarities between a complexon mixture and one of the complexons, inspired by a similar result for graphon mixup [16].

Theorem 1 *Consider a set of simplicial complexes $\{(K_i, y_i)\}_{i=1}^T$ from which we estimate a set of complexons $\{\hat{W}_i\}_{i=1}^T$. Let the convex combination $W_{\text{new}} = \sum_{i=1}^T \gamma_i \hat{W}_i$ for $\sum_{i=1}^T \gamma_i = 1$ denote a complexon mixture from* (4) *or* (5)*, and let $\mathcal{F}_{y_j}$ be the discriminative simplicial complex set for class y_j. For any finite $F \in \mathcal{F}_{y_j}$, we present the following convergence result on the homomorphism density difference for the complexon mixture W_{new} and the estimate $\hat{W}_j$. As $\gamma_j \to 1$ or $\rho_{\text{fus}}(\hat{W}_j, \sum_{i \neq j} \frac{\gamma_i}{1-\gamma_j} \hat{W}_i) \to 0$, we have that*

$$|t(F, W_{\text{new}}) - t(F, \hat{W}_j)| \to 0. \tag{7}$$

For example, when $\gamma_i = 1/T$ for every $i = 1, 2, \ldots, T$, then as $\hat{W}_j$ approaches the mean of the remaining complexons $\frac{1}{T-1}\sum_{i \neq j} \hat{W}_i$ with respect to ρ_{fus}, we have convergence as in (7)*.*

Method		Vietoris-Rips	MNIST
Data mixup	Label mixup		
None	None	0.631 ± 0.167	0.782 ± 0.051
Linear	Linear	0.709 ± 0.051	0.802 ± 0.111
	Sigmoid	$\mathbf{0.719 \pm 0.084}$	0.687 ± 0.088
	Logit	0.594 ± 0.146	0.705 ± 0.033
	Cvx. clust.	0.669 ± 0.193	0.805 ± 0.057
Cvx. clust.	Linear	0.688 ± 0.196	0.804 ± 0.110
	Sigmoid	0.688 ± 0.156	$\mathbf{0.819 \pm 0.072}$
	Logit	0.709 ± 0.064	0.817 ± 0.049
	Cvx. clust.	$\mathbf{0.738 \pm 0.057}$	$\mathbf{0.856 \pm 0.052}$

Table 1: Simplicial complex classification accuracy. The top performing methods are **bolded**.

Proof sketch. We omit a full proof of Theorem 1 for space and provide a brief description instead. By the definition of the mixture W_{new}, if either of the two conditions in the statement of Theorem 1 hold, then we have that $\rho_{\text{fus}}(W_{\text{new}}, \hat{W}_j) \to 0$. Then, it can be shown that the left-hand side of (7) is bounded above by a finite scaling of $\rho_{\text{fus}}(W_{\text{new}}, \hat{W}_j)$. Thus, if either condition holds, then the homomorphism density difference converges to 0 as in (7). ■

Note that for complexons of dimension 1, when $\gamma_i = \lambda$, $\gamma_j = 1 - \lambda$, and $\gamma_k = 0$ for every $k \neq i, j$, Theorem 1 is analogous to the result for pairwise graphon mixup in [16]. Our result generalizes that of [16] by allowing arbitrary convex combinations and any complexon dimension. Theorem 1 shows that the discriminative structure of a given class y_j grows increasingly present in the mixture W_{new} as $\hat{W}_j$ becomes more prominent in the mixture or as $\hat{W}_j$ grows closer to the remaining complexons in the set. Furthermore, since convex clustering obtains mixtures of every complexon in a set, the complexon mixtures obtained from (5) will contain the discriminative structure for every class.

4. NUMERICAL EVALUATION

We evaluate SC-MAD for generating labeled simplicial complexes to improve classification accuracy. We use a simplicial convolutional network (SCN) as the architecture for each of the following simulations [9], and we compare model prediction performance with and without data augmentation. We perform simplicial complex mixup via linear mixup (4), denoted "Linear", and convex clustering mixup (5), denoted "Cvx. clust.", as described in Section 3. For both methods, we let $\lambda \sim \text{Unif}([0, 1])$. We also compare four methods for mixup of labels [17]. We interpolate labels y_i and y_j given the mapping $g : [0, 1] \to [0, 1]$ as

$$y_{\text{new}} = (1 - g(\lambda))\, y_i + g(\lambda) y_j.$$

For $a > 0$, we consider "Linear" mixup $g(\lambda) = \lambda$; "Sigmoid" mixup $g(\lambda) = 1/(1 + \exp\{-a(2\lambda - 1)\})$; "Logit" mixup $g(\lambda) = \log(\lambda/(1-\lambda))/2a + 1/2$; and "Cvx. clust.", convex clustering label mixup as introduced in [17].

Synthetic data. Consider two classes of Vietoris-Rips complexes, where each complex is formed from i.i.d. points sampled from one of two shapes in $\mathbb{R}^2$, a circle and a figure eight. We perform simplicial complex classification to identify from which shape each complex is sampled. We present the shape classification accuracy for each method in the column of Table 1 denoted "Vietoris-Rips". The first row of Table 1 corresponds to the original dataset with no data augmentation. The column "Data mixup" indicates the simplicial complex mixup method and "Label mixup" the label mixup method.

In all cases but one, data augmentation via mixup improves prediction performance. We observe the greatest increase in classification accuracy when using convex clustering for both data and labels, as expected due to the more informative sampling of new labeled simplicial complexes. We emphasize the practicality of convex clustering for mixup as we achieve superior performance without requiring a specified mixup function for data or labels, nor do we require a user-defined sampling mechanism for the mixup parameter λ [13]. We thus demonstrate the viability of the complexon for interpolating in the higher-order simplicial complex space. With this choice of interpolation space, we reap the advantages of mixup for improving performance even for such complex structures.

Image data. We also evaluate our proposed mixup on the MNIST image dataset [32]. Any image can be represented as a superpixel graph, where each node corresponds to a cluster of pixels denoting meaningful regions and each edge connects nodes that are adjacent in pixel space [8]. To encode richer visual information, we add triangles for every clique of three nodes in the superpixel graph, resulting in a set of simplicial complexes modeling related regions within each image. In Fig. 1b and d, we show simplicial complex representations of two handwritten digits in the MNIST dataset, while in Fig. 1c, we present the simplicial complex sampled from a complexon mixture obtained via linear mixup of the original two images with $\lambda = 0.5$. We obtain a mixed superpixel simplicial complex that exhibits structural interpolation rather than mere pixel-wise value mean. In particular, the mixture in Fig. 1c not only mixes pixel values by interpolating simplex features but also changes how image regions, represented by nodes, are connected, modifying which regions are relevant to which.

A comparison of our results for simplicial complex classification on a subset of three classes of MNIST images is shown in the column of Table 1 denoted "MNIST". As image classification is well understood, superpixel network classification serves as a useful benchmark for comparing simplicial complex-based learning methods. Convex clustering for both images and labels results in the greatest increase in classification accuracy over the original superpixel dataset. This demonstrates the power of convex clustering for providing informative synthetic samples for real-world multiclass data. Moreover, almost all mixup methods achieve superior performance relative to the original dataset, including those that apply different methods for label and image mixup. This motivates future investigation in pursuing optimal ways to mixup data and labels.

5. CONCLUSION

In this work, we presented simplicial complex mixup via complexons, the limit object of convergent simplicial complex sequences. With the continuous complexon, we were able to exploit the efficiency of linear pairwise mixup along with the effectiveness of convex clustering mixup for discrete, irregular simplicial complexes. The success of our method for simplicial complexes implies the practicality of exploring limit objects for other data types to perform useful tasks typically limited to Euclidean data, without needing domain expertise or computationally intensive approaches. Furthermore, we theoretically validated our ability to manipulate simplicial complexes while preserving structural characteristics, so the ubiquitous use of graphs in many applications can be naturally extended to simplicial complexes. We may more easily adopt these higher-order networks for other useful fields that graphs already occupy, such as social network analysis.

6. REFERENCES

[1] S. Barbarossa and S. Sardellitti, "Topological signal processing over simplicial complexes," *IEEE Trans. Signal Process.*, vol. 68, pp. 2992–3007, 2020.

[2] A. Hatcher, *Algebraic Topology*. Cambridge University Press, 2002.

[3] M. T. Schaub, Y. Zhu, J.-B. Seby, T. M. Roddenberry, and S. Segarra, "Signal processing on higher-order networks: Livin' on the edge... and beyond," *Signal Process.*, vol. 187, p. 108149, 2021.

[4] T. M. Roddenberry, F. Frantzen, M. T. Schaub, and S. Segarra, "Hodgelets: Localized spectral representations of flows on simplicial complexes," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 5922–5926, 2022.

[5] T. M. Roddenberry and S. Segarra, "HodgeNet: Graph neural networks for edge data," in *Asilomar Conf. Signals, Syst., and Computers*, pp. 220–224, 2019.

[6] L. Kanari, P. Dłotko, M. Scolamiero, R. Levi, J. Shillcock, K. Hess, and H. Markram, "A topological representation of branching neuronal morphologies," *Neuroinformatics*, vol. 16, no. 1, pp. 3–13, 2018.

[7] T. Roman, A. Nayyeri, B. T. Fasy, and R. Schwartz, "A simplicial complex-based approach to unmixing tumor progression data," *BMC Bioinformatics*, vol. 16, no. 1, p. 254, 2015.

[8] C. W. J. Goh, C. Bodnar, and P. Liò, "Simplicial attention networks," *arXiv:2204.09455*, 2022.

[9] S. Ebli, M. Defferrard, and G. Spreemann, "Simplicial neural networks," *arXiv:2010.03633*, 2020.

[10] T. M. Roddenberry, N. Glaze, and S. Segarra, "Principled simplicial neural networks for trajectory prediction," in *Intl. Conf. on Mach. Learn. (ICML)*, vol. 139, pp. 9020–9029, PMLR, 2021.

[11] D. M. Cinque, C. Battiloro, and P. Di Lorenzo, "Pooling strategies for simplicial convolutional networks," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 1–5, 2023.

[12] A. Hernández-Garcia and P. König, "Data augmentation instead of explicit regularization," *arXiv:1806.03852*, 2018.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Intl. Conf. on Learning Representations (ICLR)*, 2018.

[14] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *Advances in Neural Info. Process. Syst.*, vol. 32, 2019.

[15] L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou, "How does mixup help with robustness and generalization?," in *Intl. Conf. on Learning Representations (ICLR)*, 2021.

[16] X. Han, Z. Jiang, N. Liu, and X. Hu, "G-Mixup: Graph data augmentation for graph classification," in *Intl. Conf. on Mach. Learn. (ICML)*, vol. 162, pp. 8230–8248, PMLR, 2022.

[17] M. Navarro and S. Segarra, "GraphMAD: Graph mixup for data augmentation using data-driven convex clustering," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 1–5, 2023.

[18] Y. Wang, W. Wang, Y. Liang, Y. Cai, and B. Hooi, "Mixup for node and graph classification," in *Proc. of the Web Conf. (WWW)*, pp. 3663–3674, ACM, 2021.

[19] T. Wei, Y. You, T. Chen, Y. Shen, J. He, and Z. Wang, "Augmentations in hypergraph contrastive learning: Fabricated and generative," in *Advances in Neural Info. Process. Syst.*, vol. 35, pp. 1909–1922, 2022.

[20] Y. You, T. Chen, Y. Shen, and Z. Wang, "Graph contrastive learning automated," in *Intl. Conf. on Mach. Learn. (ICML)*, vol. 139, pp. 12121–12132, PMLR, 2021.

[21] L. Lovász, *Large Networks and Graph Limits*. American Math. Society, 2012.

[22] H. Xu, D. Luo, L. Carin, and H. Zha, "Learning graphons via structured Gromov-Wasserstein barycenters," *AAAI Conf. on Artif. Intell.*, vol. 35, no. 12, pp. 10505–10513, 2021.

[23] T. M. Roddenberry and S. Segarra, "Limits of dense simplicial complexes," *J. Mach. Learn. Res. (JMLR)*, vol. 24, no. 225, pp. 1–42, 2023.

[24] K. Pelckmans, J. De Brabanter, B. De Moor, and J. Suykens, "Convex clustering shrinkage," in *Stat. and Optimization of Clustering Wrkshp. (PASCAL)*, 2005.

[25] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert, "Clusterpath: An algorithm for clustering using convex fusion penalties," in *Intl. Conf. on Mach. Learn. (ICML)*, p. 1, 2011.

[26] F. Lindsten, H. Ohlsson, and L. Ljung, "Clustering using sum-of-norms regularization: With application to particle filter output computation," in *IEEE Wrkshp. Statistical Signal Process. (SSP)*, pp. 201–204, 2011.

[27] J. R. Munkres, *Topology*. Prentice Hall, Inc, 2000.

[28] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *Intl. Conf. on Mach. Learn. (ICML)*, vol. 97, pp. 6438–6447, PMLR, 2019.

[29] S. Chan and E. Airoldi, "A consistent histogram estimator for exchangeable graph models," in *Intl. Conf. on Mach. Learn. (ICML)*, vol. 32, pp. 208–216, PMLR, 2014.

[30] P. J. Bickel and A. Chen, "A nonparametric view of network models and Newman–Girvan and other modularities," *Proc. of the Nat. Acad. of Sciences (PNAS)*, vol. 106, no. 50, pp. 21068–21073, 2009.

[31] J. Yang, C. Han, and E. Airoldi, "Nonparametric estimation and testing of exchangeable graph models," in *Intl. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 33, pp. 1060–1067, PMLR, 2014.

[32] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.