Leveraging Topological Guidance for Improved Knowledge Distillation

Eun Som Jeon [1] **Rahul Khurana** [1] **Aishani Pathak** [1] **Pavan Turaga** [1]

Editors: *S. Vadgama, E.J. Bekkers, A. Pouplin, S.O. Kaba, H. Lawrence, R. Walters, T. Emerson, H. Kvinge, J.M. Tomczak, S. Jegelka*

Abstract

Deep learning has shown its efficacy in extracting useful features to solve various computer vision tasks. However, when the structure of the data is complex and noisy, capturing effective information to improve performance is very difficult. To this end, topological data analysis (TDA) has been utilized to derive useful representations that can contribute to improving performance and robustness against perturbations. Despite its effectiveness, the requirements for large computational resources and significant time consumption in extracting topological features through TDA are critical problems when implementing it on small devices. To address this issue, we propose a framework called Topological Guidance-based Knowledge Distillation (TGD), which uses topological features in knowledge distillation (KD) for image classification tasks. We utilize KD to train a superior lightweight model and provide topological features with multiple teachers simultaneously. We introduce a mechanism for integrating features from different teachers and reducing the knowledge gap between teachers and the student, which aids in improving performance. We demonstrate the effectiveness of our approach through diverse empirical evaluations.

1. Introduction

In recent years, deep learning has been widely deployed into various applications, such as image recognition (Xie et al., 2020; He et al., 2019), activity recognition (Zheng et al., 2016; Wang et al., 2016), semantic segmentation (Minaee et al., 2021), and so on. Deep learning is proficient in extracting features and performing various computer vision tasks. However, it has challenges in grasping useful features from the complex structure of the data, which limits further advancements (Najafabadi et al., 2015).

To address these issues, topological data analysis (TDA) has emerged as a solution, which is excellent at analyzing the topology of data to apprehend its arrangement (Adams et al., 2017; Wang et al., 2021). Since TDA reveals patterns that may not be extracted or magnified through traditional statistical methods, many research endeavors aim to adopt these attributes of TDA to enhance the efficacy of deep learning. Specifically, TDA is excellent in capturing inherent and invariant features, which are robust to noise and perturbation (Adams et al., 2017; Seversky et al., 2016). TDA characterizes the shape of complex data, using the persistence of connected components and high-dimensional holes by the persistent homology (PH) algorithm. This persistence information can be represented as a persistence image (PI). To utilize TDA in fusion of machine learning, PI has been widely used since it can be easily transformed and treated as a general image (Edelsbrunner & Harer, 2022). Despite various benefits of TDA, significant computational resources and time is required for TDA feature computation. Many applications have explored the use of TDA features with machine learning (Munch, 2017), however in most cases, simple fusion methods do not result in compact models. Som *et al.* (Som et al., 2020) introduced PI-net to solve this problem, however the burden of increased network size cannot be alleviated even at test-time.

Knowledge distillation (KD) has been addressed as a promising approach that leverages a power of a teacher (large model) to generate a student (small model) (Hinton et al., 2015). KD has further benefits in improving generalizability of a student model. In KD, a variety of strategies can be adopted to generate a compact model. For instance, not only one teacher model but multiple teachers can be utilized to transfer more diverse and strong knowledge to a student model (Gou et al., 2021). An approach that involves utilizing two teachers can be adopted to leverage the power of topological knowledge. In detail, two teachers are trained –

[1]Geometric Media Lab, Arizona State University, Tempe, AZ 85281, USA. Correspondence to: Eun Som Jeon <ejeon6@asu.edu>.

Proceedings of the Geometry-grounded Representation Learning and Generative Modeling at the 41st International Conference on Machine Learning, Vienna, Austria. PMLR Vol Number, 2024.
Copyright 2024 by the author(s).

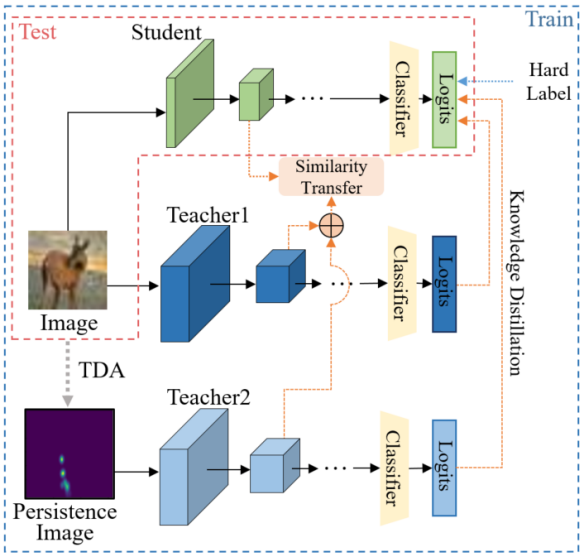


Figure 1. An overview of Topological Guidance based Knowledge Distillation (TGD). Two teachers are trained with different representations from the raw image and persistence image data, respectively. A student utilizes the original image data alone.

one on the original data and the other one on the PI – both of which are leveraged to generate a student model. This strategy has proven beneficial in time-series data analysis (Jeon et al., 2022). However, sufficient research has not been conducted on the effectiveness of such methods in KD based image analysis leveraging topological features. Additionally, when the statistical characteristics of knowledge from the two models are significantly different, there are considerable challenges and performance degradation in combining and utilizing the two sets of information (Zhu & Wang, 2021; Tan et al., 2018; Gou et al., 2021).

In this paper, we propose a framework, Topological Guidance based Knowledge Distillation (TGD), using topological knowledge in distillation for image classification task. We devise a strategy to integrate knowledge from two teachers trained with different modalities: raw image data and persistence images. An overview of the TGD is shown in Figure 1. Firstly, PI is extracted from the raw image data through TDA. The extracted PI is then used to train a teacher model. Secondly, two teacher models are employed to provide useful information to train a student model. Logits and features from intermediate layers are utilized. When features of intermediate layers are transferred, similarity maps are utilized, facilitating the integration of information with different characteristics into a single entity. Additionally, we adopt an annealing strategy that reduces knowledge gap between teachers and students while preserving the weights that the student model needs to possess for its task (Jeon et al., 2022). Finally, a student model is distilled, which uses solely the raw image data in test-time.

The contributions of this paper are as follows:

- We introduce a novel framework in knowledge distillation, using topological knowledge to generate a compact model for image classification tasks.
- We devise a technique to integrate features from intermediate layers of teachers and a strategy to reduce the knowledge gap between teachers and student.
- We demonstrate the effectiveness of leveraging topological features in KD empirically with various evaluations such as various combinations of teachers and students and feature visualizations.

Our main goal is not to outperform all the latest methods in vision, but to explore how topological guidance can be utilized in KD to improve the performance and to investigate the behavior of the distilled model along with empirical testing on image analysis.

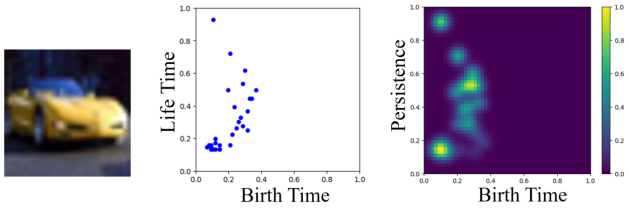
2. Background

2.1. Topological Feature

TDA algorithms are applied to the data to extract topological features, which are robust to noises or perturbations and encodes the shape of complex data (Adams et al., 2017; Wang et al., 2021). Persistent homology is a fundamental tool in TDA that helps in understanding the shape and structure of data, which involves constructing a filtration (Edelsbrunner & Harer, 2022), typically based on a distance function and tracking variations of n-dimensional holes represented by assortments of points, edges, and triangles through a dynamic thresholding process called filtration. In filtration, the appearance and disappearance of these holes are described in the persistence feature, summarized in a persistence diagram (PD), which records the birth and death times as x and y coordinates of planar scatter points (Adams et al., 2017; Edelsbrunner & Harer, 2022).

Since PDs can have a high dimensionality with complex structures or a large number of points that can vary, using PDs directly in machine learning is challenging. To address this problem, persistence image (PI) has been widely used, which is one of the ways to encode geometric information via the lifespan of homological structures present in the data. This representation can be easily integrated into machine learning (Barnes et al., 2021; Edelsbrunner & Harer, 2022). Specifically, the points within the PD are projected onto a two-dimensional grid $\rho : \mathbb{R} \to \mathbb{R}^2$. The grid points are then assigned values determined by a weighted sum of Gaussian smoothing, which is centered around the scattered points within the PD. The grid is represented as a matrix and can be treated as regular image data called PI, as depicted in Figure 2. This allows for the application of convolutional

neural networks (CNNs) and machine learning algorithms, and offers a more manageable format for analysis and visualization.



(a) Image Data (b) Persistence Diagram (c) Persistence Image

Figure 2. PD and its corresponding PI. Lifetime points in PD appears bright colors in PI.

Previous studies showed that topological features complement features of the raw data to achieve improved performance (Som et al., 2020). However, additional processing for TDA and concatenation in networks increase computational time and resources at inference-time, which poses difficulty in implementing the process on small devices having limited computational resources and power. To alleviate this issue, we propose a framework based on KD to infuse topological features into a small model that uses solely the raw image data at test-time.

2.2. Knowledge Distillation

Conventional Knowledge Distillation. Knowledge distillation obtains a smaller model by utilizing the learned knowledge of a larger model, which was first introduced by Buciluǎ *et al.* (Buciluǎ et al., 2006) and explored more by Hinton *et al.* (Hinton et al., 2015). In KD, soft labels are utilized for knowledge transfer from a teacher to a student, which provide richer supervision signals and reduce overfitting. This also leads to better transferability of learned representations. The loss function of conventional KD for training a student is:

$$\mathcal{L}_{\mathcal{KD}} = \lambda \mathcal{L}_{\mathcal{CE}} + (1-\lambda)\mathcal{L}_{\mathcal{K}}, \tag{1}$$

where, $\mathcal{L}_{\mathcal{CE}}$ is the standard cross entropy loss, $\mathcal{L}_{\mathcal{K}}$ is KD loss, and λ is a hyperparameter; $0 < \lambda < 1$. The difference between the output of the softmax layer for a student network and the ground-truth label is minimized by the cross-entropy loss:

$$\mathcal{L}_{\mathcal{CE}} = \mathcal{H}(softmax(l_S), y_g), \tag{2}$$

where, $\mathcal{H}(\cdot)$ is a cross entropy loss function, l_S is the logits of a student, and y_g is a ground truth label. The gap between outputs of student and teacher are minimized by KL-divergence loss:

$$\mathcal{L}_{\mathcal{K}} = \tau^2 KL(z_T, z_S), \tag{3}$$

where, $z_T = softmax(l_T/\tau)$ and $z_S = softmax(l_S/\tau)$ are softened outputs of a teacher and student, respectively, and τ is a hyperparameter; $\tau > 1$. To obtain the best performance, we adopt early stopping for KD (ESKD) which improves the efficacy of KD (Cho & Hariharan, 2019).

Feature-based Knowledge Distillation. Features from intermediate layers of a network can be utilized in knowledge transfer (Gou et al., 2021; Zagoruyko & Komodakis, 2017; Tung & Mori, 2019). Zagoruyko *et al.* (Zagoruyko & Komodakis, 2017) suggested activation-based attention transfer (AT), which is computed by a sum of squared attention mapping function, and calculating statistics across the channel dimension. Tung *et al.* (Tung & Mori, 2019) introduced similarity-preserving knowledge distillation, matching similarity within a mini-batch of samples between a teacher and a student. Since the size of a similarity map is determined by the size of a mini-batch, the size of the extracted similarity maps from the teacher and student is the same even if they generate different sizes of features. In details, the similarity map $M \in \mathbb{R}^{b\times b}$ is obtained as follows:

$$M = F \cdot F^{\top}; F \in \mathbb{R}^{b\times chw}, \tag{4}$$

where F is reshaped features from an intermediate layer of a model, b is the size of a mini-batch, and c, h, and w are the number of channels, height, and width of the output, respectively. These feature transfer methods are popularly used; however, these are to match knowledge with similar characteristics in a uni-modal manner.

Utilizing Multiple Teachers. Not only one teacher, but multi-teacher distillation has been widely utilized to provide more diverse knowledge in training process (Reich et al., 2020; Liu et al., 2020; Gou et al., 2021). In some cases, the data utilized for training a student cannot be used during testing. Also, teachers trained with different modalities or representations can be utilized in distillation. Thoker and Gall (Thoker & Gall, 2019) train a student with paired samples from two modalities for action recognition. Jeon *et al.* (Jeon et al., 2022) explored to train a student model with two teachers trained with different representations for wearable sensor data analysis. With this insight, we develop a framework and explore to utilize topological features involving two teachers for image data analysis.

3. Proposed Method

In this section, we describe our proposed method – TGD. Firstly, PI is extracted from an image through TDA. The extracted PI is utilized to train a teacher model. Secondly, we train a student model in KD with two teachers trained on different representations, the raw image data and PI. Then, to provide more useful knowledge to a student, features from teachers are integrated by considering correlations of each teacher's features. To reduce knowledge gap between teachers and student, an annealing strategy is applied.

3.1. Persistence Image Extraction

To compute PIs, Scikit-TDA python library (Saul & Tralie, 2019) and the Ripser package are used for generating PDs, as explained in Som *et. al.* (Som et al., 2020). Firstly, image data is normalized in range in $[0, 1]$. To compute level-set filtration PDs, image data is reshaped to row- and column-wise signals, considering different order of context which can extract different topological features (Barnes et al., 2021). By filtration, PDs summarize the different peak and local minima intensities in the data. Specifically, each channel of data is transformed and used to generate a PI, where the PI implies birth-time vs. lifetime information. We utilize row- and column-wise transforms separately for creating images with channels of (R_r, G_r, B_r) and (R_c, G_c, B_c) to collect diverse knowledge, and all created PIs are concatenated in an image. Then, six channels (P_{Rr}, P_{Gr}, P_{Br}, P_{Rc}, P_{Gc}, P_{Bc}) of PI implying persistence knowledge is created. The total dimension size of one PI is $g \times g \times c$, where g and c are a constant value and the number of channels for a sample. The created PI is utilized to train a teacher model that acts as a pre-trained model to transfer topological features to a student model in KD process.

3.2. Utilizing Multiple Teachers

Knowledge Transfer with Logits. Knowledge of logits from two teachers are transferred individually, thus additional process including concatenation or hidden layers is not required. KD loss for logit knowledge of two teachers is explained as follows:

$$\mathcal{L}_{\mathcal{KD}l} = \tau^2 \left(\alpha KL(z_{T_1}, z_S) + (1-\alpha) KL(z_{T_2}, z_S)\right), \quad (5)$$

where, α is a constant to balance the effects of different teachers, and z_{T_1} and z_{T_2} are softened outputs of teachers trained with the raw image data and PIs, respectively.

Knowledge Transfer with Intermediate Features. To transfer sufficient knowledge from two teachers, we utilize features from intermediate layers additionally. Since PI and the raw image data have different statistical characteristics in semantic information, it is more effective to align and convey the information by using the correlation between samples rather than using spatial information. We utilize similarities, as explained in equation 4, to easily integrate information from two teachers and provide topological features to student in distillation. Figure 3 shows an example of similarity maps obtained from different intermediate layers of two WRN16-3 teachers. Note, Teacher1 and Teacher2 denote models learned with the raw image data and PI, respectively. High values represent high similarities. This implies similar patterns can be created when two samples belong to the same category. Since two models are trained with different representations, their highlighted patterns are different. We merge the maps from two teachers with weighted summation as follows:

$$M_{T_m}^{(l)} = \alpha M_{T_1}^{(l^{T_1})} + (1-\alpha) M_{T_2}^{(l^{T_2})}, \quad (6)$$

where, $M_{T_m}^{(l)} \in \mathbb{R}^{b \times b}$ is the generated map from the similarity maps of two teachers M_{T_1} and M_{T_2} in a layer pair (l^{T_1} and l^{T_2}). By merging the maps, the similarities include topological features which can complement the original features to improve the performance. The loss that encourages the student to mimic teachers is:

$$\mathcal{L}_m = \frac{1}{b^2|L|} \sum_{(l,l^S)\in L} \left(\left\| \widetilde{M_{T_m}^{(l)}} - \widetilde{M_S^{(l^S)}} \right\|_F^2 \right), \quad (7)$$

where $\widetilde{M_{T_m}^{(l)}}$ and $\widetilde{M_S^{(l^S)}}$ are normalized map for a merged teacher and a student, $\|\cdot\|_F$ is the Frobenius norm (Tung & Mori, 2019), and L collects the layer pairs (l and l^S). l^{T_1}, l^{T_2}, and l^S, can be selected with the same depth or the end of the same block of networks. Since a student model uses the raw image data only, the gap between the merged features of teachers and the feature of the student can be generated, which makes degradation. To alleviate this problem, an annealing strategy (Jeon et al., 2022) is used, which initializes the student model with weight values of a model trained from scratch. The final loss function is as follows:

$$\mathcal{L}_{\text{TGD}} = \lambda \mathcal{L}_{\mathcal{CE}} + (1-\lambda)\mathcal{L}_{\mathcal{KD}l} + \gamma \mathcal{L}_m, \quad (8)$$

where γ is a hyperparameter.

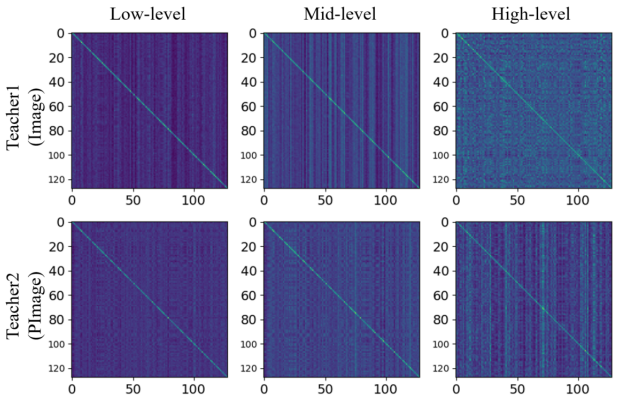


Figure 3. An illustration of similarities for two teachers, trained with the raw image and persistence image, respectively.

4. Experiments

4.1. Dataset and Experimental Settings

Datasets. The CIFAR-10 (Krizhevsky & Hinton, 2009) dataset consists of 60k images distributed among 10 classes, with each class including 5k and 1k images for training and testing, respectively. Each image is a 32×32 sized RGB image. The experiments on CIFAR-10 allows us to evaluate

Table 1. Details of teacher and student network architectures. ResNet (He et al., 2016) and WideResNet (Zagoruyko & Komodakis, 2016) are denoted by ResNet (depth) and WRN (depth)-(channel multiplication), respectively. FLOPs and the number of trainable parameters correspond to one teacher model.

DB	Setup	Compression type	Teacher1 & Teacher2	Student	FLOPs		# of params		Compression ratio
					Teachers	Student	Teachers	Student	
CIFAR-10	(A)	Channel	WRN16-3	WRN16-1	224.63M	27.24M	1.50M	0.18M	5.81%
	(B)	Depth	WRN28-1	WRN16-1	56.07M	27.24M	0.37M	0.18M	24.32%
	(C)	Depth+Channel	WRN16-3	WRN28-1	224.63M	56.07M	1.50M	0.37M	12.33%
	(D)	Different architecture	ResNet44	WRN16-1	99.34M	27.24M	0.66M	0.18M	13.64%
CINIC-10	(A)	Channel	WRN16-3	WRN16-1	224.63M	27.24M	1.50M	0.18M	5.81%
	(B)	Depth	WRN28-1		56.07M		0.37M		24.32%
	(C^a)	Depth+Channel	WRN28-3		480.98M		3.29M		2.74%
	(D)	Different architecture	ResNet44		99.34M		0.66M		13.64%

Table 2. Accuracy (%) on CIFAR-10 with various knowledge distillation methods.

Setup	Method									
	Teacher1	Student	KD	AT	SP	RKD	VID	AFDS	Base	TGD
(A)	87.63 ±0.09	84.07 ±0.08	85.18 ±0.14	85.59 ±0.08	85.55 ±0.05	85.35 ±0.06	85.28 ±0.19	–	85.60 ±0.16	**86.03** ±0.05
(B)	85.73 ±0.06	84.07 ±0.08	85.34 ±0.15	85.63 ±0.06	85.70 ±0.07	85.34 ±0.10	84.91 ±0.25	85.40 ±0.19	85.47 ±0.17	**86.06** ±0.14
(C)	87.63 ±0.09	85.73 ±0.06	86.38 ±0.11	86.63 ±0.10	86.44 ±0.05	86.16 ±0.21	86.35 ±0.18	–	86.86 ±0.11	**87.12** ±0.06
(D)	86.15 ±0.11	84.07 ±0.08	85.36 ±0.10	**85.91** ±0.09	84.69 ±0.12	85.43 ±0.08	85.05 ±0.09	85.27 ±0.17	85.53 ±0.08	85.86 ±0.04

Table 3. Accuracy (%) on CINIC-10 with various knowledge distillation methods. TGD outperforms RKD (Park et al., 2019).

Setup	Method								
	Teacher1	Student	KD	AT	SP	VID	AFDS	Base	TGD
(A)	75.27 ±0.12	71.87 ±0.09	74.20 ±0.07	74.32 ±0.11	74.25 ±0.09	74.31 ±0.06	–	74.43 ±0.26	**74.66** ±0.04
(B)	73.41 ±0.12		74.57 ±0.06	74.51 ±0.13	74.81 ±0.10	73.75 ±0.08	74.45 ±0.05	74.71 ±0.10	**74.88** ±0.02
(C^a)	76.91 ±0.03		74.18 ±0.06	74.21 ±0.10	74.95 ±0.16	73.89 ±0.16	–	74.75 ±0.05	**75.04** ±0.06
(D)	74.12 ±0.20		74.36 ±0.07	74.58 ±0.05	74.29 ±0.24	74.30 ±0.12	74.47 ±0.07	74.55 ±0.09	**74.78** ±0.07

our model's efficacy with less time consumption. We extend our experiments on CINIC-10 (Darlow et al., 2018) that augments CIFAR-10 formatting but includes a larger set of 270k images whose scale closer to ImageNet. The images are evenly split into each 'train', 'validate', and 'test' sets, with ten classes of 9k images per class. The size of the images is 32×32 as well.

Experimental Settings. In generating PIs by TDA, by referring to the previous study (Som et al., 2020), we set birth-time range and Gaussian function parameter as [0, 0.3] and 0.01. The threshold for life-time is set to 0.02. we set g and c of PI as 50 and 6, respectively.

For experiments, we set the batch size b as 128, the total epochs as 200 using SGD with momentum 0.9, and a weight decay of 1×10^{-4}. The initial learning rate lr is set to 0.1 that is decayed by a factor of 0.2 at epochs 40, 80, 120, and 160. Empirically, we set KD hyperparameters λ, τ, and γ as ($\lambda = 0.9$, $\tau = 4$, $\gamma = 3000$) and ($\lambda = 0.6$, $\tau = 16$, $\gamma = 2000$) for CIFAR-10 and CINIC-10, respectively, referred to previous studies (Cho & Hariharan, 2019; Tung & Mori, 2019; Jeon et al., 2023).

We compare with KD based baselines including conventional KD (Hinton et al., 2015), attention transfer (AT) (Zagoruyko & Komodakis, 2017), relational knowledge distillation (RKD) (Park et al., 2019), variational information distillation (VID) (Ahn et al., 2019), similarity-preserving knowledge distillation (SP) (Tung & Mori, 2019), attentive feature distillation and selection (AFDS) (Wang et al., 2020), and multi-teacher based distillation using topological features in KD (Base) (Jeon et al., 2022). AT and SP are utilized with KD. For all baseline methods, the same hyperparameter settings are used as those specified in their

papers, and their author-provided code is used for evaluation. α of Base is 0.9, and TGD is 0.99 as a default setting. All experiments were repeated three times, and the averaged accuracy and the standard deviation of performance are reported. More details are explained in appendix.

4.2. Analysis on Teacher-Student Combinations

In this section, we show analysis on various combinations including different capacity of teachers and architectural styles of teacher-student networks.

4.2.1. Effect of Teacher Capacity

We explore the performance of various methods on different types of teacher-student combinations, where the teachers have different capacity. Note, Teacher1 and Teacher2 denote models learned with the raw image data and PI, respectively, and Student denotes a model trained from scratch. As explained in Table 1, we set four different setups for combinations which consist of same or different structures. We utilize Wide-ResNet (WRN) (Zagoruyko & Komodakis, 2016) to construct various compression types of teachers and a student.

As explained in Table 2, in most of cases, TGD outperforms baselines. Base is an approach using topological features in KD. Compared to KD, Base achieves better performance. For TGD, compared to setup (d), (a) and (b) show better performance, which implies that when teachers have similar architectures to the student, a better student can be distilled. For setup (b), TGD distills a student which is even better than its teachers. Furthermore, (b) of TGD shows better results than (a) and (d) even if their teachers are larger and better than teachers of (b).

In Table 3, TGD shows the best in all cases. Setup (b) of TGD achives better performance than (a) and (d) cases, which implying that a larger or better teacher does not always generate a superior student, as studied in prior works (Cho & Hariharan, 2019). Also, if channel of networks for teachers and student is similar, a better student can be distilled compared to other combinations. We discuss about Teacher2 in Section 4.3.2.

4.2.2. Different Combinations of Teachers and Student

To analyze the performance with more diverse combinations of teacher-student, we conduct experiments using heterogeneous architectures. Also, we construct teachers with different depth or channel of networks to investigate the interaction and effects between the two teachers.

Heterogeneous Architectures of Teacher-Student. To explore the effectiveness on heterogeneous teachers and students combinations, we construct combiations with different architectures using WRN (Zagoruyko & Komodakis, 2016), ResNet (He et al., 2016), and MobileNetV2 (M.NetV2) (Sandler et al., 2018). We applied the same settings as in the experiments of the previous section.

Table 4. Accuracy (%) with various knowledge distillation methods for different structure of teachers and students on CIFAR-10. Numbers in brackets denote the number of trainable parameters and accuracy for classification task.

Teacher1	WRN 28-1 (0.4M, 85.84)	WRN 16-8 (11.0M, 89.50)	vgg13 (9.4M, 88.56)	WRN 16-3 (1.5M, 88.15)	M.Net V2 (0.6M, 89.61)
Student	vgg8 (3.9M, 85.35±0.07)		ResNet20 (0.3M, 85.08±0.13)		WRN28-1 (0.4M, 85.73±0.06)
KD	86.79 ±0.04	86.59 ±0.17	85.17 ±0.04	85.69 ±0.10	87.86 ±0.15
AT	87.05 ±0.11	87.18 ±0.12	85.45 ±0.32	86.31 ±0.15	88.80 ±0.09
SP	87.11 ±0.22	86.73 ±0.06	84.94 ±0.05	86.33 ±0.09	88.84 ±0.10
Base	86.97 ±0.07	86.93 ±0.09	85.56 ±0.20	86.32 ±0.20	88.05 ±0.11
TGD	**88.03** ±0.07	**87.28** ±0.04	**85.64** ±0.13	**86.48** ±0.05	**88.91** ±0.08

Table 5. Accuracy (%) with various knowledge distillation methods for different structure of teachers and students on CINIC-10.

(Teacher1, Student)	Student	KD	AT	SP	Base	TGD
(WRN16-3, ResNet20)	72.64 ±0.13	74.99 ±0.11	74.95 ±0.19	75.16 ±0.05	74.61 ±0.04	**75.28** ±0.03
(WRN28-3, ResNet20)		74.89 ±0.07	75.05 ±0.15	75.39 ±0.10	74.90 ±0.15	**75.47** ±0.03

In Table 4, TGD outperforms baselines on CIFAR-10. For WRN28-1 teachers and vgg8 student case, the distilled student shows even better performance than its teacher. For vgg13 teachers and ResNet20 student, SP shows even worse than a model learned from scratch. Compared to baselines, TGD achieves better performance, implying topological features help improving performance in KD. In Table 5, TGD performs better than baselines on CINIC-10 and similar tendency of results on CIFAR-10. These results also corroborate that better teacher does not guarantee to generate better student (Cho & Hariharan, 2019).

Analysis on Different Teachers. To investigate the effect of each teacher on distillation, we construct Teacher1 and Teacher2 with different depth or channel of WRN. As shown in Figure 4, when the network capacity of Teacher2 is smaller than that of Teacher1, a better student is distilled, which shows that topological features act as complementary features to those from the raw image data. For (16-3, 16-1) and (16-8, 16-3) cases, (16-3, 16-1) shows better results and

even better than (16-3, 16-3). This presents that (16-3, 16-1) case generates knowledge that is well matched with the student and stronger than the one from other combinations, which alleviates performance degradation issues that may arise due to knowledge gaps.

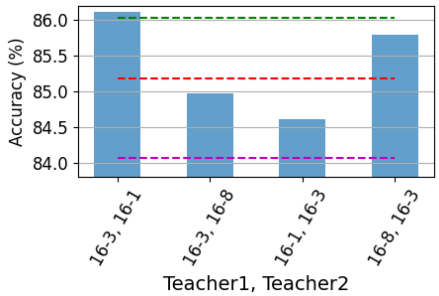


Figure 4. Accuracy (%) of students (WRN16-1) distilled by TGD with various combinations of teachers on CIFAR-10. Teacher1 and Teacher2 consist of different (depth)-(channel) of WRN. Green, red, and magenta dashed lines denote TGD (16-3, 16-3), KD (16-3 Teacher1), and Student (WRN16-1), respectively.

4.3. Ablations and Sensitivity Analysis

In this section, we investigate sensitivity for α, robustness on noise, and evaluate with feature visualization and model reliability.

4.3.1. Effect of α hyperparameter

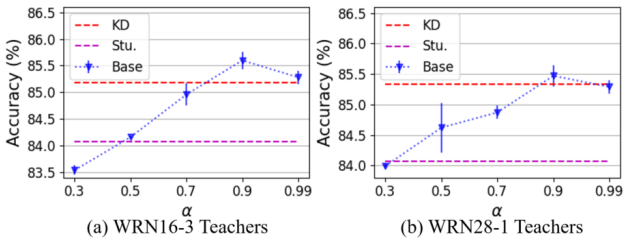


Figure 5. Accuracy (%) of students (WRN16-1) for various methods with different α on CIFAR-10.

We investigate performance of Base on various α hyperparameters. Note, Stu. denotes student trained from scratch. As illustrated in Figure 5, over 0.7 shows better results than KD using a single teacher trained with the raw image data. This implies that relying on Teacher1 more than Teacher2 is effective in distilling a superior student. This tendency is also the same on CINIC-10, which is different from using topological features on time-series data: their optimal α is vary across datasets (e.g. 0.7 or 0.3) (Jeon et al., 2022; 2024). The fact that a high α indicates good results implies that Teacher1 provides stronger information than Teacher2, which is well matched with the student model, since Teacher1 and the student are trained with the same representations and possess similar statistical characteristics. However, using excessive α does not provide the best, which implies topological features indeed act as complement features in distillation to improve performance. With this observations on Base, we utilized high α values which are larger than 0.7. For experiments of previous section, TGD uses 0.99 which is high α and shows the best results. This is because using an annealing strategy encourages a student to preserve features of the raw image data, which are better matched with Teacher1. By leveraging topological features, TGD outperforms baselines including Base. More results are described in appendix.

4.3.2. Analysis on Persistence Image

We use sublevel-set filtration to create PI from an image through TDA, which is simpler than other methods but useful in topological feature extraction (Barnes et al., 2021). As explained in (Barnes et al., 2021), coordinate transforms can affect to extracting topological features. To collect diverse and richer features, multi-scale or multiple coordinate transforms can be leveraged. In our experiments, we used row- and column-wise transforms which collect topological features differently and generate 6 channels of PI. Since datasets in our experiments have complicated patterns (e.g. complex background and multiple channels with diverse region or size of targets), using PI solely cannot show good results. In most cases, performance was close to 35% and 33% in terms of classification accuracy for CIFAR-10 and CINIC-10, respectively. The performance of this model itself is not very good, but when it is included in KD process, it has the advantage of providing useful information that complements features from the original data and helps improving performance, which can be observed in empirical evaluations. Note, Base1 denotes using row-wise transform to generate 3 channels (P_{Rr}, P_{Gr}, P_{Br}) of PIs, and Base2 denotes utilizing multi-wise transforms to create 6 channels of PIs. As shown in Figure 6, Base1 outperforms conventional KD that uses a single teacher trained with the original image data. Using multi-wise transforms, Base2, helps distilling a better student. Thus, providing more diverse topological information can generate a superior student. Also, these results represent the compatibility of topological features in distillation for performance improvement.

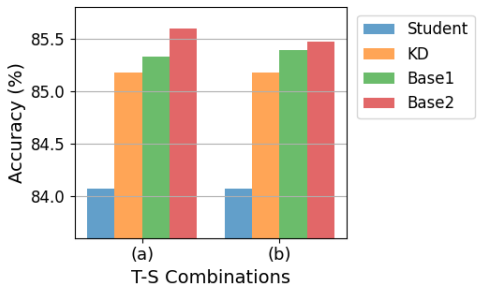


Figure 6. Accuracy (%) of students (WRN16-1) for various methods with setup (a) and (b) on CIFAR-10.

4.3.3. Robustness to Noise

Topological features have shown an excellent ability to withstand noise and perturbations. To explore this, we evaluate student models on a different level of noises for testing data. To inject noises, we utilize Gaussian noise with different levels. Specifically, we apply randomly chosen Gaussian kernel standard deviation from 0.01 to the selected parameter of σ. The kernel size is set as 5×5. The levels of noises are defined by σ as follows; Level 1 (0.5), Level 2 (1.0), Level 3 (1.5), Level 4 (2.0).

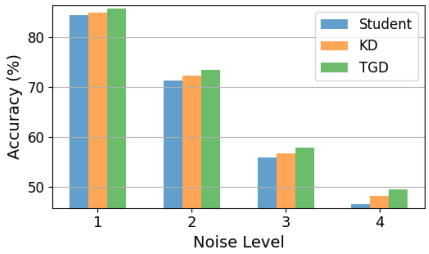


Figure 7. Accuracy (%) of students (WRN16-1) for various corruption severity levels on CIFAR-10. WRN16-3 teachers are utilized.

As shown in Figure 7, as the level of noise increases, the performance of baselines deteriorates significantly, but TGD can withstand the noise much better. This represents that topological features aid in distilling a superior student to withstand noise.

4.3.4. Visualization of Models

To study the behavior of models and characteristics of extracted features intuitively, we visualize features with diverse methods such as similarity maps and activation maps.

Analysis of Feature Map. To explore similarity maps of different methods, we visualize the similarities of high-level intermediate layers that provides more distinguishable maps between methods intuitively, as shown in Figure 8. Student models distilled from diverse methods are used for visualization. MergedT denotes the similarities of integrated features from two teachers, which includes topological features. Student and KD present similar patterns showing column-wise contrasts, which rely on image data alone. TGD shows block-wise patterns that are similar to MergedT, which differs from Student and KD. This represents that TGD encourages a student to obtain topological features, which enables to obtain improved performance. More details are provided in appendix.

Analysis of Activation Map. We visualize the activation maps of various methods by Grad-CAM (Selvaraju et al., 2017) to analyze the coarse localization map of the important regions of each model with various intermediate layers. WRN16-3 teachers and WRN16-1 student are utilized for activation map visualization. In Figure 9, each method focuses on different locations across different intermediate layers.

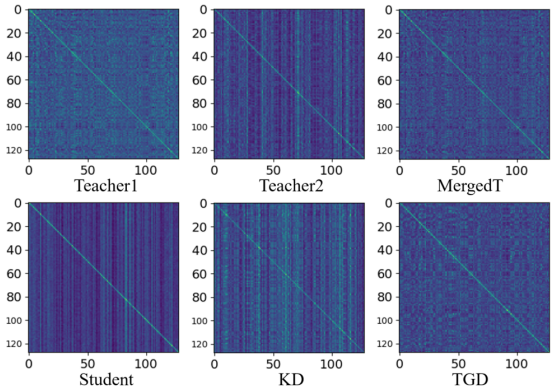


Figure 8. An illustration of similarities for various methods on CIFAR-10. WRN16-3 teachers and WRN16-1 student are utilized.

Compared to other methods, TGD focuses on whole area of a target object, which is recognizable intuitively in maps from the high-level layer. We also visualize maps of high-level layer on different input data, as shown in Figure 10. Compared to other methods, TGD distinctly focuses more on the target area with high weight and less on background regions, which indicates that TGD has better classification ability. More results are provided in appendix.

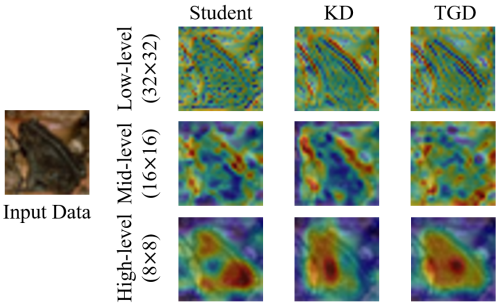


Figure 9. Activation maps for various methods, on a frog image.

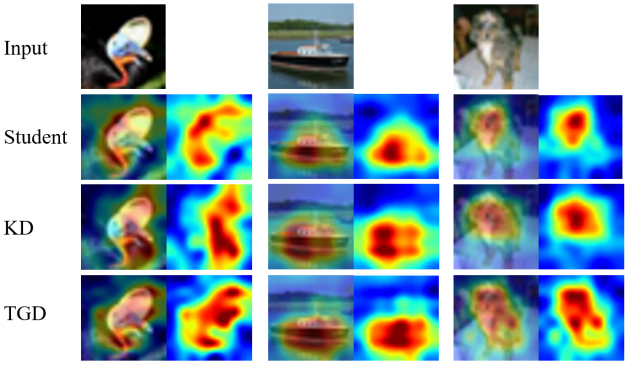


Figure 10. Activation maps of high-level layer for various methods on bird, ship, and dog images.

4.3.5. Model Reliability

To investigate the generalizability of models, we computed expected calibration error (ECE) (Guo et al., 2017) and

negative log likelihood (NLL) (Guo et al., 2017). ECE is to measure calibration errors, implying the reliability of a model. NLL represents the probabilistic quality of the model.

As explained in Table 6, for both setups, TGD shows lower ECE and NLL compared to baselines, which implies topological features aid in improving not only for accuracy but also for generalizability.

Table 6. ECE (%) and NLL (%) for various knowledge distillation methods on CIFAR-10. The results (ECE, NLL) for WRN16-3 and WRN28-1 teachers (Teacher1) are (1.469%, 44.42%) and (2.108%, 64.38%), respectively. (2.273%, 70.49 %) for WRN16-1 Student.

Method	Setup (a)		Setup (b)	
	ECE	NLL	ECE	NLL
KD	2.035	62.26	2.188	67.21
AT	1.978	60.48	2.156	67.14
TGD	**1.865**	**56.05**	**1.940**	**60.12**

4.4. Processing Time

We measure the processing time of various models on CIFAR-10 testing set (10k samples). The total processing time is explained in Table 7. A student (WRN16-1) of TGD takes much less time than teachers on both CPU and GPU. Creating PI (6 channels) takes more than 30k seconds on the CPU, which is not efficient in inference time as well. As described in the prior section, the student by TGD outperforms a model learned from scratch by 1.96% in classification accuracy. These findings clearly highlight the essential necessity of using a compact model for implementation on small devices with limited computational resources and the effectiveness of TGD.

Table 7. Processing time on CIFAR-10 testing set.

Method	Teacher1	Teacher2	TGD
	WRN16-3	WRN16-3	WRN16-1
GPU (sec)	67.83	10280 (PI on CPU) +90.11 (model)	**60.81**
CPU (sec)	263.31	10280 (PI on CPU) +449.48 (model)	**90.20**

5. Discussion

Based on the empirical results, we explored the effectiveness of TGD with various combinations of teachers and students. Also, we investigated characteristics of model behaviors by visualization of similarity maps from intermediate layers.

The focus of this paper is to leverage multiple teachers in KD for transferring topological features to a student, which is to obtain a small-sized and superior model. Utilizing multiple teachers can increase the computational cost in KD training process, however a single distilled model from our approach, TGD, has the advantage of not requiring additional data or layers at test-time after learning once. Also, TGD does not include hidden layers in knowledge transfer process, which does not require much computational cost for fusing different features to utilize multiple teachers. Recently, methods such as the teacher selection strategy (Shang et al., 2023) have been studied to save resources during training time. Reducing computing resources by using multiple teachers trained with different representations requires further exploration.

Teacher2 models trained from scratch with PI only show much worse accuracy in classification tasks compared to Teacher1 and Student models. The performance of Teacher2 is explained in more detail in appendix. However, the network model of Teacher2 is utilized as a teacher in KD to create a superior student by synergizing with Teacher1. The student possession of topological features is observed in similarity maps from intermediate layers. Not only are the transforms in filtration considered in this paper, but there are also various methods to create PI. Specifically, we used simplicial homology in this paper, which is the standard approach in many applications (Barnes et al., 2021). The other popular methods such as cubical homology (Kaczynski et al., 2004) and multiple density areas (Barnes et al., 2021) can be utilized to extract useful features for the same purpose. Additionally, there is still much room for improving performance with a more advanced Teacher2, which can be analyzed with empirical experimentation. This can be more explored in a future work.

6. Conclusion

In this paper, we present a framework for leveraging topological features in KD with multiple teachers and feature similarities for image data analysis. We demonstrated the effectiveness of utilizing topological features in KD based on the proposed method, TGD, under various evaluations, including different combinations of teachers and students, feature and activation maps visualization, and resistance to noise, with empirical testing on classification task.

In future work, more advanced ways to compute persistence features, including transform-based approaches, can be explored in improving performance, such as using cubical persistent homology (Kaczynski et al., 2004) in filtration. Also, more challenging test-conditions can be explored to highlight where TDA features provide robustness in the context of computer vision applications.

Acknowledgements

This work was supported by NSF grant 2323086.

References

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18, 2017.

Ahn, S., Hu, S. X., Damianou, A., Lawrence, N. D., and Dai, Z. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9163–9171, 2019.

Barnes, D., Polanco, L., and Perea, J. A. A comparative study of machine learning methods for persistence diagrams. *Frontiers in Artificial Intelligence*, 4:681174, 2021.

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 535–541, 2006.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Edelsbrunner, H. and Harer, J. L. *Computational topology: an introduction*. American Mathematical Society, 2022.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *Proceedings of the NeurIPS Deep Learning and Representation Learning Workshop*, volume 2, 2015.

Jeon, E. S., Choi, H., Shukla, A., Wang, Y., Buman, M. P., and Turaga, P. Topological knowledge distillation for wearable sensor data. In *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, pp. 837–842, 2022. doi: 10.1109/IEEECONF56349.2022.10052019.

Jeon, E. S., Choi, H., Shukla, A., and Turaga, P. Leveraging angular distributions for improved knowledge distillation. *Neurocomputing*, 518:466–481, 2023.

Jeon, E. S., Choi, H., Shukla, A., Wang, Y., Lee, H., Buman, M. P., and Turaga, P. Topological persistence guided knowledge distillation for wearable sensor data. *Engineering Applications of Artificial Intelligence*, 130:107719, 2024.

Kaczynski, T., Mischaikow, K., and Mrozek, M. *Cubical Homology*, pp. 39–92. Springer New York, New York, NY, 2004. ISBN 978-0-387-21597-6. doi: 10.1007/0-387-21597-2_2. URL https://doi.org/10.1007/0-387-21597-2_2.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, Toronto, Ontario, 2009.

Liu, Y., Zhang, W., and Wang, J. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

Munch, E. A user's guide to topological data analysis. *J. Learn. Anal.*, 4(2), 2017. doi: 10.18608/JLA.2017.42.6. URL https://doi.org/10.18608/jla.2017.42.6.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. Deep learning applications and challenges in big data analytics. *Journal of big data*, 2:1–21, 2015.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3967–3976, 2019.

Reich, S., Mueller, D., and Andrews, N. Ensemble Distillation for Structured Prediction: Calibrated, Accurate, Fast—Choose Three. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5583–5595, 2020.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear

bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 2018.

Saul, N. and Tralie, C. Scikit-tda: Topological data analysis for python, 2019. URL https://doi.org/10.5281/zenodo.2533369.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Seversky, L. M., Davis, S., and Berger, M. On time-series topological data analysis: New data and opportunities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 59–67, 2016.

Shang, R., Li, W., Zhu, S., Jiao, L., and Li, Y. Multi-teacher knowledge distillation based on joint guidance of probe and adaptive corrector. *Neural Networks*, 164:345–356, 2023.

Som, A., Choi, H., Ramamurthy, K. N., Buman, M. P., and Turaga, P. Pi-net: A deep learning approach to extract topological persistence images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 834–835, 2020.

Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., and Liu, T.-Y. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*, 2018.

Thoker, F. M. and Gall, J. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 6–10. IEEE, 2019.

Tung, F. and Mori, G. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, 2019.

Wang, A., Chen, G., Shang, C., Zhang, M., and Liu, L. Human activity recognition in a smart home environment with stacked denoising autoencoders. In *Web-Age Information Management: WAIM 2016 International Workshops, MWDA, SDMMW, and SemiBDMA, Nanchang, China, June 3-5, 2016, Revised Selected Papers 17*, pp. 29–40. Springer, 2016.

Wang, K., Gao, X., Zhao, Y., Li, X., Dou, D., and Xu, C.-Z. Pay attention to features, transfer learn faster cnns. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–14, 2020.

Wang, Y., Behroozmand, R., Johnson, L. P., Bonilha, L., and Fridriksson, J. Topological signal processing and inference of event-related potential response. *Journal of Neuroscience Methods*, 363:109324, 2021. ISSN 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2021.109324.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12, 2016.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the International Conference on Learning and Representations (ICLR)*, pp. 1–13, 2017.

Zheng, Y., Liu, Q., Chen, E., Ge, Y., and Zhao, J. L. Exploiting multi-channels deep convolutional neural networks for multivariate time series classification. *Frontiers of Computer Science*, 10:96–112, 2016.

Zhu, Y. and Wang, Y. Student customized knowledge distillation: Bridging the gap between student and teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5057–5066, 2021.

Appendix

We provide additional experimental settings and results. Also, more details and our findings are discussed. For reproducibility, the source codes, models, etc., are available at https://github.com/jeunsom/TGD.

A. Additional Experiments

A.1. Experimental Settings

For λ and τ, we referred to previous studies (Cho & Hariharan, 2019; Tung & Mori, 2019) to choose the popular parameters in KD (Cho & Hariharan, 2019; Tung & Mori, 2019; Jeon et al., 2023).

Since our method uses similarity maps which can be obtained from outputs of intermediate layers, additional techniques including more hidden layers or interpolations are not used. Also, no augmentation method is applied for CIFAR-10 and CINIC-10.

Life-time threshold denotes points in PD are discarded if their values are less than the threshold.

The all experiments were executed on a desktop equipped with a 2.00 GHz CPU (Intel® Xeon(R) CPU E5-26200 0), 16 GB of memory, and an NVIDIA GeForce GTX 980 graphic card (2048 NVIDIA® CUDA® cores and 4 GB of memory).

A.2. Effect of α Hyperparameter

In Figure 11, results of different methods with various α on CINIC-10 are illustrated. When α is larger than 0.7, the distilled student outperforms baselines. This implies higher weights on Teacher1 generate a superior student. This may because Teacher1's statistical characteristics are more matched with the student, where two models are trained on the same representations of data. Also, this results show that topological features are not stronger but indeed act as complement features to improve the performance. For TGD, 0.99 α shows the best in most of cases since an annealing strategy encourages a student to preserve statistical characteristics of features on the raw image.

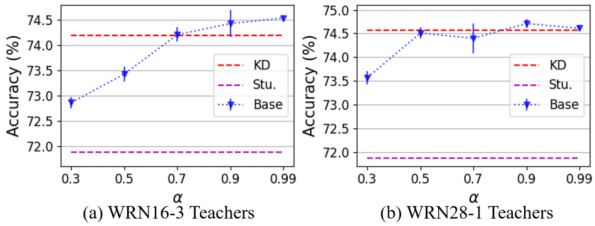


Figure 11. Accuracy (%) of students (WRN16-1) for various methods on CINIC-10.

A.3. Visualization of Models

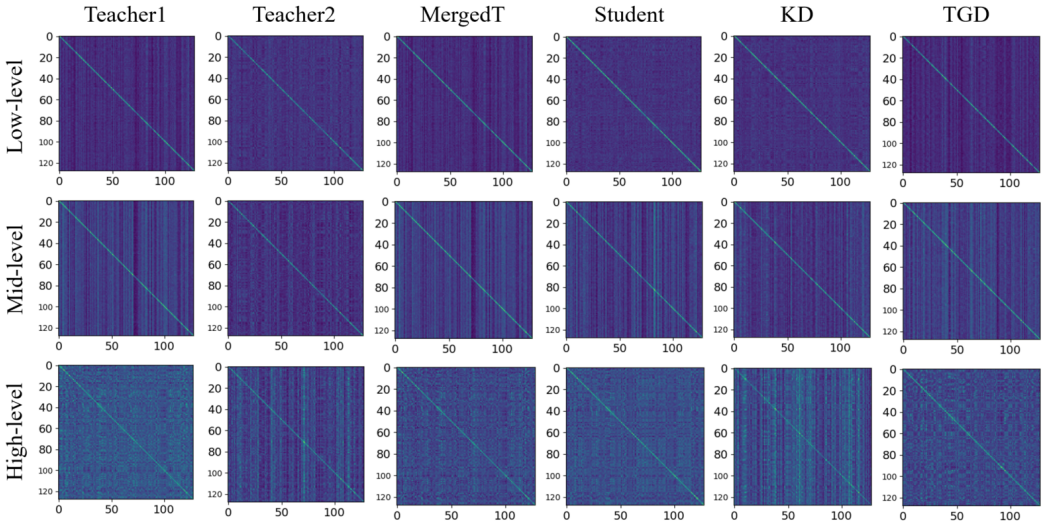


Figure 12. An illustration of similarities for various methods on CIFAR-10. WRN16-3 teachers and WRN16-1 student are utilized.

Analysis of Feature Map. More results from various intermediate layers are illustrated in Figure 12. Compared to low-level, similarities of high-level shows more highlighted patterns and more dissimilar characteristics are shown between different methods. Since Teacher1 and Student are trained from scratch with the image data, they possess similar characteristics. However, KD and Student of high-level have different patterns. This shows the effects of KD. However, TGD differs from KD since TGD is trained with MergedT providing topological features in KD learning process. These results represent that a student distilled by TGD possesses topological features, which is superior than using the raw image data alone in training process.

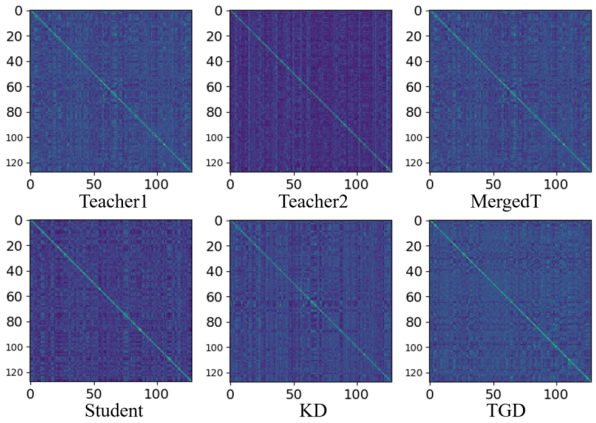


Figure 13. An illustration of similarities for various methods on CINIC-10. WRN16-3 teachers and WRN16-1 student are utilized.

Additionally, we visualize the similarities of student models on CINIC-10 in Figure 13. Student and KD present contrast patterns compared to TGD. TGD shows brighter patterns on correlated points between different samples. Since TGD is trained with MergedT, their patterns and characteristics are more similar. This implies that a distilled student of TGD produces topological features that complements the features from the raw image data.

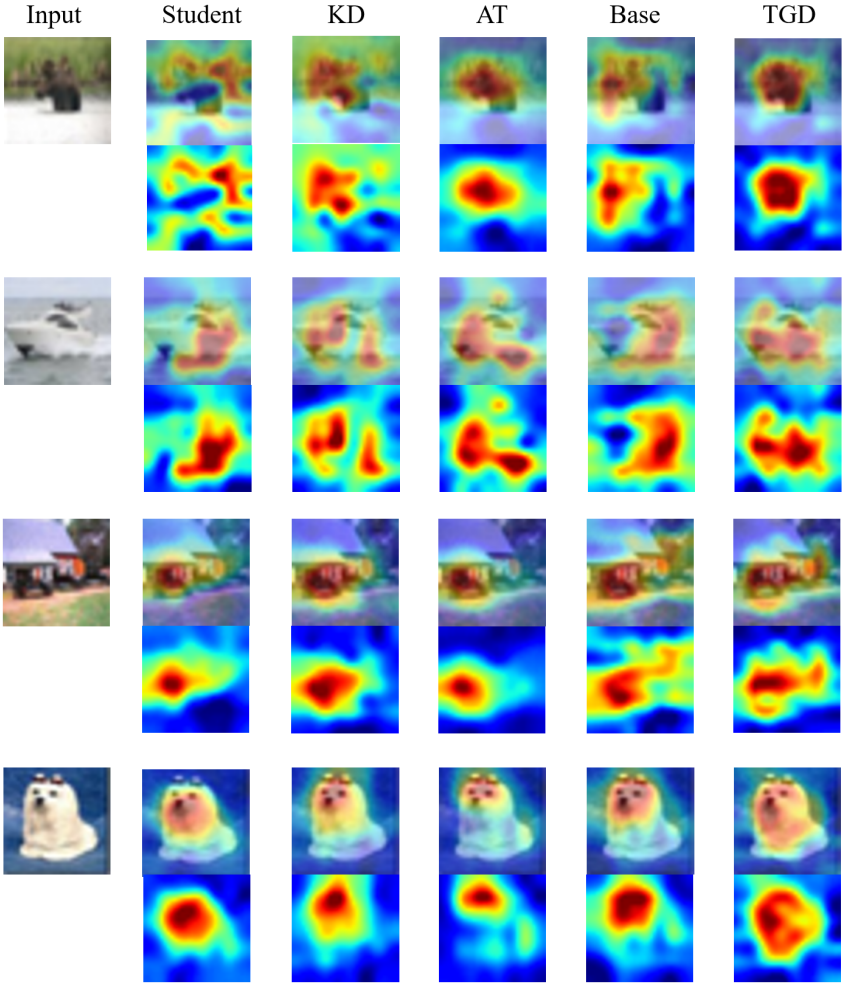


Figure 14. Activation maps of high-level layer for various methods. Labels of input data are deer, ship, truck, and dog.

Analysis of Activation Map. We provide more activation maps of high-level intermediate layer on different input data. WRN16-3 teachers and WRN16-1 student are utilized. As illustrated in Figure 14, TGD focuses more on the target area with high weight and less on the background area compared to other methods. This implies that TGD has better discrimination ability between target and background regions, leading to better classification performance. Thus, based on TGD, topological features guide a student to obtain better discrimination ability, improving performance in image analysis.

A.4. Analysis of PI and Teacher2

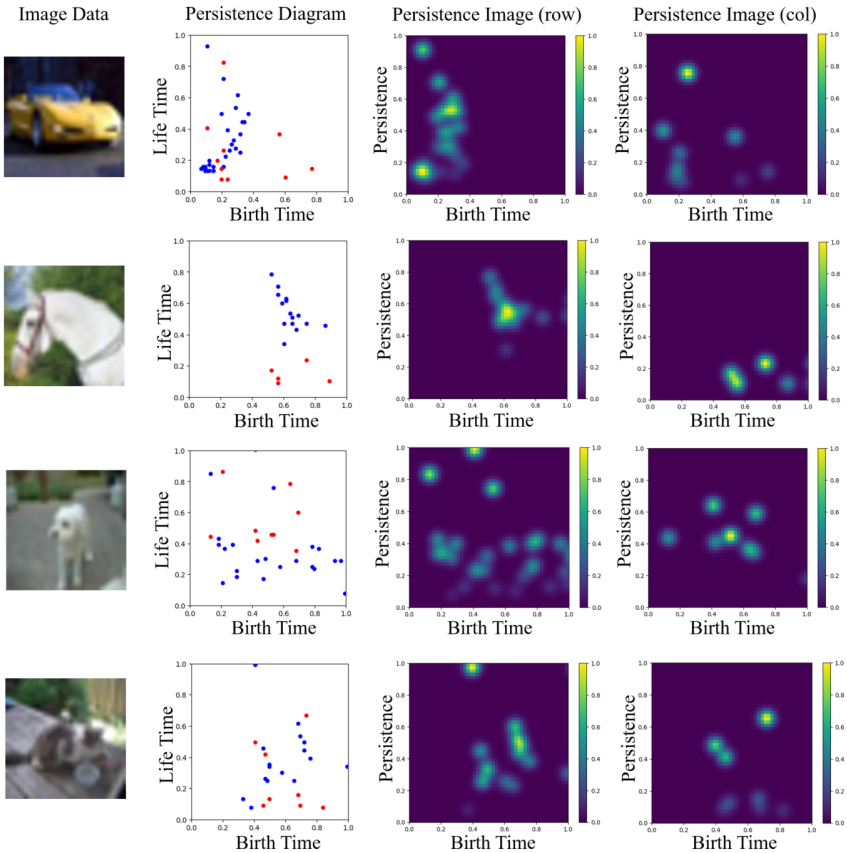
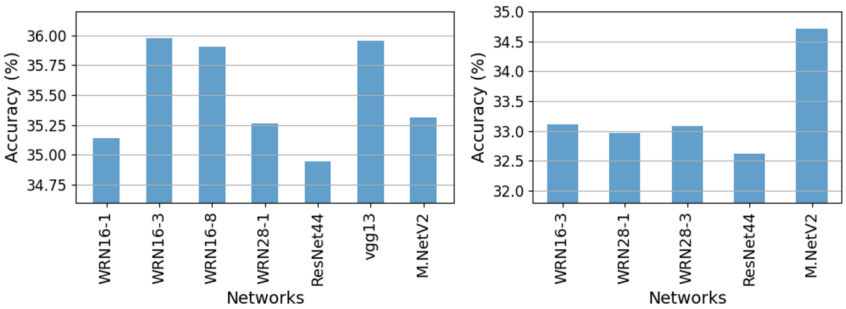


Figure 15. PD and its corresponding PI. Lifetime points in PD appears bright colors in PI. Red and blue denote points from row- and column-wise transforms, respectively.

We illustrate more examples of images and their corresponding PD and PI of (P_{Rr} and P_{Rc}) in Figure 15. We visualize the results on different transforms of image for filtration. As shown in the figure, PIs for row- and column-wise transforms are different, which can be observed intuitively. As explained results of Base on leveraging single or multiple transforms, using more diverse topological information is more useful in distillation.



(a) CIFAR10 (b) CINIC10

Figure 16. Accuracy (%) for various network models trained from scratch with PI.

The models (Teacher2) trained from scratch with PI achieves approximately 35% and 33% in overall cases of classification task for CIFAR-10 and CINIC-10, respectively, as shown in Figure 16. To train models, 6 channels of PIs are utilized. As explained in the manuscript, using PI solely to train a model does not show good results, which differs from time-series data analysis (Jeon et al., 2022; 2024). However, this can be combined in KD process and utilized to improve the performance while this provides complementary features, topological features.

A.5. Robustness to Noise

We investigate the robustness to noise on students distilled with WRN28-1 teachers by various methods, as illustrated in Figure 17. For noise injection, the settings are the same as explained in the manuscript. In all noise levels, TGD shows the best accuracy. This implies that topological features help the student to obtain better resilience to noise.

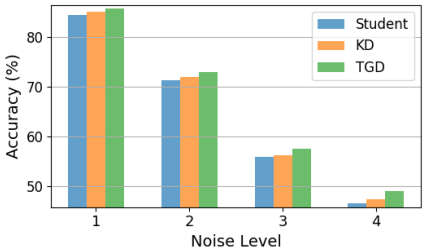


Figure 17. Accuracy (%) of students (WRN16-1) for various corruption severity levels on CIFAR-10. WRN28-1 teachers are utilized.