2025 International Conference on Machine Learning and Applications (ICMLA)

Multitask Contrastive Learning using Task-Wise Training and Partitioned Embedding Space

M Shifat Hossain[1,a], Sumit Kumar Jha[2,b], Hao Zheng[1,c], Rickard Ewetz[2,d]
[1]*University of Central Florida*, Orlando, FL
[2]*University of Florida*, Gainesville, FL
[[a]mshifat.hossain,[c]hao.zheng]@ucf.edu, [[b]sumit.jha,[d]rewetz]@ufl.edu

***Abstract*—Many real-world computer vision tasks require learning to associate multiple properties of different modalities with the same image. Multi-task learning enables a single model to learn these properties simultaneously by leveraging shared knowledge across related tasks to enhance generalization with single-modal data. On the other hand, contrastive learning effectively captures robust multi-modal features by aligning similar representations and distinguishing dissimilar ones. However, state-of-the-art methods struggle with combining these two learning approaches due to the difficulty in optimizing both shared and task-specific objectives. In this paper, we introduce a Multi-Task Contrastive Learning (*MTCL*) framework that partitions the embedding space to support both classification and regression tasks within a multi-task paradigm. By batching samples with tasks and structuring the embedding space to accommodate diverse task-specific requirements, our method retains the advantages of contrastive learning while addressing the unique challenges of multi-task learning. We evaluate our approach on three benchmark multi-task datasets—Zappos50K, CUB200, and MEDIC. We also introduce a multi-task Vehicles dataset that includes orientation. On the benchmark datasets, our model shows** 24.5%, 17.2%, **and** 30.0% **increase in overall classification accuracy compared to the SOTA methods.**

***Index Terms*—Multi-task learning, Contrastive learning, Multi-objective learning, embedding space partitioning.**

I. Introduction

In computer vision, Multi-task learning (MTL) is particularly important for assigning multiple attributes or properties to images or objects within an image, enabling models to handle complex tasks more effectively. MTL is an optimization framework that has gained prominence in machine learning by improving model generalization through the simultaneous training of multiple related tasks [1]. By sharing a common embedding space across tasks, MTL exploits task commonalities and benefits from inductive transfer, where learning one task enhances performance on others. This approach has demonstrated significant performance improvements over single-task learning across various domains, including computer vision, natural language processing, and speech recognition [2], [3].

Common embedding spaces integrate multi-modal data (e.g., text, images, audio) by projecting them into a shared latent space, enabling tasks like cross-modal retrieval and fusion [4], [5]. Contrastive learning enhances this process by aligning similar data points and distinguishing dissimilar ones through methods like SimCLR and CLIP, effectively learning structured, cross-modal representations without requiring explicit supervision [6], [7]. These techniques have advanced tasks requiring robust alignment and representation learning across diverse modalities.

Despite the individual successes of MTL and contrastive learning, integrating MTL with contrastive learning presents challenges, as MTL predicts multiple independent tasks while contrastive learning aligns and separates representations. Standard contrastive frameworks struggle when a single sample exhibits multiple task-relevant properties, complicating joint optimization. Methods like Multi-Task Contrastive Learning (MTCon) [8], Conditional Similarity Networks (CSN) [9], and Similarity Condition Embedding Networks (SCE-Net) [10] attempt to bridge this gap using task-specific embeddings or gated composite spaces. However, MTCon is computationally intensive for large-scale tasks, while CSN and SCE-Net may fail to generalize to complex datasets. These challenges underscore the need for more efficient frameworks to unify multi-task and contrastive learning effectively.

In this paper, we propose Multi-Task Contrastive Learning (MTCL) framework for that can learn multiple tasks with a multi-modal dataset while making the model light on computational resources. Our approach introduces task-wise trainable partitioned embedding space that facilitates the contrastive learning method in a multi-task objective, improving both generalization and task performance across multi-modal data. This method also ensures the axes in the embedding space are independent within the tasks and can be independently assessed for separate task-specific queries. We further explore multi-objective tasks to show that our method performs equally well with regression tasks. Our results demonstrate that this approach significantly enhances performance on multiple tasks compared to state-of-the-art multi-task contrastive learning techniques. Our main contributions are outlined as follows:

1) We conduct a comprehensive case study comparing the performance of single-task and contrastive multi-task hypothesis models. The study reveals that contrastive learning, when applied without any architectural modifications, fails to surpass the performance of single-task baselines.
2) We propose an approach that explicitly partitions the embedding space for each task and leverages contrastive learning across task-specific labels. This method enables the model to effectively learn shared representations

1946-0759/25/$31.00 ©2025 IEEE
DOI 10.1109/ICMLA66185.2025.00101

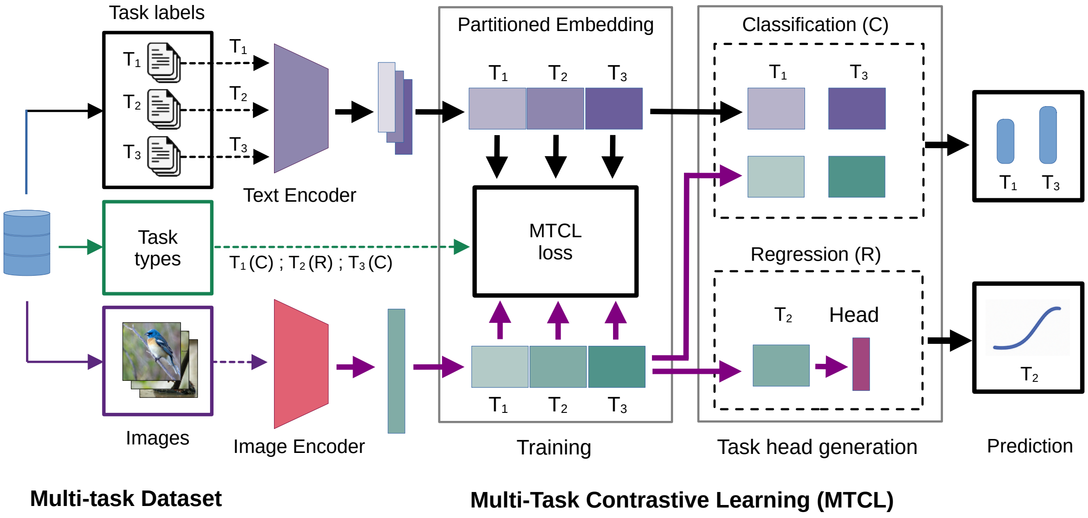

Fig. 1. Overview of the multi-task contrastive learning (MTCL) using task-wise training and partitioned embedding space study. The MTCL framework consists of three parts: batch preparation with task text labels, partitioning of the embedding space, and task agnostic extension of the current framework.

while maintaining task-specific embedding spaces.

3) Additionally, our proposed framework supports multi-objective training, allowing it to perform equally well on regression tasks as on classification tasks. This approach significantly reduces task complexity in real-world scenarios by enabling the training of multi-task contrastive models across datasets with mixed task types.
4) We evaluated our proposed MTCL framework using three benchmark datasets: Zappos50K, CUB200, and MEDIC, along with a custom multi-task vehicle dataset. Our model consistently outperformed state-of-the-art methods, achieving an average accuracy improvement of 30.8% and a lower MAE on the regression task compared to the single-task baseline.

The remaining structure of the paper is outlined as follows: Section II reviews background and related works, Section III provides a motivating case study, and Section IV details our methodology. The findings from our experiments are presented in Section V, and the paper concludes with a discussion of potential future research in Section VI.

II. Background

a) Problem Definition: A conventional approach to learn a multi-task dataset with a Multi-Task Learning (MTL) framework requires the model to train with different task target values for each input sample. Let $D = (x_i, y_i)_{i=1}$ be a multimodal dataset, where x_i represents an input and y_i is the corresponding target for a specific task. The dataset is associated with N distinct tasks, denoted T_1,T_2,...,T_N. Then the objective function for the MTL model can be defined by a weighted sum of individual task losses over the entire dataset D. If the MTL model is denoted as $f_{MTL}(\theta_{MTL})$, where θ_{MTL} is the model parameters, the objective function is as follows,

$$\theta^*_{MTL} = \underset{\theta_{MTL}}{argmin} \sum_{k=1}^{N} \alpha_k . L(f_{MTL}(x_i, y_i^k)) \tag{1}$$

However, existing MTL frameworks face significant challenges when applied to complex scenarios involving multimodal inputs (e.g., image, text, etc.) and diverse learning objectives (classification and regression) within a single model. As depicted in Figure 1, a multimodal multi-objective dataset D can contain multiple inputs (images and text labels) and heterogeneous task types (classification and regression). According to the illustration tasks T_1 and T_3 are classification tasks ($T_1(C)$, $T_3(C)$), while task T_2 is a regression task ($T_2(R)$). Effectively handling such complexity within a unified, resource-constrained architecture remains an open problem.

This work aims to develop a novel multi-task learning framework that efficiently integrates diverse learning objectives within a single model, thereby substantially reducing overall parameter count and computational footprint, without compromising task-specific performance.

b) Contrastive Learning: Contrastive learning is a self-supervised learning framework that learns meaningful representations by contrasting positive pairs of data samples against negative ones. The primary goal is to map similar data points closer in the latent space while pushing dissimilar data points apart [11], [12].

A popular contrastive learning objective is the InfoNCE loss [12], which can be formulated as:

$$\mathcal{L}_{\text{InfoNCE}}(z_i, z_j) = -\log\left(\frac{\exp(sim_{\cos}(z_i, z_j))}{\sum_{k=1}^{N}\exp(sim_{\cos}(z_i, z_k))}\right) \tag{2}$$

Here, z_i and z_j represent the embedding vectors of a positive pair, and $sim_{\cos}$ denotes the cosine similarity function.

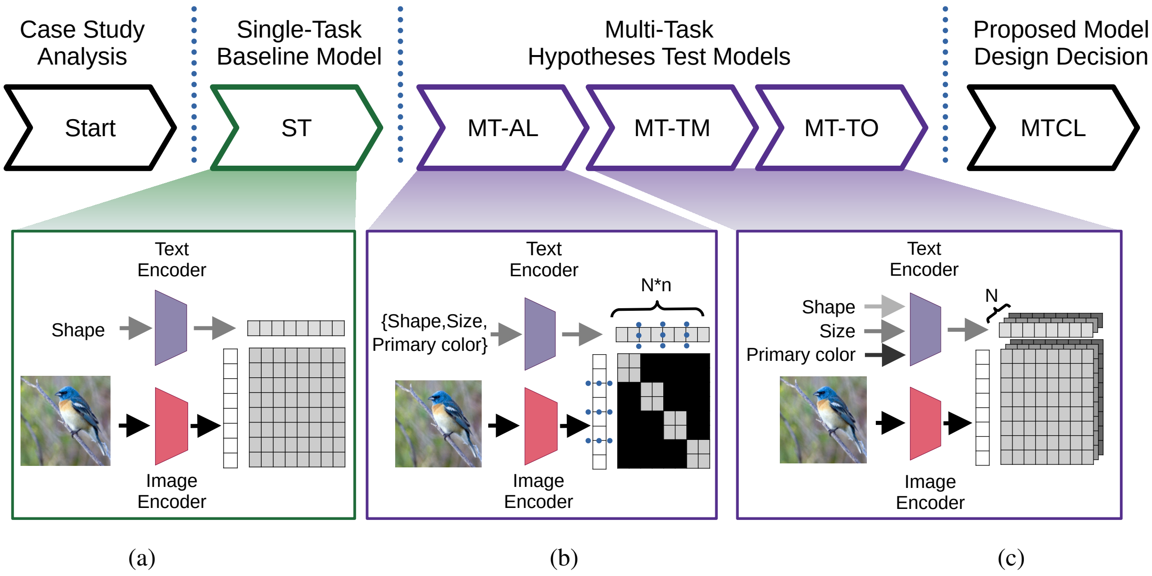

Fig. 2. The overall case study process and architectures for (a) Single-Task Baseline model (b) Multi-Task Aggregated Labels (MT-AL), and (c) Multi-Task Time Multiplexed (MT-TM) and Multi-Task time-multiplexed Task wise Optimization (MT-TO) hypotheses tests. N and n represent number of tasks and length of a partition, respectively.

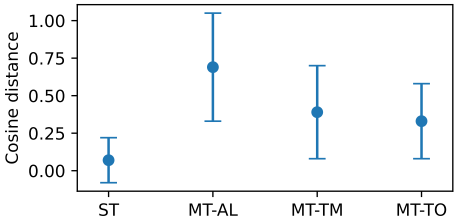

Fig. 3. Mean cosine distance of concepts/properties in the embedding space with the spread for Single-Task (ST) baseline, MT-AL, MT-TM, and MT-TO hypotheses models. Smaller distance with a shorter spread indicates better model performance.

The denominator sums over the similarities of z_i with both positive and negative samples, effectively penalizing the model for assigning high similarity to negative pairs.

Contrastive learning has been extensively employed in VLMs [6], [13], [14], large language models (LLMs) [15], [16], and other multi-modal architectures [17]–[20]. In particular, models like CLIP [6] and ALIGN [21] leverage contrastive learning to align text and image representations, enabling strong performance on multi-modal tasks such as image-text retrieval. In LLMs, contrastive learning has been adapted to align semantic embeddings across different modalities, facilitating multi-modal understanding and improving contextual reasoning [22]–[24].

c) Contrastive Multi-Task Learning: The MTCon [8] framework integrates multi-task and contrastive learning by introducing a Multi-Task Contrastive Loss that leverages multiple similarity metrics to learn robust embeddings. It employs task uncertainty weighting to enhance generalization across in-domain and out-of-domain tasks. The Conditional Similarity Network (CSN) [9] focuses on learning embeddings across multiple semantic subspaces by using masks to disentangle features for distinct notions of similarity, outperforming specialized networks while maintaining interpretability. Finally, SCE-Net [10] builds on CSN by learning similarity conditions and their contributions as latent variables without explicit supervision. It dynamically assigns similarity masks to capture multiple notions of similarity in a unified embedding space, achieving superior generalization and performance across various datasets. While these studies effectively employ contrastive learning for multi-task settings, a notable limitation is their focus on unimodal (image-only) data, primarily learning image-image similarities. Crucially, they do not inherently support the integration of multimodal inputs (e.g., image-text pairs) nor do they directly address diverse task objectives encompassing both classification and regression, which are critical for many real-world applications.

III. Case Study

In the previous section, we investigated how multi-task learning works in general terms. In this section, we present a case study to analyze the principles of contrastive learning in a multi-task setting through a series of step-by-step experiments. These baseline and hypothesis experiments integrate multiple tasks into a single model, with model architectures illustrated in Figures 2(a), (b), and (c), and corresponding results shown in Figure 3. The evaluation metric used throughout this study is the cosine distance between image and text embeddings, where lower distances indicate stronger semantic alignment and thus better model representation capabilities.

To establish an upper performance bound, we first conduct Single-Task (ST) training (Figure 2(a)), where each task is assigned to a separate model that independently maps images to their respective labels using a contrastive learning objective.

The single-task model achieves a low mean cosine distance (mean: 0.07, standard deviation (SD): 0.15), indicating strong alignment between image and text embeddings. This result confirms that learning task-specific features independently can produce highly specialized and effective representations, albeit at the cost of increased model count and compute.

Our initial multi-task approach, the Multi-Task Aggregated Labels (MT-AL) Model, processes all task labels as a concatenated input separated by delimiters, thus encoding multiple task semantics jointly into a single textual representation. To preserve task-specific granularity during training, we apply separate contrastive loss functions to different segments of the embedding vector. Specifically, the final embedding is partitioned into N non-overlapping segments, each of length n, where N denotes the number of tasks and n the length allocated per task segment. This strategy ensures that each task still supervises a distinct region of the shared representation space. However, the MT-AL model exhibits relatively high embedding distances (mean: 0.69, SD: 0.36), suggesting poor alignment between image and text representations.

To improve performance, we introduce the Multi-Task Time-Multiplexed (MT-TM) Model, which batches image samples with their corresponding textual labels and feeds the labels into the text encoder in a temporally partitioned or sequential manner. In this configuration, each image-label sample corresponds to a single task, allowing the model to associate task-specific supervision with its own forward pass. However, similar to the MT-AL setup, the model performs optimization in a joint fashion—aggregating losses from all tasks and updating the shared encoder weights in a single backward pass. This design enables task-specific inference while retaining a common representation space across tasks. It results in a slightly decreased cosine distance in the embedding space (mean: 0.39, SD: 0.31), indicating improved but still suboptimal task alignment.

We further refine this approach with the Multi-Task Time-Multiplexed Task-wise Optimization (MT-TO) model, which uses the same input structure and architecture as MT-TM but modifies the optimization strategy. In MT-TO, the model computes and applies gradients independently for each task by separating the optimization process on a per-task basis, rather than aggregating all losses before backpropagation. This task-isolated optimization allows the encoder to better respect task boundaries and avoid conflicting gradient signals that could arise from heterogeneous task objectives. As a result, MT-TO achieves substantially better alignment performance, with a mean cosine distance of 0.33 and a standard deviation of 0.25, approaching the single-task baseline. These findings highlight the importance of both temporal task separation and independent optimization for effective multi-task contrastive learning, a direction we further elaborate upon in the following section with our proposed partitioned embedding space framework.

IV. Methodology

In this section, we detail our proposed multi-task contrastive learning framework, developed from insights gained by the models described in the case study section. The proposed model is specifically designed to achieve enhanced parameter and computational efficiency in multimodal, multi-objective settings. We introduce an embedding space partitioning technique that strategically allocates embedding dimensions for each task label while tasks are time-multiplexed to a single label encoder. The subsequent subsections will first elaborate on the model architecture, which accommodates both multimodal inputs and the proposed partitioning scheme, followed by a thorough exposition of the Multi-Task Contrastive Learning (MTCL) loss function tailored for joint classification and regression objectives.

A. Model Architecture

The proposed multitask contrastive model follows the standard CLIP architecture and is comprised of two separate encoders—an image encoder and a text encoder. The image encoder processes the visual component of a multitask sample, while the text encoder encodes task-specific textual labels. These encoders are jointly optimized to align representations of paired inputs in a shared latent space.

The image encoder used in this framework is a ResNet101 model, which follows the visual encoder design from the original CLIP implementation. This encoder processes the input image and outputs a latent vector representation, which is then used for similarity comparison against task embeddings. Conversely, the text encoder is a modified version of the CLIP text transformer, designed to support multiple tasks simultaneously. To achieve this, we modify the embedding vector size of the text encoder to be $1/N$th of the image embedding vector, where N is the number of tasks in the given dataset. Then, for training/evaluation purposes, the image embedding vector is partitioned into N equal parts. These parts of the image embedding vector and the individual text embedding vectors are considered the task-specific embedding vectors. The resulting multimodal architecture is trained using our proposed Multi-objective Multi-Task Contrastive Learning (MTCL) loss, which enables simultaneous optimization over heterogeneous task objectives.

B. Multi-Objective MTCL Loss

Our proposed MTCL loss function is designed to support simultaneous learning of both classification and regression tasks within a single unified contrastive framework. This enables the model to handle multitask datasets that consist of mixed objectives—e.g., combining classification and regression. Based on the nature of the task and the dataset annotations, the loss function dynamically configures the optimization objective for each task partition. The following subsections describe the loss formulations for classification and regression settings.

TABLE I
TASK COMPLEXITY OF THE FOUR GIVEN DATASETS. C AND N SYMBOLS DENOTE CLASSIFICATION AND NUMERIC DATA, RESPECTIVELY.

Vehicles				Zappos50K			CUB200			MEDIC			
Vehicle (C)	Azimuth (N)	Color (C)	Background (C)	Category (C)	Closure (C)	Gender (C)	Shape (C)	Size (C)	Primary Color (C)	Damage Severity (C)	Informative (C)	Humanitarian (C)	Disaster Type (C)
8	72	8	4	4	18	4	14	5	15	3	2	4	7

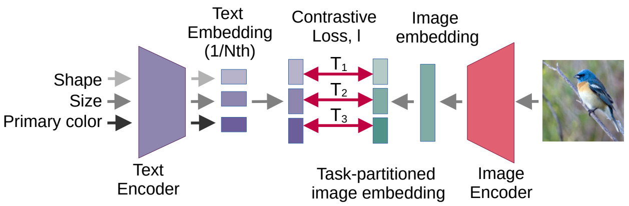

Fig. 4. Multi-task partitioned embedding space for learning classification tasks using contrastive loss.

a) Classification: To train the image and text encoders for classification tasks, we adopt a contrastive learning approach using the InfoNCE loss. In this setup, the text encoder generates embedding vectors for each class, which are then shortened or padded such that the dimensionality of the image and text embeddings match. Let N denote the number of tasks; then the image embedding vector is partitioned into N task-specific segments. Each partition corresponds to the embedding subspace allocated for a particular task, enabling contrastive alignment to occur independently across task-specific regions of the latent space. The classification training process is illustrated in Figure 4.

The contrastive loss is computed per task between the corresponding image embedding partition and the text label embedding. The following equations represent the process:

$$Z_{Lj} = [z_i]_n \tag{3}$$

$$Z_I = [z_i]_m \tag{4}$$

$$Z_{Ij} = [z_i]_{m/N*(j-1)+1:m/N*j} \tag{5}$$

$$l_j = \mathcal{L}_{\text{InfoNCE}}(Z_{Ij}, Z_{Lj}) \tag{6}$$

Here, Z_{Lj} is the label embedding vector for task j, Z_I is the full image embedding, and Z_{Ij} denotes the j-th partition of the image embedding, corresponding to task j. The dimensions n and m represent the size of the label and image embeddings, respectively. N is the number of tasks, and $j \in \{1, 2, \ldots, N\}$ is the task index.

b) Regression: In the case of regression, the goal is not to align embeddings in a contrastive sense, but rather to predict continuous scalar values from the embedding vector. To this end, the image embedding is partitioned as in the classification case, and each partition is passed through a dedicated linear projection layer that maps it to a scalar value.

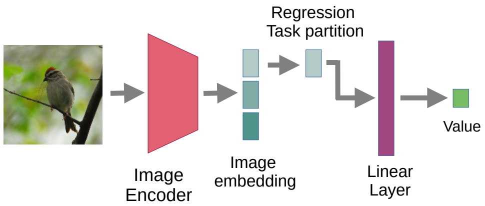

Fig. 5. Multi-task partitioned embedding space for learning classification tasks using L1 loss.

The predicted value is then compared against the true regression target using an $L_1 loss$ function, which corresponds to the Mean Absolute Error (MAE). Figure 5 illustrates the regression architecture within the broader multitask model.

The following equations define the regression-specific loss computation:

$$Z_I = [z_i]_m \tag{7}$$

$$Z_{Ij} = [z_i]_{m/N*(j-1)+1:m/N*j} \tag{8}$$

$$v_I = Linear_{m,1}(Z_{Ij}) \tag{9}$$

$$l_j = \mathcal{L}_{\text{L1}}(v_I, v_T) \tag{10}$$

Here, Z_I and Z_{Ij} denote the full and task-specific image embeddings, respectively, as before. The function $Linear_{m,1}(\cdot)$ is a linear transformation layer mapping the embedding to a scalar prediction. v_I is the predicted regression output, and v_T is the corresponding ground truth target for task j. The index j again identifies the task, and N denotes the total number of tasks.

This formulation enables the proposed model to handle multiple tasks of mixed types in a unified, end-to-end trainable framework by isolating task interactions at the embedding and loss computation levels.

V. RESULTS AND DISCUSSION

In this section, we present the evaluation of the proposed multi-task contrastive learning framework described in Section IV.

A. Datasets

All the experiments on the proposed method are performed with the Zappos50k [25], [26], CUB200 [27], MEDIC [28], and an in-house Vehicles dataset.

The Zappos50k is a large shoe catalog database containing about 50,000 images. This dataset contains three tasks ('Category', 'Closure', and 'Gender'), each containing 4, 18, and 4 classes, respectively. The CUB200 dataset contains about 12,000 images of birds with their metadata properties. The tasks are 'Shape', 'Size', and 'Primary Color', containing 14, 5, and 15 classes, respectively. The MEDIC dataset is a humanitarian response image classification dataset containing four different tasks ('Damage Severity', 'Informative', 'Humanitarian', and 'Disaster Type'). The number of classes in these tasks is 3, 2, 4, and 7, respectively.

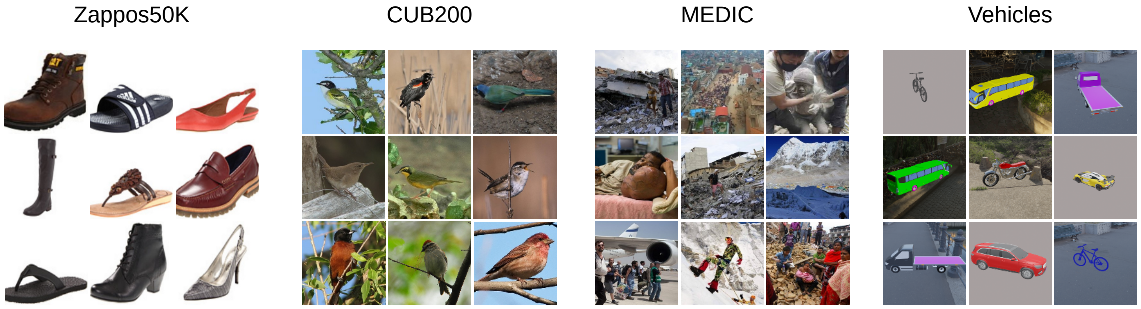

Fig. 6. Image input samples from the benchmark (Zappos50K, CUB200, and MEDIC) and in-house (Vehicles) datasets.

TABLE II
MAX ACCURACY Acc AND WIDTH OF THE EMBEDDING SPACE W IN A RUN. T_1 TO T_4 REFER TO THE TASKS IN A DATASET. THE TASKS FOR THE CORRESPONDING DATASETS ARE - CUB200:{SHAPE, SIZE, PRIMARY COLOR}, ZAPPOS50K:{CATEGORY, CLOSURE, GENDER}, MEDIC:{DAMAGE SEVERITY, INFORMATIVE, HUMANITARIAN, DISASTER TYPE}, VEHICLES:{VEHICLE, AZIMUTH, COLOR, BACKGROUND}.

Dataset	T_1		T_2		T_3		T_4	
	Acc (%)	W	Acc (%)	W	Acc (%)	W	Acc (%)	W
CUB200	55.2	512	64.0	32	58.8	128	-	-
Zappos50k	90.6	64	80.0	128	73.6	64	-	-
MEDIC	69.3	64	85.3	16	79.3	16	59.3	512
Vehicles	100.0	512	75.7	256	99.9	512	100.0	512

In our in-house Vehicles dataset we tried to estimate the car type, the car's angular rotation corresponding to the camera (azimuth), the color of the car, and the type of background. The three tasks - vehicle type, vehicle color, and vehicle background are classification tasks. And the task azimuth can be considered both a classification or a regression task. To consider the azimuth as a classification problem, we have divided Euler angle 0 to 360 with a 5-degree step size. Resulting in 8, 8, 4, and 72 classes for the corresponding tasks. Table I shows the complexity of each task for all the datasets used in this study. Figure 6 shows some image input samples of the datasets used in this study. The task class labels and values for all the datasets are given in the supplementary document.

B. Environment Setup

This complete system is implemented using the PyTorch deep learning framework [29]. All experiments are conducted on a high-performance computing setup equipped with four NVIDIA A40 GPUs, each offering 48GB of VRAM, allowing for efficient parallel training and large batch sizes suitable for multitask contrastive objectives.

The proposed method is evaluated quantitatively and benchmarked against several state-of-the-art out-of-distribution (OOD) detection models. These models include Conditional Similarity Networks (CSN) [9], SCE-Net [10], and Multi-task Contrastive Learning (MTCon) [8]. All three models represent recent advances in OOD detection under multitask settings and employ contrastive learning techniques to align modality representations across tasks. The inclusion of these baselines enables a comprehensive evaluation of our method's performance and its ability to generalize across heterogeneous tasks.

TABLE III
TASK-WISE PERFORMANCE ON ZAPPOS50K

	Category	Closure	Gender	Overall
ST Baseline	89.6	75.3	66.6	77.2
MT-AL Test	58.4	34.9	42.5	45.3
MT-TO Test	73.8	67.9	60.0	67.2
CSN	83.3	72.1	69.2	74.9
SCE-Net	86.2	**75.3**	71.3	77.6
MTCon	62.8	42.8	54.5	53.4
MTCL	**87.4**	72.7	**73.5**	**77.9**

TABLE IV
TASK-WISE PERFORMANCE ON CUB200

	Shape	Size	Primary Color	Overall
ST Baseline	49.2	52.8	60.4	54.1
MT-AL Test	26.8	54.8	32.8	38.1
MT-TO Test	52.0	51.6	43.2	48.9
CSN	45.1	48.2	25.2	39.5
SCE-Net	**48.3**	51.5	28.8	42.9
MTCon	46.4	48.0	24.8	39.7
MTCL	47.6	**64.8**	**55.2**	**55.9**

C. Multi-task contrastive learning

Tables III, IV, V, and VI present the performance evaluation results for the Zappos50k, CUB200, MEDIC, and Vehicles datasets, respectively. Across all datasets, the proposed method consistently outperformed state-of-the-art approaches. The single-task baseline represents the upper performance bound for models trained on individual tasks in isolation. Notably,

TABLE V
TASK-WISE PERFORMANCE ON MEDIC

	Damage Severity	Informative	Humanitarian	Disaster Type	Overall
ST Baseline	72.7	80.0	77.3	62.0	73.0
MT-AL Test	60.7	70.0	62.7	45.3	59.7
MT-TO Test	76.7	84.7	80.7	66.7	77.2
CSN	75.1	76.3	70.5	70.0	73.0
SCE-Net	77.3	77.5	72.1	71.1	74.5
MTCon	43.3	68.0	52.7	36.0	50.0
MTCL	**81.3**	**89.3**	**84.0**	**65.3**	**80.0**

TABLE VI
TASK-WISE PERFORMANCE ON MULTI-TASK VEHICLES DATASET

	Vehicle	Azimuth	Color	Background	Overall
ST Baseline	100.0	44.1	100.0	100.0	86.0
MT-AL Test	48.2	4.4	83.6	89.2	56.4
MT-TO Test	97.6	37.0	99.2	100.0	83.5
MTCon	20.4	3.2	33.3	90.1	36.8
MTCL	**99.2**	**58.7**	**98.4**	**100.0**	**89.1**

TABLE VII
MIXED TASK TYPE EXPERIMENT FOR THE LOSS AGNOSTICITY ON VEHICLES DATASET. R AND C INDICATES REGRESSION AND CLASSIFICATION TASKS, RESPECTIVELY. THE VALUE OF C IS THE HIGHER THE BETTER, FOR R THE LOWER THE BETTER.

Task Type	Vehicle C (%)	Azimuth R (MAE)	Color C (%)	Background C (%)
Single-Task Baseline	100.0	20.6	100.0	100.0
MTCL	**100.0**	**18.8**	**99.7**	**100.0**

the proposed method achieved performance close to the single-task baseline.

In some instances, we can see results better than the single-task baselines. Specifically, improvements over the single-task baseline were observed in the following cases: Azimuth for Vehicles; Gender for Zappos50k; Shape and Size for CUB200; and Damage Severity, Informative, Humanitarian, and Disaster Type for MEDIC. These results suggest that learning multiple tasks simultaneously enables the model to leverage inter-task dependencies, resulting in enhanced performance compared to models trained on individual tasks.

D. Multi-Objective MTCL

To test the multi-objective support of our proposed framework, the model is modified to include an additional linear layer designed to handle regression tasks. This layer operates on the specific partition of the embedding vector associated with the regression objective and projects it to a single scalar output. This output is then compared with the target values using an appropriate regression loss function.

For our experiments, we employ the $L1loss$ (Mean Absolute Error, MAE) as the regression loss function due to its robustness and interpretability. The model is evaluated on the USD dataset within the proposed multi-task contrastive learning (MTCL) framework, where the azimuth angle prediction task is framed as a regression problem. By treating azimuth as a continuous target rather than a discrete classification problem, we reduce the output complexity from a 72-neuron classification layer (one per azimuth bin) to a single scalar output, simplifying the learning objective while preserving fidelity. This reduction in dimensionality not only lowers computational overhead but also improves numerical stability and convergence behavior.

The regression-based formulation leads to improved performance in terms of prediction accuracy. Specifically, the MAE achieved for the regression version of the azimuth task is 18.8 (lower is better), compared to the classification baseline which achieves only 58.7% accuracy—barely above the midpoint baseline of 50.0%. These results validate the effectiveness of the proposed loss-agnostic framework in handling continuous-valued targets more efficiently than traditional classification-based approaches. The results also demonstrate that the MTCL framework supports a seamless blend of different objective types within a single encoder-decoder architecture, thereby improving flexibility and task precision.

As detailed in Table VII, the regression-based azimuth task in our multi-task framework not only surpasses its single-task baseline but also demonstrates that joint learning with contrastive objectives and shared representations significantly enhances precision for low-complexity continuous prediction tasks, ultimately leading to higher accuracy with reduced training complexity across mixed task types.

E. Effect of embedding width

Since all tasks in a multi-task learning framework share a common embedding space, the embedding width directly impacts model performance. Table II reports the maximum accuracy for each dataset and the corresponding embedding width yielding optimal performance. A detailed task-wise analysis is provided in the supplementary document.

From Table II, tasks such as Shape and Primary Color in CUB200, Closure in Zappos50K, Disaster Type in MEDIC, and all tasks in the Vehicles dataset require a larger embedding width for optimal performance. In contrast, tasks like Size in CUB200, Category and Gender in Zappos50K, and Damage Severity, Informative, and Humanitarian classification in MEDIC perform better with a smaller embedding width. This trend may correlate with dataset complexity, as shown in Table I, where higher task complexity often necessitates a larger embedding space.

However, as illustrated in the Accuracy-Embedding Width plots in the supplementary document, most tasks exhibit a peak in performance at a specific width, followed by a decline and gradual recovery. Notably, the Size task in CUB200 shows a continuous performance drop with increasing embedding width, possibly due to the need for additional training epochs to optimize the larger parameter space effectively.

VI. Conclusion

In this paper, we introduced Multi-Task Contrastive Learning (MTCL) framework that effectively integrates multi-task learning and contrastive learning to balance shared and task-specific objectives. MTCL's partitioned embedding space facilitates independent task supervision while maintaining a unified representation, leading to stable training and effective feature disentanglement. This approach yielded significant performance gains across diverse benchmarks: 24.5% on Zappos50K, 17.2% on CUB200, and 30.0% on MEDIC in classification accuracy over state-of-the-art methods. These performance gains underscore the ability of MTCL to generalize across diverse task domains, from fine-grained product recognition to medical condition classification. Furthermore, MTCL effectively supports a multi-objective training mechanism, including classification and regression, demonstrating superior precision and reduced model complexity on an in-house 3D multi-task dataset. These results validate MTCL's efficacy in addressing multi-task learning trade-offs and highlight its potential for practical applications such as healthcare, autonomous driving, and e-commerce, where classification and regression tasks often coexist. Future work will explore extending MTCL to larger-scale datasets and more complex multi-task scenarios, further refining its adaptability and scalability.

Acknowledgement

This material is based upon work supported by the U.S. Department of Energy under grant DE-SC0024428, and by the Defense Advanced Research Projects Agency (DARPA) under Cooperative agreement FA8750-23-2-0501. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of those supporting this work.

References

[1] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, Jan. 2018.

[2] N. Vaessen and D. A. v. Leeuwen, "Towards Multi-task Learning of Speech and Speaker Recognition," in *INTERSPEECH 2023*, 2023, pp. 4898–4902, iSSN: 2958-1796.

[3] S. Chen, Y. Zhang, and Q. Yang, "Multi-Task Learning in Natural Language Processing: An Overview," *ACM Comput. Surv.*, vol. 56, no. 12, pp. 295:1–295:32, Jul. 2024.

[4] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One Embedding Space To Bind Them All," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 15 180–15 190.

[5] F. Wang, Y. Zhou, S. WANG, V. Vardhanabhuti, and L. Yu, "Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 33 536–33 549.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021, _eprint: 2103.00020.

[7] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673.

[8] E. Mu, J. Guttag, and M. Makar, "Multitask Contrastive Learning," Oct. 2023.

[9] A. Veit, S. Belongie, and T. Karaletsos, "Conditional Similarity Networks," Apr. 2017, arXiv:1603.07810 [cs].

[10] R. Tan, M. I. Vasileva, K. Saenko, and B. A. Plummer, "Learning Similarity Conditions Without Explicit Supervision," Aug. 2019, arXiv:1908.08589 [cs].

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," 2015, pp. 815–823.

[12] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *ArXiv*, vol. abs/1807.03748, 2018.

[13] C. Jiang, W. Ye, H. Xu, S. Huang, F. Huang, and S. Zhang, "Vision Language Pre-training by Contrastive Learning with Cross-Modal Similarity Regulation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 660–14 679.

[14] L. Liu, X. Sun, T. Xiang, Z. Zhuang, L. Yin, and M. Tan, "Contrastive Vision-Language Alignment Makes Efficient Instruction Learner," *ArXiv*, vol. abs/2311.17945, 2023.

[15] P. Teterwak, X. Sun, B. A. Plummer, K. Saenko, and S.-N. Lim, "CLAMP: Contrastive LAnguage Model Prompt-tuning," Mar. 2024, arXiv:2312.01629 [cs].

[16] J. Zhang, H. Gao, P. Zhang, B. Feng, W. Deng, and Y. Hou, "LA-UCL: LLM-Augmented Unsupervised Contrastive Learning Framework for Few-Shot Text Classification," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 10 198–10 207.

[17] X. Yuan, Z. L. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal Contrastive Training for Visual Representation Learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6991–7000, 2021.

[18] R. Nakada, H. I. Gulluk, Z. Deng, W. Ji, J. Zou, and L. Zhang, "Understanding Multimodal Contrastive Learning and Incorporating Unpaired Data," Mar. 2023, arXiv:2302.06232 [cs].

[19] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 9564–9576.

[20] P. Hager, M. J. Menten, and D. Rueckert, "Best of Both Worlds: Multimodal Contrastive Learning With Tabular and Imaging Data," 2023, pp. 23 924–23 935.

[21] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," Jun. 2021, arXiv:2102.05918 [cs] version: 2.

[22] Y. Sun, Z. Zhang, and J. Ortiz, "A dual contrastive framework," 2024.

[23] K. E. Ak, J. Mohta, D. Dimitriadis, S. Manchanda, Y. Xu, and M. Shen, "Aligning Vision Language Models with Contrastive Learning," in *Computer Vision – ECCV 2024 Workshops*, A. Del Bue, C. Canton, J. Pont-Tuset, and T. Tommasi, Eds. Cham: Springer Nature Switzerland, 2025, pp. 32–45.

[24] S. Ma, L. Wang, S. Hou, and B. Yan, "Aligned with llm: a new multi-modal training paradigm for encoding fmri activity in visual cortex," 2024. [Online]. Available: https://arxiv.org/abs/2401.03851

[25] A. Yu and K. Grauman, "Fine-Grained Visual Comparisons with Local Learning," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 192–199, iSSN: 1063-6919.

[26] ——, "Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5571–5580, iSSN: 2380-7504.

[27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "CUB200," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[28] F. Alam, T. Alam, M. A. Hasan, A. Hasnat, M. Imran, and F. Ofli, "MEDIC: A Multi-Task Learning Dataset for Disaster Image Classification," *arXiv preprint arXiv:2108.12828*, 2021, _eprint: 2108.12828.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," 2017.