Injecting Imbalance Sensitivity for Multi-Task Learning

Zhipeng Zhou [1], **Liu Liu** [2,†], **Peilin Zhao** [2] and **Wei Gong** [1,†]
[1]University of Science and Technology of China
[2]Tencent AI Lab
zzp1994@mail.ustc.edu.cn, {leonliuliu, masonzhao}@tencent.com, weigong@ustc.edu.cn,

Abstract

Multi-task learning (MTL) has emerged as a promising approach for deploying deep learning models in real-life applications. Recent studies have proposed optimization-based learning paradigms to establish task-shared representations in MTL. However, our paper empirically argues that these studies, specifically gradient-based ones, primarily emphasize the conflict issue while neglecting the potentially more significant impact of imbalance/dominance in MTL. In line with this perspective, we enhance the existing baseline method by injecting imbalance-sensitivity through the imposition of constraints on the projected norms. To demonstrate the effectiveness of our proposed IMbalance-sensitive Gradient (`IMGrad`) descent method, we evaluate it on multiple mainstream MTL benchmarks, encompassing supervised learning tasks as well as reinforcement learning. The experimental results consistently demonstrate competitive performance.

1 Introduction

Real-life scenarios often involve the need to handle multiple distinct tasks concurrently, typically achieved by designing task-specific models to ensure satisfactory performance. However, this approach becomes impractical as the number of tasks grows, as it would require significant computational resources and memory. To address this challenge and establish an efficient multi-task learning (MTL) framework, recent research has focused on developing a single model capable of performing well on all target tasks.

Currently, research on MTL can be broadly categorized into two frameworks: architecture-based [Liu *et al.*, 2019; Ye and Xu, 2022; Gao *et al.*, 2019; Chen *et al.*, 2023] and optimization-based approaches [Sener and Koltun, 2018; Yu *et al.*, 2020a; Liu *et al.*, 2021a; Zhou *et al.*, ; Liu *et al.*, 2023]. The former emphasizes the design of efficient parameter sharing architectures for multiple tasks, whereas the latter typically employs a fixed architecture and focuses on developing optimization strategies to extract task-shared representations. In this paper, we exclusively introduce and compare our method with optimization-based approaches, as our proposed method falls within this framework.

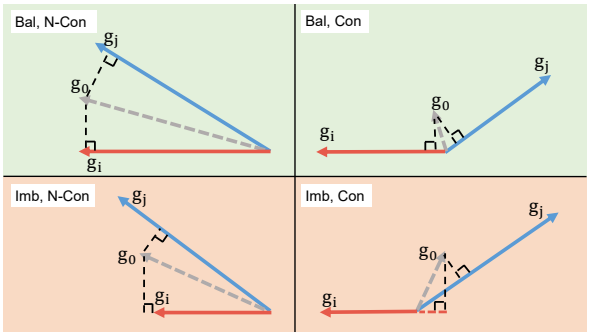


Figure 1: Illustration of imbalance and conflicting issue in multi-task learning. 'Bal' and 'Imb' represent balanced and imbalanced, while 'N-Con' and 'Con' represent non-conflicting and conflicting.

In the realm of optimization-based methods, particularly those involving gradient manipulation, a shared paradigm is commonly followed, where task gradients are combined to achieve Pareto optimality for individual tasks. Despite the high performance demonstrated by these methods, the literature has predominantly overlooked the significance of the inherent imbalance nature among individuals (see **Definition 2**). This oversight can be attributed to the greater emphasis placed on addressing the conflict issue. However, it is important to note that the conflict issue alone may not be the fundamental obstacle hindering optimization in MTL. As illustrated in Figure 1, a naïve linear scalarization (LS) strategy (g_{mean}) effectively improves all individuals when they are balanced, regardless of conflicts. But it proves ineffective when both imbalance and conflict coexist, underscoring the importance of addressing conflicts that arise solely from imbalances. Furthermore, imbalanced task gradients can introduce optimization preferences and lead to imbalanced progress even in the absence of conflicts [Liu *et al.*, 2023]. Although previous solutions, such as IMTL [Liu *et al.*, 2021b] and Nash-MTL [Navon *et al.*, 2022] illustrated in Table 1, have somewhat mitigated the imbalance/dominance issue, they neither explicitly provide evidence to demonstrate the importance of the imbalance issue nor consider both conflict and imbalance issues simultaneously.

[†] Corresponding authors. Work done when Z. Zhou works as an intern in Tencent AI Lab.

	GD	GradDrop	MGDA	PCGrad	IMTL	CAGrad	Nash-MTL	MoCo	IMGrad
Conflict-averse	✗	✗	✓	✓	✗	✓	✗	✓	✓
Imbalance-sensitive	✗	✗	✗	✗	✓	✗	✓	✗	✓

Table 1: Conflict-averse and imbalance-sensitive comparison for mainstream optimization-based MTL. Note that those which are imbalance-sensitive mean that their solution can tackle the imbalance issue.

In this paper, we begin by empirically highlighting the significance of the imbalance issue in MTL and elucidate the advantages of incorporating imbalance sensitivity into baseline methods as our primary motivation. Subsequently, we enhance the well-established baseline method by injecting imbalance sensitivity through the imposition of constraints on the projected norms. Convergence and speedup analysis are provided in the **Appendix** [1]. In a nutshell, we summarize our contributions as three-fold:

- We place significant emphasis on and empirically identify that the primary challenge in optimization-based MTL lies more in the aspect of imbalance rather than conflict. To the best of our knowledge, we are the first to explicitly assert this claim.
- To introduce the imbalance sensitivity into the existing paradigm, we integrate the projected norm constraint into the objectives. This incorporation allows for a dynamic equilibrium between Pareto property (see **Definition 3**) and convergence (two decoupled objectives), thereby enhancing the combined gradients and optimization trajectories.
- The extensive experimental results present compelling evidence that `IMGrad` consistently enhances its baselines and surpasses the current advanced gradient manipulation methods in a diverse range of evaluations, e.g., supervised learning tasks, and reinforcement learning benchmarks.

2 Related Work

Currently, MTL approaches can be broadly categorized into two groups: architecture-based and optimization-based methods. Architecture-based approaches encompass various paradigms, including hard parameter sharing [Heuer *et al.*, 2021; Kokkinos, 2017], soft parameter sharing [Yang and Hospedales, 2016; Gao *et al.*, 2019], modulation and adapters [He *et al.*, 2021; Liu *et al.*, 2022], and mixture of experts (MoE) [Chen *et al.*, 2023; Fan *et al.*, 2022], etc. On the other hand, optimization-based MTL methods primarily focus on learning paradigms rather than structural designs or parameter sharing strategies. These methods aim to optimize all individual tasks to extract task-shared representations.

One classical optimization-based MTL approach is MGDA [Sener and Koltun, 2018], which seeks a combined gradient with minimal norm using the Frank-Wolfe algorithm [Jaggi, 2013]. PCGrad [Yu *et al.*, 2020a] addresses the conflict issue by projecting individual gradients onto orthogonal directions with respect to others. CAGrad [Liu *et al.*, 2021a] considers preserving both the Pareto property and global optimization, ultimately striving for a balance between the two objectives using a hyper-parameter. Nash-MTL [Navon *et al.*, 2022] negotiates to reach an agreement on a joint direction of parameter update, enabling all individual tasks to achieve more balanced progress. MoCo [Fernando *et al.*, 2023] tackles the problem of biased gradient directions in previous solutions by developing tracking parameters for correction. Our method falls within the realm of optimization-based MTL, with a specific focus on addressing the issue of imbalance-sensitivity, which is largely lacking in the aforementioned solutions.

Discussion with Counterparts: To the best of our knowledge, IMTL [Liu *et al.*, 2021b], Nash-MTL [Navon *et al.*, 2022], and FAMO [Liu *et al.*, 2023] are three recent works that explicitly consider the imbalance issue. However, all three works fail to provide evidence demonstrating the importance of the imbalance issue. Moreover, none of these approaches possess conflict-averse properties. Thus, there is still room for improvement. Although Nash-MTL appears to be designed to avoid conflicts, its practical implementation does not achieve this goal. Please refer to the **Appendix** for more discussion.

3 Preliminary

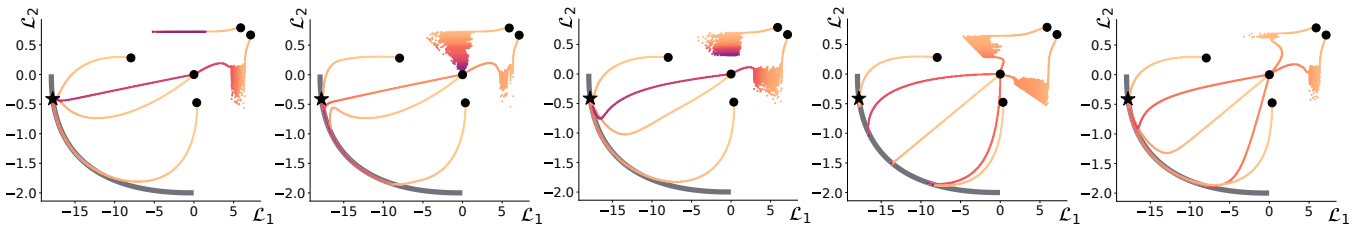
3.1 Setup of Optimization-based MTL

As mentioned, optimization-based MTL approaches operate under the assumption that the model consists of a task-shared backbone network alongside task-specific branches. Consequently, the primary objective of these approaches is to devise gradient combination strategies that optimize the backbone network to yield benefits across all tasks. Let us consider a scenario where there are $K \geq 2$ tasks available, each associated with a differentiable loss function $\mathcal{L}_i(\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ represents the task-shared parameters. The goal of optimization-based MTL is to search for the optimal $\boldsymbol{\Theta}^* \in \mathbb{R}^m$ that minimizes the losses for all tasks. However, it is widely recognized that a simplistic linear scalar strategy, $\mathcal{L}_0(\boldsymbol{\Theta}) = \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_i(\boldsymbol{\Theta})$, fails to achieve satisfactory performance due to the conflict and imbalance issue.

3.2 Pareto Concept

Formally, let us assume the weighted loss as $\mathcal{L}_{\boldsymbol{\omega}} = \sum_{i=1}^{K}\omega_i\mathcal{L}_i(\boldsymbol{\Theta})$, where $\boldsymbol{\omega} \in \boldsymbol{\mathcal{W}}$ and $\boldsymbol{\mathcal{W}}$ represents the probability simplex on $[K]$. A point $\boldsymbol{\Theta'}$ is said to Pareto dominate $\boldsymbol{\Theta}$ if and only if $\forall i, \mathcal{L}_i(\boldsymbol{\Theta'}) \leq \mathcal{L}_i(\boldsymbol{\Theta})$. Consequently, the Pareto optimal situation arises when no $\boldsymbol{\Theta'}$ can be found that satisfies $\forall i, \mathcal{L}_i(\boldsymbol{\Theta'}) \leq \mathcal{L}_i(\boldsymbol{\Theta})$ for the given point $\boldsymbol{\Theta}$. All points that meet these conditions are referred to as Pareto sets, and their solutions are known as Pareto fronts. Another concept, known as Pareto stationary, requires $\min_{\boldsymbol{\omega}\in\boldsymbol{\mathcal{W}}}\|\boldsymbol{g}_{\boldsymbol{\omega}}\| = 0$, where $\boldsymbol{g}_{\boldsymbol{\omega}}$

[1] Refer to https://arxiv.org/abs/2503.08006 for the Appendix.



(a) LS (b) PCGrad (c) CAGrad (d) Nash-MTL (e) IMGrad

Figure 2: Comparison of MTL approaches on the imbalanced synthetic two-task benchmark. • and ⋆ represent the starting point and global optimum, respectively, and grey line ▬ represents the Pareto front. Two objectives are extremely imbalanced weighted, i.e., $(0.9{*}\mathcal{L}_1, 0.1{*}\mathcal{L}_2)$. Please refer to the **Appendix** for more optimization trajectories under various pre-defined task weights.

represents the weighted gradient $\boldsymbol{\omega}^\top \boldsymbol{G}$, and $\boldsymbol{G}$ is the gradients matrix whose each row is an individual gradient. We also provide some definitions here for ease of description.

Definition 1 (**Gradient Similarity**)**.** Denote ϕ_{ij} as the angle between two task gradients $\boldsymbol{g_i}$ and $\boldsymbol{g_j}$, then we define the gradient similarity as $\cos\varphi_{ij}$ and the gradients as conflicting when $\cos\phi_{ij} < 0$.

Definition 2 (**Imbalance of Individuals**)**.** Assume the gradient owns the maximal norm in $\boldsymbol{G}$ is $\boldsymbol{g}_{max}$, and the corresponding minimal one is $\boldsymbol{g}_{min}$. We define the imbalance ratio of $\boldsymbol{G}$ as $r = \frac{\|\boldsymbol{g}_{max}\|}{\|\boldsymbol{g}_{min}\|}$. If $r > 1$, we call it's imbalanced.

Definition 3 (**Pareto Property**)**.** For each training step, the combined optimization direction strives to promote all individuals simultaneously (or at the very least, not cause detriment), i.e. for $\forall i$, the gradient similarity between $\boldsymbol{g_i}$ and the combined gradient $\boldsymbol{g_\omega}$ satisfies $\cos\phi_{\boldsymbol{\omega}i} \geq 0$. When this condition is not met, it is referred to as **Pareto failure**.

4 Motivation and Observation

A substantial body of previous studies [Sener and Koltun, 2018; Liu *et al.*, 2021a; Yu *et al.*, 2020a; Navon *et al.*, 2022] have primarily focused on addressing the conflict issue rather than the imbalance issue. In this section, we aim to provide empirical insights into the significance of imbalance and elucidate how imbalance-sensitivity can bring benefits to current popular optimization-based MTL paradigms. Based on these insights, we naturally deduce our design in the next section.

4.1 Why Does Imbalance Matter More?

To begin, we conducted experiments on the CityScapes dataset [Cordts *et al.*, 2016] to statistically analyze the imbalance ratios of representative optimization-based MTL methods (e.g., PCGrad [Yu *et al.*, 2020a], CAGrad [Liu *et al.*, 2021a], Nash-MTL [Navon *et al.*, 2022]). The results of these experiments are presented in the **Appendix**. From the depicted results, it is evident that all the methods exhibit significant imbalance during training, which poses a substantial challenge when attempting to optimize all individuals simultaneously, thereby underscoring the importance of addressing the imbalance issue.

Secondly, to demonstrate the higher priority of imbalance issue, we show the toy example results that present imbalance and conflict among gradients in the following cases:

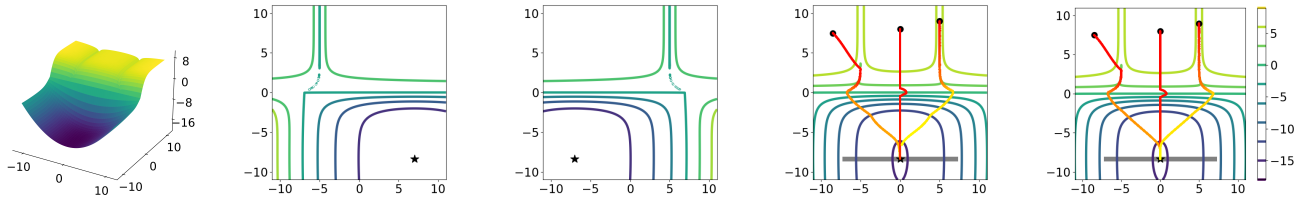
- Conflict (✔); Imbalance (✗): In Figure 3, we manually create scenarios where conflict exists but imbalance is absent. By closely examining the center trajectories in Figure 3 (d)(e), we observe that all methods can easily reach the optimal point when imbalance is absent, regardless of the presence of conflicts. This observation suggests that the sole existence of conflicts has limited impact on optimization, emphasizing the importance of addressing the imbalance issue.
- Conflict (✗); Imbalance (✔): Simulating an optimization trajectory without conflicts among individuals can indeed be challenging. Therefore, we adopt the setting from Nash-MTL [Navon *et al.*, 2022] to handcraft an imbalance-dominated optimization scenario. The resulting trajectories are depicted in Figure 2. It is evident that all the compared approaches fail to converge at the desired global optimum from all initial starts under the extreme imbalance circumstances, though most of them reach the Pareto front. Additionally, the trajectories at the sides in Figure 3 (d)(e) also highlight the issue of progress hindered by imbalance. Specifically, CAGrad fails to reach the global optimum compared to IMGrad despite undergoing the same number of optimization steps.

4.2 The Impacts of the Imbalance Issue

In Table 1, we list and compare mainstream optimization-based MTL approaches. The table focuses on two key properties: conflict-averse and imbalance-sensitive properties. It is observed that most MTL approaches possess the conflict-averse property due to their design nature. However, only a few approaches are imbalance-sensitive [2], and currently, there are no methods that possess both properties simultaneously. Furthermore, we analyze two imbalance-deduced issues that occur and impede past solutions during optimization: Pareto failure and imbalanced individual progress.

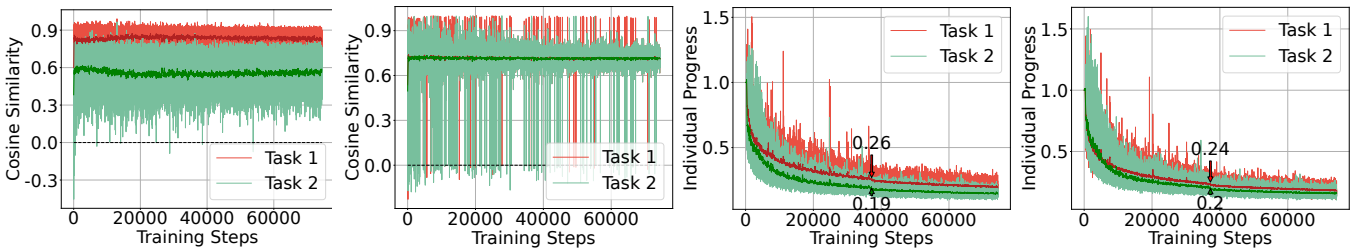
Pareto Failure: As shown in Figure 4 (a)(b), CAGrad exhibits a certain probability of failing to preserve the conflict issue due to its inherent compromise between conflict-averse and convergence. This compromise is inevitably influenced by the issue of imbalance. As illustrated in Figure 6, CAGrad tends to prioritize the combined gradient that deviates from the

[2]We provide imbalance-sensitive analysis for IMTL, Nash-MTL, and FAMO in the **Appendix**.



(a) Multi-Task Objective (b) Task 1 Objective (c) Task 2 Objective (d) CAGrad (e) `IMGrad`

Figure 3: Comparison of MTL approaches on the toy examples. We use the tool provided CAGrad to generate the synthetic toy examples with two objective shown in (b) and (c). In this case, both objective are equally weighted.



(a) CAGrad (b) Nash-MTL (c) CAGrad ($\Delta m\% = 9.97$) (d) Nash-MTL ($\Delta m\% = 8.20$)

Figure 4: Individual gradient similarity and progress analysis of MTL algorithms on CityScapes. (a)-(c) show the gradient similarities between individuals and the combined gradient; (d)-(e) present the progress of individuals during optimization.

individual with the least norm when encountering imbalanced scenarios, leading to potential conflicts. Surprisingly, although Nash-MTL imposes a strong constraint for the Pareto property, i.e., $\forall i, -\varphi_i(\boldsymbol{\omega}) \leq 0, \varphi_i(\boldsymbol{\omega}) = \log(\omega_i) + \log(\boldsymbol{g_i}^\top \boldsymbol{G}\boldsymbol{\omega}), \boldsymbol{G} = [\boldsymbol{g_1}, \boldsymbol{g_2}, ..., \boldsymbol{g_K}]$, it often fails to achieve such a guarantee. This failure can be attributed to the presence of negative terms in $\boldsymbol{g_i}^\top \boldsymbol{G}$, indicating conflicts between $\boldsymbol{g_i}$ and $\boldsymbol{g_j}$. Consequently, this leads to infeasible errors in the *cvxpy* [Diamond and Boyd, 2016] implementation, and the Nash-MTL algorithm chooses to skip the current step when such errors occur. As a result, Nash-MTL frequently encounters Pareto failures due to the co-existence of imbalance and conflict, as depicted in Figure 4 (b). `IMGrad` demonstrates a tendency to acquire a combined gradient that effectively preserves the Pareto property as the imbalance ratio increases.

Imbalanced Individual Progress: We employ an individual progress metric proposed by [Chen *et al.*, 2018], which is defined as follows:

$$r_i(t) = \mathcal{L}_i(t)/\mathcal{L}_i(0) \tag{1}$$

where $\mathcal{L}_i(t)$ represents the individual loss value at t time. As depicted in Figure 4 (c)(d), Nash-MTL demonstrates a narrower gap in terms of individual progress compared to CAGrad. This can be attributed to the more balanced combination employed by Nash-MTL, as indicated by the cosine similarity in (a)(b). Consequently, Nash-MTL exhibits superior overall performance, characterized by a smaller $\Delta m\%$. Specifically, $\Delta m\%$ is widely adopted to evaluate the overall degradation compared to independently trained models, which are considered as the reference oracles. Its formal definition can be found in the **Performance Evaluation** section.

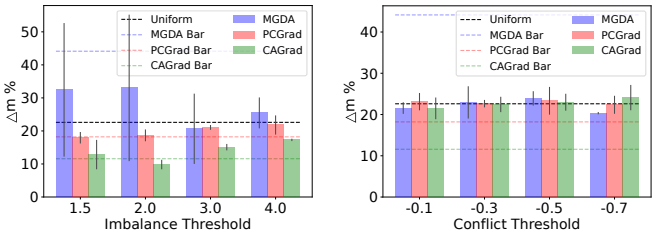
Unfortunately, none of the above methods get rid of both Pareto failure and imbalanced individual progress, primarily due to their limited focus on the imbalance issue.

4.3 Benefits of Integrating Imbalance-Sensitivity

The toy results depicted in Figure 2 and Figure 3 demonstrate that among the methods evaluated, only our proposed `IMGrad`, which incorporates imbalance-sensitivity, consistently arrives at the optimal point from all initial starts.

To further elucidate the advantages of imbalance-sensitivity in optimization-based MTL, we have implemented a naïve method called *Adaptive Threshold*. This baseline selectively applies optimization-based MTL approaches only when the imbalance ratio surpasses a specific threshold. The results of this implementation on CityScapes are presented in Figure 5 (a). It is evident that all baselines exhibit varying performance as the imbalance ratio fluctuates, emphasizing the significance of imbalance-sensitivity. Notably, all baselines outperform their respective vanilla versions under specific threshold conditions, providing additional evidence of the effectiveness of injecting imbalance-sensitivity.

Additionally, we have conducted a series of control group experiments to further support our findings. Similarly, we only apply optimization-based MTL when the gradient similarity falls below a certain threshold. As depicted in Figure 5 (b), all baselines demonstrate relatively stable performance compared to those in (a) and fail to outperform the vanilla version, except for MGDA (which itself performs worse than LS). This outcome further reinforces the claim that imbalance matters more.



(a) Imbalance Sensitivity (b) Conflict Sensitivity

Figure 5: Imbalance and conflict sensitivity examination.

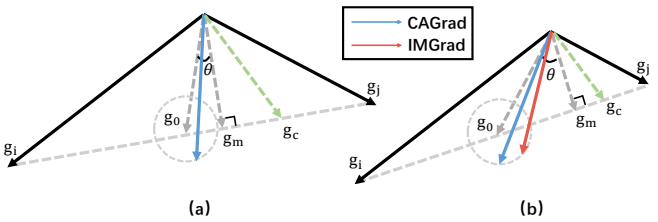


Figure 6: Multi-objective optimization Comparison between CAGrad and IMGrad. Here we suppose the angles between $\boldsymbol{g_i}$ and $\boldsymbol{g_j}$ in (a) and (b) are same. $\boldsymbol{g_m}$ can be obtained via MGDA.

5 Principal Design

In this section, taking CAGrad as the baseline, we present the principal design of IMGrad, encompassing its formulation in the objective function and the practical implementation. And we provide convergence and speedup analysis in the **Appendix**.

5.1 Injecting Imbalance-Sensitivity

As a widely adopted baseline, CAGrad strikes a balance between Pareto property and globe convergence, and its dual objective is formulated as follows:

$$\max_{\boldsymbol{d}\in\mathbb{R}^m}\min_{\boldsymbol{\omega}\in\mathcal{W}}\boldsymbol{g_\omega}^\top\boldsymbol{d}\quad \text{s.t.}\ \|\boldsymbol{d}-\boldsymbol{g_0}\|\le c\,\|\boldsymbol{g_0}\| \tag{2}$$

where $\boldsymbol{d}$ represents the combined gradient, while $\boldsymbol{g_0}$ denotes the averaged gradient, and c is the hyper-parameter.

To alleviate the imbalance-deduced Pareto failures or individual progress issue as illustrated in Figure 4, a logical approach is to maximize the projected norm of the combined gradient across all individuals. To achieve this, we incorporate a stronger constraint ($\boldsymbol{g_i}^\top\boldsymbol{d}-\|\boldsymbol{g_i}\|^2$) into Eqn. 2, which encourages projected norms that surpass individual norms. This formulation is reflected in our objective presented in Eqn. 3, and subsequently, we derive the corresponding Lagrangian equations in Eqn. 4.

$$\max_{\boldsymbol{d}\in\mathbb{R}^m}\min_{\boldsymbol{\omega}\in\mathcal{W}}\boldsymbol{g_\omega}^\top\boldsymbol{d}-\mu(\boldsymbol{g_\omega}^\top\boldsymbol{d}-\|\boldsymbol{g_\omega}\|^2)\ \text{ s.t.}\ \|\boldsymbol{d}-\boldsymbol{g_0}\|\le c\,\|\boldsymbol{g_0}\| \tag{3}$$

$$\max_{\boldsymbol{d}\in\mathbb{R}^m}\min_{\lambda\ge0,\boldsymbol{\omega}\in\mathcal{W}}\boldsymbol{g_\omega}^\top\boldsymbol{d}-\lambda(\|\boldsymbol{d}-\boldsymbol{g_0}\|^2-\phi)/2 \tag{4}$$

$$-\mu(\boldsymbol{g_\omega}^\top\boldsymbol{d}-\|\boldsymbol{g_\omega}\|^2),\quad \lambda>0,\mu>0$$

The strong duality property holds for the aforementioned objective, as supported by convex programming principles and the fulfillment of Slater's condition. Consequently, we interchange the positions of the minimum and maximum operators:

$$\min_{\lambda\ge0,\boldsymbol{\omega}\in\mathcal{W}}\max_{\boldsymbol{d}\in\mathbb{R}^m}(1-\mu)\boldsymbol{g_\omega}^\top\boldsymbol{d} \tag{5}$$

$$-\frac{\lambda}{2}(\|\boldsymbol{d}-\boldsymbol{g_0}\|^2-\phi)+\mu\,\|\boldsymbol{g_\omega}\|^2$$

With λ,ω fixing, the optimal $\boldsymbol{d}$ is achieved when $\boldsymbol{d}=\boldsymbol{g_0}+\frac{(1-\mu)\boldsymbol{g_\omega}}{\lambda}$. Substitude the optimal $\boldsymbol{d}$ into Eqn. 5, yielding the following problem:

$$\min_{\lambda\ge0,\boldsymbol{\omega}\in\mathcal{W}}(1-\mu)\boldsymbol{g_\omega}^\top\boldsymbol{g_0}+\mu\,\|\boldsymbol{g_\omega}\|^2 \tag{6}$$
$$+\frac{(1-\mu)^2}{2\lambda}\,\|\boldsymbol{g_\omega}\|^2+\frac{\lambda}{2}\phi$$

After optimizing out the λ we have

$$\min_{\boldsymbol{\omega}\in\mathcal{W}}(1-\mu)\boldsymbol{g_\omega}^\top\boldsymbol{g_0}+\mu\,\|\boldsymbol{g_\omega}\|^2+(1-\mu)\sqrt{\phi}\,\|\boldsymbol{g_\omega}\| \tag{7}$$

where $\lambda=(1-\mu)\,\|\boldsymbol{g_\omega}\|/\phi^{1/2}$, and finally we have the optimization objective in Eqn. 8. By solving this objective, we can obtain $\boldsymbol{g_\omega}$ and have $\boldsymbol{d}=\boldsymbol{g_0}+\frac{\phi^{1/2}}{\|\boldsymbol{g_\omega}\|}\boldsymbol{g_\omega}$.

$$\min_{\boldsymbol{\omega}\in\mathcal{W}}(1-\mu)(\underbrace{\boldsymbol{g_\omega}^\top\boldsymbol{g_0}+\sqrt{\phi}\,\|\boldsymbol{g_\omega}\|}_{\text{CAGrad}})+\mu\,\underbrace{\|\boldsymbol{g_\omega}\|^2}_{\sim\text{MGDA}} \tag{8}$$

Upon careful examination of Eqn. 8, it becomes evident that the final objective can be decomposed into two distinct components: CAGrad and MGDA. As depicted in Figure 6 (a), the gradient obtained by solving the practical objective in Eqn. 10, denoted as $\boldsymbol{g_c}$ (represented by the green dotted line), predominantly resides within the region bounded by $\boldsymbol{g_m}$ and $\boldsymbol{g_j}$. However, in the case of an extreme imbalance scenario, as illustrated in Figure 6 (b), the corresponding $\boldsymbol{g_c}$ tends to lean towards the dominant gradient $\boldsymbol{g_i}$, thereby increasing the risk of conflicting with $\boldsymbol{g_j}$ and resulting in Pareto failures. When confronted with such a situation characterized by varying imbalances, it is desirable for μ to adaptively adjust $\boldsymbol{g_c}$ to consistently avoid Pareto failures while still promoting individual progress when the imbalance is less pronounced. Consequently, we establish a connection between μ and the gradient imbalances, effectively controlling the constraint ($\boldsymbol{g_i}^\top\boldsymbol{d}-\|\boldsymbol{g_i}\|^2$) adaptively based on the imbalance circumstances.

Multiple alternatives exist for quantifying the imbalance ratio among individuals [3]. We here choose to compute $\cos\theta$ to represent the imbalance ratio (see negative correlation between imbalance ratio and $\cos\theta$ in the **Appendix**), where θ denotes the angle between $\boldsymbol{g_0}$ and $\boldsymbol{g_m}$. As a result, Eqn. 8 can be re-written as:

$$\min_{\boldsymbol{\omega}\in\mathcal{W}}(1-\cos\theta)(\boldsymbol{g_\omega}^\top\boldsymbol{g_0}+\sqrt{\phi}\,\|\boldsymbol{g_\omega}\|)+\cos\theta\,\|\boldsymbol{g_\omega}\|^2 \tag{9}$$

[3] Please refer to the **Appendix** for additional alternatives.

Simplification: As a matter of fact, CAGrad itself contains decoupled components in its practical objective:

$$\min_{\boldsymbol{\omega}\in\mathcal{W}} \underbrace{\boldsymbol{g}_{\boldsymbol{\omega}}^{\top}\boldsymbol{g}_{\mathbf{0}}}_{\text{Push Away from } g_0} + \underbrace{\sqrt{\phi}\,\|\boldsymbol{g}_{\boldsymbol{\omega}}\|}_{\sim\text{MGDA}} \tag{10}$$

where $\boldsymbol{g}_{\boldsymbol{\omega}}^{\top}\boldsymbol{g}_{\mathbf{0}}$ tends to push away from $\boldsymbol{g}_{\mathbf{0}}$ and $\sqrt{\phi}\,\|\boldsymbol{g}_{\boldsymbol{\omega}}\|$ plays the role of MGDA does. Thus we can simplify the Eqn. 9 as:

$$\min_{\boldsymbol{\omega}\in\mathcal{W}} (1-\cos\theta)\boldsymbol{g}_{\boldsymbol{\omega}}^{\top}\boldsymbol{g}_{\mathbf{0}} + \cos\theta\sqrt{\phi}\,\|\boldsymbol{g}_{\boldsymbol{\omega}}\|^2 \tag{11}$$

5.2 Augment Nash-MTL with Imbalance Sensitivity

As stated in the previous **Pareto Failure** analysis, while Nash-MTL effectively addresses the imbalance issue and appears to be naturally conflict-averse, its implementation often leads to frequent Pareto failures. To address this problem, let's first examine its decoupled objective:

$$\min_{\omega} \underbrace{\sum_i \boldsymbol{g}_{\boldsymbol{i}}^{\top}\boldsymbol{G}\boldsymbol{\omega}}_{\text{Push Away from } g_0} + \underbrace{\varphi(\boldsymbol{\omega})}_{\text{Strike balance among individuals}} \tag{12}$$
$$\text{s.t. } \forall i, -\varphi_i(\boldsymbol{\omega}) \le 0, \;\; \omega_i > 0$$

where $\varphi_i(\boldsymbol{\omega}) = \log(\omega_i) + \log(\boldsymbol{g}_{\boldsymbol{i}}^{\top}\boldsymbol{G}\boldsymbol{\omega}), \boldsymbol{G} = [\boldsymbol{g_1}, \boldsymbol{g_2}, ..., \boldsymbol{g_K}]$. $\sum_i \boldsymbol{g}_{\boldsymbol{i}}^{\top}\boldsymbol{G}\boldsymbol{\omega}$ tends to push away from $\boldsymbol{g}_{\mathbf{0}}$ and $\varphi(\boldsymbol{\omega})$ strives balance among individuals. Intuitively, we expect to preserve the Pareto property when encountering extremely imbalanced scenarios; therefore, $\sum_i \boldsymbol{g}_{\boldsymbol{i}}^{\top}\boldsymbol{G}\boldsymbol{\omega}$ should be given more weight:

$$\min_{\omega} (1-\cos\theta)\sum_i \boldsymbol{g}_{\boldsymbol{i}}^{\top}\boldsymbol{G}\boldsymbol{\omega} + \cos\theta\varphi(\boldsymbol{\omega}) \tag{13}$$

With the proper assumption of H-Lipschitz on gradients, we can still avoid Pareto failure with the derived weights among individuals from the last step. In a word, we augment Nash-MTL by injecting imbalance sensitivity to reduce Pareto failures. Please refer to the **Appendix** for more details.

5.3 Implementation

We implement our approach with Python 3.8, PyTorch 1.4.0 and cvxpy 1.3.1, while all experiments are carried out on Tesla V100 GPUs [4]. We follow the setting and general implementation of [Liu *et al.*, 2021a], and the toy example generation is borrowed from [Navon *et al.*, 2022; Senushkin *et al.*, 2023]. See more implementation details in the **Appendix**.

6 Performance Evaluation

Following the evaluation protocol in [Navon *et al.*, 2022] and taking it as the baseline, we conduct experiments under the supervised learning and reinforcement learning scenarios. Specifically, two scene understanding and one image classification benchmarks are involved in supervised learning, and the classical MT10 benchmark is adopted for reinforcement learning. The examination of Pareto failures, individual task progress, a sensitivity analysis of μ, the verification of negative correlation between imbalance ratio and $\cos\theta$, speed analysis, and more visualizations are also provided in the **Appendix**, please refer them for more details.

Evaluation metric. In addition to reporting individual performance, we also incorporate a widely used metric, $\Delta m\%$ [Maninis *et al.*, 2019], which evaluates the overall degradation compared to independently trained models that are considered as the reference oracles. The formal definition of $\Delta m\%$ is given as: $\Delta m\% = \frac{1}{K}\sum_{k=1}^{K}(-1)^{\delta_k}(M_{m,k} - M_{b,k})/M_{b,k}$. $M_{m,k}$ and $M_{b,k}$ represent the metric M_k for the compared method and the independent model, respectively. The value of δ_k is assigned as 1 if a higher value is better for M_k, and 0 otherwise.

Method	Segmentation (Higher Better)		Depth (Lower Better)		Δ m% ↓
	mIoU	Pix. Acc.	Abs. Err.	Rel. Err.	
Independent	74.01	93.16	0.0125	27.77	-
LS	75.18	93.49	0.0155	46.77	22.60
RLW	74.57	93.41	0.0158	47.79	24.37
DWA	75.24	93.52	0.0160	44.37	21.43
MGDA	68.84	91.54	0.0309	33.50	44.14
GradDrop	75.27	93.53	0.0157	47.54	23.67
PCGrad	75.13	93.48	0.0154	42.07	18.21
CAGrad	75.16	93.48	0.0141	37.60	11.58
IMTL	75.33	93.49	0.0135	38.41	11.04
Nash-MTL	75.41	93.66	0.0129	35.02	6.82
MoCo	75.42	93.55	0.0149	34.19	9.90
FAMO	74.54	93.29	0.0145	32.59	8.13
`IMGrad`	75.13	93.45	0.0128	34.95	6.61

Table 2: **Scene understanding** (*CityScapes*, 2 tasks). We report MTAN model performance averaged over 3 random seeds.

6.1 Supervised Learning

Customary evaluation in supervised learning for MTL involves assessing the ability of MTL approaches to handle multiple scene understanding and classification tasks. For scene understanding tasks, we follow previous studies [Liu *et al.*, 2021a; Liu *et al.*, 2021b; Navon *et al.*, 2022] and employ a Multi-Task Attention Network (MTAN)[Liu *et al.*, 2019] as the fundamental architecture for all MTL approaches. Our experiments are conducted on two well-established datasets: NYUv2[Silberman *et al.*, 2012] and CityScapes [Cordts *et al.*, 2016]. To ensure fair comparisons, we adopt the same training strategy as described in prior works [Liu *et al.*, 2021a; Navon *et al.*, 2022]. Specifically, models are trained for 200 epochs using the Adam optimizer, with an initial learning rate of 1e-4, which decays to 5e-5 after 100 epochs. For the image classification task, we utilize a 9-layer convolutional neural network (CNN) as the backbone, with linear layers serving as task-specific heads, and conduct experiments on CelebA [Liu *et al.*, 2015]. The model is trained using the Adam optimizer for 15 epochs, with an initial learning rate of 3.0e-4 and a batch size of 256.

NYUv2. NYUv2 is a widely used indoor scene understanding dataset for MTL benchmarking, encompassing three tasks:

[4] Code is avaliable at https://github.com/zzpustc/IMGrad.

Method	Segmentation (Higher Better)		Depth (Lower Better)		Surface Normal					Δ m% ↓
					Angle Distance (Lower Better)		Within t° (Higher Better)			
	mIoU	Pix. Acc.	Abs Err	Rel Err	Mean	Median	11.25	22.5	30	
Independent	38.30	63.76	0.68	0.28	25.01	19.21	30.14	57.20	69.15	-
LS	39.29	65.33	0.55	0.23	28.15	23.96	22.09	47.50	61.08	5.46
RLW	37.17	63.77	0.58	0.24	28.27	24.18	22.26	47.05	60.62	7.67
DWA	39.11	65.31	0.55	0.23	27.61	23.18	24.17	50.18	62.39	3.49
MGDA	30.47	59.90	0.61	0.26	24.88	19.45	29.18	56.88	69.36	1.47
GradDrop	39.39	65.12	0.55	0.23	27.48	22.96	23.38	49.44	62.87	3.61
PCGrad	38.06	64.64	0.56	0.23	27.41	22.80	23.86	49.83	63.14	3.83
CAGrad	39.79	65.49	0.55	0.23	26.31	21.58	25.61	52.36	65.58	0.29
IMTL	39.35	65.60	0.54	0.23	26.02	21.19	26.20	53.13	66.24	-0.59
Nash-MTL	40.13	65.93	0.53	0.22	25.26	20.08	28.40	55.47	68.15	-4.04
MoCo	40.30	66.07	0.56	0.21	26.67	21.83	25.61	51.78	64.85	0.16
FAMO	38.88	64.90	0.55	0.22	25.06	19.57	29.21	56.61	68.98	-4.10
IMGrad	40.20	66.19	0.52	0.22	25.15	19.94	28.69	55.80	68.44	-4.57

Table 3: **Scene understanding** (*NYUv2*, 3 tasks). We report MTAN model performance averaged over 3 random seeds.

semantic segmentation, depth estimation, and surface normal prediction. The results, presented in Table 3, show that IMGrad surpasses the previous SOTA in terms of $\Delta m\%$, highlighting the effectiveness of incorporating imbalance sensitivity. IMGrad also achieves best performance on segmentation and depth tasks without much promise on other tasks.

CityScapes. The CityScapes dataset is used for MTL evaluation, focusing on semantic segmentation and depth estimation tasks. Following the previous experimental setup, we utilize a coarser version that categorizes segmentation into 7 classes. The results, presented in Table 2, indicate that IMGrad exhibits a similar trend to its performance on NYUv2 and achieves SOTA results in terms of $\Delta m\%$.

MT10		**CelebA**	
Method	Success ± SEM ↑	Method	**Δm%** ↓
LS	0.49 ± 0.070	LS	4.15
STL SAC	0.90 ± 0.032	SI	7.20
MTL SAC	0.49 ± 0.073	RLW	1.46
MH SAC	0.54 ± 0.047	DWA	3.20
SM	0.73 ± 0.043	UW	3.23
CARE	0.84 ± 0.051	MGDA	14.85
PCGrad	0.72 ± 0.022	PCGrad	3.17
CAGrad	0.83 ± 0.045	CAGrad	2.48
Nash-MTL	0.91 ± 0.031	Nash-MTL	2.84
FAMO	0.83 ± 0.050	FAMO	1.21
IMGrad	0.93 ± 0.068 **(+0.10)**	IMGrad	1.27

Table 4: **Reinforcement learning** (*MT10*, 10 tasks) and **image classification** (*CelebA*, 40-task).

CelebA. CelebA is a widely used face attributes dataset containing over 200,000 images annotated with 40 attributes. Recently, it has been adopted as a 40-task MTL benchmark to evaluate a model's ability to handle a large number of tasks. The results, presented in Table 4, are averaged over three random seeds. While IMGrad does not achieve the best performance, it consistently ranks among the top methods, underscoring the importance of imbalance sensitivity.

6.2 Reinforcement Learning

Reinforcement learning is another domain where MTL is often essential, as it seeks to acquire a policy capable of succeeding across various manipulation tasks. To evaluate the generalizability of our proposed method, we use CAGrad as the baseline and conduct experiments on the MT10 environment from the Meta-World benchmark [Yu *et al.*, 2020b]. The results, presented in Table 4, report the average success rate on the validation set over 10 random seeds. Consistent with the improvements observed in supervised learning evaluations, IMGrad enhances CAGrad by over 0.10, achieving SOTA performance on this benchmark. It is worth noting that Nash-MTL does not provide an official implementation for reinforcement learning benchmarks. As a result, we did not augment it for evaluation in this context.

7 Conclusion

In this paper, we begin by empirically demonstrating the significance of addressing the imbalance issue in optimization-based MTL. We assert that incorporating imbalance-sensitivity is crucial for avoiding Pareto failures and promoting balanced individual progress. Building upon this motivation, we propose IMGrad, a method derived from a projection norm constraint, which is further simplified as an adaptive balancer between decoupled objectives. Through extensive experiments, we validate the effectiveness of our proposed approach. We believe that our explicit emphasis on the imbalance issue, rather than the conflict issue, provides valuable insights for the future development of optimization-based MTL.

Acknowledgements

We thank anonymous reviewers for their valuable comments.

References

[Chen *et al.*, 2018] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

[Chen *et al.*, 2023] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[Diamond and Boyd, 2016] Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

[Fan *et al.*, 2022] Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, Zhangyang Wang, et al. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35:28441–28457, 2022.

[Fernando *et al.*, 2023] Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *The Eleventh International Conference on Learning Representations*, 2023.

[Gao *et al.*, 2019] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan 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*, pages 3205–3214, 2019.

[He *et al.*, 2021] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

[Heuer *et al.*, 2021] Falk Heuer, Sven Mantowsky, Saqib Bukhari, and Georg 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*, pages 997–1005, 2021.

[Jaggi, 2013] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

[Kokkinos, 2017] Iasonas Kokkinos. Ubernet: 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*, pages 6129–6138, 2017.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Liu *et al.*, 2019] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

[Liu *et al.*, 2021a] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.

[Liu *et al.*, 2021b] Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *International Conference on Learning Representations*, 2021.

[Liu *et al.*, 2022] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[Liu *et al.*, 2023] Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. *arXiv preprint arXiv:2306.03792*, 2023.

[Maninis *et al.*, 2019] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1851–1860, 2019.

[Navon *et al.*, 2022] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*, pages 16428–16446. PMLR, 2022.

[Sener and Koltun, 2018] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[Senushkin *et al.*, 2023] Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20083–20093, 2023.

[Silberman *et al.*, 2012] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation

and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

[Yang and Hospedales, 2016] Yongxin Yang and Timothy Hospedales. Deep multi-task representation learning: A tensor factorisation approach. *arXiv preprint arXiv:1605.06391*, 2016.

[Ye and Xu, 2022] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *European Conference on Computer Vision*, pages 514–530. Springer, 2022.

[Yu *et al.*, 2020a] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

[Yu *et al.*, 2020b] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[Zhou *et al.*,] Zhipeng Zhou, Liu Liu, Peilin Zhao, and Wei Gong. Pareto deep long-tailed recognition: A conflict-averse solution. In *The Twelfth International Conference on Learning Representations*.