Learning with Preserving for Continual Multitask Learning

Hanchen David Wang[*1], **Siwoo Bae**[*1], **Zirong Chen**[1], **Meiyi Ma**[1]

[1]Vanderbilt University
Nashville, TN 37235 USA
{hanchen.wang.1, siwoo.bae, zirong.chen, meiyi.ma}@vanderbilt.edu

Abstract

Artificial intelligence systems in critical fields like autonomous driving and medical imaging analysis often continually learn new tasks using a shared stream of input data. For instance, after learning to detect traffic signs, a model may later need to learn to classify traffic lights or different types of vehicles using the same camera feed. This scenario introduces a challenging setting we term Continual Multitask Learning (CMTL), where a model sequentially learns new tasks on an underlying data distribution without forgetting previously learned abilities. Existing continual learning methods often fail in this setting because they learn fragmented, task-specific features that interfere with one another. To address this, we introduce Learning with Preserving (LwP), a novel framework that shifts the focus from preserving task outputs to maintaining the geometric structure of the shared representation space. The core of LwP is a Dynamically Weighted Distance Preservation (DWDP) loss that prevents representation drift by regularizing the pairwise distances between latent data representations. This mechanism of preserving the underlying geometric structure allows the model to retain implicit knowledge and support diverse tasks without requiring a replay buffer, making it suitable for privacy-conscious applications. Extensive evaluations on time-series and image benchmarks show that LwP not only mitigates catastrophic forgetting but also consistently outperforms state-of-the-art baselines in CMTL tasks.

Code — https://github.com/AICPS-Lab/lwp
Extended version — https://arxiv.org/abs/2511.11676

1 Introduction

In critical applications such as intelligent driving, a system must continually adapt by learning new tasks from a consistent stream of sensory data. This paradigm is driven by practicality: when a new task is introduced, the cost of retrospectively annotating the entire existing dataset with the new labels is often unsustainable (Golatkar, Achille, and Soatto 2020). It is far more efficient to instead leverage the existing data stream by acquiring labels only for the new task as needed. For instance, after a model learns to detect traffic signs, it can later be taught to classify other attributes like scene types using the same camera feed (Shaheen et al. 2022; Kang, Kum, and Kim 2024). Similarly, in medical imaging, a model trained for tumor classification can be updated to recognize secondary characteristics such as tissue density or shape, all while using the same underlying patient scans (An et al. 2025; Freeman et al. 2021). The central challenge is to learn new tasks by acquiring new labels for a shared and potentially evolving input distribution.

We term this challenging real-world setting **Continual Multitask Learning (CMTL)**. CMTL combines challenges from both Multitask Learning (MTL) and Continual Learning (CL). In a typical CMTL scenario, a model is presented with a sequence of tasks, $T_1, T_2, ..., T_n$. Each task introduces a new label set applied to inputs from the same sensor/input space (though their underlying distribution may differ across tasks). This setting is distinct from standard MTL, where all tasks are known and trained on concurrently, and it presents unique challenges not fully addressed by conventional Task-Incremental Learning (Task-IL) methods (Van De Ven, Tuytelaars, and Tolias 2022). The key distinctions are summarized in Table 1.

Characteristic	MTL	Task-IL (Standard CL)	CMTL (Our Setting)
Tasks Arrive Sequentially	X	✓	✓
All Task Data Available Concurrently	✓	X	X
Tasks Share an Input Domain	✓	(Often Not)	✓
Goal: Learn Shared Representation	✓	X	✓
Goal: Mitigate Forgetting	N/A	✓	✓

Table 1: Comparison of Scenarios. CMTL uniquely requires learning new tasks sequentially on a shared data distribution without full access to past task labels.

The primary challenge in CMTL is twofold: the model must 1) retain knowledge from previous tasks to prevent catastrophic forgetting (a core CL goal), and 2) develop robust, shared representations that benefit multiple distinct tasks (a core MTL goal), all without having simultaneous access to the complete labeled data for all tasks. This is especially difficult when the underlying data distribution shifts over time (a non-stationary setting), which can exacerbate task interference.

Although CMTL can be formally categorized as a case of Task-IL, its strong emphasis on building a unified representation from a shared input domain exposes a key weak-

[*]These authors contributed equally.

ness in conventional CL methods. These approaches are primarily designed to prevent catastrophic forgetting, often by isolating task-specific knowledge (Kirkpatrick et al. 2017; Zenke, Poole, and Ganguli 2017; Ma et al. 2020). As our experiments confirm (Table 2), this strategy frequently struggles in the CMTL setting. Handling heterogeneous tasks on shared inputs requires unified representations (Jiao et al. 2025; Huang et al. 2023), yet conventional CL methods fail this condition by design, relying on parameter freezing and replay buffers that isolate task-specific knowledge.

To address these challenges, we introduce **Learning with Preserving (LwP)**, a framework designed specifically for the CMTL setting. Instead of focusing only on task outputs, LwP directly preserves the integrity of the shared representation space throughout sequential training. Its core principles are: **(i)**, a novel regularization term (the DWDP loss) to explicitly maintain the geometric structure of the model's latent space and prevent representation drift, **(ii)**, a stabilized representation space to preserve the implicit knowledge encoded in the geometric relationships between data points, and **(iii)**, a framework operates without a replay buffer, making it efficient for privacy-constrained applications.

The main contributions of this paper are: **(1)** We formally define and analyze CMTL, and we demonstrate that conventional CL methods are often ill-suited for this context. **(2)** We propose Learning with Preserving (LwP), a novel, replay-free framework whose key innovation is a Dynamically Weighted Distance Preservation (DWDP) loss function that maintains the geometric integrity of the latent representation space, mitigating catastrophic forgetting while promoting knowledge sharing. **(3)** We conduct extensive evaluations on image and time-series benchmarks, showing that LwP consistently and significantly outperforms state-of-the-art baselines and, unlike other methods, surpasses the performance of independently trained single-task models, especially in scenarios with distribution shifts.

2 Problem Formulation: Continual Multitask Learning

CMTL is a sequential learning scenario involving T tasks $\{\mathcal{T}_t\}_{t=1}^{T}$. Each task $\mathcal{T}_t$ is associated with a label space $\mathcal{Y}_t$ and involves learning a mapping $f_t : \mathcal{X} \to \mathcal{Y}_t$. At each time step t, we receive a dataset $D_t = \{(x_i, y_i^t)\}_{i=1}^{n_t}$ where the input x_i is drawn from a task-specific distribution $x_i \sim P_X^{(t)}$, and $y_i^t \in \mathcal{Y}_t$ is the corresponding label for task $\mathcal{T}_t$. Note that for time t, only label y_i^t is available.

In simpler CMTL settings, the distribution is stationary ($P_X^{(t)} = P_X^{(j)}$ for all t, j), while in more challenging settings, it can be non-stationary ($P_X^{(t)} \neq P_X^{(j)}$ for $t \neq j$). Our goal is to find a predictor $\varphi(x; \theta_s, \theta_t) : \mathcal{X} \to \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_T$ parameterized by a set of shared parameters θ_s and task-specific parameters θ_t, such that

$$\mathcal{L}(\theta_s, \{\theta_t\}_{t=1}^{T}) \coloneqq \sum_{t=1}^{T} \mathbb{E}_{(\boldsymbol{x}, y^t) \leftarrow \mathcal{D}_t} \left[\ell \left(y^t, \varphi(\boldsymbol{x}, t; \theta_s, \theta_t) \right) \right], \tag{1}$$

is minimized for some loss function $\ell(\cdot, \cdot)$.

3 Learning with Preserving

This section details our proposed framework, Learning with Preserving (LwP). We begin with a high-level overview of the architecture and training process in Section 3.1. Then we explore the theoretical motivation for our core contribution, a preservation loss designed to maintain the geometric structure of the latent space, in Section 3.2. Finally, in Section 3.3, we introduce the dynamic weighting mechanism that makes this loss effective for discriminative tasks.

3.1 Overview

We introduce **Learning with Preserving (LwP)**, a framework designed to manage CMTL scenarios by maintaining the structural integrity of the model's shared representation space across a sequence of tasks. As depicted in Figure 1, LwP uses a shared feature extractor $f_{\theta_s}(\boldsymbol{x})$ to produce a representation $\boldsymbol{z}$. For each task t, a separate, task-specific *head* $g_{\theta_t}(\boldsymbol{z})$ (e.g., a linear layer) generates the final prediction.

The training process at a given task t proceeds as follows. First, we duplicate the model from task $t-1$ and add a new, randomly initialized head, g_{θ_t}, for the new task. The parameters of the previous model (both the feature extractor $f_{\theta_s}^{[t-1]}$ and all previous heads $g_{\theta_o}^{[t-1]}$ for $o < t$) are then **frozen** to serve as a stable teacher.

The current model is trained using a composite loss function, $\mathcal{L}_{\text{lwp}}$, which consists of three key components:

1. A standard supervised loss ($\mathcal{L}_{\text{cur}}$) for the current task t, which trains the new head g_{θ_t} and fine-tunes the shared extractor f_{θ_s} on new data (x_i, y_i^t).
2. A distillation loss ($\mathcal{L}_{\text{old}}$) that preserves performance on previous tasks. The frozen teacher model generates "pseudolabels" $\tilde{y}_o$ for old tasks $o < t$, and the current model is trained to match them.
3. Our novel preservation loss ($\mathcal{L}_{\text{DWDP}}$), which is the core of LwP. This loss prevents representation drift by ensuring the geometric structure of the current latent space ($z^{[t]}$) remains consistent with the structure of the frozen latent space ($z^{[t-1]}$).

Thus, the total objective is a weighted sum of these three components:

$$\mathcal{L}_{\text{lwp}} = \lambda_{\text{c}} \mathcal{L}_{\text{cur}} + \lambda_{\text{o}} \mathcal{L}_{\text{old}} + \lambda_{\text{d}} \mathcal{L}_{\text{DWDP}} \tag{2}$$

3.2 Preserving Implicit Knowledge

In CMTL, a model must preserve not only explicit knowledge from past tasks but also the *implicitly learned knowledge* encoded in its shared representation $\boldsymbol{z}$. We define this implicit knowledge as the geometric structure of the latent space. To prevent this structure from degrading as new tasks are learned, we introduce a loss function designed to explicitly preserve it (see Figure 2).

A direct way to maintain this geometric structure is to ensure that the pairwise distances between data points in the current model's latent space, $Z' = f_{\theta_s^{[t]}}(X)$, remain close to those from the frozen previous model's space, $Z =$

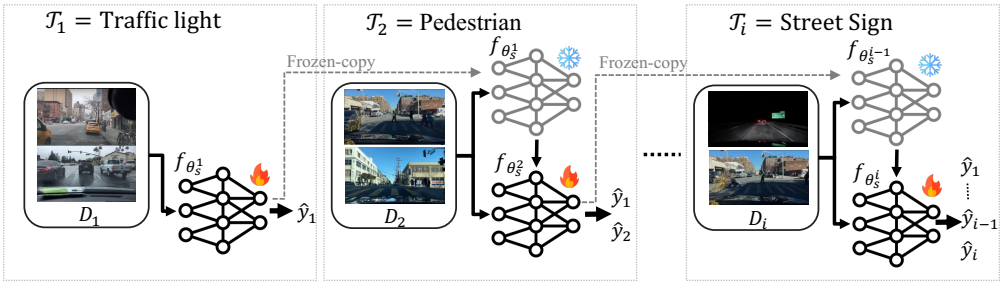


Figure 1: Overview of the LwP framework. For the first task, $\mathcal{T}_1$ (e.g., Traffic Light), the model is trained on data D_1. When learning subsequent tasks like $\mathcal{T}_2$ (Pedestrian), the model from $\mathcal{T}_1$ is frozen (grayed) as a teacher, while a new student model is fine-tuned. This generalizes to any task $\mathcal{T}_i$: the previous model $f_{\theta_s^{[i-1]}}$ acts as teacher for the student model $f_{\theta_s^{[i]}}$, which learns on new data D_i. Each task adds a new head while preserving previously learned representations. See Figure 2 for details.

$f_{\theta_s^{[t-1]}}(X)$. This leads to a family of preservation losses:

$$\mathcal{L}_{\text{pres}}(Z, Z') = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(d(\boldsymbol{z}_i, \boldsymbol{z}_j) - d(\boldsymbol{z}'_i, \boldsymbol{z}'_j) \right)^2, \quad (3)$$

where $d(\cdot, \cdot)$ is a distance or similarity function, which in our primary implementation is the squared Euclidean distance.

This formulation is closely related to kernel methods. Preserving pairwise distances implicitly preserves the structure defined by certain kernels. As we show in the appendix (Wang et al. 2025a), the change in the Gaussian kernel value is bounded by the change in the squared Euclidean distance. Therefore, minimizing the difference in distances effectively minimizes the difference between the Gram matrices $K(Z)$ and $K(Z')$, where $K_{ij} = k(\boldsymbol{z}_i, \boldsymbol{z}_j)$.

Preserving the Gram matrix ($K(Z') \approx K(Z)$) ensures that the new representation Z' is functionally equivalent to the old representation Z within the Reproducing Kernel Hilbert Space (RKHS) induced by the kernel (Yamada 2013; Schölkopf, Herbrich, and Smola 2001). This means that for any function f in this universal function space, its evaluation on a new data point $f(\boldsymbol{z}'_i)$ can be mapped to an equivalent function f' evaluated on the old data point, $f'(\boldsymbol{z}_i)$. Formally, this is because preserving the kernel matrix implies an isometry T in the RKHS such that $\phi(\boldsymbol{z}'_i) = T(\phi(\boldsymbol{z}_i))$, where ϕ is the feature map. Consequently, any learning problem defined on Z has an equivalent solution on Z', as shown in Equations 10-13.

In essence, by using a simple and efficient distance preservation loss, we ensure that the representation space remains stable in a high-dimensional feature space, preserving its capability to solve not only previously learned tasks but also potential future ones.

To formalize the alignment objective, we define the loss function $\mathcal{L}_{pres}$ as the squared Frobenius norm of the difference between the two kernel matrices:

$$\mathcal{L}_{\text{pres}}(Z, Z') = \|K(Z) - K(Z')\|_F^2 \quad (4)$$

By minimizing $\mathcal{L}_{pres}$ while keeping Z fixed, we align the images of Z and Z' under the feature map ϕ:

$$\langle \phi(\boldsymbol{z}_i), \phi(\boldsymbol{z}_j) \rangle_{\mathcal{H}} \approx \langle \phi(\boldsymbol{z}'_i), \phi(\boldsymbol{z}'_j) \rangle_{\mathcal{H}}, \quad \forall i, j. \quad (5)$$

This alignment implies that there exists an isometry $T : \mathcal{H} \to \mathcal{H}$ such that:

$$\phi(\boldsymbol{z}'_i) = T(\phi(\boldsymbol{z}_i)), \quad \forall i. \quad (6)$$

For any function $f \in \mathcal{H}$, the Riesz representation theorem states that there exists a unique element $w_f \in \mathcal{H}$ such that $f(\boldsymbol{z}) = \langle w_f, \phi(\boldsymbol{z}) \rangle_{\mathcal{H}}$. The evaluation of f at $\boldsymbol{z}'_i$ becomes:

$$f(\boldsymbol{z}'_i) = \langle w, \phi(\boldsymbol{z}'_i) \rangle_{\mathcal{H}} = \langle w, T(\phi(\boldsymbol{z}_i)) \rangle_{\mathcal{H}}. \quad (7)$$

Because T is an isometry, its adjoint T^* is also an isometry, and we can write:

$$f(\boldsymbol{z}'_i) = \langle T^* w, \phi(\boldsymbol{z}_i) \rangle_{\mathcal{H}}. \quad (8)$$

Define $w' = T^* w$ and $f'(\boldsymbol{z}) = \langle w', \phi(\boldsymbol{z}) \rangle_{\mathcal{H}}$. Then:

$$f(\boldsymbol{z}'_i) = f'(\boldsymbol{z}_i), \quad \forall i. \quad (9)$$

Thus, Z' becomes an alternative representation that is functionally equivalent to Z in terms of any operations performed within the RKHS induced by the Gaussian kernel. Now, consider a learning problem defined on Z:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{z}_i), y_i) + \Omega(f), \quad (10)$$

and the corresponding problem on Z':

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{z}'_i), y_i) + \Omega(f). \quad (11)$$

Using the relationship $f(\boldsymbol{z}'_i) = f'(\boldsymbol{z}_i)$, the loss terms satisfy $\ell(f(\boldsymbol{z}'_i), y_i) = \ell(f'(\boldsymbol{z}_i), y_i)$. Since $\|f\|_{\mathcal{H}} = \|f'\|_{\mathcal{H}}$, the regularization terms are equal: $\Omega(f) = \Omega(f')$. Thus, the risk functionals for the problems on Z and Z' are equivalent when considering f and f':

$$\frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{z}'_i), y_i) + \Omega(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f'(\boldsymbol{z}_i), y_i) + \Omega(f'). \quad (12)$$

Because the risk functionals are equivalent, the optimal solutions f^* obtained on Z' correspond to the optimal solutions f'^* on Z via the isometry T^*:

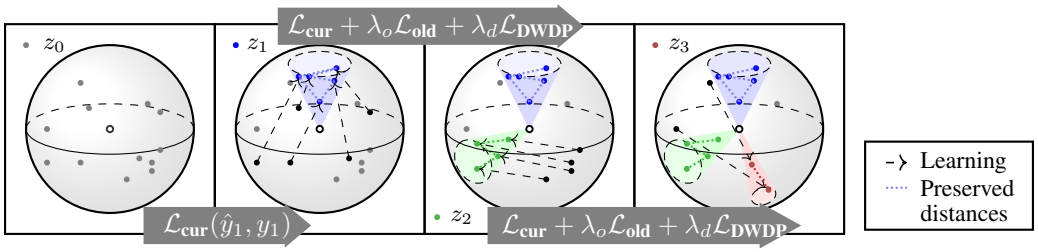


Figure 2: Visualization of latent space evolution during sequential learning. As new tasks arrive ($z_0 \to z_1 \to z_2 \to z_3$), LwP organizes data into distinct task-specific representations (colored points). Dotted lines show pairwise distances preserved by $\mathcal{L}_{\text{DWDP}}$ within each task, maintaining geometric structure while learning new representations. The first task uses only $\mathcal{L}_{\text{cur}}$; subsequent tasks add $\mathcal{L}_{\text{old}}$ and $\mathcal{L}_{\text{DWDP}}$.

$$f^*(\boldsymbol{z}_i') = f'^*(\boldsymbol{z}_i). \tag{13}$$

This means any model trained on Z can be transformed to a model on Z' with identical performance, and vice versa.

Through empirical observation, we have determined that maintaining the squared Euclidean distance leads to enhanced performance. This is likely because the non-exponentiated distance metric more effectively retains the global structure of the representation. Refer to Section 4.6 for the ablation result. Additionally, in appendix (Wang et al. 2025a), we show that the difference in RBF kernel values is bounded by the difference in the squared L^2 norm.

3.3 Dynamic Weighting

$\mathcal{L}_{\text{pres}}$ is designed to maintain the implicitly learned knowledge of the input data in the representation space. However, in scenarios where there are distinct classes or labels, this loss can conflict with other objectives, such as separating distinct classes.

To address this issue, we introduce the Dynamically Weighted Distance Preservation (DWDP) Loss, $\mathcal{L}_{\text{DWDP}}$, which applies a dynamic mask m_{ij} to deactivate preservation for pairs with different labels, preventing conflicts with separation objectives, as illustrated in Figure 2. Unlike PODNet (Douillard et al. 2020), which preserves spatial features uniformly across all pairs, RKD (Park et al. 2019a), which maintains all pairwise distances regardless of class, or Asadi et al. (Asadi et al. 2023), which preserves only prototype-sample distances, LwP uses dynamic per-batch masking to preserve complete intra-class pairwise structure while avoiding inter-class conflicts. The dynamic mask m_{ij} is defined as follows:

$$m_{ij} = \begin{cases} 1, & \text{if } y_i^{[t]} = y_j^{[t]}, \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $y^{[t]}$ represents the labels of the current task. Thus, the DWDP loss is given by:

$$\mathcal{L}_{\text{DWDP}} = \frac{1}{N^2} \sum_{i,j=1}^{N} m_{ij} \left(\Delta d_{ij}\right)^2, \tag{15a}$$

$$\Delta d_{ij} = d(\boldsymbol{z}_i, \boldsymbol{z}_j) - d(\boldsymbol{z}_i', \boldsymbol{z}_j'). \tag{15b}$$

Consequently, this modification alleviates the objective conflict issue at the cost of reducing the scope for preservation to intraclass sets of the current task. Our detailed pseudo-code algorithm is presented in appendix (Wang et al. 2025a).

4 Evaluation

4.1 Experiment Setup

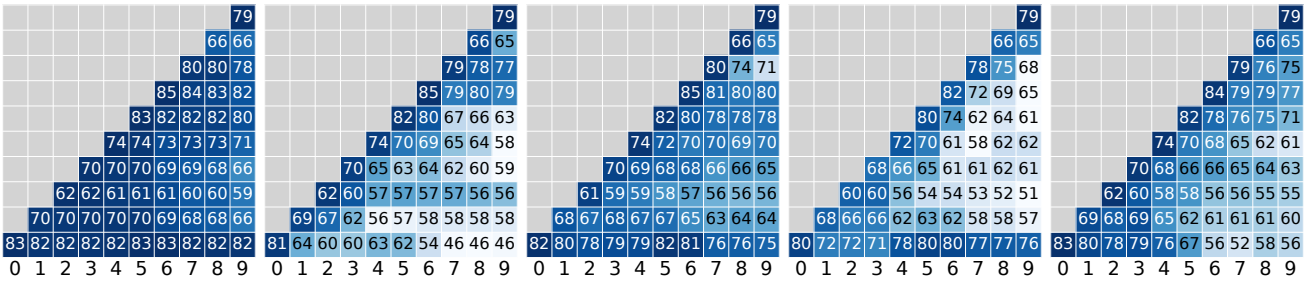
We validate our approach through a series of experiments designed to demonstrate LwP's effectiveness. First, we establish LwP's superior performance and its enhanced robustness to distribution shifts against state-of-the-art CL baselines (Sec. 4.2, 4.3). We then investigate the underlying reason for this success, showing how LwP mitigates catastrophic forgetting by preserving the geometric structure of the representation space (Sec. 4.4). Finally, we validate our specific design choices and scalability with ablation studies and tests on larger models and higher-resolution inputs (Sec. 4.5, 4.6). The appendix provides further details on our experimental setup, model architectures, and baselines, along with additional results on MTL comparisons, accuracy progression over time, and evaluation on training tasks from the MTL to the CMTL setting (Wang et al. 2025a).

Dataset We utilize four datasets from two distinct modalities for our task-incremental learning experiments: BDD100K (Yu et al. 2020a) (object detection in driving scenes), CelebA (Liu et al. 2018) (facial attribute recognition), PhysiQ (Wang and Ma 2023; Wang et al. 2025b, 2024) (IMU-based exercise quality HAR), and FairFace (Karkkainen and Joo 2021) (facial attribute recognition). Each dataset is structured into a series of tasks. Details are in appendix (Wang et al. 2025a).

Baselines Our primary emphasis is on CL baselines since integrating many MTL methods into CMTL often requires substantial modifications to accommodate the incremental characteristics of CMTL. For CL, we compare against Online Bias Correction (OBC) (Chrysakis and Moens 2023), Dual View Consistency (DVC) (Gu et al. 2022), Dark Experience Replay (DER) (Buzzega et al. 2020), DERPP (Boschini et al. 2022), Function Distance Regularization (FDR) (Benjamin, Rolnick, and Kording 2019), Experience Replay (ER) (Robins 1995; Ratcliff 1990), Gradient-based Sample

Method Type	Model	No Shift				Shift Scenarios (BDD100k)			
		BDD100k (3 Tasks)	**CelebA** (10 Tasks)	**PhysiQ** (3 Tasks)	**FairFace** (3 Tasks)	**Weather Shift**	**Scene Shift**	**Time-of-Day Shift**	**Combined Shift**
STL	-	75.123 ± 6.543	72.230 ± 7.297	87.167 ± 10.102	64.435 ± 3.660	76.760 ± 5.210	76.787 ± 5.183	76.418 ± 5.567	76.751 ± 5.180
Naive FT	-	75.572 ± 6.382	70.068 ± 9.941	81.588 ± 15.429	56.291 ± 2.221	75.281 ± 5.893	73.797 ± 8.804	76.417 ± 5.609	76.029 ± 5.895
CL	LwF	76.645 ± 6.577	64.626 ± 10.806	69.952 ± 21.090	61.034 ± 6.162	76.794 ± 5.552	77.499 ± 5.215	76.031 ± 6.062	**76.938 ± 5.006**
	oEWC	74.873 ± 8.375	69.666 ± 9.019	82.640 ± 12.166	63.604 ± 3.122	73.529 ± 9.222	77.224 ± 5.057	75.885 ± 5.782	74.999 ± 6.449
	ER	69.933 ± 9.112	67.598 ± 7.452	76.798 ± 16.347	63.220 ± 4.730	72.287 ± 7.359	68.533 ± 9.071	68.331 ± 8.110	67.372 ± 11.000
	SI	76.601 ± 5.277	68.735 ± 10.545	83.727 ± 11.828	63.359 ± 3.451	75.848 ± 6.193	77.893 ± 4.239	74.818 ± 5.745	74.567 ± 7.218
	GSS	75.434 ± 4.066	71.680 ± 8.468	85.741 ± 10.950	64.230 ± 3.918	74.049 ± 6.793	74.582 ± 5.554	74.332 ± 5.541	73.520 ± 7.924
	FDR	76.779 ± 6.024	69.514 ± 8.917	71.859 ± 18.687	63.709 ± 3.151	76.098 ± 6.564	73.623 ± 11.168	75.588 ± 6.654	76.200 ± 5.847
	DER	77.183 ± 5.055	70.703 ± 8.388	84.796 ± 11.168	64.114 ± 3.484	76.748 ± 5.519	76.166 ± 6.507	76.727 ± 5.252	75.943 ± 6.326
	DERPP	76.677 ± 5.751	67.693 ± 9.425	82.838 ± 13.775	63.806 ± 3.694	76.582 ± 6.071	77.409 ± 5.078	75.846 ± 6.411	68.581 ± 14.210
	DVC	72.683 ± 4.982	71.441 ± 7.640	85.100 ± 10.381	63.848 ± 3.193	72.011 ± 7.061	69.830 ± 7.905	70.019 ± 6.902	70.661 ± 7.100
	OBC	76.993 ± 5.118	70.829 ± 8.267	83.999 ± 11.377	63.872 ± 3.449	72.270 ± 14.364	76.661 ± 5.988	74.835 ± 6.853	73.732 ± 8.306
CMTL	**LwP**	**78.299 ± 3.828**	**73.484 ± 8.019**	**88.242 ± 12.010**	**66.482 ± 3.138**	**77.937 ± 4.041**	**78.198 ± 3.842**	**76.820 ± 5.331**	74.004 ± 11.268

Table 2: Accuracy Comparison Across Models, Datasets, and Distribution Shifts



(a) **LwP** (b) LwF (c) OBC (d) DER (e) DERPP

Figure 3: Selected matrices showcasing the accuracy progression for the dataset CelebA as adding tasks (0-9) in x-axis.

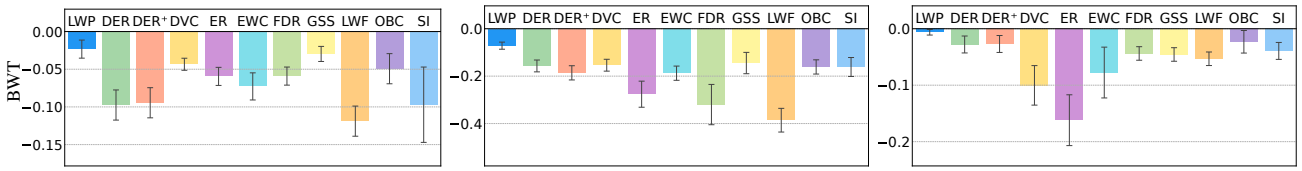
Selection (GSS) (Aljundi et al. 2019), online Elastic Weight Consolidation (oEWC) (Kirkpatrick et al. 2017), Synaptic Intelligence (SI) (Zenke, Poole, and Ganguli 2017), and Learning without Forgetting (LwF) (Li and Hoiem 2017). In addition, we compare our approach with MTL methods in appendix (Wang et al. 2025a). These include the basic MTL approach of training all tasks simultaneously with different predictors (Caruana 1997), as well as more advanced techniques like PCGrad (Yu et al. 2020b), Impartial MTL (IMTL) (Liu et al. 2021), and NashMTL (Navon et al. 2022). We also include a single-task learning (STL) baseline, where each task is learned separately, and a naive fine-tune (FT), where a previously trained model is fine-tuned on the current task. For the choice of distance metric d, we test common options such as Euclidean distance and cosine similarity, as well as loss functions designed to preserve relational knowledge, such as those proposed in RKD (Park et al. 2019b) and Co2L (Cha, Lee, and Shin 2021).

Model Architectures For the image-based datasets (BDD100K, CelebA, and FairFace), we use a ResNet structure as the shared feature extractor f_{θ_s}. For PhysiQ, which consists of time-series data from IMU sensors, we use a 3-layer 1D-CNN model more suited to that modality. This demonstrates the versatility of our LwP framework across different data types and architectures. For all models, each task is handled by a separate linear projection layer (head) applied to the shared representation $\boldsymbol{z}$. We evaluate additional architectures and image sizes in Section 4.5.

4.2 Comprehensive Performance in CMTL

In this experiment, we evaluate the performance of our LwP against several state-of-the-art CL methods. All methods, except for LwF and LwP, are provided with a buffer size of 512 for the CelebA and FairFace datasets, and 46 for the PhysiQ dataset, corresponding to approximately 2-3% of the training set for each dataset. Each model is trained five times using different random seeds. The standard training protocol consists of 20 epochs, with a batch size of 256 for image-based datasets and 32 for PhysiQ, coupled with early stopping. For PhysiQ, we only compare the average accuracy across the final task iteration due to the training instability caused by a smaller dataset size. Table 2 reports the average test accuracy, along with the standard deviation over five runs for each method and dataset. Fig. 3 visualizes the progression of task accuracy in task iterations (left to right). Additional results are in appendix (Wang et al. 2025a).

Table 2 highlights that LwP consistently achieves superior performance across all three benchmarks and is the only method to exceed the Single Task Learning (STL) baseline. This suggests that other continual learning methods likely experience significant task interference. Furthermore, our approach is modality-agnostic, as evidenced by LwP's ability to generalize across different domains. This is demonstrated by the results on the PhysiQ dataset from the IMU sensor domain, which underscores LwP's robustness against



(a) CelebA (b) PhysiQ (c) BDD100k

Figure 4: Selected backward transfer diagrams for the benchmark datasets.

challenges unique to non-image-based tasks. While high final accuracy is a primary goal, the strength of a continual learning method is also measured by its ability to retain knowledge from past tasks. LwP's superior accuracy, particularly its success in surpassing the strong STL baseline, suggests it is more effective at mitigating the task interference that degrades the performance of other methods. This ability to minimize forgetting is analyzed more directly using the Backward Transfer metric in Section 4.4.

4.3 Robustness to Non-Stationary Task Distributions

For real-world applications, such as autonomous driving, a model must be robust not only to test-time shifts but also to scenarios where the training data itself is non-stationary. We leveraged the BDD100k dataset's rich annotations to design a CMTL protocol that explicitly simulates this challenge.

Instead of training on the whole, mixed dataset, we created sequential tasks defined by shifting environmental conditions. For instance, in the Weather Shift scenario, the model was trained sequentially on distinct tasks, where the data for each task came from a specific weather condition (e.g., Task 0: clear, Task 1: rainy, etc.). This protocol directly evaluates the model's ability to handle non-stationary input distributions ($P_X^{(t)} \neq P_X^{(t+1)}$).

The final three columns of Table 2 show the average accuracy of all models trained under these non-stationary protocols. While most models experience a significant performance drop when faced with compounding distribution shifts between tasks, LwP maintains a more significant advantage over the baselines. This demonstrates that LwP learns more generalizable and fundamentally robust representations. By preserving the core latent structure, our method is less susceptible to catastrophic forgetting induced by shifting data distributions, a critical advantage for deploying intelligent systems in dynamic, real-world environments.

4.4 Mitigating Forgetting by Preserving Representation Structure

We demonstrate that LwP effectively mitigates catastrophic forgetting in CMTL settings through both empirical evaluation and controlled demonstrations on a constructed dataset. By preserving the structure of the learned representation space, LwP enables the model to retain performance on previously learned tasks, even as it acquires new ones. This structured preservation also enhances the model's future adaptability: as it accumulates experience across tasks, previously learned features are retained in a generalized form,

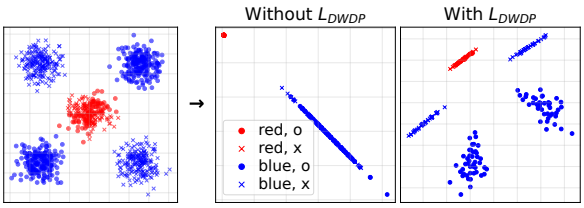


Figure 5: Impact of $\mathcal{L}_{\text{DWDP}}$ on a two-task example (concentric circles, then XOR). After Task 1 (left), learning Task 2 without preservation (middle) loses the geometric structure needed for both tasks. With $\mathcal{L}_{\text{DWDP}}$ (right), structure is maintained for both tasks.

facilitating more efficient learning on subsequent tasks that share latent structure. This is supported by improvements in the Backward Transfer (BWT) metric across all benchmarks, as well as by the toy example in Fig. 5, where maintaining latent geometry is critical on sequential tasks.

The Backward Transfer (Lopez-Paz and Ranzato 2017) is a metric to evaluate the influence of learning the current task on the performance of previous tasks. A positive backward transfer value indicates that, on average, accuracies on the previous tasks have increased during the current task iteration and vice versa. It is defined as:

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} R_{T,i} - R_{i,i}, \tag{16}$$

where T is the index of the current task, i is an index of previous tasks ranging from 1 to $T-1$, $R_{T,i}$ is the accuracy on task i after training up to task T, and $R_{i,i}$ is the accuracy on task i after learning. As illustrated in Fig. 4, we observe that LwP outperforms all baselines in terms of BWT across all benchmarks. This result is consistent with the visualization, where LwP can maintain the accuracy of each task since its initial training.

4.5 Effect of Model Parameters and Image Sizes on Performance

Table 3 illustrates that the LwP method scales effectively with increased input resolution and model size. We find that preserving the Gaussian kernel, as shown in eq. 4, results in improved performance on larger scales, especially with respect to input resolution. In the ResNet50 benchmark utilizing a 224x224 image size, LwP notably surpasses other baselines by achieving an 85% accuracy, which is about

Method Type	Model	ResNet50 (32×32)	ResNet101 (32×32)	ResNet50 (224×224)
CL	LwF	59.277 ± 11.920	58.279 ± 11.202	60.012 ± 14.448
	oEWC	66.975 ± 10.110	67.159 ± 10.506	68.511 ± 13.352
	ER	65.335 ± 9.298	65.646 ± 8.784	65.973 ± 14.729
	SI	66.698 ± 10.030	67.456 ± 9.880	67.747 ± 13.754
	GSS	65.926 ± 13.120	65.587 ± 13.142	69.817 ± 18.771
	FDR	61.753 ± 11.943	61.720 ± 12.017	65.225 ± 15.545
	DER	62.105 ± 12.114	63.797 ± 10.774	69.859 ± 12.690
	DERPP	62.814 ± 11.071	62.957 ± 11.577	68.102 ± 13.557
	DVC	67.084 ± 10.380	65.340 ± 11.427	70.921 ± 13.823
	OBC	64.220 ± 11.237	66.058 ± 10.370	69.319 ± 13.607
CMTL	**LwP**	$\mathbf{67.388 \pm 11.125}$	$\mathbf{69.432 \pm 10.416}$	$\mathbf{85.064 \pm 5.388}$

Table 3: Accuracy Percentage Comparison Across Models on CelebA Dataset (10 Task)

Method on PhysiQ	Dynamic Weighting	w/o Dynamic Weighting
LwP (Full Model)	$\mathbf{88.2 \pm 12.0}$	86.0 ± 12.3
LwP w/o $\mathcal{L}_{\text{old}}$	87.1 ± 9.44	85.4 ± 12.1
LwP (Cosine)	85.4 ± 13.1	84.1 ± 14.4
LwP (RBF)	84.5 ± 13.7	84.8 ± 14.5
IRD (Co2L)	86.4 ± 11.5	79.9 ± 17.1
RKD	85.1 ± 13.3	85.9 ± 11.9

Table 4: Ablation study on the components of the LwP framework on the PhysiQ dataset.

15% percentage points greater than the runner-up. This suggests that, as the input allows the model to create more insightful representations, LwP becomes increasingly advantageous because it can maintain these representations. We also note that the bigger models with the same input size are not performing as well as the one with ResNet18. This is because the inputs lack sufficient information to capture generalized patterns, leading to overfitting.

4.6 Ablation Study

To validate our framework, we perform two types of ablation studies. First, we evaluate the design of the proposed DWDP loss function by selectively disabling the dynamic weighting feature and comparing different distance metrics. We include Co2L (Cha, Lee, and Shin 2021), RKD (Park et al. 2019b) (which preserves distances across all pairs), cosine similarity, and the RBF kernel (eq. 4) as alternatives. The results in Table 4 confirm our design choices: the squared Euclidean distance combined with Dynamic Weighting yields the best performance, underscoring the importance of preserving global structure while avoiding conflicts with class-separation objectives.

Second, we analyze the model's sensitivity to its key hyperparameters. As shown in Figure 6, LwP exhibits remarkable robustness. Its performance remains high and stable across a range of hyperparameter values. Crucially, its entire performance range surpasses that of the best-performing baseline, demonstrating a stable advantage that is not dependent on precise tuning.

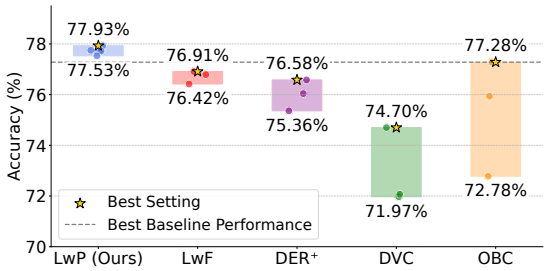


Figure 6: Accuracy on hyperparameter sensitivity analysis on the BDD100k weather shift scenario.

5 Related Work

MTL leverages shared representations across related tasks to improve generalization (Caruana 1997; Sener and Koltun 2018; Chen et al. 2022, 2023), but multiple objectives often conflict. **MGDA** (Sener and Koltun 2018) seeks Pareto-optimal solutions via convex combinations of task-specific gradients, while PCGrad (Yu et al. 2020b) reduce interference onto normal planes. Navon et al. (Navon et al. 2022) model gradient combinations as cooperative bargaining games ensuring task fairness. Methods like IMTL (Liu et al. 2021) incorporate gradient and loss balancing mechanisms. In distributed settings, heterogeneous client specifications (An, Johnson, and Ma 2024) present similar challenges. CL enables sequential task learning without catastrophic forgetting (Ratcliff 1990; Robins 1995). **MER (Riemer et al. 2018)** maximizes knowledge transfer while minimizing interference through experience replay, and HAL (Chaudhry et al. 2021) anchors past knowledge preventing representation drift. In RL settings, (Coursey, Quinones-Grueiro, and Biswas 2025) shows balancing performance and safety is critical for avoiding catastrophic forgetting. Work on pretrained representations includes (Cochran et al. 2024) for explanation generation, while (Duncker et al. 2020) demonstrates organizing recurrent dynamics enables continual learning. CMTL bridges MTL and CL to manage performance across sequential and concurrent tasks (Wu et al. 2023). Methods like MC-SGD (Mirzadeh et al. 2020) enhance CL via linear mode connectivity. **Task-free CL (Aljundi, Kelchtermans, and Tuytelaars 2019)** eliminates task boundaries during training.

6 Conclusion

We addressed Continual Multitask Learning (CMTL), a challenging setting where conventional methods fail by preserving task-specific information while neglecting broadly applicable implicit features needed for unified representations. We introduced *Learning with Preserving* (LwP), which uses dynamically weighted distance preservation to maintain representation space structure without replay buffers, valuable for privacy-sensitive domains. Experiments show LwP surpasses state-of-the-art baselines and single-task models, consistently retaining accuracy and mitigating catastrophic forgetting. Future work includes low-rank approximations to reduce DWDP's $O(N^2)$ complexity for larger-scale applications.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation under Grants 2427711 and 2443803, the Institute of Education Sciences under Award Numbers R305C240010. The opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

References

Aljundi, R.; Kelchtermans, K.; and Tuytelaars, T. 2019. Task-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11254–11263.

Aljundi, R.; Lin, M.; Goujaud, B.; and Bengio, Y. 2019. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32.

An, Z.; Johnson, T. T.; and Ma, M. 2024. Formal logic enabled personalized federated learning through property inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10882–10890.

An, Z.; Moyer, D.; Oguz, I.; Johnson, T. T.; and Ma, M. 2025. ISL: Monitoring Image Segmentation Logic in Medical Imaging Analysis. In *International Conference on Runtime Verification*, 477–496. Springer.

Asadi, N.; Davari, M.; Mudur, S.; Aljundi, R.; and Belilovsky, E. 2023. Prototype-sample relation distillation: Towards replay-free continual learning. In *International Conference on Machine Learning*, 1098–1116. PMLR.

Benjamin, A. S.; Rolnick, D.; and Kording, K. 2019. Measuring and regularizing networks in function space. ArXiv:1805.08289 [cs, stat].

Boschini, M.; Bonicelli, L.; Buzzega, P.; Porrello, A.; and Calderara, S. 2022. Class-Incremental Continual Learning into the eXtended DER-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and CALDERARA, S. 2020. Dark Experience for General Continual Learning: a Strong, Simple Baseline. In *Advances in Neural Information Processing Systems*, volume 33, 15920–15930. Curran Associates, Inc.

Caruana, R. 1997. Multitask learning. *Machine learning*, 28: 41–75.

Cha, H.; Lee, J.; and Shin, J. 2021. Co$ˆ2$L: Contrastive Continual Learning. ArXiv:2106.14413 [cs].

Chaudhry, A.; Gordo, A.; Dokania, P.; Torr, P.; and Lopez-Paz, D. 2021. Using hindsight to anchor past knowledge in continual learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6993–7001.

Chen, Z.; Li, I.; Zhang, H.; Preum, S.; Stankovic, J. A.; and Ma, M. 2022. CitySpec: An Intelligent Assistant System for Requirement Specification in Smart Cities. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, 32–39.

Chen, Z.; Li, I.; Zhang, H.; Preum, S.; Stankovic, J. A.; and Ma, M. 2023. CitySpec with shield: A secure intelligent assistant for requirement formalization. *Pervasive and Mobile Computing*, 92: 101802.

Chrysakis, A.; and Moens, M.-F. 2023. Online bias correction for task-free continual learning. *ICLR 2023 at OpenReview*.

Cochran, K.; Cohn, C.; Hastings, P.; Tomuro, N.; and Hughes, S. 2024. Using BERT to identify causal structure in students' scientific explanations. *International Journal of Artificial Intelligence in Education*, 34(3): 1248–1286.

Coursey, A.; Quinones-Grueiro, M.; and Biswas, G. 2025. On the Design of Safe Continual RL Methods for Control of Nonlinear Systems. *arXiv preprint arXiv:2502.15922*.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. PODNet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*, 86–102. Springer.

Duncker, L.; Driscoll, L.; Shenoy, K. V.; Sahani, M.; and Sussillo, D. 2020. Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in neural information processing systems*, 33: 14387–14397.

Freeman, B.; Hammel, N.; Phene, S.; Huang, A.; Ackermann, R.; Kanzheleva, O.; Hutson, M.; Taggart, C.; Duong, Q.; and Sayres, R. 2021. Iterative Quality Control Strategies for Expert Medical Image Labeling. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9(1): 60–71.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9304–9312.

Gu, Y.; Yang, X.; Wei, K.; and Deng, C. 2022. Not Just Selection, but Exploration: Online Class-Incremental Continual Learning via Dual View Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7442–7451.

Huang, T. E.; Liu, Y.; Van Gool, L.; and Yu, F. 2023. Video task decathlon: Unifying image and video tasks in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8647–8657.

Jiao, Y.; Qiu, H.; Jie, Z.; Chen, S.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2025. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3600–3610.

Kang, D.; Kum, D.; and Kim, S. 2024. Continual Learning for Motion Prediction Model via Meta-Representation Learning and Optimal Memory Buffer Retention Strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15438–15448.

Karkkainen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Li, Z.; and Hoiem, D. 2017. Learning without Forgetting. ArXiv:1606.09282 [cs, stat].

Liu, L.; Li, Y.; Kuang, Z.; Xue, J.-H.; Chen, Y.; Yang, W.; Liao, Q.; and Zhang, W. 2021. Towards Impartial Multi-task Learning. In *International Conference on Learning Representations*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018): 11.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6470–6479. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Ma, M.; Gao, J.; Feng, L.; and Stankovic, J. 2020. STLnet: Signal temporal logic enforced multivariate recurrent neural networks. *Advances in Neural Information Processing Systems*, 33: 14604–14614.

Mirzadeh, S. I.; Farajtabar, M.; Gorur, D.; Pascanu, R.; and Ghasemzadeh, H. 2020. Linear mode connectivity in multitask and continual learning. *arXiv preprint arXiv:2010.04495*.

Navon, A.; Shamsian, A.; Achituve, I.; Maron, H.; Kawaguchi, K.; Chechik, G.; and Fetaya, E. 2022. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*.

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

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019b. Relational Knowledge Distillation. ArXiv:1904.05068 [cs].

Ratcliff, R. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2): 285.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*.

Robins, A. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2): 123–146.

Schölkopf, B.; Herbrich, R.; and Smola, A. J. 2001. A generalized representer theorem. In *International conference on computational learning theory*, 416–426. Springer.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Shaheen, K.; Hanif, M. A.; Hasan, O.; and Shafique, M. 2022. Continual Learning for Real-World Autonomous Systems: Algorithms, Challenges and Frameworks. *Journal of Intelligent & Robotic Systems*, 105(1): 9.

Van De Ven, G. M.; Tuytelaars, T.; and Tolias, A. S. 2022. Three types of incremental learning. *Nature Machine Intelligence*, 4(12): 1185–1197.

Wang, H. D.; Bae, S.; Chen, Z.; and Ma, M. 2025a. Learning with Preserving for Continual Multitask Learning. *arXiv preprint arXiv:2511.11676*.

Wang, H. D.; Bae, S.; Sun, X.; Thatigotla, Y.; and Ma, M. 2025b. EXACT: A Meta-Learning Framework for Precise Exercise Segmentation in Physical Therapy. In *Proceedings of the ACM/IEEE 16th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2025)*, 1–11.

Wang, H. D.; Khan, N.; Chen, A.; Sarkar, N.; Wisniewski, P.; and Ma, M. 2024. MicroXercise: A Micro-Level Comparative and Explainable System for Remote Physical Therapy. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 73–84. IEEE.

Wang, H. D.; and Ma, M. 2023. PhysiQ: Off-site Quality Assessment of Exercise in Physical Therapy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4): 208:1–208:25.

Wu, Z.; Tran, H.; Pirsiavash, H.; and Kolouri, S. 2023. Is multi-task learning an upper bound for continual learning? In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Yamada, A. 2013. Inequalities for Gram matrices and their applications to reproducing kernel Hilbert spaces. *Project Euclid*.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020a. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2636–2645.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020b. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, 3987–3995. PMLR.