Contrastive Learning based Deep Convolutional Transform Learning for Multi-sensor Fusion

Saurabh Sahu*, Kriti Kumar*†, A Anil Kumar*, M Girish Chandra*

* TCS Research, Bangalore, India.
† IIIT Delhi, New Delhi, India.
Email: {sahu.saurabh1, kriti.kumar, achannaanil.kumar, m.gchandra}@tcs.com

***Abstract*—This paper introduces a novel approach called Contrastive Learning-based Deep Convolutional Transform Learning (CL-DCTL) for multi-sensor data fusion. The proposed framework addresses challenges posed by availability of data; further with limited labeled instances, for sensor fusion tasks. The proposed CL-DCTL integrates contrastive learning (CL) with deep convolutional transform learning (DCTL) to jointly optimize and extract robust features from multi-modal sensor data in an unsupervised setting. By leveraging DCTL encoders, the method ensures diversity in learned filters while reducing the number of trainable parameters, thus mitigating overfitting issues with limited data. Experimental results on two datasets from different domains show that the proposed CL-DCTL outperforms state-of-the-art methods, providing higher classification accuracy even with only 10% labeled data using an external classifier.**

***Index Terms*—Multi-sensor fusion, Deep convolutional transform learning, Contrastive learning, Representation learning.**

I. Introduction

Single-sensor systems often face challenges associated with data uncertainties and inability to capture complex environmental conditions, leading to sub-optimal performance. These limitations are addressed through multi-modal data fusion, where information from multiple sensors with differing physical characteristics is combined to improve accuracy and reliability [1]. This fusion enables a broader perspective on the environment and improves inferencing capabilities. As a result, multi-modal fusion is widely applied in areas such as computer vision, industrial manufacturing, medical diagnosis and robotics [2]–[4]. Of late, many applications utilize machine learning, particularly techniques based on Deep Neural Networks (DNNs) [2], [3] for performing fusion. While DNNs are effective at identifying complex patterns in data, their requirement for large labeled datasets and heavy computational resources limits their scalability [5]. Since manual labeling is both costly and time-consuming, these techniques are rendered unsuitable for real-world applications where the data is mostly unlabeled or partially labeled. The scenario becomes more challenging when the access to data itself is limited.

To address the challenge of limited/partially labeled data, semi-supervised learning techniques have been extensively studied with a recent focus on Contrastive Learning (CL) [6]–[8]. CL has emerged as a powerful self-supervised technique that learns representations by maximizing similarity within positive pairs —samples derived from the same data instance, while minimizing similarity between negative pairs—samples drawn from different data instances. While some CL based methods use labels information to aid in representation learning [6], [7], others learn representations in a fully unsupervised setting [9]–[14]. The rich representations learned from these techniques provides improved classification performance even when the classifier is trained on limited labeled data. However, most existing CL techniques are designed for single-sensor data [9]–[13], where different data augmentations (like, jitter, permutation etc.) are employed to generate different views of the original sample. The encoder network is then trained to align these views, making them invariant to the applied augmentations.

Single-sensor-based contrastive learning (CL) techniques face challenges with multimodal data due to their inability to effectively capture the heterogeneity that arises from samples generated by different sensor types [14]. Addressing this heterogeneity requires a more specialized approach than the single encoder network used in typical single-sensor CL methods. To address this, the techniques proposed in [14]–[16] employ modality-specific encoders for each sensor. The loss function in [14], [15] is designed to capture both consistent and complementary information across different modalities. The work in [16] proposes cross-modal self-supervised learning that incorporates latent masking in the intermediate embeddings produced by modality-specific encoders and subsequently create global embeddings using a cross-modal aggregator. However, the encoder architectures used in all these techniques [14]–[16] are based on Convolutional Neural Networks (CNNs), where the learned filters are not guaranteed to be unique, resulting in redundant filters to be learned, thereby increasing the number of trainable parameters. Given the limited data scenario, this may result in overfitting and hence, there is a need for learning distinct/unique filters for learning the fused representations for effective performance.

In this work, we propose a novel framework called Contrastive Learning based Deep Convolutional Transform Learning (CL-DCTL), which integrates CL with the Deep Convolutional Transform Learning (DCTL) [17] approach to efficiently extract features from limited multimodal data. This method uses a joint optimization formulation to learn robust representation from multi-modal data in a self-supervised/ unsupervised manner. The DCTL framework learns sensor-

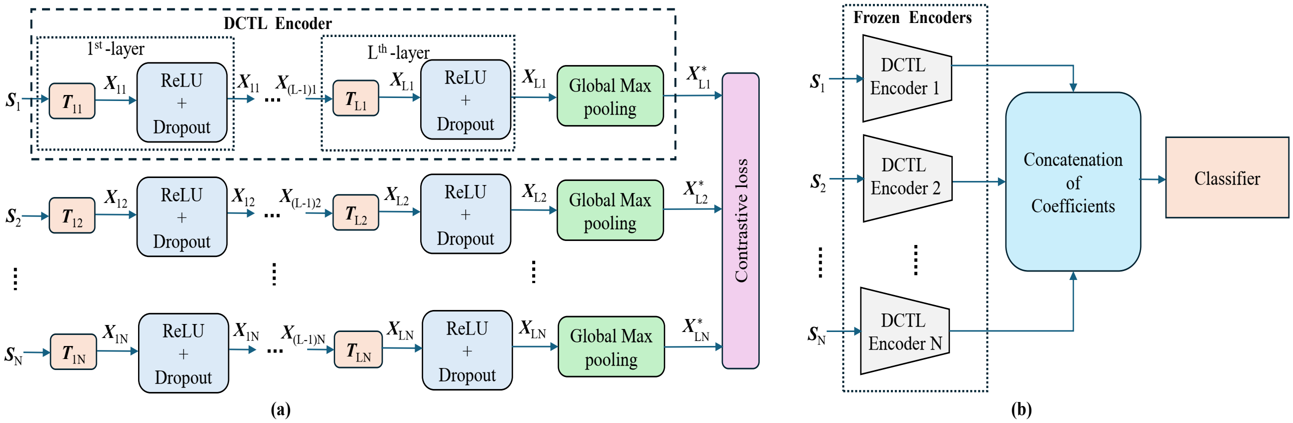

Fig. 1. (a) Block Diagram of the proposed CL-DCTL method for Fusion, (b) Classification using the learned DCTL encoder network.

specific encoders while ensuring diversity in the learned filters, resulting in fewer learnable parameters. On the other hand, CL maximizes mutual information between representations from sensor-specific DCTL encoders. To evaluate the multi-modal representation learning capability of the proposed framework, the features from the learned sensor-specific encoders are concatenated to study the performance of the classifier trained with different percentage of labeled data. Experimental results obtained with two datasets from different domains demonstrate the superior performance of the proposed method compared to state-of-the art methods, even with as low as 10% labeled data.

The paper is organized in the following manner: Section II provides background on the Convolutional Transform Learning (CTL) method, and Section III introduces the proposed CL-DCTL method for multi-sensor fusion. Section IV provides the experimental details, comparisons, and results. Finally, Section V concludes the paper.

II. Background on Convolutional Transform Learning (CTL)

In CTL, a set of M independent convolutional filters $\{\boldsymbol{t}_m\}_{m=1}^{M}$ are learned from the data samples $\{\boldsymbol{s}_k\}_{k=1}^{K}$ with K measurements of length d, in an unsupervised manner. The learned convolutional filters extract features or coefficients $\{\boldsymbol{x}_{m,k}\}_{m=1}^{M}$ using the following standard CTL formulation [18]:

$$\begin{aligned}\min_{\boldsymbol{t}_m,\boldsymbol{x}_{m,k}} \frac{1}{2}\sum_{k=1}^{K}\sum_{m=1}^{M}&(\|(\boldsymbol{t}_m * \boldsymbol{s}_k - \boldsymbol{x}_{m,k}\|_F^2 + \phi(\boldsymbol{x}_{m,k})) \\ &+ \epsilon\|\boldsymbol{T}\|_F^2 - \mu\log\det(\boldsymbol{T})\end{aligned} \tag{1}$$

where $*$ denotes the convolution operation and ϕ is a regularization function that penalizes the coefficients $\boldsymbol{x}_{m,k}$ to avoid overfitting. The matrix $\boldsymbol{T}$ is a concatenation of the filters $[\boldsymbol{t}_1|\boldsymbol{t}_2|\ldots|\boldsymbol{t}_M]$, with $\det(\boldsymbol{T})$ represents its determinant. The hyperparameters ϵ and μ are positive real numbers for the additional constraints imposed on the filters in $\boldsymbol{T}$ for effective learning. While the $\log\det(\boldsymbol{T})$ ensures that the learned filters are unique (linearly independent), the $\|\boldsymbol{T}\|_F^2$ keeps the values bounded to balance the scale.

Re-writing (1) in matrix-vector form results in:

$$\min_{\boldsymbol{T},\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{T}\cdot\boldsymbol{S}-\boldsymbol{X}\|_F^2+\Phi(\boldsymbol{X})+\epsilon\|\boldsymbol{T}\|_F^2-\mu\log\det(\boldsymbol{T}) \tag{2}$$

here, $\boldsymbol{S} = [\boldsymbol{s}_1|\boldsymbol{s}_2|\ldots|\boldsymbol{s}_K]$, $X = [\boldsymbol{x}_{1,k}|\boldsymbol{x}_{2,k}|\ldots|\boldsymbol{x}_{M,k}]_{1\leq k\leq K}$, $\boldsymbol{T}\cdot\boldsymbol{S} = \begin{pmatrix}\boldsymbol{t}_1 * \boldsymbol{s}_1 & .. & \boldsymbol{t}_M * \boldsymbol{s}_1 \\ : & : & : \\ \boldsymbol{t}_1 * \boldsymbol{s}_K & .. & \boldsymbol{t}_M * \boldsymbol{s}_K\end{pmatrix}$ and Φ imposes the regularization ϕ column-wise on $\boldsymbol{X}$. The single-layer CTL formulation in (2) can be made deep (DCTL) by cascading multiple layers of convolutional filters together to produce the coefficients. The DCTL formulation for an L-layer network is expressed as [17]:

$$\begin{aligned}\min_{\boldsymbol{T}_1,\ldots,\boldsymbol{T}_L,\boldsymbol{X}} \; &\frac{1}{2}\|(\boldsymbol{T}_L\cdots(\boldsymbol{T}_2\cdot(\boldsymbol{T}_1\cdot\boldsymbol{S}))) - \boldsymbol{X}\|_F^2 \\ &+ \Phi(\boldsymbol{X}) + \sum_{l=1}^{L}\{\epsilon\|\boldsymbol{T}_l\|_F^2 - \mu\log\det(\boldsymbol{T}_l)\}\end{aligned} \tag{3}$$

where $l = 1, 2, ..., L$ corresponds to the different layers and $\boldsymbol{X}$ represents the coefficients of the DCTL architecture at the last layer. More information on the update of convolutional filters and associated coefficients is presented in [17]. This background forms the basis for the proposed CL-DCTL method for fusion discussed in the next section.

III. Proposed Contrastive Learning based Deep Convolutional Transform Learning (CL-DCTL)

This work proposes a joint optimization formulation employing DCTL and CL for learning the features from multi-modal data in an unsupervised manner. Fig. 1(a) presents the block diagram of the proposed CL-DCTL method. For $i = 1, 2, \ldots, N$ sensors, let $\boldsymbol{S}_i = [\boldsymbol{s}_{1i}, \boldsymbol{s}_{2i}, \ldots, \boldsymbol{s}_{Ki}]$ denote the data collected from the i^{th} sensor. Each of the i^{th} sensor data is processed through a dedicated L-layer DCTL encoder network. Subsequently, contrastive loss is incorporated to maximize the similarity between coefficients/features from the same original sample across different sensors, while reducing similarity between coefficients/features from different samples.

The loss function of the proposed method is a combination of DCTL and contrastive loss, expressed as:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_{DCTL_i} + \alpha \sum_{1 \leq i \leq j \leq N} \mathcal{L}_{CL_{ij}} \tag{4}$$

where α controls the trade-off between DCTL and contrastive loss. Note that the DCTL loss is computed for each sensor, while the contrastive loss is applied to all pairs (i, j) where $i \neq j$, resulting in $N!/(N-2)!$ pairs [19].

A. DCTL loss

Using (3), the DCTL loss for the i^{th} sensor is computed as:

$$\begin{aligned}\mathcal{L}_{DCTL_i} = \min_{\substack{\boldsymbol{T}_{1i},\ldots,\boldsymbol{T}_{Li},\\ \boldsymbol{X}_{Li}}} \frac{1}{2} \|(\boldsymbol{T}_{Li} \ldots (\boldsymbol{T}_{2i} \cdot (\boldsymbol{T}_{1i} \cdot \boldsymbol{S}_i))) - \boldsymbol{X}_{Li}\|_F^2 \\ + \Phi(\boldsymbol{X}_{Li}) + \sum_{l=1}^{L} \{\epsilon \|\boldsymbol{T}_{li}\|_F^2 - \mu \log \det(\boldsymbol{T}_{li})\}\end{aligned} \tag{5}$$

where the first convolutional layer uses filters $\boldsymbol{T}_{1i}$ to generate coefficients $\boldsymbol{X}_{1i}$. These coefficients are regularized using Rectified Linear Unit (ReLU) activation function and dropout before moving through additional convolutional layers. Finally, $\boldsymbol{X}_{Li}$ is obtained that represents the coefficients/features of the L^{th}-layer of the i^{th} DCTL encoder.

B. Contrastive loss

The contrastive loss for the $(i, j)^{th}$ pair is first considered, and the same approach can be used for the remaining pairs. For the $(i, j)^{th}$ pair, contrastive loss is computed between the coefficients $\boldsymbol{X}_{Li}^*$ and $\boldsymbol{X}_{Lj}^*$ from the i^{th} and j^{th} DCTL encoder networks. Here, $\boldsymbol{X}_{Li}^*$ and $\boldsymbol{X}_{Lj}^*$ denote the output of global max pooling on $\boldsymbol{X}_{Li}$ and $\boldsymbol{X}_{Lj}$, respectively. Note that $\boldsymbol{X}_{Li}^* = [\boldsymbol{x}_{Li}^{*1}, \ldots, \boldsymbol{x}_{Li}^{*K}]$ and $\boldsymbol{X}_{Lj}^* = [\boldsymbol{x}_{Lj}^{*1}, \ldots, \boldsymbol{x}_{Lj}^{*K}]$. For each sample $\boldsymbol{x}_{Li}^{*p}$, one positive pair is created by pairing it with $\boldsymbol{x}_{Lj}^{*p}$. Two sets of $K-1$ negative pairs are generated by pairing $\boldsymbol{x}_{Li}^{*p}$ with $\boldsymbol{x}_{Li}^{*q}$ and $\boldsymbol{x}_{Lj}^{*q}$, where $q = \{1, 2, ..., K\}, q \neq p$. Now, the contrastive loss for the p^{th} sample of the $(i, j)^{th}$ pair is given as:

$$\begin{aligned}\mathcal{L}_{CL_{ij}^p} = -\log\Big(\exp(\mathrm{s}(\boldsymbol{x}_{Li}^{*p}, \boldsymbol{x}_{Lj}^{*p})/\tau) \cdot (\sum_{q=1, q\neq p}^{K} \exp(\mathrm{s}(\boldsymbol{x}_{Li}^{*p}, \boldsymbol{x}_{Li}^{*q})/\tau) \\ + \sum_{q=1}^{K} \exp(\mathrm{s}(\boldsymbol{x}_{Li}^{*p}, \boldsymbol{x}_{Lj}^{*q})/\tau))^{-1}\Big)\end{aligned} \tag{6}$$

where $s(\cdot)$ is the cosine similarity score and τ denotes a temperature parameter. The loss $\mathcal{L}_{CL_{ij}}$ of the $(i, j)^{th}$ pair is computed across all positive pairs and can be expressed as: $\sum_{p=1}^{K} \mathcal{L}_{CL_{ij}^p}$. Using the Adaptive Moment Estimation (ADAM) optimizer, the overall loss function in (4) is minimized until it converges to the empirically calculated threshold. This completes the training process, where all L-layer convolutional filters, denoted by $\boldsymbol{T}_{1i}, \boldsymbol{T}_{2i}, \ldots, \boldsymbol{T}_{Li}$ are learned for $i = 1, 2, ..., N$ sensors. Here, each of the DCTL encoder network ensures diversity in the learned convolutional filters.

In this work, the feature/coefficient extraction capabilities of the CL-DCTL framework is assessed for classification tasks with limited labeled data. As illustrated in Fig. 1(b), a supervised learning approach is used, where any existing classifier from [20] can be employed to perform classification on concatenated coefficients obtained from all frozen DCTL encoder (sensor-specific) networks.

IV. Results and discussion

This section introduces two datasets from different domains for evaluating the CL-DCTL method. It also briefly describes the baseline methods for comparison, followed by a detailed discussion of the experimental results in the subsequent sections.

A. Datasets

1) Cylindrical Roller Bearing (CRB) [21]: This is bearing fault classification dataset. It includes vibration and acoustic signals from the NU205E cylindrical roller bearing, recorded at 70 kHz under 2050 rpm and 200 N load. It features three fault types—Roller Fault (RF), Outer-race Fault (OF), and Inner-race Fault (IF), each with five defect widths. For this work, defect widths of 2.12 mm for RF, 1.97 mm for OF, and 2.03 mm for IF, are considered along with a healthy state for classification.

2) USC-HAD [22]: This is a human activity detection dataset. This data was collected using the MotionNode sensing platform, with a 3-axis accelerometer and 3-axis gyrometer placed on the front of the right hip of each subject. It includes data from 14 subjects (7 female, 7 male, aged 21 to 49, mean age 30.1), sampled at 100 Hz. Each subject performed 12 activities: walking forward, right, left, downstairs, upstairs, running forward, sitting, jumping, sleeping, standing, and using an elevator (up and down), each repeated five times.

B. Baseline Methods

Different state-of-the-art methods with single and multiple sensors are used to evaluate the performance of the proposed CL-DCTL method. They are broadly classified into:

- **Single-sensor contrastive learning methods:** These methods consider only single-sensor data: SimCLR-TS [11], TS-TCC [13], SemiTime [6], and TS-TFC [7]. Both SimCLR-TS and TS-TCC learn representations using only unlabeled data, with SimCLR-TS employing a 1-dimensional CNNs as its encoder and TS-TCC utilizing a transformer-based encoder. In contrast, SemiTime and TS-TFC utilize label information for representation learning with CNNs based encoders.
- **Multi-sensor contrastive learning method:** Cosmo [14] is a feature fusion contrastive learning method designed to extract consistent information from multimodal time-series data using separate encoder. We also provide late fusion results employing one of the single-sensor based techniques, SimCLR-TS, to highlight the importance of joint multi-modal learning in the CL-DCTL framework. SimCLR-TS is considered since it does not utilize label

information for representation learning similar to the proposed method. In late fusion, features from separate SimCLR-TS trained on different sensor data are concatenated and fed into an external classifier.

Additionally, comparison with the recent CroSSL [16], a self-supervised learning technique that utilizes latent masking to learn efficient global embeddings from multimodal sensor data is presented.

C. Experimental details and Results

The proposed CL-DCTL formulation with $N = 2$ is used for both datasets, as they both have two sensors. In the CRB dataset, data is collected using vibration and acoustic sensors, while the USC-HAD dataset utilizes a 3-axis accelerometer and a 3-axis gyrometer. Classification performance is evaluated through accuracy scores and the results are presented in Table I and II for the respective datasets. Both datasets are divided into class-balanced training and test sets, with varying percentages of labeling applied to the training data. It is to be noted that since the proposed method does not require label information, the DCTL encoders are learned using the entire training data, while the classifier is learned only on the *labeled instances* of the training data.

The best results of the CL-DCTL method are achieved with a 3-layer DCTL encoder with 32, 64, and 96 convolutional filters, ReLU activation, and a 0.1 dropout rate. ADAM optimizer with a learning rate of 0.001 and a batch size of 64 is used. Hyper-parameters for CL-DCTL are optimized through a grid search to determine the best values for each dataset. More information on the implementation details specific to the two datasets is given below.

1) CRB dataset: The raw sensor data from both the sensors is initially segmented into non-overlapping windows of 4096 samples and then normalized using min-max normalization. For the 3-layer DCTL encoder network, kernel filter sizes of 24, 16, and 8 are applied to layers 1, 2, and 3, respectively, for each sensor. The optimal hyperparameters are $\alpha = 1$, $\mu = \sigma = 10^{-4}$, and $\tau = 0.1$. The training is carried out for 500 epochs after which convergence is observed.

TABLE I
Results with Cylindrical Roller Bearing dataset

Methods	**Training data**		
	10%	**30%**	**50%**
SimCLR-TS (S_1)	0.633	0.813	0.852
SimCLR-TS (S_2)	0.706	0.810	0.841
TS-TCC (S_1)	0.543	0.728	0.797
TS-TCC (S_2)	0.684	0.777	0.815
SemiTime (S_1)	0.631	0.817	0.879
SemiTime (S_2)	0.696	0.853	0.883
TS-TFC (S_1)	0.550	0.598	0.850
TS-TFC (S_2)	0.541	0.750	0.899
SimCLR-TS (Late Fusion)	0.755	0.873	0.905
Cosmo	0.750	0.897	0.942
CroSSL	0.730	0.889	0.939
Proposed CL-DCTL	**0.948**	**0.950**	**0.951**

2) USC-HAD dataset: The raw sensor data for both the sensors is divided into 2-second time window for all the 3 axes and concatenated, resulting in 600 samples for each sensor. The data is normalized using Z-score normalization. Here, the data samples from 10 subjects are used for training, while samples from the remaining four subjects are used for testing. For the 3-layer DCTL encoder network, kernel filter sizes of 8, 16, and 24 are applied to layers 1, 2, and 3, respectively. The optimal hyperparameter values of $\alpha = 0.1$, $\mu = \sigma = 10^{-4}$ and $\tau = 1$ are used to generate the results. Here, the training is carried out for 40 epochs after which convergence is observed.

D. Results Discussion

Table I summarizes the CRB classification results obtained using five-fold cross-validation with a linear classifier trained on 10%, 30%, and 50% of labeled data. Here, $\boldsymbol{S}_1$ refers to vibration sensor data, and $\boldsymbol{S}_2$ refers to acoustic sensor data. Table II shows the USR-HAR classification results obtained from different methods using 10%, 20%, 30%, and 40% of labeled data, where $\boldsymbol{S}_1$ denotes 3-axis accelerometer data and $\boldsymbol{S}_2$ denotes 3-axis gyrometer data. For this dataset, a 2-layer MLP (Multi-Layer Perceptron) classifier with [48, 12] hidden neurons is used instead of a linear classifier for improved classification results. The best-performing methods in both tables are highlighted in bold. Results indicate that the proposed CL-DCTL framework consistently outperforms other methods, demonstrating its effectiveness for multi-modal representation learning.

TABLE II
Results with USC-HAR dataset

Methods	**Training data**			
	10%	**20%**	**30%**	**40%**
SimCLR-TS (S_1)	0.293	0.477	0.482	0.484
SimCLR-TS (S_2)	0.324	0.335	0.365	0.370
TS-TCC (S_1)	0.454	0.483	0.544	0.607
TS-TCC (S_2)	0.410	0.455	0.472	0.509
SemiTime (S_1)	0.418	0.500	0.534	0.566
SemiTime (S_2)	0.456	0.549	0.532	0.546
TS-TFC (S_1)	0.482	0.497	0.522	0.575
TS-TFC (S_2)	0.427	0.519	0.514	0.540
SimCLR-TS (Late Fusion)	0.430	0.485	0.522	0.542
Cosmo	0.502	0.530	0.617	**0.670**
CroSSL	0.437	0.569	0.602	0.633
Proposed CL-DCTL	**0.510**	**0.599**	**0.633**	0.665

It can be observed from Tables I and II that multi-sensor based methods which fuse data from both sensors, show improved performance compared to single-sensor based methods. It can be seen that late fusion of SimCLR-TS outperforms its respective single sensor method for both the datasets, but shows poor performance compared to Cosmo, CroSSL (except for 10% labeled data in Table I) and CL-DCTL. This is due to its inability to capture multi-modal correlations that CL-DCTL, CroSSL and Cosmo handle efficiently through the joint optimization. When compared against Cosmo and CroSSL, the proposed CL-DCTL demonstrates superior performance especially for lower % of labeled data while a comparable performance is observed for higher % of labeled data (≥40%)

for both the datasets. The results demonstrate the fact that diversity promoting DCTL based encoders exploit the complex relationship between multi-modal data in a better way that enables a robust classifier to be learned even with 10% labeled data.

To analyze the computational complexity of the proposed method, we compare the proposed CL-DCTL with the recent Cosmo on the CRB dataset. As shown in Table I, the CL-DCTL method (with $\approx$ 80k trainable parameters) outperforms Cosmo [14] (with 1.5 times more trainable parameters), achieving $\approx$ 20% improvement in classification accuracy even with only 10% labeled data. This highlights the ability of the proposed CL-DCTL to learn effective representations using less number of trainable parameters by enforcing uniqueness in the learned filters. This efficiency not only reduces memory usage but also enhances training and inference speeds, making CL-DCTL highly practical for real-world applications.

Note that an ablation study is also conducted to evaluate the significance of the uniqueness constraint (i.e., the $\log\det(\boldsymbol{T}_{li})$ term in (5)) on the filters in the DCTL loss. The removal of this constraint resulted in a decrease in the accuracy of the proposed method from 94.8% to 72.7% for the case with 10% labeled CRB data, despite maintaining a fixed number of trainable parameters. This demonstrates the advantage of incorporating the uniqueness constraint in the formulation which ensures that the learned filters are linearly independent. The diversity in the learned filters enable effective feature extraction from multi-modal data, resulting in improved performance.

V. Conclusion

This paper presents a novel CL-DCTL framework that combines CL with DCTL to learn robust representations from multi-modal data in an unsupervised manner. The effectiveness of the CL-DCTL framework in feature extraction facilitates robust classifier learning with limited data. Experimental results demonstrate superior classification performance, even with just 10% labeled data, compared to baseline single and multi-sensor approaches. By learning with fewer parameters, the proposed method demonstrates scalability and potential for real-world applications, particularly in scenarios with limited data. Future research could focus on extending the CL-DCTL framework to include various sensor modalities with different dimensionalities.

References

[1] M. L. Fung, M. Z. Chen, and Y. H. Chen, “Sensor fusion: A review of methods and applications,” in *2017 29th Chinese Control And Decision Conference (CCDC)*. IEEE, 2017, pp. 3853–3860.

[2] Q. Tang, J. Liang, and F. Zhu, “A comparative review on multi-modal sensors fusion based on deep learning,” *Signal Processing*, p. 109165, 2023.

[3] X. Xu, Z. Tao, W. Ming, Q. An, and M. Chen, “Intelligent monitoring and diagnostics using a novel integrated model based on deep learning and multi-sensor feature fusion,” *Measurement*, vol. 165, p. 108086, 2020.

[4] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, “Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions,” *Information Fusion*, vol. 91, pp. 424–444, 2023.

[5] H. Yuan, S. Chan, A. P. Creagh, C. Tong, A. Acquah, D. A. Clifton, and A. Doherty, “Self-supervised learning for human activity recognition using 700,000 person-days of wearable data,” *NPJ digital medicine*, vol. 7, no. 1, p. 91, 2024.

[6] H. Fan, F. Zhang, R. Wang, X. Huang, and Z. Li, “Semi-supervised time series classification by temporal relation prediction,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3545–3549.

[7] Z. Liu, Q. Ma, P. Ma, and L. Wang, “Temporal-frequency co-training for time series semi-supervised learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8923–8931.

[8] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, “Self-supervised learning: Generative or contrastive,” *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[9] X. Yang, Z. Zhang, and R. Cui, “Timeclr: A self-supervised contrastive learning framework for univariate time series representation,” *Knowledge-Based Systems*, vol. 245, p. 108606, 2022.

[10] M. N. Mohsenvand, M. R. Izadi, and P. Maes, “Contrastive representation learning for electroencephalogram classification,” in *Machine Learning for Health*. PMLR, 2020, pp. 238–253.

[11] J. Pöppelbaum, G. S. Chadha, and A. Schwung, “Contrastive learning based self-supervised time-series analysis,” *Applied Soft Computing*, vol. 117, p. 108397, 2022.

[12] L. Cui, X. Tian, Q. Wei, and Y. Liu, “A self-attention based contrastive learning method for bearing fault diagnosis,” *Expert Systems with Applications*, vol. 238, p. 121645, 2024.

[13] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, “Time-series representation learning via temporal and contextual contrasting,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 2352–2359, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/324

[14] X. Ouyang, X. Shuai, J. Zhou, I. W. Shi, Z. Xie, G. Xing, and J. Huang, “Cosmo: contrastive fusion learning with small data for multimodal human activity recognition,” in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 324–337.

[15] S. Liu, T. Kimura, D. Liu, R. Wang, J. Li, S. Diggavi, M. Srivastava, and T. Abdelzaher, “Focal: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 309–47 338, 2023.

[16] S. Deldari, D. Spathis, M. Malekzadeh, F. Kawsar, F. D. Salim, and A. Mathur, “Crossl: Cross-modal self-supervised learning for time-series through latent masking,” in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 152–160.

[17] J. Maggu, A. Majumdar, E. Chouzenoux, and G. Chierchia, “Deep convolutional transform learning,” in *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27*. Springer, 2020, pp. 300–307.

[18] J. Maggu, E. Chouzenoux, G. Chierchia, and A. Majumdar, “Convolutional transform learning,” in *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part III 25*. Springer, 2018, pp. 162–174.

[19] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[20] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, “Supervised machine learning: A review of classification techniques,” *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[21] A. Kumar and R. Kumar, “Vibration and acoustic data for defect cases of the cylindrical roller bearing (nbc: Nu205e),” *IEEE Dataport*, 2022.

[22] M. Zhang and A. A. Sawchuk, “Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors,” in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 1036–1043.