Neural Networks xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Episodic task agnostic contrastive training for multi-task learning☆

Fan Zhou [a,1], Yuyi Chen [a,1], Jun Wen [b], Qiuhao Zeng [c], Changjian Shui [d], Charles Ling [c], Shichun Yang [a,*], Boyu Wang [c,e,*]

[a] School of Transportation Science and Engineering, Beihang University, No. 37 Xueyuan Road, Beijing, 100083, China
[b] Department of Biomedical Informatics, Harvard Medical School, Boston, 02115, MA, USA
[c] Department of Computer Science, Western University, 1151 Richmond St, London, N6A 3K7, Ontario, Canada
[d] Department of Electrical and Computer Engineering, McGill University, Montreal, H3A 0G4, Quebec, Canada
[e] Vector Institute, 661 University Ave Suite 710, Toronto, M5G 1M1, Ontario, Canada

ARTICLE INFO

Article history:
Received 6 July 2022
Received in revised form 8 February 2023
Accepted 15 February 2023
Available online xxxx

Keywords:
Multi-task learning
Meta learning
Contrastive learning

ABSTRACT

Learning knowledge from different tasks to improve the general learning performance is crucial for designing an efficient algorithm. In this work, we tackle the Multi-task Learning (MTL) problem, where the learner extracts the knowledge from different tasks simultaneously with limited data. Previous works have been designing the MTL models by taking advantage of the transfer learning techniques, requiring the knowledge of the task index, which is not realistic in many practical scenarios. In contrast, we consider the scenario that the task index is not explicitly known, under which the features extracted by the neural networks are task agnostic. To learn the task agnostic invariant features, we implement model agnostic meta-learning by leveraging the episodic training scheme to capture the common features across tasks. Apart from the episodic training scheme, we further implemented a contrastive learning objective to improve the feature compactness for a better prediction boundary in the embedding space. We conduct extensive experiments on several benchmarks compared with several recent strong baselines to demonstrate the effectiveness of the proposed method. The results showed that our method provides a practical solution for real-world scenarios, where the task index is agnostic to the learner and can outperform several strong baselines, achieving state-of-the-art performances.

1. Introduction

General machine learning methodologies usually aim to solve individual problems, where the learning model is generally trained and tested on a single dataset. This paradigm assumes that the training and testing data are from the same data distribution. Although impressive progress in many applications has been achieved in recent years with the help of deep neural networks, a large amount of labelled data is still required to ensure successful model training. Obtaining the labelled data can be expensive in many practical scenarios. For example, collecting and annotating labels can be very prohibitive when designing an intelligent healthcare system, developing a perception model for intelligent vehicles or designing a prediction model of multiple objects. In many scenarios, we can only have several datasets of relatively small size or with limited labelled data from each dataset. We need to design a learning model that can leverage the knowledge from each of them.

To this end, Multi-task Learning (MTL) aims to simultaneously learn the shared knowledge among different tasks so that one can reduce the label annotations. MTL has been adopted in many research areas, including computer vision (e.g. Georgescu et al., 2021; Yu, Kumar et al., 2020), natural language processing (e.g. Chen, Zhang, & Yang, 2021), healthcare applications (e.g. Gupta et al., 2022; Li, Carlson et al., 2018; Moeskops et al., 2016; Nie et al., 2016) , and autonomous driving systems (e.g. Yu, Chen et al., 2020) *etc*. The goal of MTL considered in our work is to learn from limited data from several tasks so that the model can improve the overall learning performances on all the tasks (see Fig. 1).

Previous works (e.g. Shui, Abbasi, Robitaille, Wang, & Gagné, 2019; Zhou, Chaib-draa & Wang, 2021; Zhou, Shui et al., 2021) have investigated the MTL problems in the context of representation learning aspects on minimizing the generalization errors.

☆ This document is the results of the research project funded by the National key R&D Program of China (No. 2021YFB2501300, No. 2022YFB3206600), National Natural Science Foundation of China (No. U22A202101), Natural Sciences and Engineering Research Council of Canada (NSERC) and China Scholarship Council.

* Corresponding authors.
E-mail addresses: yangshichun@buaa.edu.cn (S. Yang), bwang@csd.uwo.ca (B. Wang).

[1] The first two authors (F. Zhou and Y. Chen) contribute equally to this work.

https://doi.org/10.1016/j.neunet.2023.02.023
0893-6080/

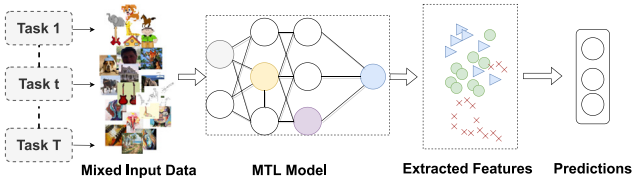


Fig. 1. Learning problems considered in our work. The model is given with the data from several tasks, aiming to learn from the mixed data (task-index-agnostic) and make the prediction for each task.

This kind of approach lies in the idea that the tasks share some commonalities, which can guide the learner to take advantage of the feature similarities across tasks to efficiently learn the cross-task information by either adversarial training (Mao, Liu, & Lin, 2020; Shui et al., 2019) or statistical distribution matching (Zhou, Chaib-draa & Wang, 2021).

Although the performance was improved, this stream of approaches still requires an explicit task index when training the model, *i.e.*, when training the model, they need to know from which task the data is coming. In practice, on the contrary, we cannot have distinguishable dataset collections, *i.e.,* we may have mixed data from different tasks (task agnostic data, as shown in Fig. 1). For example, one may have a large ensemble of data collected from different task scenarios and wants to train a model on this pile of data but test it on a specific task. Considering a mobile robot equipped with a machine learning-based vision system, it may need to operate under different working conditions *e.g.*, indoor environment with lighting illumination or outdoor environment in the midnight. Thus, the model should be trained beforehand with various data from different working conditions when designing the robot. When deploying the robot, it will be typically operated in certain task conditions (*e.g.*, running indoors for two hours). To this end, the model must learn data from different tasks while operating under certain tasks. If we apply some existing MTL algorithms (e.g. Shui et al., 2019; Zhou, Chaib-draa & Wang, 2021), which require the task index for training, the cost for pair-wised task training will be expensive since it will request pair-wised training. Furthermore, Zhou, Shui et al. (2021) attempted to relax the task index, but the performance dropped off. Therefore, we need to build a MTL model that can learn from task-agnostic data that does not require the task index and can perform well on certain tasks.

To this end, we implement the episodic training scheme originating from model agnostic meta-learning method (Finn, Abbeel, & Levine, 2017) to learn the mixed data, allowing the model to adapt quickly to new data distributions. The intrinsic idea of adopting this method lies in simulating the data distribution shifts across tasks through the episodic training scheme and leveraging the bi-level optimization process to improve the model. This approach has shown improved performances in many aspects of transfer learning, including domain adaptation (Volpi et al., 2018) and domain generalization (Li, Yang, Song & Hospedales, 2018).

However, leveraging the shared knowledge using the meta-learning approach remains problematic. For example, Dou, de Castro, Kamnitsas, and Glocker (2019) has demonstrated that this kind of approach may lead to some feature misalignment problems, which refers to the situation where the features are overlapped with each other in the extracted semantic feature space. Besides, Zhou, Jiang, Shui, Wang, and Chaib-draa (2021) also showed that only leveraging the invariant features across different distributions may lead the feature space to become indiscriminative. To alleviate this issue, the learner should also leverage the label similarities for better decision boundaries (Zhou, Jiang et al., 2021). One solution is to constrain the label similarity when extracting the common features across tasks and to simultaneously leverage the similarities to get the class-specific cohesion and separation feature space for all the tasks. Recent works (e.g. Dou et al., 2019; Zhou, Jiang et al., 2021) have proposed to improve learning performance by adopting metric learning objectives. However, the metric learning objectives usually require a large batch size to ensure the pair relations when training the model, leading to high computational costs.

To efficiently extract the feature similarities, we propose leveraging the data's label information by learning the task-agnostic features with a contrastive learning objective. The contrastive learning approaches have been an active research topic and have been widely studied in many different learning regimes. The core idea of contrastive learning is to leverage the pairs of feature augmentations of training examples to define a classification task for feature embeddings (Ho & Nvasconcelos, 2020). In the MTL problems, we can implement contrastive learning techniques to leverage the feature similarities for the input data from different tasks. The intuition is that if the instances belong to the same class, then the model can map them together in the feature space; conversely, if the instances are from different classes, we can push them apart. Therefore, we apply a contrastive learning objective for local feature relation persistence along with the episodic training scheme implemented for global feature extraction across tasks.

To summarize, the contributions of our work are trifold:

- We propose a new strategy to learn task-agnostic data by incorporating the episodic training scheme of meta-learning to allow the model to extract the cross-task knowledge.
- We design a mechanism to encourage the feature compactness for better prediction performance by introducing a contrastive learning objective that leverages the feature similarities for MTL.
- We then incorporate the cross-task knowledge transfer with local feature compactness and propose a novel MTL framework that globally learns cross-task data and locally exploits the similarities of classes that enhance the feature compactness.

We proposed the *Episodic Contrastive Multi-task Learning* algorithm. We conducted extensive experiments on several benchmarks, comparing against some recent strong baselines to evaluate the effectiveness of the proposed method to demonstrate the effectiveness of our algorithm.

We first evaluated our method with seven recent baselines, trained with task indexes, while ours does not require the task index. The empirical results showed that our method outperforms most baseline methods and achieves state-of-the-art performance on these benchmark datasets, especially when dealing with limited data. This demonstrates that *although our method did not take the task index information, it can still outperform the methods trained with task index*. Furthermore, in Section 5.5, we showed that when the task index of the input data is unavailable, recent strong baselines can perform worse. Thus, our method provides a more practical solution for real-world learning problems where the task index is not usually accessible to the learner. Besides, the visualized feature alignment performance further confirms the effectiveness of our method.

The rest of this article is organized as follows. Section 2 summarizes recent works most connected to our proposed work. Section 3 introduces the necessary background knowledge for this work. Section 4 presents the full methodology and algorithm of our work. Section 5 demonstrates the experimental results to show the effectiveness of the proposed algorithm.

2. Related works

Our work mainly relates to representation learning-based multitask learning, model-agnostic meta-learning, and contrastive learning.

2.1. Multi-task learning

Multi-task Learning (MTL) aims to learn several individual tasks simultaneously. Our work is mostly related to representation learning-based approaches. In this context, Maurer, Pontil, and Romera-Paredes (2016) firstly analysed the generalization risk of representation-based MTL approaches. Then, Murugesan, Liu, Carbonell, and Yang (2016), Pentina and Lampert (2017) tackled MTL with a weighted summation of loss functions. Chen, Badrinarayanan, Lee, and Rabinovich (2018) proposed a balanced joint training method over all the tasks with the same rate. Wang et al. (2023) studied the notion of performance gap, which theoretically provide new insights and motivates a novel principle for designing strategies for knowledge sharing and transfer. Shui et al. (2019), Zhou, Shui et al. (2021) investigated the generalization property by leveraging the task similarities under the adversarial training framework (Ganin et al., 2016). One drawback of Shui et al. (2019), Zhou, Shui et al. (2021) is that their theoretical results originated from the domain adaptation theory, requiring extra assumptions (e.g. Ben-David et al., 2010, the combined error across tasks is small), which may not hold in practice. Thus, Zhou, Chaib-draa and Wang (2021) leveraged the label and semantic information across tasks. However, in practical concerns, Zhou, Chaib-draa and Wang (2021) requires maintaining a matrix of the feature centroids. Besides, most of the previous work (e.g. Mao et al., 2020; Shui et al., 2019; Zhou, Chaib-draa & Wang, 2021; Zhou, Shui et al., 2021) require an explicit task index. In Section 5.5, we show that the baselines' performance became worse when the task indexes were not available. Thus, in this work, we provide an algorithm that does not have to take the task index into account, *i.e.*, task agnostic training. Furthermore, requiring task index is not practical in some real-world applications. For example, Wang et al. (2020) tackled a reinforcement learning-based navigation problem with MTL algorithms where the environment for the agent is agnostic. Gao et al. (2020) proposed a neural architecture search method for MTL problems that share similar insights of our work on handling the task agnostic data by conducting a regularization term on the model's architecture weights. At the same time, ours focused on the representation learning side of the MTL algorithm.

Furthermore, the conventional MTL problems aim to solve a fixed number of known tasks and are usually implemented as single-level optimization without the meta-learning objective. Previous work has leveraged the meta-learning methodologies for measuring prioritizes (Lin, Baweja, Kantor, & Held, 2019) of tasks and learning the task relations (Franceschi, Donini, Frasconi, & Pontil, 2017). In this work, we propose a novel episodic training scheme to learn from task-agnostic task data, allowing the model to adapt to unknown tasks without measuring task relations.

Besides, MTL showed improved performances in handling data from various tasks, provided a practical solution for real-world applications. For example, Gupta et al. (2022) tackled the multimental classification problems and presented a novel feature representation learning approach for the brain–computer interaction applications, which indicates a promising direction for applying MTL framework for intelligent health management systems. Song, Jeong, and Kim (2022) studied the obstacle detection problems for self-driving systems with a MTL framework, showing improved detection performances.

2.2. Learn transferable features using model agnostic meta learning

Meta-learning, *a.k.a.* learning to learn, aims to learn to improve the learning algorithm itself by leveraging the experience of several different learning episodes. A comprehensive survey on the recent progress of meta-learning with neural networks can be referred to Hospedales, Antoniou, Micaelli, and Storkey (2021). In this work, we consider meta-learning as a generic knowledge transfer method that can provide new perspectives for related research topics in transfer learning, *e.g.* Domain Adaptation (DA) (Volpi, Larlus, & Rogez, 2021) and Domain Generalization (DG) (Dou et al., 2019; Li, Yang et al., 2018). The core idea of learning transferable features using model agnostic meta-learning is to adopt an episodic training paradigm, *i.e.*, splitting the available data distributions into the general meta-train and meta-test subsets at each iteration so that the model can simulate the task data shift.

In the context of DG, Meta Agnostic Meta-Learning (MAML) (Finn et al., 2017) was adopted by Li, Yang et al. (2018) to back-propagate the gradient of the losses of the meta-test tasks (Dou et al., 2019). Du et al. (2020) proposed to model the shared classifier model parameters as a probabilistic meta-learning model. Gong et al. (2021) introduced a setting where the target domain is assumed as a compound of several unknown domains, which is treated as a sub-target domain. Then a meta-learning algorithm is implemented to fuse the sub-target domain together with the MAML algorithm for handling the generalization process.

In the context of DA, Volpi et al. (2021) adopted the meta-learning objective to generate some intermediate meta-domains with the randomized image manipulations to solve the DA problems. Yue et al. (2021) proposed a self-supervised learning framework for the few-shot DA problem, which not only aligns the cross-domain features but also captures the category-wise semantic structure of the source and target domain features through the self-supervised learning process.

In this work, we tackle the MTL problems and cast the problem of learning common features across tasks into the episodic training process by simulating the meta-train and meta-test process.

2.3. Contrastive learning

Contrastive learning has been an active research topic for different learning regimes, including unsupervised (Chen, Kornblith, Norouzi, & Hinton, 2020), weakly-supervised (Zheng et al., 2021), or self-supervised learning (Kim, Tack, & Hwang, 2020). The contrastive learning methods leverage the feature augmentation pairs of unlabelled training examples to define a classification task for feature embeddings. Its core idea lies in building an encoder to map the inputs generated by some data augmentations to similar features of some random inputs for the distinguishable features. Contrastive learning has recently been extended to the supervised setting (Khosla et al., 2020) to leverage the label similarities. This kind of approach has been widely studied in different aspects *e.g.*, video representation learning (Kuang et al., 2021), image captioning (Dai & Lin, 2017) and learning with noisy labels (Yi, Liu, She, McLeod, & Wang, 2022) *etc.*

In the context of transfer learning, Motiian, Piccirilli, Adjeroh, and Doretto (2017) adopts contrastive loss to encourage the data instances from the same category embedded close to each other in the feature space. Kang, Jiang, Yang, and Hauptmann (2019) studied the contrastive learning approach for the unsupervised DA problems, where the contrastive adaptation network was proposed together with a new metric to leverage the class similarities.

Since contrastive learning methodology has impressive performances in the unsupervised training regime that does not require

the labelled data, it inspires the source-free DA approaches. For example, Thota and Leontidis (2021) explored the contrastive learning methods for DA problems by training the unlabelled source and target domain data, where neither labelled data nor a pre-trained Imagenet model was required. Huang, Guan, Xiao, and Lu (2021) studied the unsupervised model adaptation, *i.e.*, the source-free UDA problems, by implementing a historical contrastive learning method to exploit the historical hypothesis trained on the sources that can learn both the instance and category level discriminative target representations for source free UDA. Wang et al. (2022) adopted a self-supervised contrastive training method to leverage the similarities of a certain category. Yang et al. (2022) studied the partial DA problems by incorporating a contrastive learning objective on top of the general adversarial training to find out the class-discriminative information across domains, which then leads to a contrastive learning-assisted alignment algorithm. Kim, Yoo, Park, Kim, and Lee (2021) proposed a self-supervised contrastive learning-based DG approach that takes advantage of the positive pairs in the contrastive objective as a regularization term to learn transferable domain features. Duboudin et al. (2021) explored the DG problems by designing a reverse contrastive loss for solving the correlated patterns across domains, which thus helps to encourage intra-class diversity to improve learning performances.

Our work lies in a similar insights of incorporating contrastive learning objectives in the aforementioned DA and DG approaches to leverage the power of contrastive pairs in addition to the knowledge-transferring techniques to enhance the feature compactness across different data distributions (Dou et al., 2019; Zhou, Jiang et al., 2021).

3. Background and preliminaries

We start by introducing some necessary notations and preliminaries, including background knowledge on meta-learning and contrastive learning. Then, we will introduce the methodology and the proposed algorithm.

3.1. Problem setup

In multi-task learning (MTL), assume we have a set of total M tasks $\{\hat{\mathcal{D}}_1, \dots \hat{\mathcal{D}}_i, \dots \hat{\mathcal{D}}_M\}_{i=1}^M$, each of which is generated by the underlying distribution $\mathcal{D}_i$ over $\mathcal{X}$ and by the underlying labelling functions $f_i : \mathcal{X} \to \mathcal{Y}$ for $\{(\mathcal{D}_i, f_i)\}_{i=1}^M$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output space, respectively. For a task i, let $\hat{\mathcal{D}}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{m_i}$ be a set of m_i training examples drawn independently from $\mathcal{D}_i$.

A multi-task learner aims to find M hypothesis: $h_1, \dots, h_M$ over the hypothesis class $\mathcal{H}$ to minimize the average expected risk of all the tasks:

$$\arg\min_{h \in \mathcal{H}} \frac{1}{M} \sum_{i=1}^{M} R_i(h_i) \tag{1}$$

where $R_i(h_i) \equiv R_i(h_i(x_i), f_i) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_i} \ell(h_i(\mathbf{x}), f_i(\mathbf{x}))$ is the expected risk of task i and ℓ is the loss of hypothesis h_i at $(\mathbf{x}, y)$. For each task i, assume that there are m_i examples, the empirical loss of h on $\hat{\mathcal{D}}_i$ is defined by $\hat{R}_i(h_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(h(\mathbf{x}_j), y_j)$.

We consider a classification model consisting of a feature extractor $F_{\boldsymbol{\psi}}$, parameterized by $\boldsymbol{\psi}$ and a task network $T_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$. The feature extractor $F_{\boldsymbol{\psi}} : \mathcal{X} \to \mathcal{Z}$ maps the input feature into a latent feature space $\mathcal{Z}$, which is lower dimensional than the input space $\mathcal{X}$. The task network $T_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathbb{R}^K$ predicts the label of the extracted features into $\mathbb{R}^K$, where K is the total number of classes in the output space $\mathcal{Y}$. The predication of the task network is estimated by the softmax operation $\mathbb{P}(\hat{y}|\mathbf{x}) = \text{softmax}(T_{\boldsymbol{\theta}}(F_{\boldsymbol{\psi}}(\mathbf{x})))$. The feature extractor and task network model parameters $(\boldsymbol{\psi}, \boldsymbol{\theta})$ can be optimized *w.r.t.* a specific task loss. In this work, we consider the classification model thus we adopt the general cross-entropy loss as task loss:

$$\ell_{task} = -\hat{y} \cdot \log(y). \tag{2}$$

3.2. Preliminaries on meta learning

Our work is built upon the model-agnostic meta learning (Finn et al., 2017) approach to extract the transferable features across tasks. Meta learning is usually cast into a bi-level optimization process. The data from different distributions $\mathcal{D}_{i=1}^M$ are usually split into *meta-train* $\mathcal{D}_{tr}^{(i)}$ and *meta-test* $\mathcal{D}_{te}^{(i)}$ set. It can be viewed as the following,

$$\begin{aligned} &\min_{i=1}^{M} \boldsymbol{\theta}^\star \circ \boldsymbol{\psi}^\star \mathcal{L}(\mathcal{D}_{te}^{(i)}) \\ &\text{s.t.} \quad \boldsymbol{\theta}^\star \circ \boldsymbol{\psi}^\star = \arg\min \mathcal{L}(\mathcal{D}_{tr}^{(i)}) \end{aligned} \tag{3}$$

The model is trained in two steps, first is to learn a base model that minimizes the risk on all the meta train sets $\mathcal{D}_{tr}$ and then to adapt to the meta test sets $\mathcal{D}_{te}$.

3.3. Preliminaries on contrastive learning

Contrastive learning has been widely studied from various aspects due to its flexibility in leveraging the similarities of input data no matter whether the data was labelled or not. The fundamental idea of contrastive learning is to learn the representations based on the data augmentations (*e.g.,* crop, resize or rotation etc.) (Yang et al., 2022). Then, with these augmentations, the contrastive learning objectives then guide the model to map an input instance (*a.k.a.* anchor) to be closer to its positive samples *i.e.*, its augmentations, and to be far away from its negative samples.

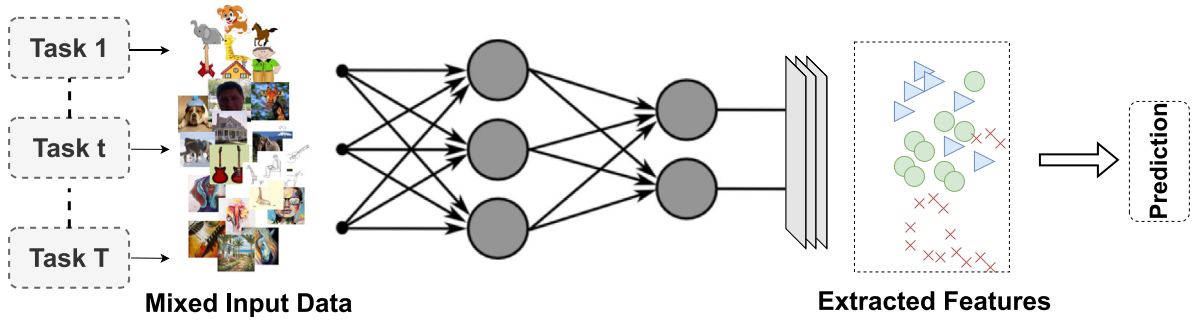
Denote by $\mathbf{x}_a$ as an input instance (anchor), $\mathbf{x}^+$, and $\mathbf{x}^-$ as the positive and negative instance, respectively. Then, for a score function, the relation of the positive and negative instances to the anchor can be summarized as

$$score\big[F_{\boldsymbol{\psi}}(\mathbf{x}_a; \boldsymbol{\psi}), F_{\boldsymbol{\psi}}(\mathbf{x}^+; \boldsymbol{\psi})\big] \gg score\big[F_{\boldsymbol{\psi}}(\mathbf{x}_a; \boldsymbol{\psi}), F_{\boldsymbol{\psi}}(\mathbf{x}^-; \boldsymbol{\psi})\big] \tag{4}$$

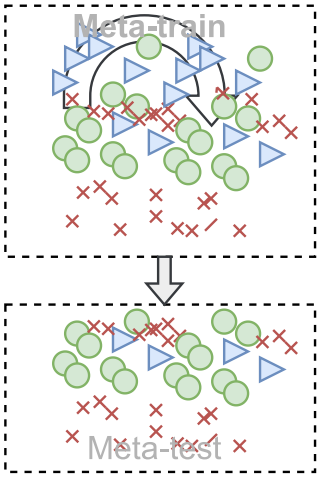
where score means a function that to measure the similarities between the two features $F_{\boldsymbol{\psi}}(\mathbf{x}; \boldsymbol{\psi})$ and $F_{\boldsymbol{\psi}}(\mathbf{x}'; \boldsymbol{\psi})$. This kind of framework then inspires a series of self-supervised contrastive learning methods. In case the labels are available in the MTL problems, we can leverage the label similarities and design the learning objective function to guide the model to bring the features from the same category close to each other while pulling the instances from different categories to far from each other, regardless of which task they come from, *i.e.*, to apply contrastive objective to the task-agnostic data to improve the feature compactness. We will elaborate on this aspect in Section 4.3.

4. Methodology

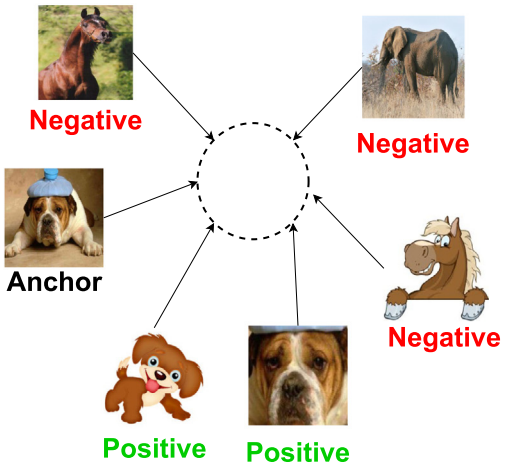
In this section, we introduce the main methodology of our work. Specifically, in Section 4.1, we illustrate an overview of our method. Then, in Section 4.2 and Section 4.3, we introduce the episodic training scheme with meta-learning and contrastive learning for similarity mining, respectively. Lastly, in Section 4.4, we present the proposed algorithm.



(a) Main of our method.



(b) Episodic Training for Feature Alignment



(c) Contrastive learning for feature similarity mining

Fig. 2. (a) The main workflow of our method. The learner takes the input data mixed from several tasks and learn to predict for each certain task. (b) The mixed data are learned through an episodic meta learning scheme which allows us to not rely on the task indexes. (c) On top of that, in order to further promote feature compactness, we also implement a contrastive learning objective which enables the similarities of instances.

4.1. Methodology overview

As aforementioned, our work tackles the multi-task learning problems to learn from the task-agnostic data, *i.e.*, the input data are mixed, and the learner has no prior knowledge about the task index of the mixed data. In order to learn the task-agnostic data, we adopt an episodic training scheme originated from model agnostic meta-learning that simulated distribution shift across tasks to learn the shareable features. In each training round, in addition to the episodic learning for the shareable features, we still need to promote the feature compactness at the category level by incorporating a contrastive learning objective. An overview of the model is illustrated in Fig. 2, and we will introduce the detailed methodology in the following sections.

4.2. Episodic training for learning multi-task features

In this work, we consider the scenario in which the data are given to the learner in a pool without knowing the task index. Regarding MTL, we assume there are M related task distributions with different statistics. As aforementioned, inputs to the model are mixed and are task agnostic. To learn from such task-agnostic data with different data distributions, we adopt the meta-learning notion to split the data into different subsets so that the learner can extract the invariant features from different data distributions. The intrinsic idea of using meta-learning to learn the task-agnostic features is to implement an episodic training scheme, which is rooted in the model-agnostic meta-learning method (Finn et al., 2017). In order to capture the data distribution shift, the model is trained with several episodes to simulate the shift across task data distributions (Li, Yang, Song, & Hospedales, 2017).

The model is trained on the limited task data and then tested with new task data. To achieve this, at each episodic iteration, input data from all the tasks are split into two subsets: *meta-train* set $\mathcal{T}$ and *meta-test* set $\hat{\mathcal{T}}$; thus we have $|\mathcal{T}|+|\hat{\mathcal{T}}|=M$. This kind of data splitting does not require indexes of the task distributions, and the input data are randomly split. Through this process, we can mimic the real cross-task data shift over different learning tasks to train a model to achieve good generalization performance on the final testing task data.

Meta-train. In line with our learning goal of MTL to learn the task-agnostic data, the model is trained on a sequence of simulated episodic subsets of data with the data shift across tasks. Specifically, for each episodic iteration, during the *meta-train* phase, the model is updated on all the $\mathcal{T}$ meta-train tasks. The loss for the general meta-train phase can be computed as

$$\mathcal{L}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\frac{1}{N_i}\sum_{j=1}^{N_i}\ell_{\boldsymbol{\theta},\boldsymbol{\psi}}(\hat{y}_j^{(i)}, y_j^{(i)}) \tag{5}$$

where $y_j^{(i)}$ refers to the label of the jth data instances in the ith task, $\hat{y}_j^{(i)}$ is the predicted label of the input instance $\mathbf{x}_j^{(i)}$, N_i is the number of total instances in task i, and ℓ is the task loss function aforementioned, *i.e.*, the cross-entropy loss in Eq. (2). Then, in the meta-train phase, the model can then be updated via the following,

$$(\boldsymbol{\psi}', \boldsymbol{\theta}') = (\boldsymbol{\psi}, \boldsymbol{\theta}) - \alpha\nabla_{\boldsymbol{\psi},\boldsymbol{\theta}}\mathcal{L}_{\mathcal{T}} \tag{6}$$

Eq. (6) leads to improved task accuracy on the meta-train sets of the input data of the prediction model $T_{\boldsymbol{\theta}} \circ F_{\boldsymbol{\psi}}$. Once the meta-train optimization is finished (*i.e.*, got the updated $\boldsymbol{\psi}', \boldsymbol{\theta}'$), we can further apply the meta-test step.

Meta-test. For each mini-batch, we also evaluate the model on *meta-test* set. The meta-test process simulates a real data shift of tasks with different statistics. The loss can be computed as,

$$\mathcal{L}_{\hat{\mathcal{T}}}(\cdot) = \frac{1}{|\hat{\mathcal{T}}|} \sum_{i=1}^{|\hat{\mathcal{T}}|} \frac{1}{N_i} \ell_{\boldsymbol{\theta}', \boldsymbol{\psi}'}(\hat{y}_j^{(i)}, y_j^{(i)}) \tag{7}$$

where $\hat{\mathcal{T}}$ is the meta-test set, $|\hat{\mathcal{T}}|$ refers to the number of *simulated* tasks in the meta-test set. Unlike Eq. (5) here $\hat{y}_j^{(i)}$ refers to the predicted label on the updated model $(\boldsymbol{\theta}', \boldsymbol{\psi}')$.

To summarize, the learning objective of the episodic training scheme, meta-train and meta-test phases at each episodic iteration

$$\min \mathcal{L}_{mt} = \min_{\boldsymbol{\theta}, \boldsymbol{\psi}} \mathcal{L}_{\mathcal{T}}(\boldsymbol{\theta}, \boldsymbol{\psi}) + \beta \mathcal{L}_{\hat{\mathcal{T}}}\big[(\boldsymbol{\theta}, \boldsymbol{\psi}) - \alpha \nabla \mathcal{L}_{\mathcal{T}}(\boldsymbol{\theta}, \boldsymbol{\psi})\big] \tag{8}$$

where α is the learning rate for the meta-train optimization and β is a coefficient to regularize the meta-test phase. This optimization can be efficiently optimized through the general gradient based optimization method (*e.g.,* SGD or Adam) for neural network.

For each training round, we also encourage feature compactness with a contrastive learning objective, which we will introduce in the next section.

4.3. Contrastive learning for feature compactness

As pointed out by previous works (Dou et al., 2019; Zhou, Jiang et al., 2021), only extracting the shareable features through the episodic training scheme is not sufficient for the multiple data distribution feature alignment and may lead to feature misalignment issues (Zhou, Chaib-draa & Wang, 2021). Thus, apart from the invariant feature learning with the episodic training scheme, we still have to constrain the learning process to promote the feature compactness (Kamnitsas et al., 2018).

To this end, in addition to performing the feature alignment across tasks, *i.e.*, global alignment across tasks with the episodic training scheme mentioned above, we further encourage the feature compactness via a contrastive learning objective as *local* sample clustering function to enhance the local feature alignment. The goal of the local contrastive learning objective is to take advantage of the power of the similarities of input instances from the mixed task data. Thus, we incorporate contrastive learning together with meta-learning. That is, *the data that come from the same category should stay close to each other in the feature space, while those from different categories should stay apart.* Therefore, after learning invariant features through the episodic training process, we further promote the feature compactness in the extracted feature space, regardless of which task the data comes from.

In line with our MTL problems, the labels of the training sample are available to the model. In this case, as pointed out by Khosla et al. (2020), the contrastive learning paradigm has the intrinsic ability to find out the hard positive or negative pairs, which can thus be incorporated with the knowledge transfer process with episodic training to improve the model's ability to distinguish the features.

The general self-supervised contrastive loss (Chen et al., 2020; Henaff, 2020) can be computed as,

$$\mathcal{L} = \sum_{i \in \mathcal{A}} \mathcal{L}_i = -\sum_{i \in \mathcal{A}} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in \mathcal{A}, a \neq i} \exp(z_i \cdot z_a/\tau)} \tag{9}$$

Algorithm 1 The proposed Episodic Contrastive Multi-task Learning algorithm

Require: Mixed, task-agnostic input data from M different tasks $\{\mathcal{D}_i\}_{i=1}^{M}$

Ensure: Neural network models and parameters: feature extractor $F_{\boldsymbol{\psi}}$ parameterized by $\boldsymbol{\psi}$ and task network $T_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$

1: **for** mini-batch of samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}$ from input data **do**
2: Compute the *Meta-train* objective via Eq. (3)
3: Update the model through Eq. (6)
4: Compute the *Meta-test* objective via Eq. (7)
5: Update the model by solving Eq. (8)
6: Extract intermidate features and compute the contrastive learning objective via Eq. (10)
7: Update $\boldsymbol{\psi}, \boldsymbol{\theta}$ by solving Eq. (12) with learning rate η:

$$\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} - \eta \frac{\partial(\mathcal{L}_{mt} + \gamma \mathcal{L}_{con})}{\partial \boldsymbol{\psi}},$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{\partial(\mathcal{L}_{mt} + \gamma \mathcal{L}_{con})}{\partial \boldsymbol{\theta}}$$

8: **end for**
9: **return** Optimal model parameters $\boldsymbol{\psi}^\star$ and $\boldsymbol{\theta}^\star$

where z is the extracted features, $\tau \in \mathbb{R}^+$ is a scalar temperature parameter. $\mathcal{A}$ is a set of anchors, which usually refers to the augmented samples in self-supervised learning. Index a is usually referred as an anchor and p is usually considered as a *positive* sample. In our MTL setting, the labels are available to the learner, and the model can learn the contrastive features under supervised training mode,

$$\mathcal{L}_{con} = \sum_{i \in \mathcal{A}} \mathcal{L}_i = \sum_{i \in \mathcal{A}} \frac{-1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in \mathcal{A}, a \neq i} \exp(z_i \cdot z_a/\tau)} \tag{10}$$

where $\mathcal{P} \equiv \{p \in \mathcal{A} : y_p = y_i\}$ is the set of the positive pairs, *i.e.*, the instances with the same labels of the anchor regardless which task they belong to. In terms of MTL problems where the labels are accessible to the learner, we can easily set the positive set as those instances that have the same labels as the anchor, regardless of which task they come from. During training, the contrastive learning objective is computed over the mixed data from all the tasks, *i.e.*, the task-agnostic data. Through employing the supervised contrastive learning process, the features from different tasks can be well aligned regardless of task indexes.

Then, at each training round, the gradients of the contrastive learning objective *w.r.t.* the extracted features can be computed by,

$$\begin{aligned} \frac{\partial \mathcal{L}_i}{\partial z_i} = & \frac{1}{\tau} \Big\{ \sum_{p \in \mathcal{P}} z_p \Big(\frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p/\tau)}{\sum_{a \in \mathcal{A}} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau)} - \frac{1}{|\mathcal{P}|} \Big) \\ & + \sum_{n \in \mathcal{N}} z_n \Big(\frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_n/\tau)}{\sum_{n \in \mathcal{N}} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_n/\tau)} \Big) \Big\} \end{aligned} \tag{11}$$

where $\mathcal{N} \equiv \{n \in \mathcal{A} : y_n \neq y_i\}$ is the set of negative samples. In our MTL setting, the negative samples refer to the instances that have different labels with the anchor. With the major components introduced above, we can summarize the full proposed method in the following section.

4.4. Full objective and proposed algorithm

The learning objective of the proposed method majorly has two components that can be summarized as,

$$\mathcal{L} = \mathcal{L}_{mt} + \gamma \mathcal{L}_{con} \tag{12}$$

Table 1
The empirical comparison results of the Multi-task Learning algorithms (in %) on the digits datasets (including MNIST, MNIST-M and SVHN) with LeNet-5 model as the feature extractor. The baseline methods are trained with task indexes while ours is task-index-agnostic.

	3K				5K				8K			
Approach	MNIST	MNIST-M	SVHN	Avg.	MNIST	MNIST-M	SVHN	Avg.	MNIST	MNIST-M	SVHN	Avg.
MTL-Uniform	93.9 ± 3.2	77.1 ± 2.6	57.3 ± 0.4	76.1	96.3 ± 1.2	79.1 ± 3.1	68.0 ± 2.9	81.1	97.7 ± 0.5	83.7 ± 2.2	71.4 ± 0.9	84.2
MTL-Weighed	89.3 ± 3.3	76.4 ± 3.1	70.2 ± 1.8	78.3	91.8 ± 2.7	74.2 ± 0.9	73.6 ± 3.1	79.8	92.3 ± 2.6	76.9 ± 3.1	74.1 ± 1.6	81.1
Adv.$\mathcal{H}$	90.1 ± 1.2	$\mathbf{81.2} \pm 1.3$	70.8 ± 0.5	80.7	91.9 ± 2.6	$\mathbf{83.7} \pm 1.4$	73.6 ± 1.6	82.9	94.9 ± 1.6	$\mathbf{85.2} \pm 0.3$	79.1 ± 0.3	86.4
Adv.W	96.8 ± 0.6	81.3 ± 0.7	69.5 ± 1.1	82.5	97.5 ± 0.2	83.4 ± 0.4	72.6 ± 1.2	84.5	98.1 ± 0.3	84.3 ± 0.4	75.4 ± 1.1	86.1
Multi-Obj.	$\mathbf{97.5} \pm 0.3$	76.9 ± 0.5	54.8 ± 0.3	76.4	$\mathbf{98.2} \pm 0.2$	80.2 ± 0.7	61.2 ± 0.8	79.9	$\mathbf{98.5} \pm 0.3$	82.8 ± 0.5	69.9 ± 0.9	83.7
AMTNN	96.9 ± 0.2	80.8 ± 1.5	77.1 ± 0.9	84.9	97.7 ± 0.1	83.6 ± 1.1	78.4 ± 0.8	86.6	98.1 ± 0.2	83.1 ± 2.1	80.2 ± 1.3	87.1
SMTL	95.4 ± 0.3	80.1 ± 0.5	$\mathbf{81.5} \pm 0.6$	85.7	95.8 ± 0.3	82.4 ± 0.4	$\mathbf{83.3} \pm 0.3$	87.2	96.0 ± 0.3	83.9 ± 0.4	$\mathbf{85.4} \pm 0.2$	88.4
Ours	$\mathbf{97.5} \pm 0.4$	80.7 ± 0.9	79.7 ± 1.1	**86.0**	97.6 ± 0.3	82.9 ± 0.5	81.9 ± 0.6	**87.5**	97.9 ± 0.2	84.4 ± 0.4	84.8 ± 0.3	**89.1**

where $\mathcal{L}_{mt}$ and $\mathcal{L}_{con}$ are the objective functions defined in Eq. (8) and Eq. (10), respectively; γ is a coefficient to regularize the contrastive learning objective.

We present the architecture of the model in Fig. 2. The model takes the input data mixed from several tasks and learns to predict each specific task. The mixed task data are learned through an episodic training scheme that allows us to not rely on the task indexes. Furthermore, in order to further promote feature compactness, we also implement a contrastive learning objective which enables the model to leverage the similarities of task instances. When deploying, the model is then tested on specific tasks. The proposed *Episodic Contrastive Multitask Learning* (EpiConMTL) method is presented in Algorithm 1. With the proposed method, we empirically demonstrate the effectiveness of our method in the next section.

5. Experiments results

In order to demonstrate the effectiveness of our proposed method, we evaluate the algorithms on several benchmarks. We first evaluate the MTL algorithms with the baselines, where we allow the baseline methods to have access to the task index information, while ours was tested under the task-index-agnostic scenario. Then, in Section 5.5, we compare the performances of the algorithms under the scenario where the task indexes are not available to the learner. Lastly, in Section 5.6, we conduct several experiments to further investigate the insights of the methodology.

5.1. Datasets

We compare our proposed algorithm against recent principled baselines on the following benchmarks:

- **Digits**: The digits benchmark considered in this work is a collection of several datasets, including *MNIST* (LeCun, Bottou, Bengio, Haffner, et al., 1998), *MNIST-M* (Ganin et al., 2016) and *SVHN* (Netzer et al., 2011). The model aims to learn these tasks simultaneously.
- **Office-31** (Saenko, Kulis, Fritz, & Darrell, 2010): It is a vision benchmark widely used in transfer learning related problems which consists of three different tasks: *Amazon*, *Dslr* and *Webcam*.
- **Office-Caltech** (Gong, Shi, Sha, & Grauman, 2012): This benchmark contains the shared classes between the Office-31 and Caltech256, including four different tasks: *Amazon* (A), *Dslr* (D), *Webcam* (W) and *Caltech* (C) for 10 different classes.
- **Office-Home** (Venkateswara, Eusebio, Chakraborty, & Panchanathan, 2017): This one is a more complex benchmark, containing four different tasks: *Art*, *Clipart*, *Product* and *Real World*, with 65 classes in each of the four tasks.

5.2. Baselines

We follow the previous work (Zhou, Chaib-draa & Wang, 2021; Zhou, Shui et al., 2021) to evaluate the performance of the algorithms when handling limited data. In this work, we consider several principled approaches:

- MTL-Uniform: The MTL model is composed of a feature extractor and a task-specific classifier. The model learns all the tasks simultaneously while optimizing the average summation loss: $\frac{1}{M}\sum_{i=1}^{M}\hat{R}_i(\boldsymbol{\psi}, \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ is a task-specific classifier for task i, *i.e.*, optimize the whole model with the loss uniformly computed from all the tasks.
- MTL-Weighted: Adapted from Murugesan et al. (2016), a MTL model is built to learn a weighted summation of losses over different tasks: $\frac{1}{T}\sum_{t=1}^{T}\hat{R}_{\boldsymbol{\alpha}_t}(\boldsymbol{\psi}, \boldsymbol{\theta}_t)$, where the weight coefficient $\boldsymbol{\alpha}_t$ for a certain task t is measured by a probabilistic interpretation.
- Adv.$\mathcal{H}$: An adversarial learning-based approach adapted from Liu, Qiu, and Huang (2017) by using the same loss functions while the adversarial objective is trained with the $\mathcal{H}$-divergence.
- Adv.W: Similar with Adv.$\mathcal{H}$ while replacing the adversarial training objective Adv.$\mathcal{H}$ with Wasserstein distance based adversarial training method as per (Shen, Qu, Zhang, & Yu, 2018).
- Multi-Obj. (Sener & Koltun, 2018): A methodology where the MTL problem was cast as solving a multi-objective problem.
- AMTNN (Zhou, Shui et al., 2021): A Multi-task Learning algorithm that takes advantage of task feature similarities via adversarial training using Wasserstein distance and can update the task relations automatically.
- SMTL: A methodology proposed in Zhou, Chaib-draa and Wang (2021), where the task distributions are matched via controlling the semantic conditional distance as well as the label distribution divergence.

Note that these baselines all require task indexes when training the model. In the following sections, we first evaluate the baseline methods under the scenario that the task index are available to the learner while our approach keeps task-index-agnostic. In this respect, we present in Table 1 to Table 4 the results of the baseline method with task index while ours is under the task-index-free setting. We demonstrate that although the baselines can have task index information while ours does not, our method can still outperform those baselines. Later in Section 5.5, we further evaluated the performance of all the baselines under the task-index-agnostic setting for a more fair comparison and to demonstrate the MTL algorithms' performances when task indexes are unavailable.

5.3. Data pre-processing and training details

In order to demonstrate the effectiveness of our method for handling limited data, we randomly selected parts of the dataset

Table 2
Empirical results comparison (accuracy in %) of the Multi-task learning algorithms on the Office-Caltech benchmark using AlexNet as the feature extractor. The baseline methods are trained with task indexes.

Method	Amazon	Caltech	Dslr	WebCam	Avg.
MTL-Uniform	84.2 ± 1.1	80.6 ± 0.8	90.8 ± 2.3	81.8 ± 0.9	84.3
MTL-Weighted	88.1 ± 0.2	81.5 ± 0.9	94.9 ± 0.2	94.2 ± 0.5	88.6
Adv.$\mathcal{H}$	81.5 ± 0.5	73.8 ± 1.8	91.4 ± 2.1	86.1 ± 1.4	83.3
Adv.W	84.9 ± 0.4	80.9 ± 0.9	94.5 ± 2.2	87.5 ± 1.5	86.9
Multi-Obj.	82.3 ± 0.7	76.7 ± 2.4	91.2 ± 1.7	86.8 ± 0.9	84.3
AMTNN	89.3 ± 0.9	84.3 ± 0.6	**98.4** ± 1.3	94.1 ± 0.7	91.7
SMTL	**90.9** ± 0.4	**85.3** ± 0.5	98.1 ± 0.8	94.2 ± 0.6	92.1
Ours	90.4 ± 0.9	84.1 ± 0.5	**98.4** ± 0.4	**95.8** ± 0.8	**92.2**

to train the model. For the Digits benchmark, we randomly select 3k, 5k, and 8k of data instances as per the evaluation protocol of Zhou, Chaib-draa and Wang (2021), Zhou, Shui et al. (2021) and select 1k instances as validation set while testing on the whole test set split of the original dataset. The image size of SVHN is 32×32 while the image size of MNIST and MNIST-M is 28×28; thus, we resize the images of these three datasets to 28×28 without any data augmentation.

For the Office-31 and Office-Home benchmark, we randomly take 5%, 10% and 20% of the total training data and test with the full test set. For Office-31 and Office-Home datasets, we follow the pre-processing protocol of Zhou, Chaib-draa and Wang (2021) to first resize the image to 256×256, then randomly resize cropping to 224×224, and finally apply the *RandomHorizontalFlip()* function of *PyTorch* for the training data.

5.4. Experiments details and test results

We first test our model on digits benchmark with LeNet-5 (LeCun et al., 1998) model as the feature extractor and a MLP as a classifier to make the prediction. We train the model with Adam (Kingma & Ba, 2014) optimizer with an initial learning rate 1×10^{-3}. The learning rate is decayed by 5% for every five epochs. Besides, for enforcing the regularization, we also enable the weight decay of the Adam optimizer with decay rate 10^{-5}. The model was trained for a total of 50 epochs with a mini-batch size of 64. The results comparing against baselines on the Digits datasets are reported in Table 1.

Then, we test our algorithm against the baselines on Office-Caltech, Office-31, and Office-Home datasets. For the empirical evaluations on the Office-Caltech dataset, we train the model with Adam optimizer with an initial learning rate 1×10^{-4}. Along with the training procedure, the learning rate is decayed 5% for every five epochs for a total of 120. The experimental results on the PACS dataset are reported in Table 2. It can be observed from the results that our algorithm can outperform the baselines. Furthermore, the improvement compared to the baseline AMTNN is not quite significant. It may be due to the small number of data in Office-Caltech; thus, the features have been well extracted, and the improvements are relatively small.

For the experiments on Office-31 and Office-Home datasets, we train the ResNet-18 model as a feature extractor whose output dimension is 512. Then, a three-layer MLP with bottleneck size 256 is implemented for the prediction. We set $\gamma = 0.1$ to regularize the contrastive learning objective. The model is trained with Adam optimizer with learning rate 2×10^{-4} as well as the learning rate decay 5% for every five epochs and with learning rate 2×10^{-4} as well as the learning rate decay 10% for every ten epochs with total 120 epochs for Office-31 and Office-Home dataset, respectively. To encourage the regularization, we set the *weight decay* of the Adam optimizer with 10^{-5}. The experimental results on the Office-31 and Office-Home datasets are illustrated in Table 3 and Table 4, respectively.

As we can observe from the table, our method can have improved results on these two benchmarks, achieving state-of-the-art performances. Besides, when the number of training samples is limited (*e.g.*, with 5% of total training data), our method can have promising improvement compared to the baseline methods.

Besides, as a remark, the proposed method did not require the task index for the model, while the baselines are all so needed. Considering that in real-world scenarios, usually, the data are task agnostic, our work provides a more realistic methodology to handle multi-task learning problems. Also, as we can observe from Table 1 to Table 4, our method can have better improvement when the number of training instances is limited (*e.g.*, 5% or 10%), this also confirms the effectiveness of our MTL algorithm when dealing the limited data.

5.5. Performance of MTL algorithms under task-index-agnostic setting

Note that the baseline results compared in Table 1 to Table 4 were obtained with task index to the learner, while ours was tested with no task index information. *Even though our method is task-index-agnostic, our method can still outperform the task-index-needed method.*

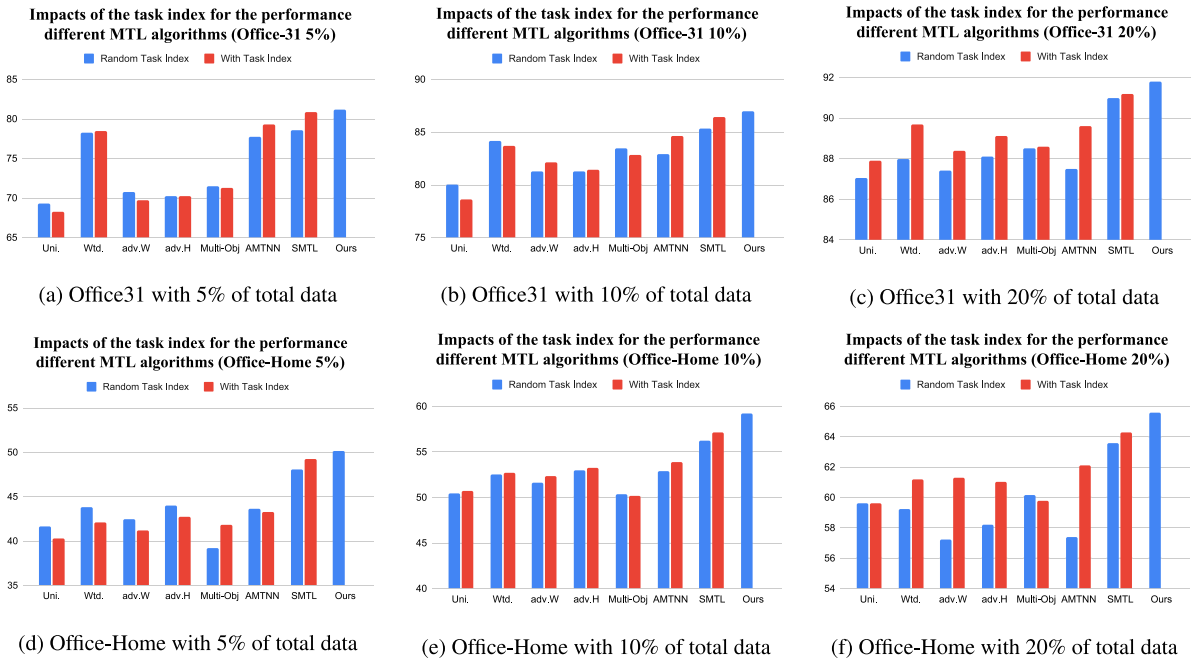
Considering that in real-world applications, the task index of the input data may not always be available. Thus, we further

Table 3
The empirical comparison results of the Multi-task Learning algorithms (accuracy in %) on the Office-31 dataset with the ResNet-18 model as the feature extractor. The baseline methods are trained with task indexes.

	5%				10%				20%			
Approach	Amazon	Dslr	Webcam	Avg.	Amazon	Dslr	Webcam	Avg.	Amazon	Dslr	Webcam	Avg.
MTL-Uniform	61.3 ± 1.3	71.8 ± 2.1	72.1 ± 1.1	68.3	73.2 ± 0.5	80.6 ± 1.4	82.1 ± 0.9	78.6	79.4 ± 0.8	91.2 ± 1.0	93.1 ± 0.8	87.9
MTL-Weighted	63.3 ± 0.2	87.4 ± 2.3	84.9 ± 0.6	78.5	70.6 ± 1.2	92.1 ± 0.9	88.4 ± 1.3	83.7	76.8 ± 0.9	96.6 ± 0.7	95.6 ± 0.5	89.7
Adv.W	66.5 ± 1.9	71.8 ± 1.1	69.9 ± 0.9	69.7	74.7 ± 1.1	85.9 ± 0.8	85.7 ± 0.8	82.1	79.3 ± 0.6	93.8 ± 0.4	92.2 ± 0.9	88.4
Adv.$\mathcal{H}$	65.8 ± 1.1	73.5 ± 0.8	71.4 ± 0.7	70.2	71.0 ± 0.9	84.1 ± 0.9	89.4 ± 0.1	81.4	79.7 ± 0.5	93.7 ± 0.7	93.7 ± 0.6	89.1
Multi-Obj.	68.9 ± 1.2	72.5 ± 1.4	72.3 ± 0.4	71.3	74.6 ± 0.9	86.8 ± 1.1	86.9 ± 0.8	82.8	79.2 ± 0.8	92.1 ± 0.6	94.7 ± 0.6	88.6
AMTNN	63.3 ± 0.6	80.1 ± 1.6	85.4 ± 0.3	79.3	71.3 ± 1.2	92.8 ± 0.9	89.6 ± 1.2	84.6	80.2 ± 0.9	94.2 ± 1.2	94.4 ± 0.9	89.6
SMTL	68.5 ± 0.6	**87.9** ± 0.8	86.5 ± 0.5	80.9	**75.7** ± 0.2	92.8 ± 0.2	90.8 ± 0.3	86.4	81.1 ± 0.2	96.5 ± 0.1	96.1 ± 0.2	91.2
Ours	**69.4** ± 0.8	87.5 ± 0.7	**86.6** ± 0.8	**81.2**	75.4 ± 0.8	**93.2** ± 0.8	**92.2** ± 0.9	**87.0**	**81.2** ± 0.5	**97.5** ± 0.6	**96.8** ± 0.5	**91.8**

Table 4
The empirical results of the Multi-task Learning algorithms (accuracy in %) on Office-home dataset with ResNet-18 as a feature extractor. The baseline methods are trained with task indexes.

	5%					10%					20%				
Approach	Art	Clipart	Product	Real-world	Avg.	Art	Clipart	Product	Real-world	Avg.	Art	Clipart	Product	Real-world	Avg.
MTL-Uniform	26.2 ± 0.3	30.1 ± 0.2	57.6 ± 0.1	47.4 ± 1.1	40.3	35.8 ± 0.7	43.3 ± 0.6	67.1 ± 0.4	56.8 ± 1.3	50.7	45.5 ± 0.8	56.1 ± 0.6	74.4 ± 0.7	62.6 ± 0.6	59.6
MTL-Weighted	26.8 ± 1.6	31.8 ± 1.8	59.2 ± 0.4	50.5 ± 1.2	42.1	38.2 ± 1.0	45.3 ± 1.6	69.1 ± 0.2	58.3 ± 0.8	52.7	47.9 ± 0.1	56.7 ± 0.9	75.6 ± 0.6	64.8 ± 0.9	61.2
Adv.W	26.8 ± 0.8	32.7 ± 0.5	58.3 ± 0.9	47.1 ± 0.4	41.2	38.5 ± 0.8	44.4 ± 0.7	67.6 ± 0.7	59.5 ± 0.9	52.3	47.9 ± 0.5	56.7 ± 0.6	75.4 ± 1.1	65.7 ± 0.8	61.3
Adv.$\mathcal{H}$	27.7 ± 1.4	32.1 ± 1.5	59.6 ± 0.7	51.1 ± 0.9	42.7	39.0 ± 0.9	45.8 ± 1.8	69.4 ± 0.4	58.8 ± 0.6	53.2	46.7 ± 0.5	56.5 ± 1.1	75.6 ± 0.4	65.1 ± 0.7	61.0
Multi-Obj.	25.6 ± 1.5	31.7 ± 1.7	58.7 ± 1.3	51.5 ± 0.9	41.8	34.6 ± 0.9	43.3 ± 1.4	66.1 ± 1.5	56.8 ± 0.7	50.2	46.2 ± 0.8	56.6 ± 0.5	74.3 ± 0.7	62.8 ± 0.6	59.8
AMTNN	32.5 ± 1.3	34.5 ± 0.9	56.3 ± 0.8	49.9 ± 1.8	43.3	41.1 ± 1.0	47.5 ± 0.8	68.4 ± 0.7	58.9 ± 0.9	53.9	48.9 ± 0.5	60.7 ± 0.4	75.4 ± 0.4	64.7 ± 0.4	62.1
SMTL	38.3 ± 0.9	40.9 ± 0.9	**62.3** ± 0.8	**55.5** ± 0.6	49.2	43.8 ± 0.6	50.4 ± 0.8	**71.3** ± 0.9	62.3 ± 0.6	57.1	51.2 ± 0.7	60.6 ± 0.8	77.9 ± 0.4	66.1 ± 0.6	64.3
Ours	**39.4** ± 0.3	**43.4** ± 0.8	62.1 ± 0.8	**55.5** ± 0.8	**50.1**	**47.6** ± 0.7	**54.5** ± 0.8	70.2 ± 0.3	**64.7** ± 0.8	**59.2**	**53.9** ± 0.4	**63.1** ± 0.6	**77.2** ± 0.5	**67.9** ± 0.6	**65.6**



(a) Office31 with 5% of total data (b) Office31 with 10% of total data (c) Office31 with 20% of total data

(d) Office-Home with 5% of total data (e) Office-Home with 10% of total data (f) Office-Home with 20% of total data

Fig. 3. Comparison of the Impacts of the task index for the performance of different MTL algorithms on Office-31 and Office-Home datasets with different data ratios.

investigate the performance for all the baselines aforementioned under the setting that the task indexes are not available. To check the performance of the MTL algorithms under the task-index-agnostic scenario and also for a more fair comparison of our method against the baselines, we further evaluate the baselines with the agnostic task index.

To this end, we assign a random task index to each input instance, and the baseline algorithms will use the randomly assigned task index to train the model. We report the test results in Fig. 3, where the testing results are compared under *random task index*, and *with task index* to denote the baselines were trained under the task-index-agnostic setting and task-index-accessible setting.

As we can observe from Fig. 3, when the task index was randomly assigned to the input instances, the baseline method *Uniform* and *Weighted* can have minor improvement with limited data (see Fig. 3(a) and Fig. 3(d)). This might be due to the randomly assigned task index can lead to mixed input data, which can improve the generalization property of these two algorithms when they only have very limited data. For example, for the baseline *Uniform*, when the input data are randomly mixed, the method will be reduced to the Empirical Risk Minimization (ERM) method with the mixed input. Recent work (Gulrajani & Lopez-Paz, 2021) has shown that ERM can have good performance when dealing with multi-source data. This coincides with our results. When we have more data (*e.g.* 10% or 20%), we can observe that the performance of the recent strong baselines *Adv.*$\mathcal{H}$, *Adv.W*, *AMTNN* and *SMTL* will decrease when the task index is not available (the random task index setting in Fig. 3). This confirms the effectiveness of our method when handling task-agnostic data. Thus, our method provides a practical solution for real-world scenarios.

5.6. Further analysis

Apart from the general experimental results on the benchmark datasets, we then further conduct several experiments to study the insights of our algorithm.

5.6.1. Impacts on different feature extractor backbones

The backbone for the feature extractor through Table 1 to Table 4 are selected by following some general selection by the transfer learning literature (e.g. Long, Cao, Wang, & Philip, 2017; Zhou, Chaib-draa & Wang, 2021; Zhou, Shui et al., 2021). In fact, different feature extractor backbones can influence the model's performance since different general backbones can influence the feature extraction. However, when conditioning on the same backbone (*e.g.*, ResNet-18 or AlexNet *etc.*) for both the baselines and proposed method, the performances of the algorithms are fairly comparable.

In this section, we further evaluate the impacts of the choice of different feature extractors. Apart from the ResNet-18 model-based evaluation, we conduct experiments on Office-31 and Office-Home with the AlexNet model to evaluate the impacts of feature extractor backbones *w.r.t.* the performance of all the algorithms. The results are presented in Table 5 and Table 6, respectively. We can observe that with very limited data (*e.g.*, 5% of training data), the adversarial training-based approaches suffered from worse performances. Compared with the baselines, our method can still outperform the baselines.

5.6.2. Feature visualization

We illustrate the t-SNE visualization of the proposed method on the Office-Caltech dataset with full data but trained with 20% of the total instances. The results are reported in Fig. 4. Compared to the pre-trained AlexNet model, our method can have a good feature alignment performance, regardless of tasks. Considering that the input data are task-agnostic, the well-aligned features showed that our method could have a good performance in extracting the task-agnostic features and have a cohesion boundary in the feature space.

5.6.3. Impacts of the temperature parameter

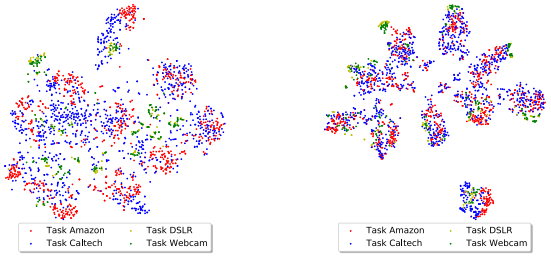
We then conduct the experiments to examine the impacts of the temperature parameter τ of the contrastive learning objective on the performance. We vary τ from 0.1 to 0.9 with 0.1 interval

Table 5
Empirical Evaluations on Office-31 dataset with AlexNet backbone.

	5%				10%				20%			
Approach	Amazon	Dslr	Webcam	Avg.	Amazon	Dslr	Webcam	Avg.	Amazon	Dslr	Webcam	Avg.
MTL-Uniform	51.2 ± 2.6	52.8 ± 2.1	60.6 ± 1.9	54.9	61.6 ± 2.1	66.5 ± 1.2	71.5 ± 1.3	66.5	72.2 ± 0.9	82.4 ± 1.2	84.8 ± 1.8	79.8
MTL-Weighted	52.2 ± 1.4	71.2 ± 3.8	73.3 ± 3.1	65.6	63.0 ± 2.5	82.5 ± 0.9	81.7 ± 2.2	75.7	72.1 ± 1.2	93.4 ± 1.3	92.6 ± 0.9	86.0
Adv.W	49.8 ± 2.1	53.7 ± 2.4	60.5 ± 1.5	54.6	61.6 ± 1.1	65.3 ± 2.0	73.7 ± 1.4	66.9	72.5 ± 0.6	82.2 ± 1.3	87.2 ± 0.5	80.6
Adv.\mathcal{H}	52.0 ± 1.3	68.7 ± 1.1	70.6 ± 1.4	63.8	62.2 ± 1.3	65.9 ± 1.1	72.9 ± 1.2	67.0	71.4 ± 0.9	83.7 ± 1.3	88.9 ± 0.9	81.3
Multi-Obj.	51.1 ± 0.2	51.3 ± 1.2	62.0 ± 1.3	54.8	61.0 ± 0.9	65.5 ± 1.2	72.2 ± 1.6	66.3	72.0 ± 0.5	81.6 ± 0.7	87.6 ± 1.7	80.4
AMTNN	27.9 ± 1.1	60.8 ± 2.1	69.5 ± 1.8	52.7	52.3 ± 1.2	73.4 ± 3.3	80.0 ± 1.8	68.6	70.6 ± 1.1	86.4 ± 0.9	86.1 ± 1.3	81.0
SMTL	**57.6** ± 2.3	76.2 ± 2.1	78.7 ± 2.1	70.8	66.8 ± 0.5	83.5 ± 0.2	87.6 ± 2.1	79.3	75.9 ± 1.1	92.1 ± 1.6	94.6 ± 0.4	87.6
Ours	57.2 ± 1.1	**76.6** ± 0.2	**79.1** ± 1.4	**70.9**	**67.5** ± 0.7	**89.4** ± 1.2	**87.9** ± 0.6	**81.6**	**76.5** ± 0.5	**95.4** ± 0.6	**93.4** ± 0.7	**88.4**

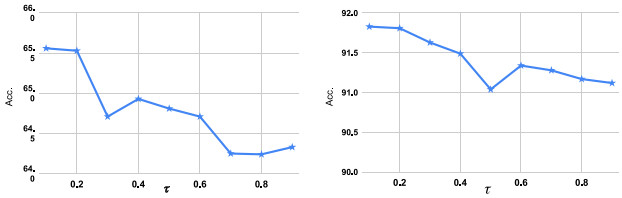
Table 6
Empirical Evaluations on Office-Home dataset with AlexNet backbone.

	5%					10%					20%				
Approach	Art	Clipart	Product	Real_World	Avg.	Art	Clipart	Product	Real_World	Avg.	Art	Clipart	Product	Real_World	Avg.
MTL-Uniform	24.2 ± 0.9	29.4 ± 0.8	47.7 ± 1.3	40.0 ± 0.5	35.3	**31.9** ± 1.2	39.6 ± 1.5	58.1 ± 1.0	49.0 ± 1.1	44.6	**37.4** ± 0.7	49.0 ± 0.7	65.3 ± 0.5	53.8 ± 0.9	51.4
MTL-Weighted	21.6 ± 2.1	27.5 ± 1.3	45.7 ± 1.2	36.2 ± 2.8	32.8	28.9 ± 2.8	37.3 ± 1.7	55.9 ± 1.4	45.5 ± 3.3	41.9	36.6 ± 1.2	46.3 ± 1.5	67.0 ± 0.4	53.1 ± 1.1	50.8
AdvW	20.1 ± 0.8	27.7 ± 0.7	46.3 ± 1.3	35.3 ± 1.3	32.4	25.6 ± 0.8	37.8 ± 0.8	57.4 ± 0.4	44.5 ± 1.9	41.3	30.3 ± 1.4	43.9 ± 0.9	63.6 ± 1.2	46.6 ± 1.4	46.1
AdvH	19.8 ± 1.8	27.4 ± 0.6	45.8 ± 1.3	35.1 ± 0.8	32.3	27.8 ± 1.9	37.2 ± 0.6	56.6 ± 1.3	43.9 ± 0.8	41.4	30.5 ± 1.6	43.3 ± 1.1	63.2 ± 1.0	46.7 ± 1.3	45.9
Multi-Obj	17.6 ± 1.2	24.1 ± 0.8	41.8 ± 1.1	31.0 ± 0.9	28.6	27.9 ± 1.9	34.5 ± 0.8	51.9 ± 1.5	45.8 ± 0.9	40.1	29.2 ± 0.7	43.8 ± 0.7	62.0 ± 1.1	45.5 ± 0.9	45.1
AMTNN	17.6 ± 0.9	25.3 ± 1.0	46.6 ± 0.7	34.1 ± 1.6	30.9	24.4 ± 2.1	29.7 ± 0.8	55.5 ± 0.9	41.2 ± 1.8	37.7	31.3 ± 2.2	30.0 ± 3.4	63.5 ± 1.6	50.3 ± 1.8	43.8
SMTL	18.3 ± 0.9	26.9 ± 1.0	45.7 ± 0.8	38.3 ± 1.2	32.3	26.2 ± 1.5	38.6 ± 0.8	54.6 ± 0.8	48.9 ± 1.1	42.1	33.8 ± 1.1	48.2 ± 0.2	65.9 ± 0.3	53.2 ± 0.2	50.3
Ours	**24.2** ± 0.6	**30.1** ± 0.2	**49.8** ± 0.4	**40.1** ± 0.6	**36.1**	30.3 ± 0.4	**40.5** ± 0.7	**59.8** ± 0.9	**49.4** ± 1.0	**45.0**	36.8 ± 0.5	**50.5** ± 0.3	**69.1** ± 0.6	**54.1** ± 0.3	**52.6**



(a) Pre-trained AlexNet model (b) Full Method

Fig. 4. t-SNE visualization of the alignment performance.



(a) Office-31 Dataset (b) Office-Home Dataset

Fig. 5. Impacts on temperature coefficient τ to the performance.

with 20% total data of Office-31 and Office-Home datasets. The results are reported in Fig. 5(a) and Fig. 5(b), respectively. As we can observe from the evaluation results, the performance on these two benchmarks slightly decreased when the value of τ increased.

5.7. Discussion of test results

In the aforementioned empirical evaluations, we first compared with several common state-of-the-art baseline MTL algorithms, which requested the task index for training. Then, we also compared our results with the random-task-index setting and further analysis. The empirical results on four benchmarks showed the proposed algorithm outperforms some strong baselines and achieves state-of-the-art performances. Besides, the empirical results showed that although our method did not take the task index information, it could still outperform the methods trained with task index, which is more suitable to a practical scenario where no task indexes are available.

6. Conclusion

Learning transferable knowledge from different task distributions is crucial for machine learning algorithms. In this work, we tackle multi-task learning problems with limited data. Specifically, we consider the learning scenario where the data from different tasks are mixed, and the task indexes are agnostic to the learner, which is a neglected issue in recent works. The *Episodic Contrastive Multitask Learning* was designed to leverage the shared knowledge from different but agnostic tasks and also enhance the feature compactness for prediction. In order to learn the task-agnostic data, we compromise the episodic training scheme of model-agnostic meta-learning to extract the shareable features across tasks. To overcome the indiscriminative features learned by the episodic training scheme, in addition to the episodic training process, we further incorporate a supervised contrastive learning objective to improve the feature compactness for better class-specific cohesion and separation of features across different tasks. The empirical results demonstrate the effectiveness of the proposed method for a more practical scenario where the input data are task-agnostic. Besides, the feature alignment evaluations also confirmed the effectiveness of this method in aligning features across tasks.

CRediT authorship contribution statement

Fan Zhou: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Yuyi Chen:** Methodology, Data curation, Writing – original draft. **Jun Wen:** Data curation, Methodology, Writing – original draft. **Qiuhao Zeng:** Validation, Writing – review & editing. **Changjian Shui:** Validation, Writing – review & editing. **Charles Ling:** Supervision, Writing – review & editing. **Shichun Yang:** Supervision, Writing – review & editing. **Boyu Wang:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work has been supported in part by the National key R&D Program of China (No. 2021YFB2501300, No. 2022YFB3206600), National Natural Science Foundation of China (No. U22A202101), the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, and the China Scholarship Council.

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1–2), 151–175.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., & Rabinovich, A. (2018). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning* (pp. 794–803). PMLR.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.

Chen, S., Zhang, Y., & Yang, Q. (2021). Multi-task learning in natural language processing: An overview. CoRR abs/2109.09138 arXiv:2109.09138.

Dai, B., & Lin, D. (2017). Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*, *30*.

Dou, Q., de Castro, D. C., Kamnitsas, K., & Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. In *Advances in neural information processing systems* (pp. 6447–6458).

Du, Y., Xu, J., Xiong, H., Qiu, Q., Zhen, X., Snoek, C. G., et al. (2020). Learning to learn with variational information bottleneck for domain generalization. In *European conference on computer vision* (pp. 200–216). Springer.

Duboudin, T., Dellandréa, E., Abgrall, C., Hénaff, G., & Chen, L. (2021). Encouraging intra-class diversity through a reverse contrastive loss for single-source domain generalization. In *Proceedings of the IEEE/CVF International conference on computer vision* (pp. 51–60).

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International conference on machine learning-volume 70* (pp. 1126–1135). JMLR. org.

Franceschi, L., Donini, M., Frasconi, P., & Pontil, M. (2017). Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning* (pp. 1165–1173). PMLR.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(1), 2030–2096.

Gao, Y., Bai, H., Jie, Z., Ma, J., Jia, K., & Liu, W. (2020). Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 11543–11552).

Georgescu, M.-I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M. (2021). Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 12742–12752).

Gong, R., Chen, Y., Paudel, D. P., Li, Y., Chhatkuli, A., Li, W., et al. (2021). Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 8344–8354).

Gong, B., Shi, Y., Sha, F., & Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on computer vision and pattern recognition* (pp. 2066–2073). IEEE.

Gulrajani, I., & Lopez-Paz, D. (2021). In search of lost domain generalization. In *International conference on learning representations*. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Gupta, A., Kumar, D., Verma, H., Tanveer, M., Javier, A. P., Lin, C.-T., et al. (2022). Recognition of multi-cognitive tasks from EEG signals using EMD methods. *Neural Computing and Applications*, 1–18.

Henaff, O. (2020). Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning* (pp. 4182–4192). PMLR.

Ho, C.-H., & Nvasconcelos, N. (2020). Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, *33*, 17081–17093.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5149–5169.

Huang, J., Guan, D., Xiao, A., & Lu, S. (2021). Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, *34*.

Kamnitsas, K., Castro, D. C., Folgoc, L. L., Walker, I., Tanno, R., Rueckert, D., et al. (2018). Semi-supervised learning via compact latent space clustering. arXiv preprint arXiv:1806.02679.

Kang, G., Jiang, L., Yang, Y., & Hauptmann, A. G. (2019). Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 4893–4902).

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., et al. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, *33*, 18661–18673.

Kim, M., Tack, J., & Hwang, S. J. (2020). Adversarial self-supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems, vol. 33* (pp. 2983–2994). Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2020/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf.

Kim, D., Yoo, Y., Park, S., Kim, J., & Lee, J. (2021). Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9619–9628).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kuang, H., Zhu, Y., Zhang, Z., Li, X., Tighe, J., Schwertfeger, S., et al. (2021). Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International conference on computer vision* (pp. 3195–3204).

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, Y., Carlson, D. E., et al. (2018). Extracting relationships by multi-domain matching. In *Advances in neural information processing systems* (pp. 6798–6809).

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. (2017). Deeper, broader and artier domain generalization. In *International conference on computer vision*.

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. M. (2018). Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on artificial intelligence*.

Lin, X., Baweja, H., Kantor, G., & Held, D. (2019). Adaptive auxiliary task weighting for reinforcement learning. *Advances in Neural Information Processing Systems*, *32*.

Liu, P., Qiu, X., & Huang, X. (2017). Adversarial multi-task learning for text classification. arXiv preprint arXiv:1704.05742.

Long, M., Cao, Z., Wang, J., & Philip, S. Y. (2017). Learning multiple tasks with multilinear relationship networks. In *Advances in neural information processing systems* (pp. 1594–1603).

Mao, Y., Liu, W., & Lin, X. (2020). Adaptive adversarial multi-task representation learning. In *International conference on machine learning*.

Maurer, A., Pontil, M., & Romera-Paredes, B. (2016). The benefit of multi-task representation learning. *Journal of Machine Learning Research*, *17*(1), 2853–2884.

Moeskops, P., Wolterink, J. M., van der Velden, B. H., Gilhuijs, K. G., Leiner, T., Viergever, M. A., et al. (2016). Deep learning for multi-task medical image segmentation in multiple modalities. In *International conference on medical image computing and computer-assisted intervention* (pp. 478–486). Springer.

Motiian, S., Piccirilli, M., Adjeroh, D. A., & Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *The IEEE international conference on computer vision*.

Murugesan, K., Liu, H., Carbonell, J., & Yang, Y. (2016). Adaptive smoothed online multi-task learning. In *Advances in neural information processing systems* (pp. 4296–4304).

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning.

Nie, L., Zhang, L., Meng, L., Song, X., Chang, X., & Li, X. (2016). Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(7), 1508–1519.

Pentina, A., & Lampert, C. H. (2017). Multi-task learning with labeled and unlabeled tasks. In *International conference on machine learning* (pp. 2807–2816).

Saenko, K., Kulis, B., Fritz, M., & Darrell, T. (2010). Adapting visual category models to new domains. In *European conference on computer vision* (pp. 213–226). Springer.

Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. In *Advances in neural information processing systems* (pp. 527–538).

Shen, J., Qu, Y., Zhang, W., & Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on artificial intelligence*.

Shui, C., Abbasi, M., Robitaille, L.-É., Wang, B., & Gagné, C. (2019). A principled approach for learning task similarity in multitask learning. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19*.

Song, T.-J., Jeong, J., & Kim, J.-H. (2022). End-to-end real-time obstacle detection network for safe self-driving via multi-task learning. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 16318–16329. http://dx.doi.org/10.1109/TITS.2022.3149789.

Thota, M., & Leontidis, G. (2021). Contrastive domain adaptation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 2209–2218).

Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on computer vision and pattern recognition (CVPR)*.

Volpi, R., Larlus, D., & Rogez, G. (2021). Continual adaptation of visual representations via domain randomization and meta-learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021* (pp. 4443–4453). Computer Vision Foundation / IEEE, URL https://openaccess.thecvf.com/content/CVPR2021/html/Volpi_Continual_Adaptation_of_Visual_Representations_via_Domain_Randomization_and_Meta-Learning_CVPR_2021_paper.html.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., & Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems* (pp. 5334–5344).

Wang, X. E., Jain, V., Ie, E., Wang, W. Y., Kozareva, Z., & Ravi, S. (2020). Environment-agnostic multitask learning for natural language grounded navigation. In *European conference on computer vision* (pp. 413–430). Springer.

Wang, B., Mendez, J. A., Shui, C., Zhou, F., Wu, D., Xu, G., et al. (2023). Gap minimization for knowledge sharing and transfer. *Journal of Machine Learning Research*, *24*(33), 1–57, URL http://jmlr.org/papers/v24/22-0099.html.

Wang, R., Wu, Z., Weng, Z., Chen, J., Qi, G.-J., & Jiang, Y.-G. (2022). Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*.

Yang, C., Cheung, Y.-M., Ding, J., Tan, K. C., Xue, B., & Zhang, M. (2022). Contrastive learning assisted-alignment for partial domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*.

Yi, L., Liu, S., She, Q., McLeod, A. I., & Wang, B. (2022). On learning contrastive representations for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16682–16691).

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., et al. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2636–2645).

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

Yue, X., Zheng, Z., Zhang, S., Gao, Y., Darrell, T., Keutzer, K., et al. (2021). Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13834–13844).

Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., et al. (2021). Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10042–10051).

Zhou, F., Chaib-draa, B., & Wang, B. (2021). Multi-task learning by leveraging the semantic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 11088–11096, URL https://ojs.aaai.org/index.php/AAAI/article/view/17323.

Zhou, F., Jiang, Z., Shui, C., Wang, B., & Chaib-draa, B. (2021). Domain generalization via optimal transport with metric similarity learning. *Neurocomputing*, [ISSN: 0925-2312] http://dx.doi.org/10.1016/j.neucom.2020.09.091, URL https://www.sciencedirect.com/science/article/pii/S0925231221002009.

Zhou, F., Shui, C., Abbasi, M., Robitaille, L.-É., Wang, B., & Gagné, C. (2021). Task similarity estimation through adversarial multitask neural network. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 466–480. http://dx.doi.org/10.1109/TNNLS.2020.3028022.