

Improving Multi-task Learning via Seeking Task-based Flat Regions

Hoang Phan^{*1} Lam Tran^{*2} Quyen Tran² Ngoc N. Tran³
Tuan Truong² Nhat Ho⁴ Dinh Phung⁵ Trung Le⁵

New York University¹, VinAI Research², Vanderbilt University³,
University of Texas, Austin⁴, Monash University, Australia⁵

November 20, 2024

Abstract

Multi-Task Learning (MTL) is a widely used and powerful learning paradigm for training deep neural networks that allows learning more than one objective by a single backbone. Compared to training tasks separately, MTL significantly reduces computational costs, improves data efficiency, and potentially enhances model performance by leveraging knowledge across tasks. Hence, it has been adopted in a variety of applications, ranging from computer vision to natural language processing and speech recognition. Among them, there is an emerging line of work in MTL that focuses on manipulating the task gradient to derive an ultimate gradient descent direction to benefit all tasks. Despite achieving impressive results on many benchmarks, directly applying these approaches without using appropriate regularization techniques might lead to suboptimal solutions to real-world problems. In particular, standard training that minimizes the empirical loss on the training data can easily suffer from overfitting to low-resource tasks or be spoiled by noisy-labeled ones, which can cause negative transfer between tasks and overall performance drop. To alleviate such problems, we propose to leverage a recently introduced training method, named Sharpness-aware Minimization, which can enhance model generalization ability on single-task learning. Accordingly, we present a novel MTL training methodology, encouraging the model to find task-based flat minima for coherently improving its generalization capability on all tasks. Finally, we conduct comprehensive experiments on a variety of applications to demonstrate the merit of our proposed approach to existing gradient-based MTL methods, as suggested by our developed theory.

1 Introduction

Over the last few years, deep learning has emerged as a powerful tool for functional approximation by exhibiting superior performance and even exceeding human ability on a wide range of applications. In spite of the appealing performance, training massive independent neural networks to handle individual tasks requires not only expensive computational and storage resources but also long runtime. Therefore, multi-task learning is a more preferable approach in many situations [1, 2, 3] as they can: (i) avoid redundant features calculation for each task through their inherently shared architecture; and (ii) reduce the number of total trainable parameters by hard parameter sharing [4, 5] or soft parameter sharing [6, 7]. However, existing state-of-the-art methods following the veins of gradient-based multi-task learning [8, 9, 10, 11, 12, 13] tend to neglect geometrical properties of the loss landscape yet solely focus on minimizing the empirical error in the optimization process, which can be easily prone to the overfitting problem [14, 15].

^{*}Equal contributions.

Meanwhile, the overfitting problem of modern neural networks is often attributed to high-dimensional and non-convex loss functions, which result in complex loss landscapes containing multiple local optima. Hence, understanding the loss surface plays a crucial role in training robust models, and developing flat minimizers remains one of the most effective approaches [16, 14, 17, 18]. To be more specific, recent studies [19, 20] show that the obtained loss landscape from directly minimizing the empirical risk can consist of many sharp minimums, thus yielding poor generalization capacity when being exposed to unseen data. Moreover, this issue is apparently exacerbated in optimizing multiple objectives simultaneously, as in the context of multi-task learning. Certainly, sharp minima of each constituent objective might appear at different locations, which potentially results in large generalization errors on the associated task. To this end, finding a common flat and low-loss valued region for all tasks is desirable for improving the current methods of multi-task learning.

Contribution. To further address the above desideratum, we propose a novel MTL training method, incorporating the recently introduced optimization sharpness-aware minimization (SAM) [21] into existing gradient manipulation strategies in multi-task learning to further boost their performance. Guiding by the generalization error in Theorem 1, the proposed approach not only orients the model to the joint low empirical loss value across tasks but also encourages the model to reach the task-based flat regions. Importantly, our approach is model-agnostic and compatible with current gradient-based MTL approaches (see Figure 1 for the overview of our approach). By using our proposed framework, the gradient conflict across tasks is mitigated significantly, which is the goal of recent gradient-based MTL studies in alleviating negative transfer between tasks. Finally, we conduct comprehensive experiments on a variety of applications to demonstrate the merit of our approach for improving not only task performance but also model robustness and calibration. Last but not least, to the best of our knowledge, ours is the first work to improve multi-task learning by investigating the geometrical properties of the model loss landscape.

2 Related work

2.1 Multi-task learning

In multi-task learning (MTL), we often aim to jointly train one model to tackle multiple different but correlated tasks. It has been proven in prior work [22, 2, 23, 24] that it is not only able to enhance the overall performance but also reduce the memory footprint and fasten the inference process. Previous studies on MTL often employ a hard parameter-sharing mechanism along with light-weight task-specific modules to handle multiple tasks.

Pareto multi-task learning. Originated from Multiple-gradient descent algorithm (MGDA), a popular line of gradient-based MTL methods aim to find Pareto stationary solutions, from which we can not further improve model performance on any particular task without diminishing another [8]. Moreover, recent studies suggest exploring the whole Pareto front by learning diverse solutions [25, 26, 27, 28], or profiling the entire Pareto front with hyper-network [29, 30]. While these methods are theoretically grounded and guaranteed to converge to Pareto-stationary points, the experimental results are often limited and lack comparisons under practical settings.

Loss and gradient balancing. Another branch of preliminary work in MTL capitalizes on the idea of dynamically reweighting loss functions based on gradient magnitudes [31], task homoscedastic uncertainty [32], or difficulty prioritization [33] to balance the gradients across tasks. More recently, PCGrad [9] developed a gradient manipulation procedure to avoid conflicts among

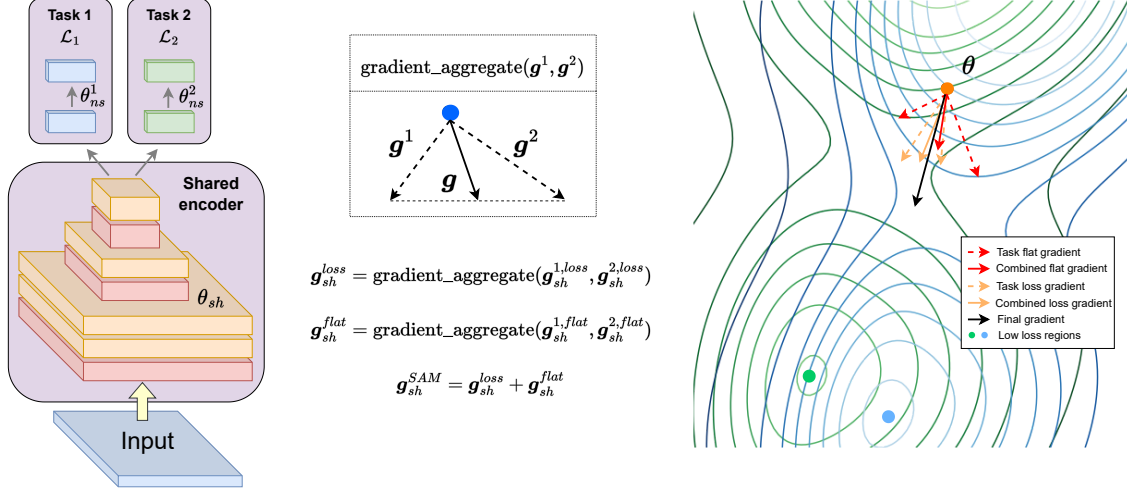


Figure 1: We demonstrate our framework in a 2-task problem. For the shared part, *task-based flat gradients* (red dashed arrows) steer the model to escape sharp areas, while *task-based loss gradients* (orange dashed arrows) lead the model into their corresponding low-loss regions. In our method, we aggregate them to find the combined flat gradient g_{sh}^{flat} and combined loss gradient g_{sh}^{loss} , respectively. Finally, we add those two output gradients to target the joint low-loss and flat regions across the tasks. Conversely, updating task-specific non-shared parts is straightforward and much easier since there is the involvement of one objective only.

tasks by projecting random task gradients on the normal plane of the other. Similarly, [10] proposes a provably convergent method to minimize the average loss, and [11] calculates loss scaling coefficients such that the combined gradient has equal-length projections onto individual task gradients.

2.2 Flat minima

Flat minimizer has been found to improve generalization ability of neural networks because it enables models to find wider local minima, by which they will be more robust against shifts between train and test losses [34, 35, 36]. This relationship between generalization ability and the width of minima is theoretically and empirically studied in many studies [37, 38, 39, 40], and subsequently, a variety of methods seeking flat minima have been proposed [41, 42, 43, 44]. For example, [43, 45, 46] analyze the impacts of different training factors, such as batch-size, learning rate, covariance of gradient, dropout, on the flatness of found minima. Additionally, several schemes pursue wide local minima by adding regularization terms to the loss function [41, 47, 48, 42], e.g., softmax output’s low entropy penalty, [41], distillation losses [47, 48].

Recently, SAM [21], which seeks flat regions by explicitly minimizing the worst-case loss around the current model, has received significant attention due to its effectiveness and scalability compared to previous methods. Particularly, it has been exploited in a variety of tasks and domains [49, 50, 51, 52, 53, 54, 55]. A notable example is the improvement that SAM brings to meta-learning bi-level optimization in [50]. Another application of SAM is in federated learning (FL) [51] in which the authors achieved tighter convergence rates than existing FL works, and proposed a generalization bound for the global model. In addition, SAM shows its generalization ability in vision models [54], language models [53] and domain generalization [49]. However, existing studies have only focused

on single task problems. In this work, we leverage SAM’s principle to develop theory and devise practical methods, allowing seeking flat minima in gradient-based multi-task learning models.

3 Preliminaries

Conventional training methods that focus on minimizing the empirical loss can be easily prone to overfitting problems (i.e., the validation error no longer decreases, but the training loss keeps declining), thus, restricting model generalization performance. In an attempt to alleviate such phenomena, [21] proposed to minimize the worst-case loss in a neighborhood of the current model parameter given by:

$$\min_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}), \quad (1)$$

where $\|\cdot\|_2$ denotes the l_2 norm and ρ represents the radius of the neighborhood. We assume \mathcal{L} is differentiable up to the first order with respect to $\boldsymbol{\theta}$. The optimization problem (1) is referred to as *sharpness aware minimization* (SAM).

To solve problem (1), [21] proposed to first find the solution for the inner maximization by approximating $\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon})$ via a first-order Taylor expansion w.r.t $\boldsymbol{\epsilon}$ around 0, which is as follows:

$$\boldsymbol{\epsilon}^* = \arg \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) \approx \arg \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \approx \rho \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})\|_2}.$$

Putting into words, the worst-case perturbation is approximated as the scaled gradient of the loss w.r.t the current parameter $\boldsymbol{\theta}$. Then, the gradient w.r.t this perturbed model is computed to update $\boldsymbol{\theta}$:

$$\boldsymbol{g}^{\text{SAM}} := \nabla_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) \approx \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon})|_{\boldsymbol{\theta} + \boldsymbol{\epsilon}^*} \quad (2)$$

4 Our Proposed Framework

This section describes our proposed framework to improve existing methods on gradient-based MTL. We first recall the goal of multi-task learning, then establish the upper bounds for the general loss of each task. Subsequently, we rely on these upper bounds to devise the proposed framework for improving the model generalization ability by guiding it to a flatter region of each task.

4.1 Multi-task learning setting

In multi-task learning, we are given a data-label distribution \mathcal{D} from which we can sample a training set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i^1, \dots, y_i^m)_{i=1}^n\}$, where \boldsymbol{x}_i is a data example and y_i^1, \dots, y_i^m are the labels of the tasks 1, 2, ..., m respectively.

The model for each task $\boldsymbol{\theta}^i = [\boldsymbol{\theta}_{sh}, \boldsymbol{\theta}_{ns}^i]$ consists of the shared part $\boldsymbol{\theta}_{sh}$ and the individual non-shared part $\boldsymbol{\theta}_{ns}^i$. We denote the general loss for task i as $\mathcal{L}_{\mathcal{D}}^i(\boldsymbol{\theta}^i)$, while its empirical loss over the training set \mathcal{S} as $\mathcal{L}_{\mathcal{S}}^i(\boldsymbol{\theta}^i)$. Existing works in MTL, typically MGDA [8], PCGrad [9], CAGrad [10], and IMTL [11], aim to find a model that simultaneously minimizes the empirical losses for all tasks:

$$\min_{\boldsymbol{\theta}_{sh}, \boldsymbol{\theta}_{ns}^{1:m}} [\mathcal{L}_{\mathcal{S}}^1(\boldsymbol{\theta}^1), \dots, \mathcal{L}_{\mathcal{S}}^m(\boldsymbol{\theta}^m)], \quad (3)$$

by calculating gradient \mathbf{g}^i for i -th task ($i \in [m]$). The current model parameter is then updated by the unified gradient $\mathbf{g} = \text{gradient_aggregate}(\mathbf{g}^1, \mathbf{g}^2, \dots, \mathbf{g}^m)$, with the generic operation `gradient_aggregate` is to combine multiple task gradients, as proposed in gradient-based MTL studies.

Additionally, prior works only focus on minimizing the empirical losses and do not concern the general losses which combat overfitting. Inspired by SAM [21], it is desirable to develop sharpness-aware MTL approaches wherein the task models simultaneously seek low loss and flat regions. However, this is challenging since we have multiple objective functions in (3) and each task model consists of a shared and an individual non-shared parts. To address the above challenge, in Theorem 1, we develop upper bounds for the task general losses in the context of MTL which signifies the concepts of sharpness for the shared part and non-shared parts and then rely on these new concepts to devise a novel MTL framework via seeking the task-based flat regions.

4.2 Theoretical development

We first state our main theorem that bounds the generalization performance of individual tasks by the empirical error on the training set:

Theorem 1. (Informally stated) *For any perturbation radius $\rho_{sh}, \rho_{ns} > 0$, under some mild assumptions, with probability $1 - \delta$ (over the choice of training set $\mathcal{S} \sim \mathcal{D}$) we obtain*

$$[\mathcal{L}_{\mathcal{D}}^i(\boldsymbol{\theta}^i)]_{i=1}^m \leq \max_{\|\boldsymbol{\epsilon}_{sh}\|_2 \leq \rho_{sh}} \left[\max_{\|\boldsymbol{\epsilon}_{ns}^i\|_2 \leq \rho_{ns}} \mathcal{L}_{\mathcal{S}}^i(\boldsymbol{\theta}_{sh} + \boldsymbol{\epsilon}_{sh}, \boldsymbol{\theta}_{ns}^i + \boldsymbol{\epsilon}_{ns}^i) + f^i(\|\boldsymbol{\theta}^i\|_2^2) \right]_{i=1}^m, \quad (4)$$

where $f^i : \mathbb{R}_+ \rightarrow \mathbb{R}_+, i \in [m]$ are strictly increasing functions.

Theorem 1 establishes the connection between the generalization error of each task with its empirical training error via worst-case perturbation on the parameter space. The formally stated theorem and proof are provided in the appendix. Here we note that the worst-case shared perturbation $\boldsymbol{\epsilon}_{sh}$ is commonly learned for all tasks, while the worst-case non-shared perturbation $\boldsymbol{\epsilon}_{ns}^i$ is tailored for each task i . Theorem 1 directly hints us an initial and direct approach.

Additionally, [21] invokes the [56] PAC-Bayesian generalization bound, hence is only applicable to the 0-1 loss in the binary classification setting. In terms of the theory contribution, we employ a more general PAC-Bayesian generalization bound [57] to tackle more general losses in MTL. Moreover, our theory development requires us to handle multiple objectives, each of which consists of the non-shared and shared parts, which is certainly non-trivial.

4.3 Initial and direct approach

A straight-forward approach guided by Theorem 1 is to find the non-shared perturbations $\boldsymbol{\epsilon}_{ns}^i, i \in [m]$ independently for the non-shared parts and a common shared perturbation for the shared part. Driven by this theoretical guidance, we propose the following updates.

Update the non-shared parts. Based on the upper bounds in Theorem 1, because the non-shared perturbations $\boldsymbol{\epsilon}_{ns}^i, i \in [m]$ are independent to each task, for task i , we update its non-shared

part θ_{ns}^i :

$$\begin{aligned}\epsilon_{ns}^i &= \rho_{ns} \frac{\nabla_{\theta_{ns}^i} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)}{\|\nabla_{\theta_{ns}^i} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)\|_2}, \\ \mathbf{g}_{ns}^{i, \text{SAM}} &= \nabla_{\theta_{ns}^i} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i + \epsilon_{ns}^i), \\ \theta_{ns}^i &= \theta_{ns}^i - \eta \mathbf{g}_{ns}^{i, \text{SAM}}, \quad \text{where } \eta > 0 \text{ is the learning rate.}\end{aligned}\tag{5}$$

Update the shared part. Updating the shared part θ_{sh} is more challenging because its worst-cased perturbation ϵ_{sh} is shared among the tasks. To derive how to update θ_{sh} w.r.t. all tasks, we first discuss the case when we update this w.r.t. task i without caring about other tasks. Specifically, this task’s SAM shared gradient is computed as:

$$\begin{aligned}\epsilon_{sh}^i &= \rho_{sh} \frac{\nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)}{\|\nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)\|_2}, \\ \mathbf{g}_{sh}^{i, \text{SAM}} &= \nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}^i, \theta_{ns}^i),\end{aligned}$$

then we have a straight-forward updating strategy:

$$\begin{aligned}\mathbf{g}_{sh}^{\text{SAM}} &= \text{gradient_aggregate}(\mathbf{g}_{sh}^{1, \text{SAM}}, \dots, \mathbf{g}_{sh}^{m, \text{SAM}}), \\ \theta_{sh} &= \theta_{sh} - \eta \mathbf{g}_{sh}^{\text{SAM}}.\end{aligned}$$

According to our analysis in Section 4.4, each $\mathbf{g}_{sh}^{i, \text{SAM}} = \mathbf{g}_{sh}^{i, \text{loss}} + \mathbf{g}_{sh}^{i, \text{flat}}$ is constituted by two components: (i) $\mathbf{g}_{sh}^{i, \text{loss}}$ to navigate to the task low-loss region and (ii) $\mathbf{g}_{sh}^{i, \text{flat}}$ to navigate to the task-based flat region. However, a direct gradient aggregation of $\mathbf{g}_{sh}^{i, \text{SAM}}, i \in [m]$ can be negatively affected by the gradient cancelation or conflict because it aims to combine many individual elements with different objectives. In this paper, we go beyond this initial approach by deriving an updating formula to decompose SAM gradient into two components, each serving its own purpose, and then combining their corresponding task gradients simultaneously. We also compare our method against the naive approach in Section 5.3.

4.4 Our proposed approach

The non-shared parts are updated normally as in Equation (5). It is more crucial to investigate how to update the shared part more efficiently. To better understand the SAM’s gradients, we analyze their characteristics by deriving them as follows:

$$\begin{aligned}\mathbf{g}_{sh}^{i, \text{SAM}} &= \nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}^i, \theta_{ns}^i) \stackrel{(1)}{\approx} \nabla_{\theta_{sh}} [\mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)] + \langle \epsilon_{sh}^i, \nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i) \rangle \\ &= \nabla_{\theta_{sh}} \left[\mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i) + \rho_{sh} \left\langle \frac{\nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)}{\|\nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)\|_2}, \nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i) \right\rangle \right] \\ &= \nabla_{\theta_{sh}} [\mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i) + \rho_{sh} \|\nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)\|_2]\end{aligned}\tag{6}$$

where in $\stackrel{(1)}{\approx}$, we apply the first-order Taylor expansion and $\langle \cdot, \cdot \rangle$ represents the dot product.

It is obvious that following the negative direction of $\mathbf{g}_{sh}^{i, \text{SAM}}$ will minimize the loss $\mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)$ and the gradient norm $\|\nabla_{\theta_{sh}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh}, \theta_{ns}^i)\|_2$ of task i , hence leading the model to the low-valued region for the loss of this task and its flatter region with a lower gradient norm magnitude.

Moreover, inspired from the derivation in Equation (6), we decompose the gradient $\mathbf{g}_{sh}^{i,\text{SAM}} = \mathbf{g}_{sh}^{i,\text{loss}} + \mathbf{g}_{sh}^{i,\text{flat}}$ where we define $\mathbf{g}_{sh}^{i,\text{loss}} := \nabla_{\theta_{sh}} \mathcal{L}_S^i(\theta_{sh}, \theta_{ns}^i)$ and $\mathbf{g}_{sh}^{i,\text{flat}} := \mathbf{g}_{sh}^{i,\text{SAM}} - \mathbf{g}_{sh}^{i,\text{loss}}$. As aforementioned, the purpose of the negative gradient $-\mathbf{g}_{sh}^{i,\text{loss}}$ is to orient the model to minimize the loss of the task i , while $-\mathbf{g}_{sh}^{i,\text{flat}}$ navigates the model to the task i 's flatter region.

Therefore, the SAM gradients $\mathbf{g}_{sh}^{i,\text{SAM}}, i \in [m]$ constitute two components with different purposes. To mitigate the possible confliction and interference of the two components when aggregating, we propose to aggregate the low-loss components solely and then the flat components solely. Specifically, to find a common direction that leads the joint low-valued losses for all tasks and the joint flatter region for them, we first combine the gradients $\mathbf{g}_{sh}^{i,\text{loss}}, i \in [m]$ and the gradients $\mathbf{g}_{sh}^{i,\text{flat}}, i \in [m]$, then add the two aggregated gradients, and finally update the shared part as:

$$\begin{aligned}\mathbf{g}_{sh}^{\text{loss}} &= \text{gradient_aggregate}(\mathbf{g}_{sh}^{1,\text{loss}}, \dots, \mathbf{g}_{sh}^{m,\text{loss}}), \\ \mathbf{g}_{sh}^{\text{flat}} &= \text{gradient_aggregate}(\mathbf{g}_{sh}^{1,\text{flat}}, \dots, \mathbf{g}_{sh}^{m,\text{flat}}), \\ \mathbf{g}_{sh}^{\text{SAM}} &= \mathbf{g}_{sh}^{\text{loss}} + \mathbf{g}_{sh}^{\text{flat}}; \quad \theta_{sh} = \theta_{sh} - \eta \mathbf{g}_{sh}^{\text{SAM}},\end{aligned}$$

Finally, the key steps of our proposed framework are summarized in Algorithm 1 and the overall schema of our proposed method is demonstrated in Figure 1.

Algorithm 1 Sharpness minimization for multi-task learning

Input: Model parameter $\theta = [\theta_{sh}, \theta_{ns}^{1:m}]$, perturbation radius $\rho = [\rho_{sh}, \rho_{ns}]$, step size η and a list of m differentiable loss functions $\{\mathcal{L}^i\}_{i=1}^m$.

Output: Updated parameter θ^*

- 1: **for** task $i \in [m]$ **do**
- 2: Compute gradient $\mathbf{g}_{sh}^{i,\text{loss}}, \mathbf{g}_{ns}^i \leftarrow \nabla_{\theta} \mathcal{L}^i(\theta)$
- 3: Worst-case perturbation direction
 $\epsilon_{sh}^i = \rho_{sh} \cdot \mathbf{g}_{sh}^{i,\text{loss}} / \|\mathbf{g}_{sh}^{i,\text{loss}}\|$ and $\epsilon_{ns}^i = \rho_{ns} \cdot \mathbf{g}_{ns}^i / \|\mathbf{g}_{ns}^i\|$
- 4: Approximate SAM's gradient
 $\mathbf{g}_{sh}^{i,\text{SAM}} = \nabla_{\theta_{sh}} \mathcal{L}^i(\theta_{sh} + \epsilon_{sh}^i, \theta_{ns}^i)$ and $\mathbf{g}_{ns}^{i,\text{SAM}} = \nabla_{\theta_{ns}^i} \mathcal{L}^i(\theta_{sh}, \theta_{ns}^i + \epsilon_{ns}^i)$
- 5: Compute flat gradient
 $\mathbf{g}_{sh}^{i,\text{flat}} = \mathbf{g}_{sh}^{i,\text{SAM}} - \mathbf{g}_{sh}^{i,\text{loss}}$
- 6: **end for**
- 7: Calculate combined update gradients:

$$\begin{aligned}\mathbf{g}_{sh}^{\text{loss}} &= \text{gradient_aggregate}(\mathbf{g}_{sh}^{1,\text{loss}}, \mathbf{g}_{sh}^{2,\text{loss}}, \dots, \mathbf{g}_{sh}^{m,\text{loss}}) \\ \mathbf{g}_{sh}^{\text{flat}} &= \text{gradient_aggregate}(\mathbf{g}_{sh}^{1,\text{flat}}, \mathbf{g}_{sh}^{2,\text{flat}}, \dots, \mathbf{g}_{sh}^{m,\text{flat}})\end{aligned}$$

- 8: Calculate shared gradient update $\mathbf{g}_{sh}^{\text{SAM}} = \mathbf{g}_{sh}^{\text{loss}} + \mathbf{g}_{sh}^{\text{flat}}$
 - 9: Update model parameter
 $\theta^* = [\theta_{sh}, \theta_{ns}^{1:m}] - \eta[\mathbf{g}_{sh}^{\text{SAM}}, \mathbf{g}_{ns}^{1:m,\text{SAM}}]$
-

5 Experiments

In this section, we compare our proposed method against other state-of-the-art methods of multi-task learning in different scenarios, ranging from image classification to scene understanding problems. Refer to the appendix for the detailed settings used for each dataset and additional experiments.

Datasets and Baselines. Our proposed method is evaluated on four MTL benchmarks including Multi-MNIST [25], CelebA [58] for visual classification, and NYUv2 [59], CityScapes [60] for scene understanding. We show how our framework can boost the performance of gradient-based MTL methods by comparing *vanilla* MGDA [8], PCGrad [9], CAGrad [10] and IMTL [11] to their flat-based versions F-MGDA, F-PCGrad, F-CAGrad and F-IMTL. We also add single task learning (STL) baseline for each dataset.

5.1 Image classification

Multi-MNIST. Following the protocol of [8], we set up three Multi-MNIST experiments with the ResNet18 [61] backbone, namely: MultiFashion, MultiMNIST and MultiFashion+MNIST. In each dataset, two images are sampled uniformly from the MNIST [62] or Fashion-MNIST [63], then one is placed on the top left and the other is on the bottom right. We thus obtain a two-task learning that requires predicting the categories of the digits or fashion items on the top left (task 1) and on the bottom right (task 2) respectively.

Table 1: Evaluation of different methods on three Multi-MNIST datasets. Rows with flat-based minimizers are shaded. Bold numbers denote higher accuracy between flat-based methods and their baselines. * denotes the highest accuracy (except for STL, since it unfairly exploits multiple neural networks). We also use arrows to indicate that the higher is the better (\uparrow) or vice-versa (\downarrow).

Method	MultiFashion			MultiMNIST			MultiFashion+MNIST		
	Task 1 \uparrow	Task 2 \uparrow	Average \uparrow	Task 1 \uparrow	Task 2 \uparrow	Average \uparrow	Task 1 \uparrow	Task 2	Average \uparrow
STL	87.10 \pm 0.09	86.20 \pm 0.06	86.65 \pm 0.02	95.33 \pm 0.08	94.16 \pm 0.04	94.74 \pm 0.06	98.40 \pm 0.02	89.42 \pm 0.03	93.91 \pm 0.02
MGDA	86.76 \pm 0.09	85.78 \pm 0.36	86.27 \pm 0.22	95.62 \pm 0.02	94.49 \pm 0.10	95.05 \pm 0.06	97.24 \pm 0.04	88.19 \pm 0.13	92.72 \pm 0.07
F-MGDA	88.12 \pm 0.11	87.35 \pm 0.11	87.73 \pm 0.09	96.37 \pm 0.06	94.99 \pm 0.06	95.68 \pm 0.00	97.30 \pm 0.09	89.26 \pm 0.14	93.28 \pm 0.03
PCGrad	86.93 \pm 0.17	86.20 \pm 0.14	86.57 \pm 0.12	95.71 \pm 0.03	94.41 \pm 0.02	95.06 \pm 0.02	97.12 \pm 0.16	88.45 \pm 0.08	92.78 \pm 0.11
F-PCGrad	88.17 \pm 0.14	87.35 \pm 0.27	87.76 \pm 0.07	96.49 \pm 0.05	95.34 \pm 0.10	95.92 \pm 0.07	97.65 \pm 0.06	89.35 \pm 0.07*	93.50 \pm 0.01
CAGrad	86.99 \pm 0.17	86.04 \pm 0.15	86.51 \pm 0.16	95.62 \pm 0.05	94.39 \pm 0.04	95.01 \pm 0.04	97.19 \pm 0.06	88.18 \pm 0.14	92.68 \pm 0.04
F-CAGrad	88.19 \pm 0.19*	87.45 \pm 0.13	87.82 \pm 0.10*	96.54 \pm 0.02	95.36 \pm 0.04*	95.95 \pm 0.01*	97.82 \pm 0.05*	89.26 \pm 0.22	93.54 \pm 0.13*
IMTL	87.35 \pm 0.22	86.45 \pm 0.09	86.90 \pm 0.15	95.93 \pm 0.09	94.63 \pm 0.13	95.28 \pm 0.02	97.47 \pm 0.06	88.46 \pm 0.11	92.97 \pm 0.03
F-IMTL	88.1 \pm 0.10	87.5 \pm 0.04*	87.80 \pm 0.06	96.55 \pm 0.07*	95.16 \pm 0.05	95.85 \pm 0.05	97.59 \pm 0.12	88.99 \pm 0.08	93.29 \pm 0.02

As summarized in Table 1, we can see that seeking flatter regions for all tasks can improve the performance of all the baselines across all three datasets. Especially, flat-based methods achieve the highest score for each task and for the average, outperforming STL by 1.2% on MultiFashion and MultiMNIST. We conjecture that the discrepancy between gradient update trajectories to classify digits from MNIST and fashion items from FashionMNIST has resulted in the fruitless performance of baselines, compared to STL on MultiFashion+MNIST. Even if there exists dissimilarity between tasks, our best obtained average accuracy when applying our method to CAGrad is just slightly lower than STL ($< 0.4\%$) while employing a single model only.

Interestingly, our proposed MTL training method also helps improve model calibration performance by mitigating the over-confident phenomenon of deep neural networks. As can be seen from Figure

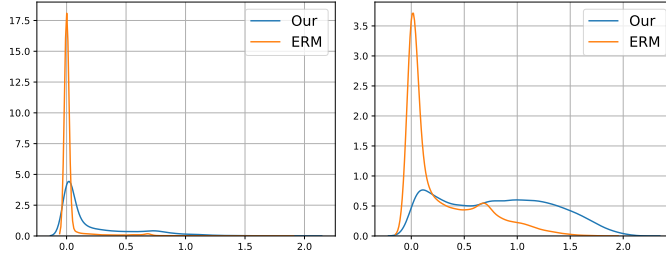


Figure 2: Entropy distributions of ResNet18 on in-domain set (left) and out-of-domain set (right).

2, our method produces high-entropy predictions that represent its uncertainty, ERM-based method outputs high confident predictions on both in and out-of-domain data. More details about model calibration improvement can be found in the appendix.

CelebA. CelebA [64] is a face dataset, which consists of 200K celebrity facial photos with 40 attributes. Similar to [8], each attribute forms a binary classification problem, thus a 40-class multi-label classification problem is constructed.

Table 2 shows the average errors over 40 tasks of the methods with linear scalarization (LS) and Uncertainty weighting (UW) [32] being included to serve as comparative baselines. The best results in each pair and among all are highlighted using **bold** font and *, respectively. When the number of tasks is large, flat region seeking still consistently shows its advantages and the lowest average accuracy error is achieved by F-CAGrad. Interestingly, when the optimizer is aware of flat minima, the gaps between PCGrad, IMTL and CAGrad, (8.23, 8.24 vs 8.22), are smaller than those using conventional ERM training, (8.69, 8.88 and 8.52). This might be due to the better aggregation of tasks’ gradients, which means that the conflict between these gradients is likely to be reduced when the shared parameters approach the common flat region of all tasks.

Table 2: Mean of error per category of MTL algorithms in multi-label classification on CelebA dataset.

Method	Average error ↓
STL	8.77
LS	9.99
UW	9.66
MGDA	9.96
F-MGDA	9.22
PCGrad	8.69
F-PCGrad	8.23
CAGrad	8.52
F-CAGrad	8.22*
IMTL	8.88
F-IMTL	8.24

5.2 Scene Understanding

Two datasets used in this sub-section are NYUv2 [59] and CityScapes [60]. NYUv2 is an indoor scene dataset that contains 3 tasks: 13-class semantic segmentation, depth estimation, and surface normal prediction. In CityScapes, there are 19 classes of street-view images, which are coarsened into 7 categories to create two tasks: semantic segmentation and depth estimation. For these two experiments, we additionally include several recent MTL methods, namely, scale-invariant (SI), random loss weighting (RLW), Dynamic Weight Average (DWA) [2], GradDrop [65], and Nash-MTL [13] whose results are taken from [13]. Details of each baseline can be found in the appendix. Also following the standard protocol used in [2, 10, 13], Multi-Task Attention Network [2] is employed on top of the SegNet architecture [66], our presented results are averaged over the last 10 epochs to align with previous work.

Evaluation metric. In this experiment, we have to deal with different task types rather than one only as in the case of image classification. Since each of them has its own set of metrics. We thus

mark the overall performance of comparative methods by reporting their relative task improvement [67] throughout this section. Let M_i and S_i be the metrics obtained by the main and the

Table 3: Test performance for two-task CityScapes: semantic segmentation and depth estimation.* denotes the best score for each task’s metrics.

	Segmentation		Depth		
Method	mIoU \uparrow	Pix Acc \uparrow	Abs Err \downarrow	Rel Err \downarrow	$\Delta m\%$ \downarrow
STL	74.01	93.16	0.0125	27.77	
LS	75.18	93.49	0.0155	46.77	22.60
SI	70.95	91.73	0.0161	33.83	14.11
RLW	74.57	93.41	0.0158	47.79	24.38
DWA	75.24	93.52	0.0160	44.37	21.45
UW	72.02	92.85	0.0140	30.13*	5.89
GradDrop	75.27	93.53	0.0157	47.54	23.73
Nash-MTL	75.41	93.66	0.0129	35.02	6.82
MGDA	68.84	91.54	0.0309	33.50	44.14
F-MGDA	73.77	93.12	0.0129	27.44*	0.67*
PCGrad	75.13	93.48	0.0154	42.07	18.29
F-PCGrad	75.77	93.67	0.0144	39.60	13.65
CAGrad	75.16	93.48	0.0141	37.60	11.64
F-CAGrad	76.02	93.72	0.0134	34.64	7.25
IMTL	75.33	93.49	0.0135	38.41	11.10
F-IMTL	76.63*	93.76*	0.0124*	31.17	1.87

single-task learning (STL) model, respectively, the relative task improvement on i -th task is mathematically given by: $\Delta_i := 100 \cdot (-1)^{l_i} (M_i - S_i)/S_i$, where $l_i = 1$ if a lower value for the i -th criterion is better and 0 otherwise. We depict our results by the average relative task improvement $\Delta m\% = \frac{1}{m} \sum_{i=1}^m \Delta_i$.

CityScapes. In Table 3, the positive effect of seeking flat regions is consistently observed in all metrics and baselines. In particular, the relative improvements of MGDA and IMTL are significantly boosted, achieving the highest and second-best $\Delta m\%$ scores, respectively. The segmentation scores of PCGrad, CAGrad and IMTL even surpass STL. Intriguingly, MGDA biases to the depth estimation objective, leading to the predominant performance on that task, similar patterns appear in [11] and the below NYUv2 experiment.

NYUv2. Table 4 shows each task’s results and the relative improvements over STL of different methods. Generally, the flat-based versions obtain comparable or higher results on most of the metrics, except for MGDA at the segmentation task, in which F-MGDA notably decreases the $mIoU$ score. However, it does significantly help other tasks, which contributes to the overall MGDA’s relative improvement, from 1.38% being worse than STL to 0.33% being higher. Remarkably, F-CAGrad and F-IMTL outperform their competitors by large margins across all tasks, resulting in the top two relative improvements, 3.78% and 4.77% .

5.3 Ablation study

So far, our proposed technique has shown state-of-the-art performances under different settings, we now investigate in more detailed how it affects conventional training by inspecting loss surfaces and model robustness. Similar patterns are observed in other experiments and given in the appendix.

Task conflict. To empirically confirm that tasks’ gradients are less conflicted when the model is driven to the flat regions, we measure the gradient conflict and present the result in Figure 3.

Table 4: Test performance for three-task NYUv2 of Segnet [66]: semantic segmentation, depth estimation, and surface normal. Using the proposed procedure in conjunction with gradient-based multi-task learning methods consistently advances their overall performance.

	Segmentation		Depth		Surface Normal					$\Delta m\%$ \downarrow
	mIoU \uparrow	Pix Acc \uparrow	Abs Err \downarrow	Rel Err \downarrow	Angle Distance \downarrow		Within t° \uparrow			
					Mean	Median	11.25	22.5	30	
STL	38.30	63.76	0.6754	0.2780	25.01	19.21	30.14	57.20	69.15	0.00
LS	39.29	65.33	0.5493	0.2263	28.15	23.96	22.09	47.50	61.08	5.59
SI	38.45	64.27	0.5354	0.2201	27.60	23.37	22.53	48.57	62.32	4.39
RLW	37.17	63.77	0.5759	0.2410	28.27	24.18	22.26	47.05	60.62	7.78
DWA	39.11	65.31	0.5510	0.2285	27.61	23.18	24.17	50.18	62.39	3.57
UW	36.87	63.17	0.5446	0.2260	27.04	22.61	23.54	49.05	63.65	4.05
GradDrop	39.39	65.12	0.5455	0.2279	27.48	22.96	23.38	49.44	62.87	3.58
Nash-MTL	40.13	65.93	0.5261*	0.2171	25.26	20.08	28.4	55.47	68.15	−4.04
MGDA	30.47	59.90	0.6070	0.2555	24.88	19.45	29.18	56.88	69.36	1.38
F-MGDA	26.42	58.78	0.6078	0.2353	24.34*	18.45*	31.64*	58.86*	70.50*	−0.33
PCGrad	38.06	64.64	0.5550	0.2325	27.41	22.80	23.86	49.83	63.14	3.97
F-PCGrad	40.05	65.42	0.5429	0.2243	27.38	23.00	23.47	49.35	62.74	3.14
CAGrad	39.79	65.49	0.5486	0.2250	26.31	21.58	25.61	52.36	65.58	0.20
F-CAGrad	40.93*	66.68*	0.5285	0.2162	25.43	20.39	27.99	54.82	67.56	−3.78
IMTL	39.35	65.60	0.5426	0.2256	26.02	21.19	26.2	53.13	66.24	−0.76
F-IMTL	40.42	65.61	0.5389	0.2121*	25.03	19.75	28.90	56.19	68.72	−4.77*

While the percentage of gradient conflict of ERM increases to more than 50%, ours decreases and approaches 0%. This reduction in gradient conflict is also the goal of recent gradient-based MTL methods in mitigating negative transfer between tasks [9, 68, 3].

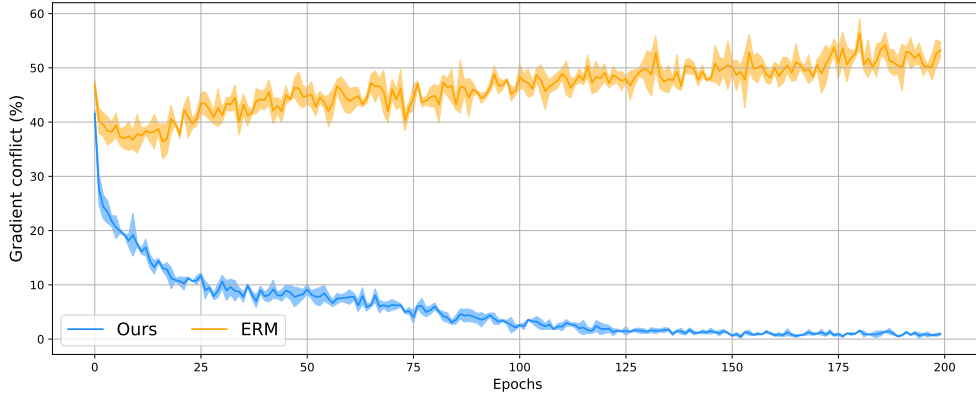


Figure 3: Proportion of conflict between per-task gradients ($\mathbf{g}^{1,\text{loss}} \cdot \mathbf{g}^{2,\text{loss}} < 0$) on MultiFashion+MNIST dataset.

Model robustness against noise. To verify that SAM can orient the model to the common flat and low-loss region of all tasks, we measure the model performance within a r -radius Euclidean ball. To be more specific, we perturb parameters of two converged models by ϵ , which lies in a r -radius ball and plot the accuracy of the perturbed models of each task as we increase r from 0 to

1000. At each value of r , 10 different models around the r -radius ball of the converged model are sampled.

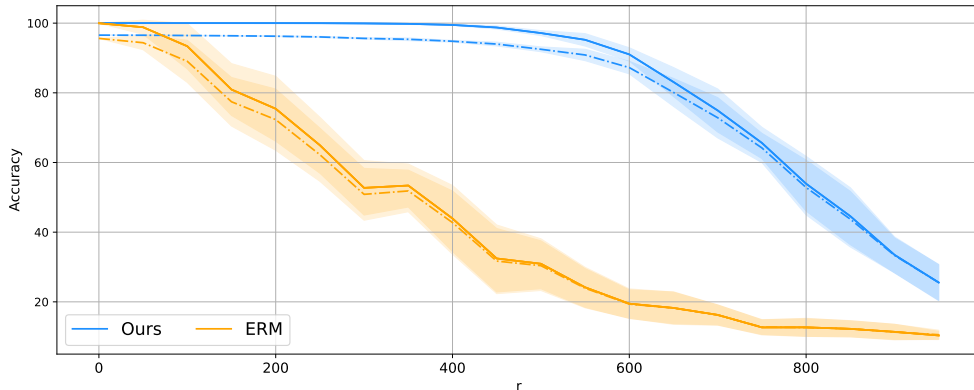


Figure 4: Accuracy within r -radius ball. Solid/dashed lines denote performance on train/test sets, respectively.

In Figure 4, the accuracy of the model trained using our method remains at a high level when noise keeps increasing until $r = 800$. This also gives evidence that our model found a region that changes slowly in loss. By contrast, the naively trained model loses its predictive capabilities as soon as the noise appears and becomes a dummy classifier that attains 10% accuracy in a 10-way classification.

Aggregation strategies comparison. Table 5 provides a comparison between the **direct** aggregation on $\{\mathbf{g}_{sh}^{i, \text{SAM}}\}_{i=1}^m$ and **individual** aggregation on $\{\mathbf{g}_{sh}^{i, \text{flat}}\}_{i=1}^m$ and $\{\mathbf{g}_{sh}^{i, \text{loss}}\}_{i=1}^m$ (our method).

Table 5: Two aggregation strategies on CityScapes.

	Segmentation		Depth		
Method	mIoU \uparrow	Pix Acc \uparrow	Abs Err \downarrow	Rel Err \downarrow	$\Delta m\%$ \downarrow
ERM	68.84	91.54	0.0309	33.50	44.14
Ours (direct)	68.93	91.41	0.0130	31.37	6.43
Ours (individual)	73.77	93.12	0.0129	27.44*	0.67

Compared to the naive approach, in which per-task SAM gradients are directly aggregated, our decomposition approach consistently improves performance by a large margin across all tasks. This result reinforces the rationale behind separately aggregating low-loss directions and flat directions.

Visualization of the loss landscapes. Following [69], we plot the loss surfaces at convergence after training Resnet18 from scratch on the MultiMNIST dataset. Test loss surfaces of checkpoints that have the highest validation accuracy scores are shown in Figure 5.

We can clearly see that the solution found by our proposed method not only mitigates the test loss sharpness for both tasks but also can intentionally reduce the test loss value itself, in comparison to traditional ERM. This is a common behavior when using flat minimizers as the gap between train and test performance has been narrowed [44, 14].

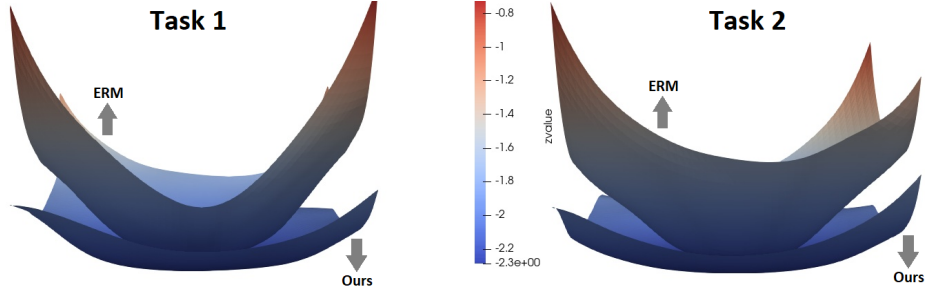


Figure 5: Visualization of test loss surfaces with standard ERM training and when applying our method. The coordinate plane axes are two random sampled orthogonal Gaussian perturbations.

6 Conclusion

In this work, we have presented a general framework that can be incorporated into current multi-task learning methods following the gradient balancing mechanism. The core ideas of our proposed method are the employment of flat minimizers in the context of MTL and proving that they can help enhance previous works both theoretically and empirically. Concretely, our method goes beyond optimizing per-task objectives solely to yield models that have both low errors and high generalization capabilities. On the experimental side, the efficacy of our method is demonstrated on a wide range of commonly used MTL benchmarks, in which ours consistently outperforms comparative methods.

References

- [1] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in *European conference on computer vision*, pp. 94–108, Springer, 2014.
- [2] S. Liu, E. Johns, and A. J. Davison, “End-to-end multi-task learning with attention,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1871–1880, 2019.
- [3] Z. Wang, Y. Tsvetkov, O. Firat, and Y. Cao, “Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models,” in *International Conference on Learning Representations*, 2020.
- [4] I. Kokkinos, “Ubertnet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6129–6138, 2017.
- [5] F. Heuer, S. Mantowsky, S. Bukhari, and G. Schneider, “Multitask-centernet (mcn): Efficient and diverse multitask learning using an anchor free approach,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 997–1005, 2021.
- [6] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, “Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3205–3214, 2019.
- [7] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, “Latent multi-task architecture learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4822–4829, 2019.
- [8] O. Sener and V. Koltun, “Multi-task learning as multi-objective optimization,” *Advances in neural information processing systems*, vol. 31, 2018.

- [9] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, “Gradient surgery for multi-task learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.
- [10] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, “Conflict-averse gradient descent for multi-task learning,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 18878–18890, 2021.
- [11] L. Liu, Y. Li, Z. Kuang, J.-H. Xue, Y. Chen, W. Yang, Q. Liao, and W. Zhang, “Towards impartial multi-task learning,” in *International Conference on Learning Representations*, 2020.
- [12] A. Javaloy and I. Valera, “Rotograd: Gradient homogenization in multitask learning,” in *International Conference on Learning Representations*, 2021.
- [13] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi, G. Chechik, and E. Fetaya, “Multi-task learning as a bargaining game,” *arXiv preprint arXiv:2202.01017*, 2022.
- [14] J. Kaddour, L. Liu, R. Silva, and M. J. Kusner, “A fair comparison of two popular flat minima optimizers: Stochastic weight averaging vs. sharpness-aware minimization,” *arXiv preprint arXiv:2202.00661*, vol. 1, 2022.
- [15] Y. Zhao, H. Zhang, and X. Hu, “Penalizing gradient norm for efficiently improving generalization in deep learning,” *arXiv preprint arXiv:2202.03599*, 2022.
- [16] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, “On large-batch training for deep learning: Generalization gap and sharp minima,” in *5th International Conference on Learning Representations, ICLR 2017*, 2017.
- [17] Z. Li, Z. Wang, and J. Li, “Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability,” *arXiv preprint arXiv:2207.12678*, 2022.
- [18] K. Lyu, Z. Li, and S. Arora, “Understanding the generalization benefit of normalization layers: Sharpness reduction,” *arXiv preprint arXiv:2206.07085*, 2022.
- [19] H. He, G. Huang, and Y. Yuan, “Asymmetric valleys: Beyond sharp and flat local minima,” *Advances in neural information processing systems*, vol. 32, 2019.
- [20] Y. Zheng, R. Zhang, and Y. Mao, “Regularizing neural networks via adversarial model perturbation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8156–8165, 2021.
- [21] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, “Sharpness-aware minimization for efficiently improving generalization,” in *International Conference on Learning Representations*, 2021.
- [22] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [23] X. Liu, P. He, W. Chen, and J. Gao, “Multi-task deep neural networks for natural language understanding,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, Association for Computational Linguistics, 2019.
- [24] S. Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv preprint arXiv:1706.05098*, 2017.
- [25] X. Lin, H.-L. Zhen, Z. Li, Q.-F. Zhang, and S. Kwong, “Pareto multi-task learning,” *Advances in neural information processing systems*, vol. 32, 2019.
- [26] X. Liu, X. Tong, and Q. Liu, “Profiling pareto front with multi-objective stein variational gradient descent,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 14721–14733, 2021.
- [27] D. Mahapatra and V. Rajan, “Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization,” in *International Conference on Machine Learning*, pp. 6597–6607, PMLR, 2020.

- [28] D. Mahapatra and V. Rajan, “Exact pareto optimal search for multi-task learning: Touring the pareto front,” *arXiv preprint arXiv:2108.00597*, 2021.
- [29] X. Lin, Z. Yang, Q. Zhang, and S. Kwong, “Controllable pareto multi-task learning,” *arXiv preprint arXiv:2010.06313*, 2020.
- [30] A. Navon, A. Shamsian, G. Chechik, and E. Fetaya, “Learning the pareto front with hypernetworks,” in *International Conference on Learning Representations*, 2021.
- [31] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, “Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks,” in *International conference on machine learning*, pp. 794–803, PMLR, 2018.
- [32] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.
- [33] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, “Dynamic task prioritization for multitask learning,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 270–287, 2018.
- [34] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, “Fantastic generalization measures and where to find them,” in *ICLR*, OpenReview.net, 2020.
- [35] H. Petzka, M. Kamp, L. Adilova, C. Sminchisescu, and M. Boley, “Relative flatness and generalization,” in *NeurIPS*, pp. 18420–18432, 2021.
- [36] G. K. Dziugaite and D. M. Roy, “Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data,” in *UAI*, AUAI Press, 2017.
- [37] S. Hochreiter and J. Schmidhuber, “Simplifying neural nets by discovering flat minima,” in *NIPS*, pp. 529–536, MIT Press, 1994.
- [38] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, “Exploring generalization in deep learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [39] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, “Sharp minima can generalize for deep nets,” in *International Conference on Machine Learning*, pp. 1019–1028, PMLR, 2017.
- [40] S. Fort and S. Ganguli, “Emergent properties of the local geometry of neural loss landscapes,” *arXiv preprint arXiv:1910.05929*, 2019.
- [41] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. E. Hinton, “Regularizing neural networks by penalizing confident output distributions,” in *ICLR (Workshop)*, OpenReview.net, 2017.
- [42] P. Chaudhari, A. Choromańska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. T. Chayes, L. Sagun, and R. Zecchina, “Entropy-sgd: biasing gradient descent into wide valleys,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, 2017.
- [43] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, “On large-batch training for deep learning: Generalization gap and sharp minima,” in *ICLR*, OpenReview.net, 2017.
- [44] P. Izmailov, D. Podoprikin, T. Garipov, D. P. Vetrov, and A. G. Wilson, “Averaging weights leads to wider optima and better generalization,” in *UAI*, pp. 876–885, AUAI Press, 2018.
- [45] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. J. Storkey, “Three factors influencing minima in sgd,” *ArXiv*, vol. abs/1711.04623, 2017.
- [46] C. Wei, S. Kakade, and T. Ma, “The implicit and explicit regularization effects of dropout,” in *International conference on machine learning*, pp. 10181–10192, PMLR, 2020.
- [47] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep mutual learning,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018.

- [48] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, “Be your own teacher: Improve the performance of convolutional neural networks via self distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3713–3722, 2019.
- [49] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, “Swad: Domain generalization by seeking flat minima,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 22405–22418, 2021.
- [50] M. Abbas, Q. Xiao, L. Chen, P.-Y. Chen, and T. Chen, “Sharp-maml: Sharpness-aware model-agnostic meta learning,” *arXiv preprint arXiv:2206.03996*, 2022.
- [51] Z. Qu, X. Li, R. Duan, Y. Liu, B. Tang, and Z. Lu, “Generalized federated learning via sharpness aware minimization,” *arXiv preprint arXiv:2206.02618*, 2022.
- [52] D. Caldarola, B. Caputo, and M. Ciccone, “Improving generalization in federated learning by seeking flat minima,” in *European Conference on Computer Vision*, pp. 654–672, Springer, 2022.
- [53] D. Bahri, H. Mobahi, and Y. Tay, “Sharpness-aware minimization improves language model generalization,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Dublin, Ireland), pp. 7360–7371, Association for Computational Linguistics, May 2022.
- [54] X. Chen, C.-J. Hsieh, and B. Gong, “When vision transformers outperform resnets without pre-training or strong data augmentations,” *arXiv preprint arXiv:2106.01548*, 2021.
- [55] V.-A. Nguyen, T.-L. Vuong, H. Phan, T.-T. Do, D. Phung, and T. Le, “Flat seeking bayesian neural networks,” *Advances in Neural Information Processing Systems*, 2023.
- [56] D. A. McAllester, “Pac-bayesian model averaging,” in *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999.
- [57] P. Alquier, J. Ridgway, and N. Chopin, “On the properties of variational approximations of gibbs posteriors,” *Journal of Machine Learning Research*, vol. 17, no. 236, pp. 1–41, 2016.
- [58] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.
- [59] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgb-d images,” in *European conference on computer vision*, pp. 746–760, Springer, 2012.
- [60] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.
- [61] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [62] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [63] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [64] Z. Liu, P. Luo, X. Wang, and X. Tang, “Large-scale celebfaces attributes (celeba) dataset,” *Retrieved August*, vol. 15, no. 2018, p. 11, 2018.
- [65] Z. Chen, J. Ngiam, Y. Huang, T. Luong, H. Kretzschmar, Y. Chai, and D. Anguelov, “Just pick a sign: Optimizing deep multitask models with gradient sign dropout,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 2039–2050, 2020.
- [66] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

- [67] K.-K. Maninis, I. Radosavovic, and I. Kokkinos, “Attentive single-tasking of multiple tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1851–1860, 2019.
- [68] S. Zhu, H. Zhao, P. Wang, H. Deng, J. Xu, and B. Zheng, “Gradient deconfliction via orthogonal projections onto subspaces for multi-task learning,” 2022.
- [69] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” *Advances in neural information processing systems*, vol. 31, 2018.
- [70] B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection,” *Annals of Statistics*, pp. 1302–1338, 2000.
- [71] J.-A. Désidéri, “Multiple-gradient descent algorithm (mgda) for multiobjective optimization,” *Comptes Rendus Mathématique*, vol. 350, no. 5-6, pp. 313–318, 2012.
- [72] B. Lin, F. Ye, Y. Zhang, and I. W. Tsang, “Reasonable effectiveness of random weighting: A litmus test for multi-task learning,” *arXiv preprint arXiv:2111.10603*, 2021.
- [73] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [74] J. Kwon, J. Kim, H. Park, and I. K. Choi, “Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks,” 18–24 Jul 2021.
- [75] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.
- [76] G. W. Brier *et al.*, “Verification of forecasts expressed in terms of probability,” *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.
- [77] M. P. Naeini, G. Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using bayesian binning,” in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [78] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift,” *Advances in neural information processing systems*, vol. 32, 2019.
- [79] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in neural information processing systems*, vol. 30, 2017.
- [80] A. Malinin and M. Gales, “Predictive uncertainty estimation via prior networks,” *Advances in neural information processing systems*, vol. 31, 2018.

Supplement to “Improving Multi-task Learning via Seeking Task-based Flat Regions”

Due to space constraints, some details were omitted from the main paper. We therefore include additional theoretical developments (section A) and experimental results (section C) in this appendix.

A Our Theory Development

This section contains the proofs and derivations of our theory development to support the main submission.

We first start with the following theorem, which is inspired by the general PAC-Bayes in [57].

Theorem 2. *With the assumption that adding Gaussian perturbation will raise the test error: $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) \leq \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta} + \epsilon)]$. Let T be the number of parameter $\boldsymbol{\theta}$, and N be the cardinality of \mathcal{S} , then the following inequality is true with the probability $1 - \delta$:*

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) \leq \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta} + \epsilon)] + \frac{1}{\sqrt{N}} \left[\frac{1}{2} + \frac{T}{2} \log \left(1 + \frac{\|\boldsymbol{\theta}\|^2}{T\sigma^2} \right) + \log \frac{1}{\delta} + 6 \log(N + T) + \frac{L^2}{8} \right]$$

where L is the upper-bound of the loss function.

Proof. We use the PAC-Bayes theory for $P = \mathcal{N}(\mathbf{0}, \sigma_P^2 \mathbb{I}_T)$ and $Q = \mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbb{I}_T)$ are the prior and posterior distributions, respectively.

By using the bound in [57], with probability at least $1 - \delta$ and for all $\beta > 0$, we have:

$$\mathbb{E}_{\boldsymbol{\theta} \sim Q} [\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})] \leq \mathbb{E}_{\boldsymbol{\theta} \sim Q} [\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})] + \frac{1}{\beta} \left[\text{KL}(Q \| P) + \log \frac{1}{\delta} + \Psi(\beta, N) \right],$$

where we have defined:

$$\Psi(\beta, N) = \log \mathbb{E}_P \mathbb{E}_{\mathcal{S}} \left[\exp \left\{ \beta (\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})) \right\} \right]$$

Note that the loss function is bounded by L , according to Hoeffding’s lemma, we have:

$$\Psi(\beta, N) \leq \frac{\beta^2 L^2}{8N}.$$

By Cauchy inequality:

$$\frac{1}{\sqrt{N}} \left[\frac{T}{2} \log \left(1 + \frac{\|\boldsymbol{\theta}\|^2}{T\sigma^2} \right) + \frac{L^2}{8} \right] \geq \frac{L}{2\sqrt{N}} \sqrt{T \log \left(1 + \frac{\|\boldsymbol{\theta}\|^2}{T\sigma^2} \right)} \geq L,$$

which means that the theorem is proved since the loss function is upper bounded by L , following assumptions.

Now, we only need to prove the theorem under the case: $\|\boldsymbol{\theta}\|^2 \leq T\sigma^2 \left[\exp \frac{4N}{T} - 1 \right]$.

We need to specify P in advance since it is a prior distribution. However, we do not know in advance the value of $\boldsymbol{\theta}$ that affects the KL divergence term. Hence, we build a family of distribution P as follows:

$$\mathfrak{P} = \left\{ P_j = \mathcal{N}(\mathbf{0}, \sigma_{P_j}^2 \mathbb{I}_T) : \sigma_{P_j}^2 = c \exp \left(\frac{1-j}{T} \right), c = \sigma^2 \left(1 + \exp \frac{4N}{T} \right), j = 1, 2, \dots \right\}.$$

Set $\delta_j = \frac{6\delta}{\pi^2 j^2}$, the below inequality holds with probability at least $1 - \delta_j$:

$$\mathbb{E}_{\boldsymbol{\theta} \sim Q} [\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})] \leq \mathbb{E}_{\boldsymbol{\theta} \sim Q} [\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})] + \frac{1}{\beta} \left[\text{KL}(Q \| P_j) + \log \frac{1}{\delta_j} + \frac{\beta^2 L^2}{8N} \right].$$

Or it can be written as:

$$\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta} + \boldsymbol{\epsilon})] \leq \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta} + \boldsymbol{\epsilon})] + \frac{1}{\beta} \left[\text{KL}(Q \| P_j) + \log \frac{1}{\delta_j} + \frac{\beta^2 L^2}{8N} \right].$$

Thus, with probability $1 - \delta$ the above inequalities hold for all P_j . We choose:

$$j^* = \left\lfloor 1 + T \log \left(\frac{\sigma^2 (1 + \exp\{4N/T\})}{\sigma^2 + \|\boldsymbol{\theta}\|^2/T} \right) \right\rfloor.$$

Since $\frac{\|\boldsymbol{\theta}\|^2}{T} \leq \sigma^2 [\exp \frac{4N}{T} - 1]$, we get $\sigma^2 + \frac{\|\boldsymbol{\theta}\|^2}{T} \leq \sigma^2 \exp \frac{4N}{T}$, thus j^* is well-defined. We also have:

$$\begin{aligned} T \log \frac{c}{\sigma^2 + \|\boldsymbol{\theta}\|^2/T} &\leq j^* \leq 1 + T \log \frac{c}{\sigma^2 + \|\boldsymbol{\theta}\|^2/T} \\ \Rightarrow \log \frac{c}{\sigma^2 + \|\boldsymbol{\theta}\|^2/T} &\leq \frac{j^*}{T} \leq \frac{1}{T} + \log \frac{c}{\sigma^2 + \|\boldsymbol{\theta}\|^2/T} \\ \Rightarrow -\frac{1}{T} + \log \frac{\sigma^2 + \|\boldsymbol{\theta}\|^2/T}{c} &\leq \frac{-j^*}{T} \leq \log \frac{\sigma^2 + \|\boldsymbol{\theta}\|^2/T}{c} \\ \Rightarrow e^{-1/T} \frac{\sigma^2 + \|\boldsymbol{\theta}\|^2/T}{c} &\leq e^{-j^*/T} \leq \frac{\sigma^2 + \|\boldsymbol{\theta}\|^2/T}{c} \\ \Rightarrow \sigma^2 + \frac{\|\boldsymbol{\theta}\|^2}{T} &\leq c e^{\frac{1-j^*}{T}} \leq e^{\frac{1}{T}} \left(\sigma^2 + \frac{\|\boldsymbol{\theta}\|^2}{T} \right) \\ \Rightarrow \sigma^2 + \frac{\|\boldsymbol{\theta}\|^2}{T} &\leq \sigma_{P_{j^*}}^2 \leq e^{\frac{1}{T}} \left(\sigma^2 + \frac{\|\boldsymbol{\theta}\|^2}{T} \right). \end{aligned}$$

Hence, we have:

$$\begin{aligned} \text{KL}(Q \| P_{j^*}) &= \frac{1}{2} \left[\frac{T\sigma^2 + \|\boldsymbol{\theta}\|^2}{\sigma_{P_{j^*}}^2} - T + T \log \frac{\sigma_{P_{j^*}}^2}{\sigma^2} \right] \\ &\leq \frac{1}{2} \left[\frac{T\sigma^2 + \|\boldsymbol{\theta}\|^2}{\sigma^2 + \|\boldsymbol{\theta}\|^2/T} - T + T \log \frac{e^{1/T} (\sigma^2 + \|\boldsymbol{\theta}\|^2/T)}{\sigma^2} \right] \\ &\leq \frac{1}{2} \left[1 + T \log \left(1 + \frac{\|\boldsymbol{\theta}\|^2}{T\sigma^2} \right) \right]. \end{aligned}$$

For the term $\log \frac{1}{\delta_{j^*}}$, use the inequality $\log(1 + e^t) \leq 1 + t$ for $t > 0$:

$$\begin{aligned}
\log \frac{1}{\delta_{j^*}} &= \log \frac{(j^*)^2 \pi^2}{6\delta} = \log \frac{1}{\delta} + \log \left(\frac{\pi^2}{6} \right) + 2 \log(j^*) \\
&\leq \log \frac{1}{\delta} + \log \frac{\pi^2}{6} + 2 \log \left(1 + T \log \frac{\sigma^2(1 + \exp(4N/T))}{\sigma^2 + \|\boldsymbol{\theta}\|^2/T} \right) \\
&\leq \log \frac{1}{\delta} + \log \frac{\pi^2}{6} + 2 \log \left(1 + T \log(1 + \exp(4N/T)) \right) \\
&\leq \log \frac{1}{\delta} + \log \frac{\pi^2}{6} + 2 \log \left(1 + T \left(1 + \frac{4N}{T} \right) \right) \\
&\leq \log \frac{1}{\delta} + \log \frac{\pi^2}{6} + \log(1 + T + 4N).
\end{aligned}$$

Choosing $\beta = \sqrt{N}$, with probability at least $1 - \delta$ we get:

$$\begin{aligned}
&\frac{1}{\beta} \left[\text{KL}(Q \| P_{j^*}) + \log \frac{1}{\delta_{j^*}} + \frac{\beta^2 L^2}{8N} \right] \\
&\leq \frac{1}{\sqrt{N}} \left[\frac{1}{2} + \frac{T}{2} \log \left(1 + \frac{\|\boldsymbol{\theta}\|^2}{T\sigma^2} \right) + \log \frac{1}{\delta} + 6 \log(N + T) \right] + \frac{L^2}{8\sqrt{N}}.
\end{aligned}$$

Thus the theorem is proved. \square

Back to our context of multi-task learning in which we have m tasks with each task model: $\boldsymbol{\theta}^i = [\boldsymbol{\theta}_{sh}^i, \boldsymbol{\theta}_{ns}^i]$, we can prove the following theorem.

Theorem 3. *With the assumption that adding Gaussian perturbation will rise the test error: $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}^i) \leq \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}^i + \epsilon)]$. Let T_i be the number of parameter $\boldsymbol{\theta}^i$ and N be the cardinality of \mathcal{S} . We have the following inequality holds with probability $1 - \delta$ (over the choice of training set $\mathcal{S} \sim \mathcal{D}$):*

$$[\mathcal{L}_{\mathcal{D}}^i(\boldsymbol{\theta}^i)]_{i=1}^m \leq [\mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^i + \epsilon)] + f^i(\|\boldsymbol{\theta}^i\|_2^2)]_{i=1}^m, \quad (7)$$

where

$$f^i(\|\boldsymbol{\theta}^i\|_2^2) = \frac{1}{\sqrt{N}} \left[\frac{1}{2} + \frac{T_i}{2} \log \left(1 + \frac{\|\boldsymbol{\theta}\|^2}{T_i \sigma^2} \right) + \log \frac{1}{\delta} + 6 \log(N + T_i) + \frac{L^2}{8} \right].$$

Proof. The result for the base case $m = 1$ can be achieved by using Theorem 2 where $\xi = \delta$ and f^1 is defined accordingly. We proceed by induction, suppose that Theorem 3 is true for all $i \in [n]$ with probability $1 - \delta/2$, which also means:

$$[\mathcal{L}_{\mathcal{D}}^i(\boldsymbol{\theta}^i)]_{i=1}^n \leq [\mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^i + \epsilon)] + f^i(\|\boldsymbol{\theta}^i\|_2^2)]_{i=1}^n.$$

Using Theorem 2 for $\boldsymbol{\theta}^{n+1}$ and $\xi = \delta/2$, with probability $1 - \delta/2$, we have:

$$\mathcal{L}_{\mathcal{D}}^{n+1}(\boldsymbol{\theta}^{n+1}) \leq \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^{n+1} + \epsilon)] + f^{n+1}(\|\boldsymbol{\theta}^{n+1}\|_2^2).$$

Using the inclusion-exclusion principle, with probability at least $1 - \delta$, we reach the conclusion for $m = n + 1$.

We next prove the result in the main paper. Let us begin by formally restating the main theorem as follows:

Theorem 4. For any perturbation radius $\rho_{sh}, \rho_{ns} > 0$, with probability $1 - \delta$ (over the choice of training set $\mathcal{S} \sim \mathcal{D}$) we obtain:

$$[\mathcal{L}_{\mathcal{D}}^i(\theta^i)]_{i=1}^m \leq \max_{\|\epsilon_{sh}\|_2 \leq \rho_{sh}} \left[\max_{\|\epsilon_{ns}^i\|_2 \leq \rho_{ns}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i) + f^i(\|\theta^i\|_2^2) \right]_{i=1}^m, \quad (8)$$

where $f^i(\|\theta^i\|_2^2)$ is defined the same as in Theorem 3.

Proof. Theorem 3 gives us

$$\begin{aligned} [\mathcal{L}_{\mathcal{D}}^i(\theta^i)]_{i=1}^m &\leq \left[\mathbb{E}_{\epsilon \sim N(0, \sigma^2 \mathbb{I})} [\mathcal{L}_{\mathcal{S}}^i(\theta^i + \epsilon)] + f^i(\|\theta^i\|_2^2) \right]_{i=1}^m \\ &= \left[\int \mathbb{E}_{\epsilon_{ns}^i} [\mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i)] p(\epsilon_{sh}) d\epsilon_{sh} + f^i(\|\theta^i\|_2^2) \right]_{i=1}^m \\ &= \mathbb{E}_{\epsilon_{sh}} \left[\mathbb{E}_{\epsilon_{ns}^i} [\mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i)] + f^i(\|\theta^i\|_2^2) \right]_{i=1}^m, \end{aligned}$$

where $p(\epsilon_{sh})$ is the density function of Gaussian distribution; ϵ_{sh} and ϵ_{ns}^i are drawn from their corresponding Gaussian distributions.

We have $\epsilon_{ns}^i \sim N(0, \sigma^2 \mathbb{I}_{ns})$ with the dimension $T_{i,ns}$, therefore $\|\epsilon_{ns}^i\|_2^2$ follows the Chi-square distribution. As proven in [70], we have for all i :

$$P\left(\|\epsilon_{ns}^i\|_2^2 \geq T_{i,ns}\sigma^2 + 2\sigma^2\sqrt{T_{i,ns}t} + 2t\sigma^2\right) \leq e^{-t}, \forall t > 0$$

$$P\left(\|\epsilon_{ns}^i\|_2^2 < T_{i,ns}\sigma^2 + 2\sigma^2\sqrt{T_{i,ns}t} + 2t\sigma^2\right) > 1 - e^{-t}$$

for all $t > 0$.

Select $t = \ln(\sqrt{N})$, we derive the following bound for the noise magnitude in terms of the perturbation radius ρ_{ns} for all i :

$$P\left(\|\epsilon_{ns}^i\|_2^2 \leq \sigma^2(2\ln(\sqrt{N}) + T_{i,ns} + 2\sqrt{T_{i,ns}\ln(\sqrt{N})})\right) > 1 - \frac{1}{\sqrt{N}}. \quad (9)$$

Moreover, we have $\epsilon_{sh} \sim N(0, \sigma^2 \mathbb{I}_{sh})$ with the dimension T_{sh} , therefore $\|\epsilon_{sh}\|_2^2$ follows the Chi-square distribution. As proven in [70], we have:

$$P\left(\|\epsilon_{sh}\|_2^2 \geq T_{sh}\sigma^2 + 2\sigma^2\sqrt{T_{sh}t} + 2t\sigma^2\right) \leq e^{-t}, \forall t > 0$$

$$P\left(\|\epsilon_{sh}\|_2^2 < T_{sh}\sigma^2 + 2\sigma^2\sqrt{T_{sh}t} + 2t\sigma^2\right) > 1 - e^{-t}$$

for all $t > 0$.

Select $t = \ln(\sqrt{N})$, we derive the following bound for the noise magnitude in terms of the perturbation radius ρ_{sh} :

$$P\left(\|\epsilon_{sh}\|_2^2 \leq \sigma^2(2\ln(\sqrt{N}) + T_{sh} + 2\sqrt{T_{sh}\ln(\sqrt{N})})\right) > 1 - \frac{1}{\sqrt{N}}. \quad (10)$$

By choosing σ less than $\frac{\rho_{sh}}{\sqrt{2 \ln N^{1/2} + T_{sh} + 2\sqrt{T_{sh} \ln N^{1/2}}}}$ and $\min_i \frac{\rho_{ns}}{\sqrt{2 \ln N^{1/2} + T_{i,ns} + 2\sqrt{T_{i,ns} \ln N^{1/2}}}}$, and referring to (9,10), we achieve both:

$$P(\|\epsilon_{ns}^i\| < \rho_{ns}) > 1 - \frac{1}{N^{1/2}}, \forall i,$$

$$P(\|\epsilon_{sh}\| < \rho_{sh}) > 1 - \frac{1}{N^{1/2}}.$$

Finally, we finish the proof as:

$$\begin{aligned} [\mathcal{L}_{\mathcal{D}}^i(\theta^i)]_{i=1}^m &\leq \mathbb{E}_{\epsilon_{sh}} [\mathbb{E}_{\epsilon_{ns}^i} [\mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i)] + f^i(\|\theta^i\|_2)]_{i=1}^m \\ &\leq \max_{\|\epsilon_{sh}\| < \rho_{sh}} \left[\max_{\|\epsilon_{ns}^i\| < \rho_{ns}} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i) + \frac{2}{\sqrt{N}} - \frac{1}{N} + f^i(\|\theta^i\|_2) \right]_{i=1}^m \end{aligned}$$

To reach the final conclusion, we redefine:

$$f^i(\|\theta^i\|_2) = \frac{2}{\sqrt{N}} - \frac{1}{N} + f^i(\|\theta^i\|_2).$$

Here we note that we reach the final inequality due to the following derivations:

$$\begin{aligned} \mathbb{E}_{\epsilon_{sh}} [\mathbb{E}_{\epsilon_{ns}^i} [\mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i)]]_{i=1}^m &\leq \int_{B_{sh}} \left[\int_{B_{ns}^i} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i) d\epsilon_{ns}^i + \frac{1}{\sqrt{N}} \right]_{i=1}^m d\epsilon_{sh} \\ &\quad + \int_{B_{sh}^c} \left[\int_{B_{ns}^i} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i) d\epsilon_{ns}^i + \frac{1}{\sqrt{N}} \right]_{i=1}^m d\epsilon_{sh} \\ &\leq \int_{B_{sh}} \left[\int_{B_{ns}^i} \mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i) d\epsilon_{ns}^i \right]_{i=1}^m d\epsilon_{sh} + \left(1 - \frac{1}{\sqrt{N}}\right) \frac{1}{\sqrt{N}} + \frac{1}{\sqrt{N}} \\ &\leq \max_{\|\epsilon_{sh}\| < \rho_{sh}} \left[\max_{\|\epsilon_{ns}^i\| < \rho_{ns}} [\mathcal{L}_{\mathcal{S}}^i(\theta_{sh} + \epsilon_{sh}, \theta_{ns}^i + \epsilon_{ns}^i)] \right]_{i=1}^m + \frac{2}{\sqrt{N}} - \frac{1}{N}, \end{aligned}$$

where $B_{sh} = \{\epsilon_{sh} : \|\epsilon_{sh}\| \leq \rho_{sh}\}$, B_{sh}^c is the compliment set, and $B_{ns}^i = \{\epsilon_{ns}^i : \|\epsilon_{ns}^i\| \leq \rho_{ns}\}$.

B Gradient aggregation strategies overview

This section details how the `gradient_aggregate` operation is defined according to recent gradient-based multi-task learning methods that we employed as baselines in the main paper, including MGDA [8], PCGrad [9], CAGrad [10] and IMTL [11]. Assume that we are given m vectors $\mathbf{g}^1, \mathbf{g}^2, \dots, \mathbf{g}^m$ represent task gradients. Typically, we aim to find a combined gradient vector as:

$$\mathbf{g} = \text{gradient_aggregate}(\mathbf{g}^1, \mathbf{g}^2, \dots, \mathbf{g}^m)$$

B.1 Multiple-gradient descent algorithm - MGDA

[8] apply MGDA [71] to find the minimum-norm gradient vector that lies in the convex hull composed by task gradients $\mathbf{g}^1, \mathbf{g}^2, \dots, \mathbf{g}^m$:

$$\mathbf{g} = \operatorname{argmin} \left\| \sum_{i=1}^m w_i \mathbf{g}^i \right\|^2, \text{ s.t. } \sum_{i=1}^m w_i = 1 \quad \text{and} \quad w_i \geq 0 \forall i.$$

This approach can guarantee that the obtained solutions lie on the Pareto front of task objective functions.

B.2 Projecting conflicting gradients - PCGrad

PCgrad resolves the disagreement between tasks by projecting gradients that conflict with each other, i.e. $\langle \mathbf{g}^i, \mathbf{g}^j \rangle < 0$, to the orthogonal direction of each other. Specifically, \mathbf{g}^i is replaced by its projection on the normal plane of \mathbf{g}^j :

$$\mathbf{g}_{\text{PC}}^i = \mathbf{g}^i - \frac{\mathbf{g}^i \cdot \mathbf{g}^j}{\|\mathbf{g}^j\|^2} \mathbf{g}^j.$$

Then compute the aggregated gradient based on these deconflict vectors $\mathbf{g} = \sum_i^m \mathbf{g}_{\text{PC}}^i$.

B.3 Conflict Averse Gradient Descent - CAGrad

CAGrad [10] seeks a worst-case direction in a local ball around the average gradient of all tasks, \mathbf{g}^0 , that minimizes conflict with all of the gradients. The updated vector is obtained by optimizing the following problem:

$$\max_{\mathbf{g} \in R} \min_{i \in [m]} \langle \mathbf{g}^i, \mathbf{g} \rangle \quad \text{s.t.} \quad \|\mathbf{g} - \mathbf{g}^0\| \leq c \|\mathbf{g}^0\|,$$

where $\mathbf{g}^0 = \frac{1}{m} \sum_i^m \mathbf{g}^i$ is the averaged gradient and c is a hyper-parameter.

B.4 Impartial multi-task learning - IMTL

IMTL [11] proposes to balance per-task gradients by finding the combined vector \mathbf{g} , whose projections onto $\{\mathbf{g}^i\}_{i=1}^m$ are equal. Following this, they obtain the closed-form solution for the simplex vector \mathbf{w} for reweighting task gradients:

$$\mathbf{w} = \mathbf{g}^1 \mathbf{U}^\top (\mathbf{D} \mathbf{U}^\top)^{-1}$$

where $\mathbf{u}^i = \mathbf{g}^i / \|\mathbf{g}^i\|$, $\mathbf{U} = [\mathbf{u}^1 - \mathbf{u}^2, \dots, \mathbf{u}^1 - \mathbf{u}^m]$, and $\mathbf{D} = [\mathbf{g}^1 - \mathbf{g}^2, \dots, \mathbf{g}^1 - \mathbf{g}^m]$. The aggregated vector is then calculated as $\mathbf{g} = \sum_i^m w_i \mathbf{g}^i$.

C Implementation Details

In this part, we provide implementation details regarding the empirical evaluation in the main paper along with additional comparison experiments.

C.1 Baselines

In this subsection, we briefly introduce some of the comparative methods that appeared in the main text:

- Linear scalarization (LS) minimizes the unweighted sum of task objectives $\sum_i^m \mathcal{L}^i(\boldsymbol{\theta})$.
- Scale-invariant (SI) aims toward obtaining similar convergent solutions even if losses are scaled with different coefficients via minimizing $\sum_i^m \log \mathcal{L}^i(\boldsymbol{\theta})$.
- Random loss weighting (RLW) [72] is a simple yet effective method for balancing task losses or gradients by random weights.
- Dynamic Weight Average (DWA) [2] simply adjusts the weighting coefficients by taking the rate of change of loss for each task into account.
- GradDrop [65] presents a probabilistic masking process that algorithmically eliminates all gradient values having the opposite sign w.r.t a predefined direction.

C.2 Image classification

Network Architectures. For two datasets in this problem, Multi-MNIST and CelebA, we replicate experiments from [8, 25] by respectively using the Resnet18 (11M parameters) and Resnet50 (23M parameters) [61] with the last output layer removed as the shared encoders and constructing linear classifiers as the task-specific heads, i.e. 2 heads for Multi-MNIST and 40 for CelebA, respectively.

Training Details. We train the all the models under our proposed framework and baselines using:

- Multi-MNIST: Adam optimizer [73] with a learning rate of 0.001 for 200 epochs using a batch size of 256. Images from the three datasets are resized to 36×36 .
- CelebA: Batch-size of 256 and images are resized to $64 \times 64 \times 3$. Adam [73] is used again with a learning rate of 0.0005, which is decayed by 0.85 for every 10 epochs, our model is trained for 50 epochs in total.

Regarding the hyperparameter for SAM [21], we use their adaptive version [74] where both ρ_{sh} and ρ_{ns} are set equally and extensively tuned from 0.005 to 5.

C.3 Scene understanding

Two datasets used in this problem are NYUv2 and CityScapes. Similar to [13], all images in the NYUv2 dataset are resized to 288×384 while all images in the CityScapes dataset are resized to 128×256 to speed up the training process. We follow the exact protocol in [13] for implementation. Specifically, SegNet [66] is adopted as the architecture for the backbone and Multi-Task Attention Network MTAN [2] is applied on top of it. We train each method for 200 epochs using Adam optimizer [73] with an initial learning rate of $1e-4$ and reduced it to $5e-5$ after 100 epochs. We use a batch size of 2 for NYUv2 and 8 for CityScapes. The last 10 epochs are averaged to get the final results, and all experiments are run with three random seeds.

D Additional Results

To further show the improvement of our proposed training framework over the conventional one, this section provides additional comparison results in terms of qualitative results, predictive performance, convergent behavior, loss landscape, model sharpness, and gradient norm. Please note that in these experiments, we choose IMTL and F-IMTL as two examples for *standard* and *flat-aware* gradient-based MTL training respectively. We also complete the ablation study in the main paper by providing results on all three datasets in the Multi-MNIST dataset.

D.1 Image segmentation qualitative result

In this section, we provide qualitative results of our method of the CityScapes experiment. We compare our proposed method against its main baseline by highlighting typical cases where our method excels in generalization performance. Figure 6 shows some visual examples of segmentation outputs on the test set. Note that in the CityScapes dataset, the “void” class is identified as unclear and pixels labeled as void do not contribute to either objective or score [60].

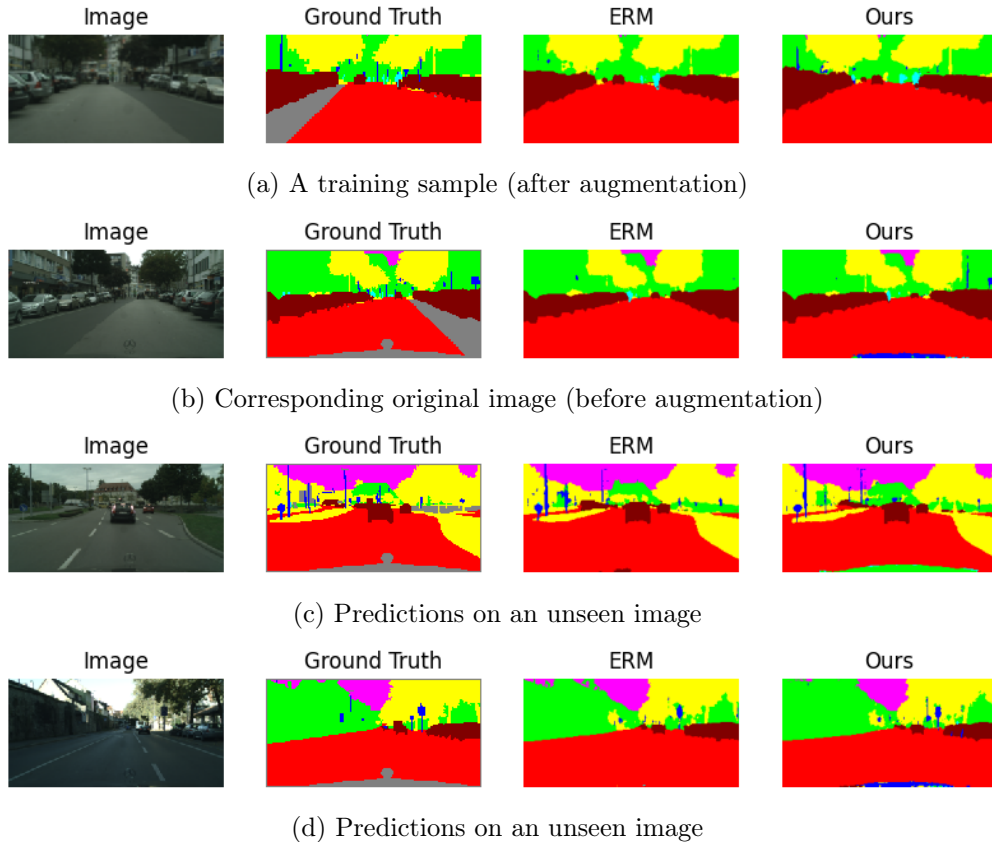


Figure 6: Semantic segmentation prediction comparison on CityScapes . From left to right are input images, ground truth, and segmentation outputs from SegNet [66] using ERM training and sharpness-aware training. Regions that are represented in gray color are ignored during training. (Best viewed in color).

While there is only a small gap between the segmentation performance of IMTL and F-IMTL, we found that a small area, which is the car hood and located at the bottom of images, is often incorrectly classified. For example, in Figure 6, the third and fourth rows compare the prediction of SegNet [66] with ERM training and with our proposed method. It can be seen that both of them could not detect this area correctly, this is because this unclear "void" class did not appear during training. Even worse, the currently employed data augmentation technique in the codebase of Nash-MTL and other recent multi-task learning methods [13, 10] consists of RandomCrop, which often unintentionally excludes edge regions. For example, Figure 6a shows an example fed to the neural network for training, which excludes the car hood and its logo, compared to the original image (Figure 6b). Therefore, we can consider this "void" class as a novel class in this experiment, since its appearance is ignored in both training and evaluation. Even though, in Figures 6c and 6d our training method is still able to distinguish between this unknown area and other nearby known classes, which empirically shows the robustness and generalization ability of our method over ERM.

D.2 Predictive performance

In this part, we provide experimental justification for an intriguing insight into the connection between model sharpness and model calibration. Empirically, we found that when a model converges to flatter minima, it tends to be more calibrated. We start by giving the formal definition of a well-calibrated classification model and three metrics to measure the calibration of a model, then we analyze our empirical results.

Consider a C -class classification problem with a test set of N samples given in the form $(x_i, y_i)_{i=1}^N$ where y_i is the true label for the sample x_i . Model outputs the predicted probability for a given sample x_i to fall into C classes, is given by

$$\hat{\mathbf{p}}(x_i) = [\hat{p}(y = 1|x_i), \dots, \hat{p}(y = C|x_i)].$$

$\hat{p}(y = c|x_i)$ is also the confidence of the model when assigning the sample x_i to class c . The predicted label \hat{y}_i is the class with the highest predicted value, $\hat{p}(x_i) := \max_c \hat{p}(y = c|x_i)$. We refer to $\hat{p}(x_i)$ as the confidence score of a sample x_i .

Model calibration is a desideratum of modern deep neural networks, which indicates that the predicted probability of a model should match its true probability. This means that the classification network should be not only accurate but also confident about its prediction, i.e. being aware of when it is likely to be incorrect. Formally stated, the *perfect calibration* [75] is:

$$P(\hat{y} = y | \hat{p} = q) = q, \forall q \in [0, 1]. \quad (11)$$

Metric. The exact computation of Equation 11 is infeasible, thus we need to define some metrics to evaluate how well-calibrated a model is.

- Brier score \downarrow (BS) [76] assesses the accuracy of a model's predicted probability by taking into account the absolute difference between its confidence for a sample to fall into a class and the true label of that sample. Formally,

$$BS = \frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C (\hat{p}(y = c|x_i) - \mathbf{1}[y_i = c])^2.$$

- Expected calibration error \downarrow (ECE) compares the predicted probability (or confidence) of a model to its accuracy [77, 75]. To compute this error, we first bin the confidence interval $[0, 1]$ into M equal bins, then categorize data samples into these bins according to their confidence scores. We finally compute the absolute value of the difference between the average confidence and the average accuracy within each bin, and report the average value over all bins as the ECE. Specifically, let B_m denote the set of indices of samples having their confidence scores belonging to the m^{th} bin. The average accuracy and the average confidence within this bin are:

$$\begin{aligned} acc(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}[\hat{y}_i = y_i], \\ conf(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}(x_i). \end{aligned}$$

Then the ECE of the model is defined as:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{N} |acc(B_m) - conf(B_m)|.$$

In short, the lower ECE neural networks obtain, the more calibrated they are.

- Predictive entropy (PE) is a widely-used measure of uncertainty [78, 79, 80] via the predictive probability of the model output. When encountering an unseen sample, a well-calibrated model is expected to yield a high PE, representing its uncertainty in predicting out-of-domain (OOD) data.

$$PE = \frac{1}{C} \sum_{c=1}^C -\hat{p}(y = c|x_i) \log \hat{p}(y = c|x_i).$$

Figures 7 and 8 plot the distribution of the model’s predicted entropy in the case of in-domain and out-domain testing, respectively. We can see when considering the flatness of minima, the model shows higher predictive entropy on both in-domain and out-of-domain, compared to ERM. This also means that our model outputs high uncertainty prediction when it is exposed to a sample from a different domain.

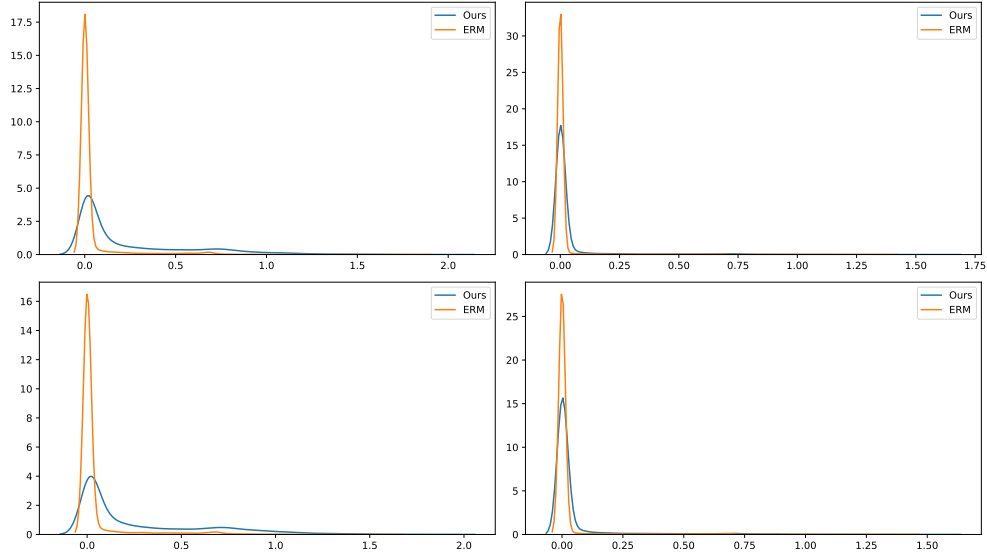


Figure 7: Histograms of predictive entropy of ResNet18 [61] on in domain dataset, train and test on MultiMNIST (left) and MultiFashion (right). We use the orange lines to denote ERM training while blue lines indicate our proposed method.

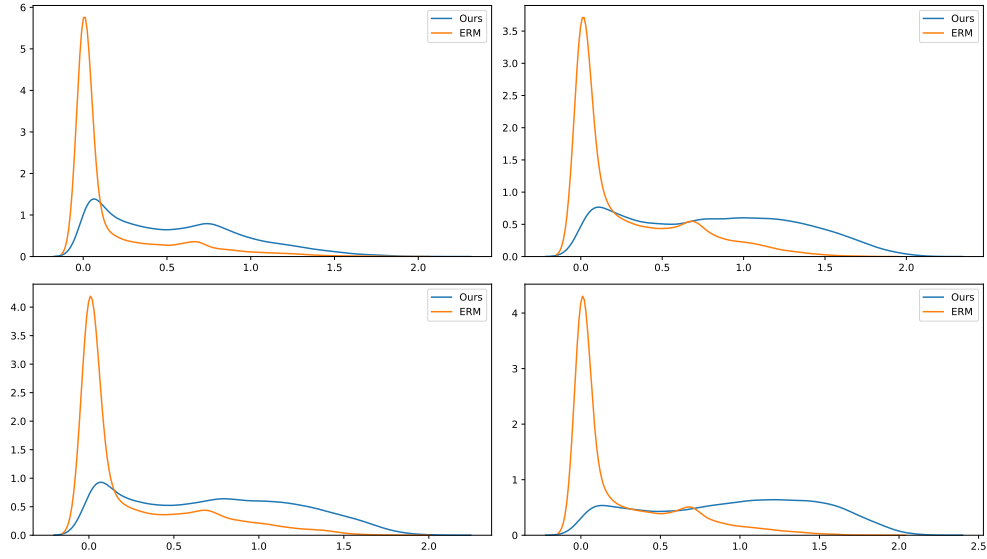


Figure 8: Out of domain: model is trained on MultiMNIST, then tested on MultiFashion (left) and vice versa (right). Models trained with ERM give over-confident predictions as their predictive entropy concentrates around 0.

Here, we calculate the results for both tasks 1 and 2 as a whole and plot their ECE in Figure 9. When we look at the in-domain prediction in more detail, our model still outperforms ERM in terms of expected calibration error. We hypothesize that considering flat minima optimizer not only lowers errors across tasks but also improves the predictive performance of the model.

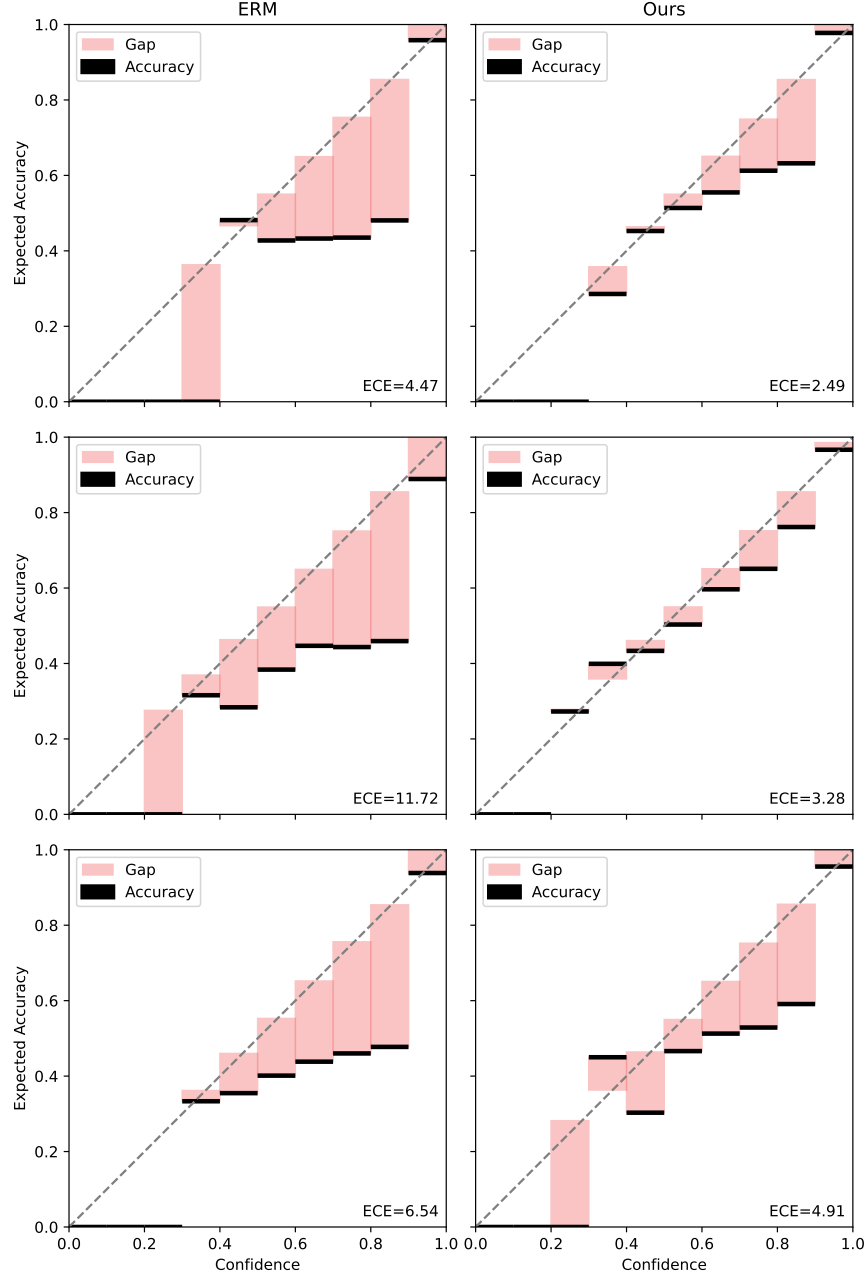


Figure 9: The predictive performance (measured by the expected calibration error) of neural networks has been enhanced by using our proposed training method (right column).

We also report the Brier score and ECE for each task in Table 6 and Table 7. As can be observed

Table 6: Brier score on Multi-Fashion, Multi-Fashion+MNIST and MultiMNIST datasets. We use the **bold** font to highlight the best results.

Dataset	Task	Multi-Fashion	Multi-Fashion+MNIST	MultiMNIST
ERM	Top left	0.237	0.055	0.082
	Bottom right	0.254	0.217	0.106
	Average	0.246	0.136	0.094
Ours	Top left	0.172	0.037	0.059
	Bottom right	0.186	0.189	0.075
	Average	0.179	0.113	0.067

from these tables, our method shows consistent improvement in the model calibration when both scores decrease over all scenarios.

Table 7: Expected calibration error on Multi-Fashion, Multi-Fashion+MNIST and MultiMNIST datasets. Here we set the number of bins equal to 10.

Dataset	Task	Multi-Fashion	Multi-Fashion+MNIST	MultiMNIST
ERM	Top left	0.113	0.027	0.039
	Bottom right	0.121	0.104	0.050
	Average	0.117	0.066	0.045
Ours	Top left	0.034	0.015	0.022
	Bottom right	0.032	0.083	0.028
	Average	0.033	0.049	0.025

D.3 Effect of choosing perturbation radius ρ .

The experimental results analyzing the sensitivity of model w.r.t ρ are given in Figure 10. We evenly picked ρ from 0 to 3.0 to run F-CAGrad on three Multi-MNIST datasets.

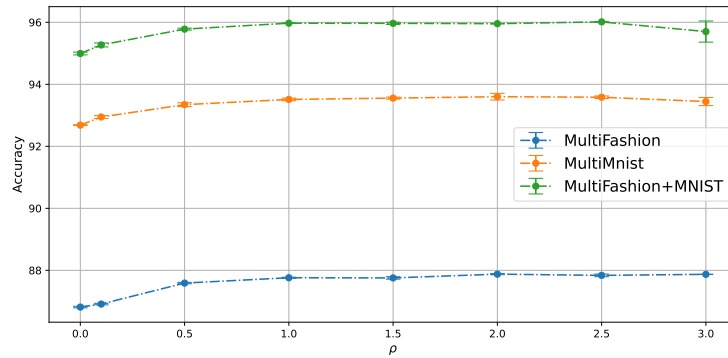


Figure 10: Average accuracy when varying ρ from 0 to 3.0 (with error bar from three independent runs).

We find that the average accuracy of each task is rather stable from $\rho = 0.5$, which means the

effect of different values of ρ in a reasonably small range is similar. It can also be noticed that the improvement tends to saturate when $\rho \geq 1.5$.

D.4 Gradient conflict.

In the main paper, we measure the percentage of gradient conflict on the MultiFashion+MNIST dataset. Here, we provide the full results on three different datasets. As can be seen from Figure 11, there is about half of the mini-batches lead to the conflict between task 1 and task 2 when using traditional training. Conversely, our proposed method significantly reduces such conflict (less than 5%) via updating the parameter toward flat regions.

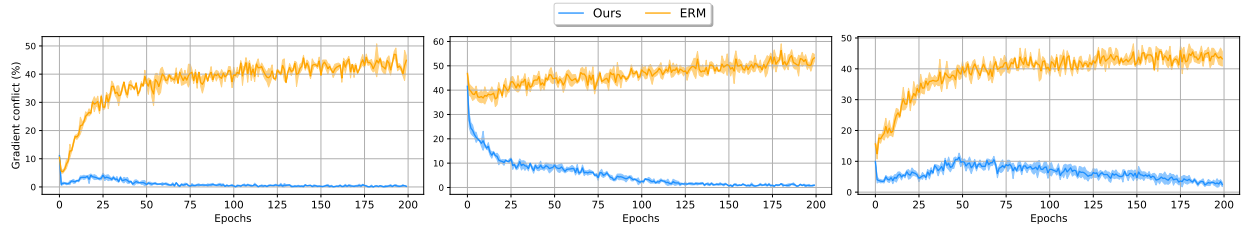


Figure 11: **Task gradient conflict proportion** of models trained with our proposed method and ERM across MultiFashion, MultiFashion+MNIST and MultiMNIST datasets (columns).

D.5 Loss landscape

Thirdly, we provide additional visual comparisons of the loss landscapes trained with standard training and with our framework across two tasks of three datasets of Multi-MNIST. As parts of the obtained visualizations have been presented in the main paper, we provide the rest of them in this subsection. The results in Figures 12 and 13 consistently show that our method obtains significantly flatter minima on both two tasks, encouraging the model to generalize well.

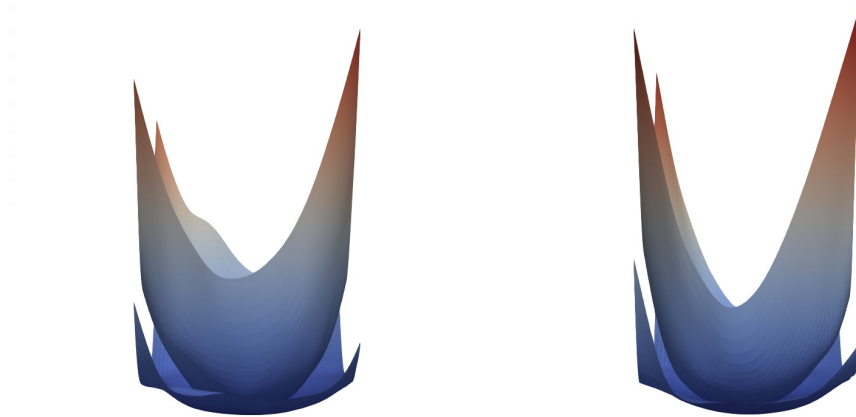


Figure 12: Loss landscapes of task 1 and task 2 on MultiFashion

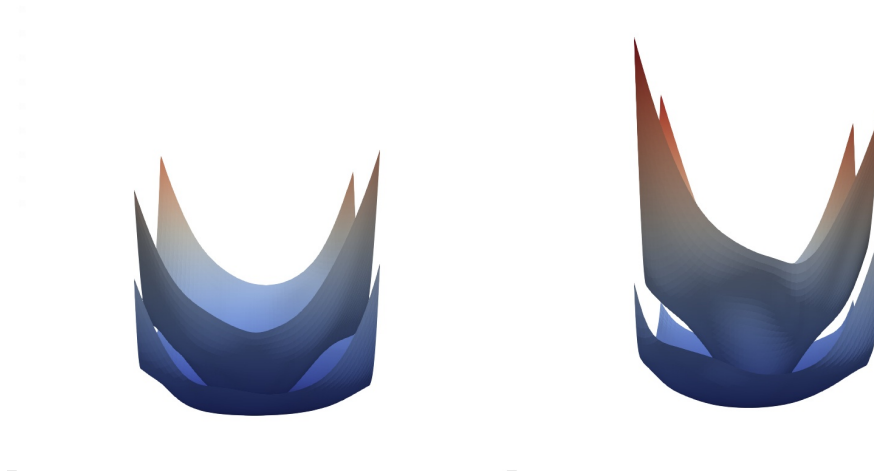


Figure 13: Loss landscapes of task 1 and task 2 on MultiFashion+MNIST

D.6 Training curves

Secondly, we compare the test accuracy of trained models under the two settings in Fig. 15. It can be seen that from the early epochs (20-th epoch), the *flat-based* method outperforms the *ERM-based* method on all tasks and datasets. Although the ERM training model is overfitted after such a long training, our model retains a high generalizability, as discussed throughout previous sections.

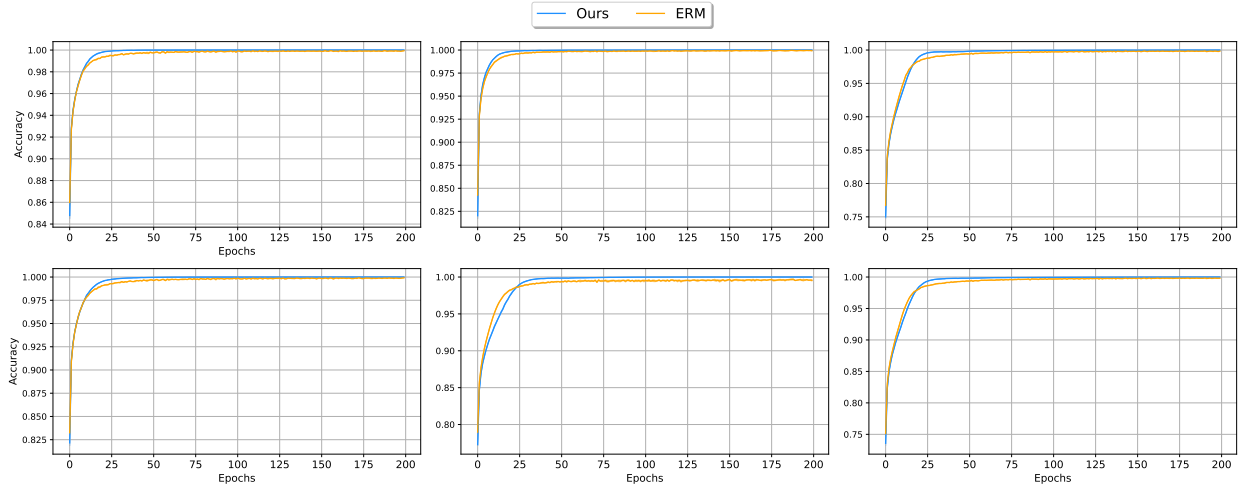


Figure 14: **Train accuracy** of models trained with our proposed method and ERM across 2 tasks (rows) of MultiFashion, MultiFashion+MNIST and MultiMNIST datasets (columns).

Furthermore, we also plot the training accuracy curves across experiments in Figure 14 to show that training accuracy scores of both ERM and our proposed method are similar and reach $\approx 100\%$ from 50-th epoch, which illustrates that the improvement is associated with generalization enhancement, not better training.

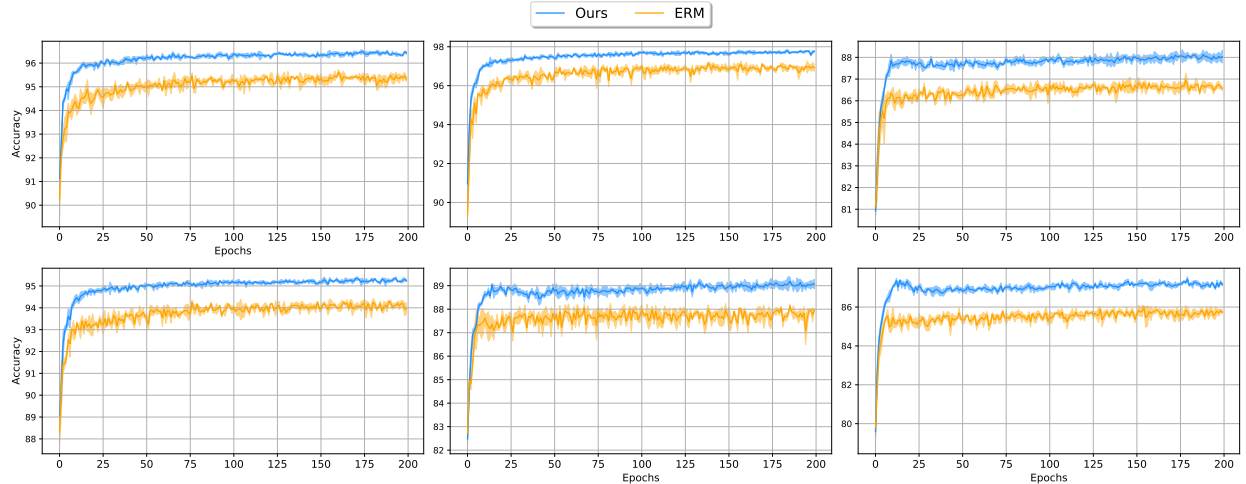


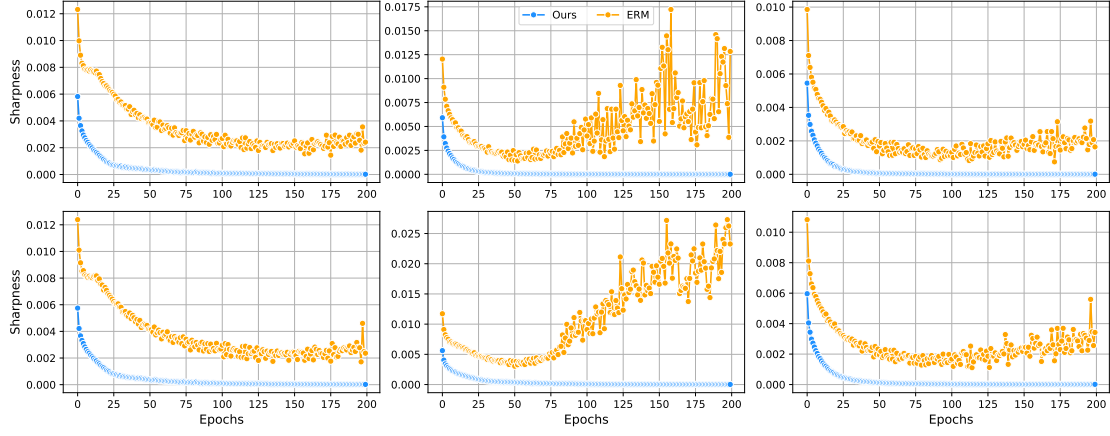
Figure 15: **Test accuracy** of models trained with our proposed method and ERM across 2 tasks (rows) of MultiFashion, MultiFashion+MNIST and MultiMNIST datasets (columns).

D.7 Model sharpness

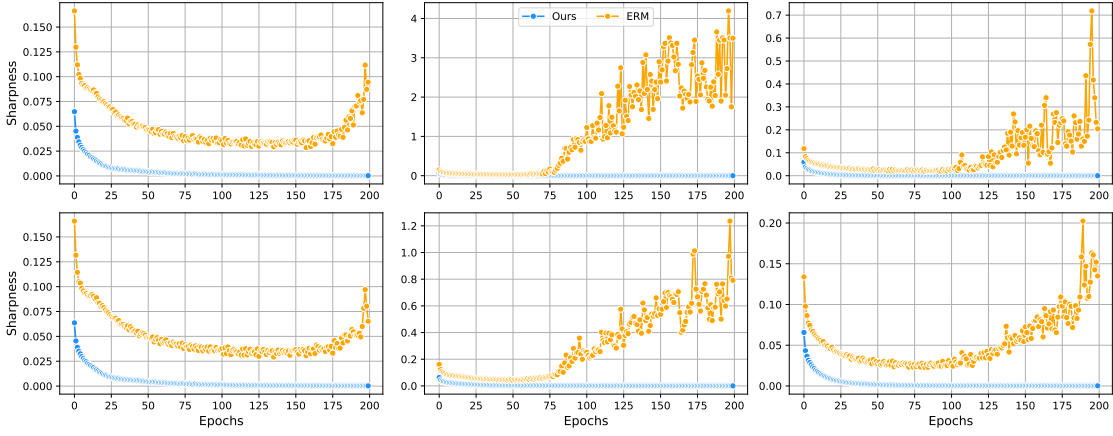
Fourthly, Figure 16 displays the evolution of ρ -sharpness of models along training epochs under conventional loss function (ERM) and worst-case loss function (ours) on training sets of three datasets from Multi-MNIST, with multiple values of ρ . We can clearly see that under our framework, for both tasks, the model can guarantee uniformly low loss value in the ρ -ball neighborhood of parameter across training process. In contrast, ERM suffers from sharp minima from certain epochs when the model witnesses a large gap between the loss of worst-case perturbed model and current model. This is the evidence for the benefit that our framework brings to gradient-based methods, which is all tasks can concurrently find flat minima thus achieving better generalization.

D.8 Gradient norm

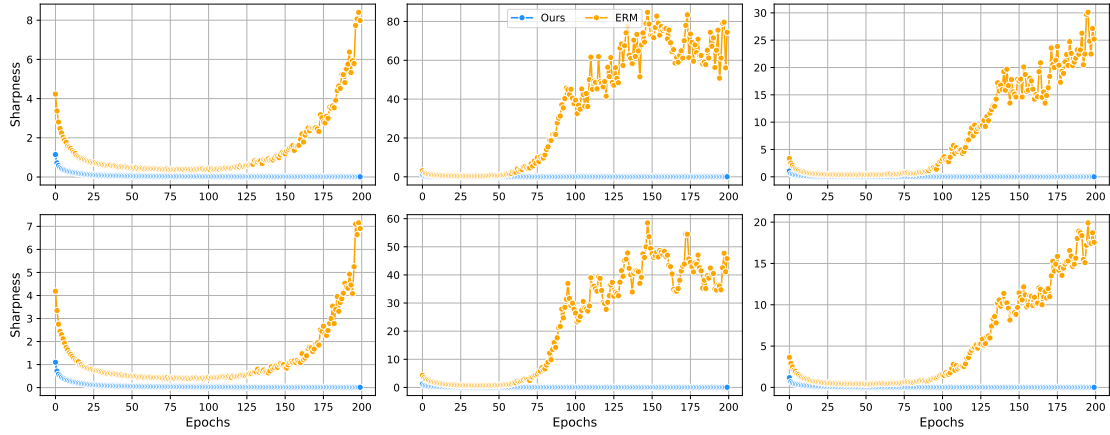
Finally, we demonstrate the gradient norm of the loss function w.r.t the worst-case perturbed parameter of each task. On the implementation side, we calculate the magnitude of the flat gradient $\mathbf{g}^{i,\text{flat}}$ for each task at different values of ρ in Figure 17. As analyzed by equation (6) from the main paper, following the negative direction of $\mathbf{g}_{sh}^{i,SAM}$ will lower the \mathcal{L}_2 norm of the gradient, which orients the model towards flat regions. This is empirically verified in Figure 17. In contrast, as the number of epochs increases, gradnorm of the model trained with ERM tends to increase or fluctuate around a value higher than that of model trained with SAM.



(a) $\rho = 0.005$

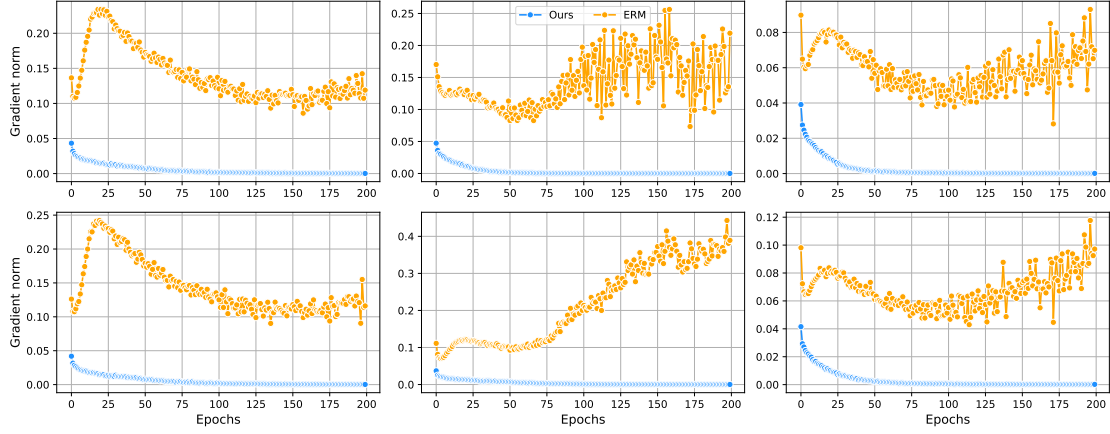


(b) $\rho = 0.05$

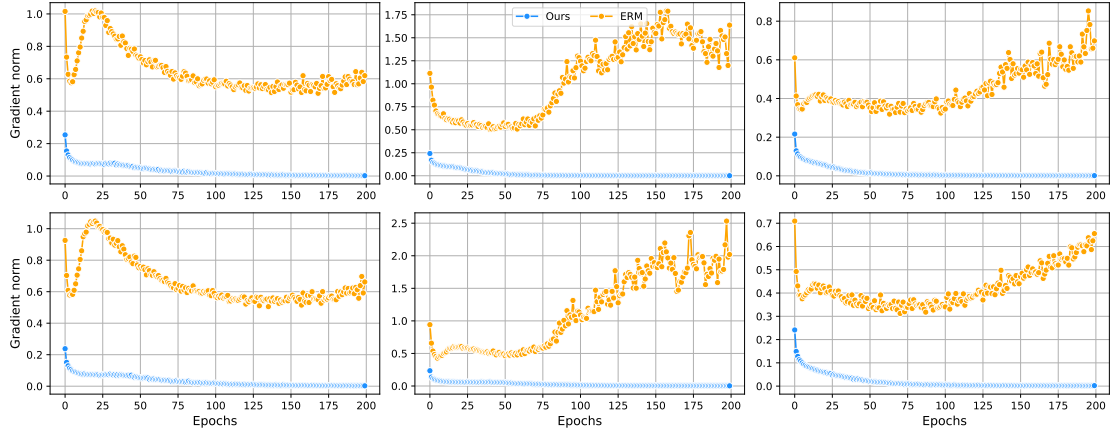


(c) $\rho = 0.5$

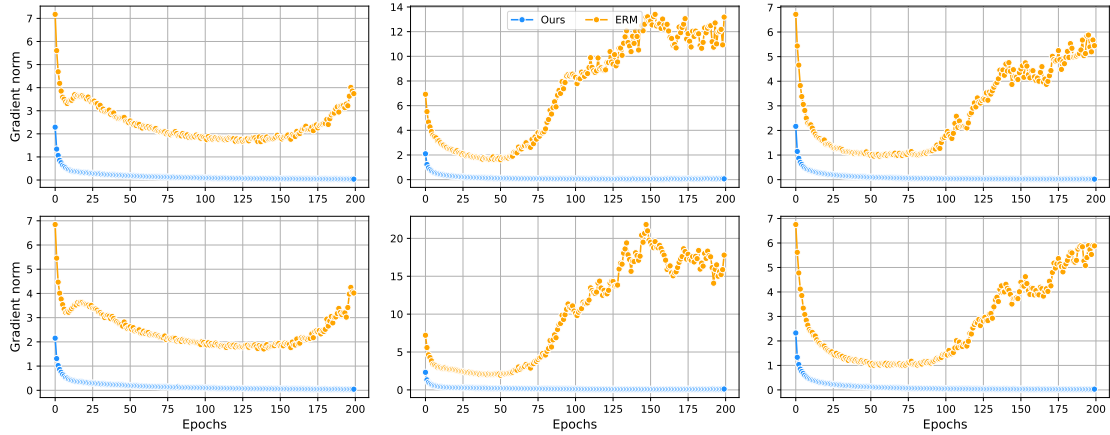
Figure 16: **Sharpness** of models trained with our proposed method and ERM with different values of ρ . For each ρ , the top and bottom row respectively represents the first and second task, and each column respectively represents each dataset in Multi-MNIST: from left to right are MultiFashion, MultiFashion+MNIST, MultiMNIST.



(a) $\rho = 0.005$



(b) $\rho = 0.05$



(c) $\rho = 0.5$

Figure 17: **Gradient magnitude** at the worst-case perturbations of models trained with our proposed method and ERM with different values of ρ . For each ρ , the top and bottom row respectively represents the first and second task, and each column respectively represents each dataset in Multi-MNIST: from left to right are MultiFashion, MultiFashion+MNIST, MultiMNIST.