TOWARDS IMPARTIAL MULTI-TASK LEARNING

Liyang Liu1, Yi Li2, Zhanghui Kuang2, Jing-Hao Xue3, Yimin Chen2, Wenming Yang1, Qingmin Liao1, Wayne Zhang2,4
1Shenzhen International Graduate School/Department of Electronic Engineering, Tsinghua University
2SenseTime Research
3Department of Statistical Science, University College London
4Qing Yuan Research Institute, Shanghai Jiao Tong University
{lulu-ly14@mails., yang.wenming@sz., liaoqm@tsinghua.edu.cn}
{liyi, kuangzhanghui, chenyimin, wayne.zhang@sensetime.com}
jinghao.xue@ucl.ac.uk

ABSTRACT

Multi-task learning (MTL) has been widely used in representation learning. However, naively training all tasks simultaneously may lead to the partial training issue, where specific tasks are trained more adequately than others. In this paper, we propose to learn multiple tasks impartially. Specifically, for the task-shared parameters, we optimize the scaling factors via a closed-form solution, such that the aggregated gradient (sum of raw gradients weighted by the scaling factors) has equal projections onto individual tasks. For the task-specific parameters, we dynamically weigh the task losses so that all of them are kept at a comparable scale. Further, we find the above gradient balance and loss balance are complementary and thus propose a hybrid balance method to further improve the performance. Our impartial multi-task learning (IMTL) can be end-to-end trained without any heuristic hyper-parameter tuning, and is general to be applied on all kinds of losses without any distribution assumption. Moreover, our IMTL can converge to similar results even when the task losses are designed to have different scales, and thus it is scale-invariant. We extensively evaluate our IMTL on the standard MTL benchmarks including Cityscapes, NYUv2 and CelebA. It outperforms existing loss weighting methods under the same experimental settings.

1 INTRODUCTION

Recent deep networks in computer vision can match or even surpass human beings on some specific tasks separately. However, in reality multiple tasks (e.g., semantic segmentation and depth estimation) must be solved simultaneously. Multi-task learning (MTL) (Caruana, 1997; Evgeniou & Pontil, 2004; Ruder, 2017; Zhang & Yang, 2017) aims at sharing the learned representation among tasks (Zamir et al., 2018) to make them benefit from each other and achieve better results and stronger robustness (Zamir et al., 2020). However, sharing the representation can lead to a partial learning issue: some specific tasks are learned well while others are overlooked, due to the different loss scales or gradient magnitudes of various tasks and the mutual competition among them. Several methods have been proposed to mitigate this issue either via gradient balance such as gradient magnitude normalization (Chen et al., 2018) and Pareto optimality (Sener & Koltun, 2018), or loss balance like homoscedastic uncertainty (Kendall et al., 2018). Gradient balance can evenly learn task-shared parameters while ignoring task-specific ones. Loss balance can prevent MTL from being biased in favor of tasks with large loss scales but cannot ensure the impartial learning of the shared parameters. In this work, we find that gradient balance and loss balance are complementary, and combining the two balances can further improve the results. To this end, we propose impartial MTL (IMTL) via simultaneously balancing gradients and losses across tasks.

For gradient balance, we propose IMTL-G(rad) to learn the scaling factors such that the aggregated gradient of task-shared parameters has equal projections onto the raw gradients of individual tasks.

*Corresponding author
We propose a novel closed-form gradient balance method, which learns task-shared parameters without any task bias; and we develop a general learnable loss balance method, where no distribution assumption is required and the scale parameters can be jointly trained with the network parameters.

We unveil that gradient balance and loss balance are complementary and accordingly propose a hybrid balance method to simultaneously balance gradients and losses.

We validate that our proposed IMTL is loss scale-invariant and is more robust against loss scale changes compared with its competitors, and we give in-depth theoretical and experimental analyses on its connections and differences with previous methods.

We extensively verify the effectiveness of our IMTL. For fair comparisons, a unified codebase will also be publicly available, where more practical settings are adopted and stronger performances are achieved compared with existing code-bases.
2 RELATED WORK

Recent advances in MTL mainly come from two aspects: network structure improvements and loss weighting developments. Network-structure methods based on soft parameter-sharing usually lead to high inference cost (review in Appendix A). Loss weighting methods find loss weights to be multiplied on the raw losses for model optimization. They employ a hard parameter-sharing paradigm (Ruder, 2017), where several light-weight task-specific heads are attached upon the heavy-weight task-agnostic backbone. There are also efforts that learn to group tasks and branch the network in the middle layers (Guo et al., 2020; Standley et al., 2020), which try to achieve better accuracy-efficiency trade-off and can be seen as semi-hard parameter-sharing. We believe task grouping and loss weighting are orthogonal and complementary directions to facilitate multi-task learning and can benefit from each other. In this work we focus on loss weighting methods which are the most economic as almost all of the computations are shared across tasks, leading to high inference speed. Task Prioritization (Guo et al., 2018) weights task losses by their difficulties to focus on the harder tasks during training. Uncertainty weighting (Kendall et al., 2018) models the loss weights as data-agnostic task-dependent homoscedastic uncertainty. Then loss weighting is derived from maximum likelihood estimation. GradNorm (Chen et al., 2018) learns the loss weights to enforce the norm of the scaled gradient for each task to be close. MGDA (Sener & Koltun, 2018) casts multi-task learning as multi-object optimization and finds the minimum-norm point in the convex hull composed by the gradients of multiple tasks. Pareto optimality is supposed to be achieved under mild conditions. GLS (Chennupati et al., 2019) instead uses the geometric mean of task-specific losses as the target loss, we will show it actually weights the loss by its reciprocal value. PCGrad (Yu et al., 2020) avoids interferences between tasks by projecting the gradient of one task onto the normal plane of the other. DSG (Lu et al., 2020) dynamically makes a task “stop or go” by its converging state, where a task is updated only once for a while if it is stopped. Although many loss weighting methods have been proposed, they are seldom open-sourced and rarely compared thoroughly under practical settings where strong performances are achieved, which motivates us to give an in-depth analysis and a fair comparison about them.

3 IMPARTIAL MULTI-TASK LEARNING

In MTL, we map a sample \( x \in \mathbb{X} \) to its labels \( \{ y_t \in \mathbb{Y}_t \}_{t \in [1, T]} \) of all \( T \) tasks through multiple task-specific mappings \( \{ f_t : \mathbb{X} \rightarrow \mathbb{Y}_t \} \). In most loss weighting methods, the hard parameter-sharing paradigm is employed, such that \( f_t \) is parameterized by heavy-weight task-shared parameters \( \theta \) and light-weight task-specific parameters \( \theta_t \). All tasks take the same shared intermediate feature \( z = f(x; \theta) \) as input, and the \( t \)-th task head outputs the prediction as \( f_t(x) = f_t(z; \theta_t) \). We aim to find the scaling factors \( \{ \alpha_t \} \) for all \( T \) task losses \( \{ L_t(f_t(x), y_t) \} \), so that the weighted sum loss \( L = \sum_t \alpha_t L_t \) can be optimized to make all tasks perform well. This poses great challenges because: 1) losses may have distinguished forms such as cross-entropy loss and cosine similarity; 2) the dynamic ranges of losses may differ by orders of magnitude. In this work, we propose a hybrid solution for both the task-shared parameters \( \theta \) and the task-specific parameters \( \theta_t \), as Fig. 2.

3.1 GRADIENT BALANCE: IMTL-G

For task-shared parameters \( \theta \), we can receive \( T \) gradients \( \{ g_t = \nabla_{\theta} L_t \} \) via back-propagation from all of the \( T \) raw losses \( \{ L_t \} \), and these gradients represent optimal update directions for individual tasks. As the parameters \( \theta \) can only be updated with a single gradient, we should compute an aggregated gradient \( g \) by the linear combination of \( \{ g_t \} \). It also implies to find the scaling factors \( \{ \alpha_t \} \) of raw losses \( \{ L_t \} \), since \( g = \sum \alpha_t g_t = \nabla_{\theta} L = \nabla_{\theta} (\sum \alpha_t L_t) \).

Motivated by the principle of balance among tasks, we propose to make the projections of \( g \) onto \( \{ g_t \} \) to be equal, as Fig. 1(d). In this way,
thus stronger stability can be achieved. Suppose the positive losses \(\theta\) manifest severe oscillations, so we further propose to

brevity. Then we should construct an appropriate scaled loss

parameters \(\theta\) idea is to compute the scaling factors as

\[ \theta = \nabla f(x; \theta) \]

Input: input sample \(x\), task-specific labels \(\{y_i\}\) and learning rate \(\eta\)
Output: task-shared/-specific parameters \(\theta/\theta_i\), scale parameters \(\{s_t\}\)

1: compute task-shared feature \(z = f(x; \theta)\)
2: for \(t = 1\) to \(T\) do
3: compute task prediction by head network \(f_t(x) = f_i^{\text{net}}(z; \theta_i)\)
4: compute raw loss by loss function \(L_i^{\text{raw}} = \ell_i^{\text{finc}}(f_t(x), y_i)\)
5: compute scaled loss \(L_t = \beta a_t L_i^{\text{raw}} - s_t\) (default \(a = b = 1\)) \(\triangleright\) loss balance
6: compute gradient of shared feature \(z; g_t = \nabla_z L_t\)
7: compute unit-norm gradient \(u_t = g_t / \|g_t\|\)
8: end for
9: compute gradient differences \(D^T = [g_1^T - g_2^T, \cdots, g_t^T - g_1^T]\)
10: compute unit-norm gradient differences \(U^T = [u_1^T - u_2^T, \cdots, u_t^T - u_1^T]\)
11: compute scaling factors for tasks 2 to \(T\): \(\alpha_{2:T} = g_t U^T (DU^T)^{-1}\) \(\triangleright\) gradient balance
12: compute scaling factors for all tasks: \(\alpha = [1 - 1\alpha_{2:T}^T, \alpha_{2:T}]\)
13: update task-shared parameters \(\theta = \theta - \eta \nabla_{\theta} (\sum_t \alpha_t L_t)\)
14: for \(t = 1\) to \(T\) do
15: update task-specific parameters \(\theta_i = \theta_i - \eta \nabla_{\theta_i} L_t\)
16: update loss scale parameter \(s_t = s_t - \eta \alpha_t L_t\)
17: end for

we treat all tasks equally so that they progress in the same speed and none is left behind. Formally, let \(u_t = g_t / |g_t|\) denote the unit-norm vector of \(\{g_t\}\) which are row vectors, then we have:

\[ g u_t^T = g u_t^T \Leftrightarrow g (u_1 - u_t)^T = 0, \forall \ 2 \leq t \leq T. \]  \[
\] (1)

The above problem is under-determined, but we can obtain the closed-form results of \(\{\alpha_t\}\) by constraining \(\sum_t \alpha_t = 1\). Assume \(\alpha = [\alpha_2, \cdots, \alpha_T], U^T = [u_1^T - u_2^T, \cdots, u_1^T - u_T^T], D^T = [g_1^T - g_2^T, \cdots, g_1^T - g_T^T]\) and \(1 = [1, \cdots, 1]\), from Eq. (1) we can obtain:

\[ \alpha = g_t U^T (DU^T)^{-1}. \]  \[
\] (IMTL-G)

The detailed derivation is in Appendix B.1. After obtaining \(\alpha\), the scaling factor of the first task can be computed by \(\alpha_1 = 1 - 1\alpha_1^T\) since \(\sum_t \alpha_t = 1\). The optimized \(\{\alpha_t\}\) are used to compute \(L = \sum_t \alpha_t L_t\), which is ultimately minimized by SGD to update the model. By now, back-propagation needs to be executed \(T\) times to obtain the gradient of each task loss with respect to the heavy-weight task-shared parameters \(\theta\), which is time-consuming and non-scalable. We replace the parameter-level gradients \(\{g_t = \nabla_{\theta} L_t\}\) with feature-level gradients \(\{\nabla_z L_t\}\) to compute \(\{\alpha_t\}\). This implies to achieve gradient balance with respect to the last shared feature \(z\) as a surrogate of task-shared parameters \(\theta\), since it is possible for the network to back-propagate this balance all the way through the task-shared backbone starting from \(z\). This relaxation allows us to do back propagation through the backbone only once after obtaining \(\{\alpha_t\}\), and thus the training time can be dramatically reduced.

\section{Loss Balance: IMTL-L}

For the task-specific parameters \(\{\theta_i\}\), we cannot employ IMTL-G described above, because \(\nabla_{\theta_i} L_t = 0, \forall t \neq \tau\), and thus only the gradient of the corresponding task \(\nabla_{\theta_i} L_t\) can be obtained for each \(\theta_i\). Instead we propose to balance the losses among tasks by forcing the scaled losses \(\{\alpha_t L_t\}\) to be constant for all tasks, without loss of generality, we take the constant as 1. Then the most direct idea is to compute the scaling factors as \(\{\alpha_t = 1/L_t\}\), but they are sensitive to outlier samples and manifest severe oscillations, so we further propose to learn to scale losses via gradient descent and thus stronger stability can be achieved. Suppose the positive losses \(\{L_t > 0\}\) are to be balanced, we first introduce a mapping function \(h : \R \rightarrow \R^+\) to transform the arbitrarily-ranged learnable scale parameters \(\{s_t\}\) to positive scaling factors \(\{h(s_t) > 0\}\), hereafter we abandon the subscript \(t\) for brevity. Then we should construct an appropriate scaled loss \(g(s)\) so that both network parameters \(\theta\) and scale parameter \(s\) can be optimized by minimizing \(g(s)\). On one hand, we balance different
tasks by encouraging the scaled losses $h(s) L(\theta)$ to be 1 for all tasks, so the optimality $s^*$ of $s$ is achieved when $h(s) L(\theta) = 1$, or equivalently:

$$f(s) = h(s) L(\theta) - 1 = 0, \text{ if } s = s^*.$$  \hfill (3)

One may expect to minimize $|f(s)| = |h(s) L(\theta) - 1|$ to find $s^*$, however when $h(s) L(\theta) < 1$, the gradient with respect to $\theta$, $\nabla_\theta f(s) = -h(s) \nabla g L(\theta)$, is in the opposite direction. On the other hand, assume our scaled loss $g(s)$ is a differentiable convex function with respect to $s$, then its minimum is achieved if and only if $s = s^*$, where the derivative of $g(s)$ is zero:

$$g'(s) = 0, \text{ if } s = s^*.$$  \hfill (4)

From Eq. (3) and (4) we find that the values of $f(s)$ and $g'(s)$ are both 0 when $s = s^*$, we can then regard $f(s)$ as the derivative of $g(s)$, which is our target scaled loss and used to optimize both the network parameters $\theta$ and loss scale parameter $s$, then we have:

$$g'(s) = f(s) \iff g(s) = \int f(s) \, ds = L(\theta) \int h(s) \, ds - s.$$  \hfill (5)

From Eq. (3) and (5), we notice that both $h(s)$ and $\int h(s) \, ds$ denote loss scales, so we have $\int h(s) \, ds = Ch(s)$, where $C > 0$ is a constant. According to ordinary differential equation, $\int h(s) \, ds$ must be the exponential function: $\int h(s) \, ds = ba^s$ with $a > 1, b > 0$ (see Appendix B.2). We then have $g''(s) = ka^s$, $k > 0$, which is always positive and verifies our assumption about the convexity of $g(s)$. Also note that the gradient of $g(s)$ with respect to $\theta$, $\nabla_\theta g(s) = \int h(s) \, ds \nabla g L(\theta) = ba^s \nabla g L(\theta)$, is in the appropriate direction since $ba^s > 0$. As an instantiation, we set $\int h(s) \, ds = e^s$ ($a = e, b = 1$), then

$$g(s) = e^s L(\theta) - s,$$  \hfill (IMTL-L).

From Eq. (6) we find that the raw loss is scaled by $e^s$, and $-s$ acts as a regularization to avoid the trivial solution $s = -\infty$ while minimizing the scaled loss $g(s)$. As for implementation, the task losses $\{L_t\}$ are scaled by $\{e^{s_t}\}$, and the scaled losses $\{e^{s_t} L_t - s_t\}$ are used to update both the network parameters $\theta$, $\{\theta_t\}$ and the scale parameters $\{s_t\}$.

### 3.3 Hybrid Balance: IMTL

We have introduced IMTL-G/IMTL-L to achieve gradient/loss balance, and both of them produce scaling factors to be applied on the raw losses. They can be used solely, but we find them complementary and able to be combined to improve the performance. In IMTL-G, even if the raw losses are multiplied by arbitrary (maybe different among tasks) positive factors, the direction of the aggregated gradient $g$ stays unchanged. Because by definition $g = \sum_t \alpha_t g_t$ is the angular bisector of the gradients $\{g_t\}$, and positive scaling will not change the directions of $\{g_t\}$ and thus that of $g$ (proof in Theorem 2). So we can also obtain the scale factors $\{\alpha_t\}$ in IMTL-G with the losses that have been scaled by $\{s_t\}$ from IMTL-L. IMTL-G and IMTL-L are combined as: 1) the task-specific parameters $\{\theta_t\}$ and scale parameters $\{s_t\}$ are updated by scaled losses $\{e^{s_t} L_t - s_t\}$; 2) the task-shared parameters $\theta$ are updated by $\sum_t \alpha_t (e^{s_t} L_t)$ which is the weighted average of $\{e^{s_t} L_t\}$, with the weights $\{\alpha_t\}$ computed by $\{\nabla_z (e^{s_t} L_t)\}$ using IMTL-G. Note that the regularization terms $\{-s_t\}$ in Eq. (6) are constants with respect to $\theta$ and $z$, and thus can be ignored when computing gradients and updating parameters in IMTL-G. In this way, we achieve both gradient balance for task-shared parameters and loss balance for task-specific parameters, leading to our full IMTL as illustrated in Alg. 1.

### 4 Discussion

We draw connections between our method and previous state-of-the-arts in Fig. 3. We will show that previous methods can all be categorized as gradient or loss balance, and thus each of them can be seen as a specification of our method. However, all of them have some intrinsic biases or short-comings leading to inferior performances, which we try to overcome.

---

*Our analysis of PCGrad [Yu et al., 2020] can be found in Appendix C.3.*
GradNorm \cite{Chen:2018} balances tasks by making the norm of the scaled gradient for each task to be approximately equal. It also introduces the inverse training rate and a hyper-parameter $\gamma$ to control the strength of approaching the mean gradient norm, such that tasks which learn slower can receive larger gradient magnitudes. However, it does not take into account the relationship of the gradient directions. We show that when the angle between the gradients of each pair of tasks is identical, our IMTL-G leads to the equivalent solution as GradNorm.

**Theorem 1.** If the angle between any pair of $u_t$, $u_r$ stays constant: $u_t u_r^\top = C_1$, $\forall t \neq r$ with $C_1 < 1$, then our IMTL-G leads to the same solution as that of GradNorm: $g u_t^\top = C_2 \iff u_t u_r^\top = C_3$, in the above $u_t = g_t / \|g_t\|$, $C_1$, $C_2$ and $C_3$ are constants.

Proof in Appendix C.1. In GradNorm, if without the above constant-angle condition $u_t u_r^\top = C_1$, the projection of the aggregated gradient $g$ onto task-specific gradient, $g u_t^\top = (\sum_{r} C_3 u_r) u_t^\top = C_3 (\sum_{r} u_r) u_t^\top$, is proportional to $(\sum_{r} u_r) u_t^\top$. It tends to optimize the “majority tasks” whose gradient directions are closer to the mean direction $\sum_t u_t$, resulting in undesired task bias.

MGDA \cite{Sener:2018} finds the weighted average gradient $g = \sum_t \alpha_t g_t$ with minimum norm in the convex hull composed by $\{g_t\}$, so that $\sum_t \alpha_t = 1$ and $\alpha_t \geq 0$. It adopts an iterative method based on Frank-Wolfe algorithm to solve the multi-objective optimization problem. We note the minimum-norm point has a closed-form representation if without the constraints $\{\alpha_t \geq 0\}$. In this case, we try to minimize $g g^\top = (\sum_t \alpha_t g_t) (\sum_t \alpha_t g_r)^\top$ such that $\sum_t \alpha_t = 1$. It implies $g$ is perpendicular to the hyper-plane composed by $\{g_t\}$ as illustrated in Fig. 1(b), and thus we have:

$$g \perp (g_t - g_i) \iff g (g_t - g_i)^\top = 0, \forall 2 \leq t \leq T,$$

and can obtain $\alpha = g_1 D^\top (DD^\top)^{-1}$ (see Appendix C.2). From Eq. (7), we note that the aggregated gradient satisfies: $gg_i^\top = C$. Then the projection of $g$ onto $g_i$, $g u_i^\top = C/\|g_i\|$, is inversely proportional to the norm of $g_i$. So it focuses on tasks with smaller gradient magnitudes, which breaks the task balance. Even with $\{\alpha_t \geq 0\}$, the problem still exists (see Appendix C.2) in the original MGDA method. Through experiments, we note that finding the minimum-norm point without the constraints $\{\alpha_t \geq 0\}$ leads to similar performance as MGDA with the constraints $\{\alpha_t \geq 0\}$. In our IMTL-G, although we do not constrain $\{\alpha_t \geq 0\}$, its loss weighting scales are always positive during the training procedure as shown in Fig. 4.

Uncertainty weighting \cite{Kendall:2018} regards the task uncertainty as loss weight. For regression, it can derive $L_1$ loss from Laplace distribution: $-\log p(y \mid f(x)) = |y - f(x)| / b + \log b$, where $x$ is the data sample, $y$ is the ground-truth label, $f$ denotes the prediction model and $b$ is the diversity of Laplace distribution. $L_2$ loss can be found in Appendix C.4. For classification, it takes the cross-entropy loss as a scaled categorical distribution and introduces the following approximation:

$$-\log p(y \mid f(x)) = -\log \left[ \text{softmax}_y \left( \frac{f(x)}{\sigma^2} \right) \right] \approx -\frac{1}{\sigma^2} \log \left[ \text{softmax}_y (f(x)) \right] + \log \sigma, \quad (8)$$
GLS (Chennupati et al., 2019) calculates the target loss as the geometric mean: \( \text{loss} \), which may not be optimal, but our IMTL-L has a unified form for all kinds of losses. It introduces approximations in order to derive scaling factors for certain losses (such as cross-entropy). Lastly, as far as we know, there is no appropriate correspondence between cosine similarity and specific distributions. GLS (Chennupati et al., 2019) calculates the target loss as the geometric mean: \( L = \left( \prod_i L_i \right)^{\frac{1}{T}} \), then the gradient of \( L \) with respect to the model parameters \( \theta \) can be obtained as Appendix C.5, which can be regarded as to weigh the loss with its reciprocal value. However, as the gradient depends on the value of \( L \), so it is not scale-invariant to the loss scale changes. Moreover, we find it to be unstable when the number of tasks is large because of the geometric mean computation.

We would like to remark the differences between our IMTL-L and uncertainty weighting (Kendall et al., 2018). Firstly, our derivation is motivated by the fairness among tasks, which intrinsically differs from uncertainty weighting which is based on task uncertainty considering each task independently. Secondly, IMTL-L learns to balance among tasks without any biases, while uncertainty weighting may sacrifice classification tasks to favor regression tasks as derived above. Thirdly, IMTL-L does not depend on any distribution assumptions and thus can be generally applied to various losses including cosine similarity, which uncertainty weighting may have difficulty with. As far as we know, there is no appropriate correspondence between cosine similarity and specific distributions. Lastly, uncertainty weighting needs to deal with different losses case by case, it also introduces approximations in order to derive scaling factors for certain losses (such as cross-entropy loss) which may not be optimal, but our IMTL-L has a unified form for all kinds of losses.

In previous methods, various experimental settings have been adopted but there are no extensive comparisons. As one contribution of our work, we re-implement representative methods and present fair comparisons among them under the unified code-base, where more practical settings are adopted and stronger performances are achieved compared with existing code-bases. The implementations exactly follow the original papers and open-sourced code to ensure the correctness. We run experiments on the Cityscapes (Cordts et al., 2016), NYUv2 (Silberman et al., 2012) and CelebA (Liu et al., 2015) dataset to extensively analyze different methods. Details can be found in Appendix D.

Results on Cityscapes. From Tab. 1 we can obtain several informative conclusions. The uniform scaling baseline, which naïvely adds all losses, tends to optimize tasks with larger losses and gradient magnitudes, resulting in severe task bias. Uncertainty weighting (Kendall et al., 2018) sacrifices classification tasks to aid regression ones, leading to significantly worse results on semantic segmentation compared with our IMTL-L. GradNorm (Chen et al., 2018) is very sensitive to the choice of the hyper-parameter \( \gamma \) controlling the strength of equal gradient magnitudes, where the default \( \gamma = 1.5 \) works well on NYUv2 but performs badly on Cityscapes. We find its best option is \( \gamma = 0 \) which makes the scaled gradient norm to be exactly equal. MGDA (Sener & Koltun, 2018) focuses on tasks with smaller gradient magnitudes. So the performance of semantic segmentation is good but the other two tasks have difficulty in converging. In addition, we find our proposed closed-form variant without the hard constraints \( \{\alpha_t \geq 0\} \) achieves similar results as the original iterative method. Through the experiments we notice the closed-form solution almost always yields \( \{\alpha_t \geq 0\} \). As for PCGrad (Yu et al., 2020), it yields slightly better performance than uniform scaling because its conflict projection will have no effect when the angles between the gradients are equal or less than \( \pi/2 \). In contrast, our IMTL method, in terms of both gradient balance and loss balance, yields competitive performance and achieves the best balance among tasks. Moreover, we verify that the two balances are complementary and can be combined to further improve the performance, with the visualizations in Appendix E. Surprisingly, we find our IMTL can beat the single-task baseline where
Table 1: Comparison between IMTL and previous methods on Cityscapes, semantic segmentation, instance segmentation and disparity/depth estimation are considered. The first group of columns shows the regular results of different methods. The second group shows the results by manually multiplying the semantic segmentation loss with 10 before applying these methods. The subscript numbers show the absolute change after scaling the loss to demonstrate the robustness of various methods. The arrows indicate the values are the higher the better (↑) or the lower the better (↓). The best and runner up results for each task are bold and underlined, respectively.

<table>
<thead>
<tr>
<th>method</th>
<th>sem. mIoU↑</th>
<th>ins. L_1 ↓</th>
<th>disp. L_1 ↓</th>
<th>sem. mIoU↑↑</th>
<th>ins. L_1↓</th>
<th>disp. L_1↓</th>
<th>time s/iter↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>single-task</td>
<td>76.67</td>
<td>21.61</td>
<td>4.182</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>uniform scaling</td>
<td>58.99</td>
<td>18.13</td>
<td>3.512</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.201</td>
</tr>
<tr>
<td>loss balance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>uncertainty (Kendall et al. 2018)</td>
<td>74.91</td>
<td>16.43</td>
<td>2.895</td>
<td>74.00_{0.91}</td>
<td>16.77_{0.34}</td>
<td>2.930_{0.035}</td>
<td>1.204</td>
</tr>
<tr>
<td>GLS (Chennupati et al. 2019)</td>
<td>75.65</td>
<td>17.18</td>
<td>2.953</td>
<td>66.22_{0.43}</td>
<td>21.09_{0.91}</td>
<td>3.358_{0.405}</td>
<td>1.202</td>
</tr>
<tr>
<td>IMTL-L</td>
<td>76.89</td>
<td>16.69</td>
<td>2.944</td>
<td>75.55_{1.34}</td>
<td>17.49_{0.80}</td>
<td>2.972_{0.028}</td>
<td>1.202</td>
</tr>
<tr>
<td>gradient balance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GradNorm (γ = 0)</td>
<td>76.27</td>
<td>17.99</td>
<td>3.195</td>
<td>72.96_{3.31}</td>
<td>19.36_{3.37}</td>
<td>3.216_{0.021}</td>
<td>1.741</td>
</tr>
<tr>
<td>GradNorm (Chen et al. 2018)</td>
<td>52.17</td>
<td>19.88</td>
<td>4.098</td>
<td>54.23_{2.06}</td>
<td>20.53_{0.65}</td>
<td>4.108_{0.010}</td>
<td>1.742</td>
</tr>
<tr>
<td>MGDA (α ≥ 0)</td>
<td>76.95</td>
<td>53.19</td>
<td>6.296</td>
<td>76.36_{0.59}</td>
<td>29.06_{24.13}</td>
<td>3.377_{9.19}</td>
<td>1.777</td>
</tr>
<tr>
<td>MGDA (Sener &amp; Koltun 2018)</td>
<td>76.56</td>
<td>53.14</td>
<td>6.644</td>
<td>72.35_{4.21}</td>
<td>29.38_{23.76}</td>
<td>3.336_{3.08}</td>
<td>1.732</td>
</tr>
<tr>
<td>PCGrad (Yu et al. 2020)</td>
<td>60.50</td>
<td>17.99</td>
<td>3.450</td>
<td>66.33_{5.83}</td>
<td>17.99_{0.00}</td>
<td>3.368_{0.064}</td>
<td>2.087</td>
</tr>
<tr>
<td>IMTL-G (exact)</td>
<td>76.13</td>
<td>17.46</td>
<td>2.979</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.769</td>
</tr>
<tr>
<td>IMTL-G</td>
<td>76.52</td>
<td>16.61</td>
<td>2.997</td>
<td>76.60_{0.46}</td>
<td>17.52_{0.91}</td>
<td>3.020_{0.028}</td>
<td>1.776</td>
</tr>
<tr>
<td>hybrid balance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMTL</td>
<td>77.00</td>
<td>15.96</td>
<td>2.905</td>
<td>76.56_{0.44}</td>
<td>15.85_{0.11}</td>
<td>2.938_{0.033}</td>
<td>1.795</td>
</tr>
</tbody>
</table>

In addition, we present the real-world training time of each iteration for different methods in Tab. 1. As shown, loss balance methods are the most efficient, and our gradient balance method IMTL-G adds acceptable computational overhead, similar to that of GradNorm (Chen et al. 2018) and MGDA (Sener & Koltun 2018). It benefits from computing gradients with respect to the shared feature maps instead of the shared model parameters (the row of “IMTL-G (exact)”), which brings similar performances but adds significant complexity due to multiple (T) backward passes through the shared parameters. Our IMTL-G only needs to do backward computation on the shared parameters once after obtaining the loss weights via Eq. (2), in which the computation overhead mainly comes from the matrix multiplication rather than the matrix inverse, since the inverted matrix \( DU^\top \in \mathbb{R}^{(T-1)\times(T-1)} \) is small compared with dimension of the shared feature \( z \).

As we outperform MGDA (Sener & Koltun 2018) and PCGrad (Yu et al. 2020) significantly in terms of the objective metrics shown in Tab. 1, we further compare the qualitative results of our hybrid balance IMTL with the loss balance method uncertainty weighting (Kendall et al. 2018) and the gradient balance method GradNorm (Chen et al. 2018) considering their strong performances (see Fig. 6). For depth estimation we only show predictions at the pixels where ground truth (GT) labels exist to compare with GT, which is different from Fig. 7 where depth predictions are shown for all pixels. Consistent with results in Tab. 1, our IMTL shows visually noticeable improvements especially for the semantic and instance segmentation tasks. It is worth noting that we conduct experiments under strong baselines and practical settings which are seldom explored before, in this case changing the backbone in PSPNet (Zhao et al. 2017) from ResNet-50 to ResNet-101 can only improve mIoU of the semantic segmentation task around 0.5% according to the public code base.

Scale invariance. We are also interested in the scale invariance, which means how the results change with the loss scale. For example, in semantic segmentation, the loss scale is different if we replace the reduction method “mean” (averaged over all locations) with “sum” (summed over all locations) in the cross-entropy loss computation, or the number of the interested classes increases. The scale invariance is beneficial for model robustness. So to simulate this effect, we manually multiply the semantic segmentation loss by 10 and apply the same methods to see how the performances are affected. In the last three columns of Tab. 1 we report the absolute changes resulting from the
Table 2: Experimental results on the NYUv2 and CelebA datasets, semantic segmentation, surface normal estimation, depth estimation and multi-class classification are considered. Arrows indicate the values are the higher the better (↑) or the lower the better (↓). The best and runner up results in each column are bold and underlined, respectively.

<table>
<thead>
<tr>
<th>method</th>
<th>NYUv2</th>
<th>CelebA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mIoU↑</td>
<td>cos↑</td>
</tr>
<tr>
<td>baselines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>single-task</td>
<td>56.82</td>
<td>0.8827</td>
</tr>
<tr>
<td>uniform scaling</td>
<td>57.40</td>
<td>0.8684</td>
</tr>
<tr>
<td>loss balance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uncertainty (Kendall et al., 2018)</td>
<td>57.20</td>
<td>0.4400</td>
</tr>
<tr>
<td>GLS (Chennupati et al., 2019)</td>
<td>57.84</td>
<td>0.8762</td>
</tr>
<tr>
<td>IMTL-L</td>
<td>58.36</td>
<td>0.8864</td>
</tr>
<tr>
<td>gradient balance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GradNorm (γ = 0)</td>
<td>55.96</td>
<td>0.8818</td>
</tr>
<tr>
<td>GradNorm (Chen et al., 2018)</td>
<td>56.92</td>
<td>0.8787</td>
</tr>
<tr>
<td>MGDA (w/o {αt ≥ 0})</td>
<td>49.43</td>
<td>0.8877</td>
</tr>
<tr>
<td>MGDA (Sener &amp; Koltun, 2018)</td>
<td>49.44</td>
<td>0.8875</td>
</tr>
<tr>
<td>PCGrad (Yu et al., 2020)</td>
<td>57.48</td>
<td>0.8696</td>
</tr>
<tr>
<td>IMTL-G</td>
<td>57.00</td>
<td>0.8785</td>
</tr>
<tr>
<td>hybrid balance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMTL</td>
<td><strong>58.85</strong></td>
<td><strong>0.8888</strong></td>
</tr>
</tbody>
</table>

multiplier. Our IMTL achieves the smallest performance fluctuations and thus the best invariance, while other methods are more or less affected by the loss scale change.

Results on NYUv2. In Tab. 2 we find similar patterns as on Cityscapes, but NYUv2 is a rather small dataset, so uniform scaling can also obtain reasonable results. Note that uncertainty weighting (Kendall et al., 2018) cannot be directly used to estimate the normal surface when the cosine similarity is used as the loss, since no appropriate distribution can be found to correspond to cosine similarity. In this case, surface normal estimation owns the smallest gradient magnitude, so MGDA (Sener & Koltun, 2018) learns it best but it performs not so well for the rest two tasks. Again, our IMTL performs best taking advantage of the complementary gradient and loss balances.

Results on CelebA. To compare different methods in the many-task setting, in Tab. 2 we also conduct the multi-label classification experiments on the CelebA (Liu et al., 2015) dataset. The mean accuracy of 40 tasks is used as the final metric. Our IMTL outperforms its competitors in the scenario where the task number is large, showing its superiority. Note that in this setting, GLS (Chennupati et al., 2019) has difficulty in converging and no reasonable results can be obtained.

6 Conclusion

We propose an impartial multi-task learning method integrating gradient balance and loss balance, which are applied on task-shared and task-specific parameters, respectively. Through our in-depth analysis, we have theoretically compared our method with previous state-of-the-arts. We have also showed that those state-of-the-arts can all be categorized as gradient or loss balance, but lead to specific bias among tasks. Through extensive experiments we verify our analysis and demonstrate the effectiveness of our method. Besides, for fair comparisons, we contribute a unified code-base, which adopts more practical settings and delivers stronger performances compared with existing code-bases, and it will be publicly available for future research.

Acknowledgements

This work was supported by the Natural Science Foundation of Guangdong Province (No. 2020A1515010711), the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (No. JCYJ20200109143010272), and the Innovation and Technology Commission of the Hong Kong Special Administrative Region, China (Enterprise Support Scheme under the Innovation and Technology Fund B/E030/18).
REFERENCES


Theorem 2. Given \( g = \sum \alpha_i g_i \), \( \sum \alpha_i = 1 \) satisfying \( gu^T = C \), when \( \{ L_t \} \) are scaled by \( \{ k_t \} > 0 \) (equivalently, \( \{ g_i \} \) are scaled by \( \{ k_t \} \)), if \( g' = \sum \alpha'_i (k_t g_i) \), \( \sum \alpha'_i = 1 \) satisfies \( g'u^T = C' \), then \( g' = \lambda g \). In the above we have \( u_t = \frac{g_t}{\| g_t \|} \). 

A RELATED WORK OF NETWORK STRUCTURE

Cross-stitch Networks [Misra et al., 2016] learn coefficients to linearly combine activations from multiple tasks to construct better task-specific representations. To break the limitation of channel-wise cross-task feature fusion only, NDDR-CNN [Gao et al., 2019] proposes the layer-wise cross-channel feature aggregation as \( 1 \times 1 \) convolutions on the concatenated feature maps from multiple tasks. More generally, MTL-NAS [Gao et al., 2020] introduces cross-layer connections among tasks to fully exploit the feature sharing from both low and high layers, extending the idea in Slice Networks [Ruder et al., 2019] by leveraging neural architecture search [Zoph & Le, 2017]. The parameters of these methods increase linearly with the number of tasks. To improve the model compactness, Residual Adapters [Kebabiti et al., 2017] introduce a small amount of task-specific parameters for each layer and convolve them with the task-agnostic representations to form the task-related ones. MTAN [Liu et al., 2019] generates data-dependent attention tensors by task-specific parameters to attend to the task-shared features. Single-tasking [Maninis et al., 2019] instead applies a squeeze-and-excitation [Hu et al., 2018] module to generate attentive vectors for each task. In Task Routing [Strezoski et al., 2019], the attentive vectors are randomly sampled before training and are fixed for each image. Piggyback [Mallya et al., 2018] opts to mask parameter weights in place of the channel feature aggregation as
Proof. As we have:
\[ g = \sum_t \alpha_t \mathbf{g}_t = \sum_t \frac{\alpha_t}{k_t} \mathbf{k}_t \mathbf{g}_t \quad \text{and} \quad \mathbf{g}_t^\top \mathbf{u}_t = C, \]  
(13)
by constructing:
\[ \alpha'_t = \frac{\alpha_t}{k_t} \sum_{\tau} \frac{\alpha_{\tau}}{k_{\tau}} \quad \text{and} \quad \mathbf{g}' = \sum_t \alpha'_t (k_t \mathbf{g}_t) = \mathbf{g} / \sum_{\tau} \frac{\alpha_{\tau}}{k_{\tau}} = \lambda \mathbf{g}, \]  
(14)
we have:
\[ \sum_t \alpha'_t = 1 \quad \text{and} \quad \mathbf{g}'^\top \mathbf{u}_t = C / \sum_{\tau} \frac{\alpha_{\tau}}{k_{\tau}} = C'. \]  
(15)
From Eq. (12) we know that \{\alpha_t\} has a unique solution, and thus \mathbf{g}' satisfying IMTL-G is unique, so it must be the one given by Eq. (14), then we can prove that \mathbf{g}' and \mathbf{g} are linearly correlated.

B.2 LOSS BALANCE: IMTL-L

With the ordinary differential equation, we can derive that the form of the scale function \( \int h(s) \, ds \) in our IMTL-L must be exponential function. As we have:
\[ \int h(s) \, ds = C h(s), \ C > 0. \]  
(16)
If we set \( y = \int h(s) \, ds \), then:
\[ y = C \frac{dy}{ds} \Rightarrow \frac{dy}{y} = \frac{1}{C} ds, \]  
(17)
By taking the antiderivative:
\[ \int \frac{dy}{y} = \frac{1}{C} \int ds \Rightarrow \ln y = \frac{1}{C} s + C'. \]  
(18)
Then we have:
\[ \int h(s) \, ds = y = e^{C'} \left( e^{\frac{a}{b}} \right)^s = ba^s, \ a > 1, \ b > 0. \]  
(19)

C DETAILED DISCUSSION

C.1 CONDITIONAL EQUIVALENCE OF IMTL-G AND GRADNORM

First we introduce the following lemma.

Lemma 3. If \( \mathbf{u}_t \mathbf{u}_t^\top = C_1, \ \forall t \neq \tau \), then the solution \{\alpha_t\} of IMTL-G satisfies \{\alpha_t > 0\}.
Proof. As \( u_t = g_t / \| g_t \| \), by constructing \( g = \sum_\tau \alpha_\tau g_\tau \) where:

\[
\alpha_\tau = \frac{\| g_\tau \|^{-1}}{\sum_\tau \| g_\tau \|^{-1}},
\] (20)

then we have \( \sum_\tau \alpha_\tau = 1 \) and:

\[
gu_t^\top = \left( \sum_\tau u_\tau u_\tau^\top \right) / \sum_\tau \| g_\tau \|^{-1} = [(T - 1) C_1 + 1] / \sum_\tau \| g_\tau \|^{-1} = C_2.
\] (21)

From Eq. (12) we know the solution \( \{ \alpha_\tau \} \) of IMTL-G is unique, so it must be the one given by Eq. (20) where \( \{ \alpha_\tau > 0 \} \), so the lemma is proved.

Then we prove Theorem 1 which states that IMTL-G leads to the same solution as GradNorm when the angle between any pair of gradients \( \{ g_\tau \} \) is identical: \( u_t u_\tau^\top = C_1, \forall t \neq \tau \).

Proof. (\( \Rightarrow \) Necessity) Given constant projections in IMTL-G, we have:

\[
gu_t^\top = \left( \sum_\tau \alpha_\tau g_\tau \right) u_t^\top = C_2.
\] (22)

Recall that \( u_t = g_t / \| g_t \| \) and \( u_t u_\tau^\top = C_1, \forall t \neq \tau \). From Lemma 3 we know that \( \{ \alpha_\tau \} \) given by IMTL-G must satisfy \( \{ \alpha_\tau > 0 \} \). If we assume \( n_t = \| \alpha_t g_t \| \), then we know \( \alpha_t g_t = n_t u_t \) and:

\[
\sum_\tau n_\tau u_\tau u_\tau^\top = \sum_\tau n_\tau C_1 + n_t = C_2.
\] (23)

Now we obtain:

\[
\sum_{\tau \neq t} n_\tau C_1 + n_t = \sum_\tau n_\tau C_1 + (1 - C_1) n_t = C_2.
\] (24)

As \( C_1 < 1 \), we can then prove \( n_t = C_3, \forall t \). It implies the norm of the scaled gradient is constant, which is requested by GradNorm [Chen et al. 2018]. Moreover, we can obtain the relationship among constants from Eq. (24):

\[
C_1 TC_3 + (1 - C_1) C_3 = C_2 \Rightarrow C_3 = \frac{C_2}{(T - 1) C_1 + 1}.
\] (25)

(\( \Leftarrow \) Sufficiency) In GradNorm, \( \{ \alpha_\tau \} \) are always chosen to satisfy \( \{ \alpha_\tau > 0 \} \), so if we assume \( n_t = \| \alpha_t g_t \| \), then given the constant norm of the scaled gradient in GradNorm, we have:

\[
\alpha_t g_t = n_t u_t = C_3 u_t,
\]

where \( u_t = g_t / \| g_t \| \). As we have \( g = \sum_\tau \alpha_\tau g_\tau \) and \( u_\tau u_\tau^\top = C_1, \forall t \neq \tau \), then we obtain:

\[
gu_t^\top = \left( \sum_\tau \alpha_\tau g_\tau \right) u_t^\top = \left( \sum_\tau C_3 u_\tau \right) u_t^\top = C_3 [(T - 1) C_1 + 1] = C_2.
\] (27)

It means the projections of \( g \) onto \( \{ g_\tau \} \) is constant, which is requested by our IMTL-G.

\( \square \)

Corollary 4. In GradNorm, if the solution \( \{ \alpha_\tau \} \) satisfies \( \sum_\tau \alpha_\tau = 1 \), then its constants are given by \( C_3 = 1 / \sum_\tau \| g_\tau \|^{-1} \) and \( C_2 = [(T - 1) C_1 + 1] / \sum_\tau \| g_\tau \|^{-1} \), and its scaling factors are given by \( \{ \alpha_t = \| g_t \|^{-1} / \sum_\tau \| g_\tau \|^{-1} \} \).

Proof. By using \( \alpha_t = C_3 / \| g_t \| \) from Eq. (26), we have \( \sum_\tau C_3 / \| g_\tau \| = 1 \), then \( C_3 = 1 / \sum_\tau \| g_\tau \|^{-1} \), and also we have \( \alpha_t = \| g_t \|^{-1} / \sum_\tau \| g_\tau \|^{-1} \). As the relationship of \( C_2 \) and \( C_3 \) from Eq. (27) is given by \( C_3 [(T - 1) C_1 + 1] = C_2 \), so \( C_2 = [(T - 1) C_1 + 1] / \sum_\tau \| g_\tau \|^{-1} \). \( \square \)
C.2 Closed-form Solution of MGDA

In our relaxed MGDA [Sener & Koltun, 2018] without \( \{\alpha_t \geq 0\} \), finding \( g = \sum_t \alpha_t g_t \) with \( \sum_t \alpha_t = 1 \) such that \( g \) has minimum norm is equivalent to find the normal vector of the hyperplane composed by \( \{g_t\} \). So we let \( g \) to be perpendicular to all of \( \{g_1 - g_t\} \) on the hyperplane:

\[
g \perp (g_1 - g_t) \iff g (g_1 - g_t) = 0, \forall 2 \leq t \leq T. \tag{28}
\]

If we set \( \alpha = [\alpha_2, \ldots, \alpha_T] \) and \( G^\top = [g_2^\top, \ldots, g_T^\top] \), then we have \( \alpha_1 = 1 - 1\alpha^\top \), and Eq. (28) can be expanded as:

\[
(\sum_t \alpha_t g_t) \begin{bmatrix} g_1^\top - g_2^\top, \ldots, g_1^\top - g_T^\top \end{bmatrix} = 0 \iff \begin{bmatrix} 1 - 1\alpha^\top \end{bmatrix} \begin{bmatrix} g_1^\top \\ G \end{bmatrix} D^\top = 0, \tag{29}
\]

where \( D^\top = [g_1^\top - g_2^\top, \ldots, g_1^\top - g_T^\top] \). 1 and 0 indicates the all-one and all-zero row vector. Eq. (29) can be represented as:

\[
[(1 - 1\alpha^\top) g_1 + \alpha G] D^\top = 0 \iff \alpha (1 g_1 - G) D^\top = g_1 D^\top.
\]

As we also have \( D = 1^\top g_1 - G \), then the closed-form solution of \( \alpha \) is given by:

\[
\alpha DD^\top = g_1 D^\top \iff \alpha = g_1 D^\top (DD^\top)^{-1}. \tag{30}
\]

**Bias of MGDA.** In the main text we state that MGDA focuses on tasks with small gradient magnitudes, where we relaxed MGDA by not constraining \( \{\alpha_t \geq 0\} \). However, even with these constraints, the problem still exists. For example in the context of two tasks, assume \( g^* \) satisfying \( g = \alpha g_1 + (1 - \alpha) g_2 \) is outside the convex hull composed by \( \{g_1, g_2\} \), or equivalently \( \alpha > 1 \), MGDA clamps \( \alpha \) to \( \alpha = 1 \) and the optimal \( g^* = g_1 \). Then the projections of \( g^* \) onto \( g_1 \) and \( g_2 \) will be \( ||g_1|| \) and \( g_1 u_2^\top \) (\( u_2 = g_2/||g_2|| \)), respectively. As \( ||g_1|| > ||g_1 u_2^\top|| \), so MGDA still focuses on tasks with smaller gradient magnitudes.

C.3 Analysis of PCGrad

PCGrad [Yu et al., 2020] mitigates the gradient conflicts by projecting the gradient of one task to the orthogonal direction of the others, and the aggregated gradient can be written as:

\[
g = \sum_t \left( g_t + \sum_{\tau} C_{t\tau} u_\tau \right), \tag{31}
\]

with \( u_t = g_t/||g_t|| \) and the coefficients:

\[
C_{t\tau} = 0, \quad C_{t\tau} = \left[ - \left( g_t + \sum_{t' < \tau} C_{t't'} u_{t'} \right) u_\tau^\top \right]_+, \quad \forall t, \tau, \tag{32}
\]

where \([.]_+\) means the ReLU operator. Note that the tasks have been shuffled before calculating the aggregated gradient \( g \) to achieve expected symmetry with respect to the task order. Eq. (31) can be represented more compactly in the matrix form:

\[
g = 1 \left( I_T + CN \right) G = \alpha G, \tag{33}
\]

where \( I_T \) is the identity matrix, \( C = \{C_{t\tau}\} \) is the coefficient matrix whose entries are given in Eq. (32) and \( N = \text{diag} \left( 1/||g_1||, \ldots, 1/||g_T|| \right) \) is the diagonal normalization matrix. In Eq. (33) we use \( G \) and \( \alpha \) to denote the raw gradients and scaling factors of all tasks. We find that PCGrad can also be regarded as loss weighting, with the loss weights given by \( \alpha = 1 \left( I_T + CN \right) \). However, it still may break the balance among tasks. For example with two tasks, assume the angle between
the gradients is $\phi$: 1) when $\pi/2 \leq \phi < \pi$, then $C = \begin{bmatrix} 0 & -g_1 g_2^\top / \|g_2\| \\ -g_1 g_2^\top / \|g_1\| & 0 \end{bmatrix}$ and the projections onto the two raw gradients are $\|g_1\| \sin^2 \phi$ and $\|g_2\| \sin^2 \phi$; 2) when $0 < \phi < \pi/2$, then $C = 0$ and the projections are $\|g_1\| \cos \phi$ and $\|g_2\| \cos \phi$. In both cases, the projections are equal if and only if $\|g_1\| = \|g_2\|$. Otherwise, the task with larger gradient magnitude will be trained more sufficiently, which may encounter the same problem as uniform scaling that naively adds all the losses despite that the loss scales are highly different.

\section{L$_2$ Loss in Uncertainty Weighting}

For regression, uncertainty weighting \cite{Kendall et al. 2018} regards the L$_2$ loss as likelihood estimation on the sample target which follows the Gaussian distribution:

$$-\log p(y \mid f(x)) = \frac{1}{2} \left( \frac{1}{\sigma^2} \|y - f(x)\|^2_2 + \log \sigma^2 \right), \quad (34)$$

where $x$ is the data sample, $y$ is the ground-truth label, $f$ denotes the prediction model and $\sigma$ is the standard deviation of Gaussian distribution. By setting $s = -\log \sigma^2$, the scaled L$_2$ loss is $L = \frac{1}{2} (e^s L_{\text{reg}} - s)$, which has a similar form as the scaled L$_1$ loss except the front factor 1/2. So uncertainty weighting has difficulty in reaching a unified form for all kinds of losses, which is less general than our IMTL-L.

\section{Gradient of Geometric Mean}

GLS \cite{Chennupati et al. 2019} computes the loss as the geometric mean, its gradient with respect to model parameters are:

$$\nabla_\theta L = \frac{1}{T} \left( \prod_i L_i \right)^{\frac{1}{T} - 1} \sum_i \left[ \prod_{\tau \neq i} L_{\tau} \right] \nabla_\theta L_i \quad (35)$$

$$= \frac{1}{T} \left( \prod_i L_i \right)^{\frac{1}{T}} \sum_i \frac{\nabla_\theta L_i}{L_i} = L \sum_i \frac{1}{L_i} \left( \nabla_\theta L_i \right). \quad (36)$$

where $L$ is the geometric mean loss and $T$ is the task number. It is equivalent to weigh the task-specific loss with its reciprocal value, except that there exists another term $L/T$ in the front where $L = (\prod_i L_i)^{\frac{1}{T}}$, so GLS is sensitive to the loss scale changes of $\{L_i\}$ and not scale-invariant.

\section{Implementation Details}

To solely compare the loss weighting methods, we fix the network structure and choose ResNet-50 \cite{He et al. 2016} with dilation \cite{Chen et al. 2017} and synchronized \cite{Peng et al. 2018} batch normalization \cite{Ioffe & Szegedy 2015} as the shared backbone and PSPNet \cite{Zhao et al. 2017} as the task-specific head, and the backbone model weights are pretrained on ImageNet \cite{Deng et al. 2009}. Following the common practice of semantic segmentation, in training we adopt augmentations as random resize (between 0.5 to 2), random rotate (between -10 to 10 degrees), Gaussian blur (with a radius of 5) and random horizontal flip. Besides, we apply strided cropping and horizontal flipping as testing augmentations. The predicted results in the overlapped region of different crops are averaged to obtain the aggregated prediction of the whole image. Only pixels with ground truth labels are included in loss and metric computation, while others are ignored. Semantic segmentation, instance segmentation, surface normal estimation and disparity/depth estimation are considered. As for the losses/metrics, semantic segmentation uses cross-entropy/MIoU, surface normal estimation adopts $(1 - \cos)$/cosine similarity and both instance segmentation and disparity/depth estimation use $L_1$ loss. We use polynomial learning rate with a power of 0.9, SGD with a momentum of 0.9 and weight decay of $10^{-4}$ as the optimizer, with the model trained for 200 epochs. After passing through the shared backbone where strided convolutions exist, the feature maps have 1/8 size as that of the
input image. Then the results predicted by PSPNet (Zhao et al., 2017) heads are up-sampled to the original image size for loss and metric computation.

For the Cityscapes dataset, the batch size is 32 (2 × 16 GPUs) with the initial learning rate 0.02. We train on the 2975 training images and validate on the 500 validation images (1024 × 2048 full resolution) where ground truth labels are provided. Three tasks are considered, namely semantic segmentation, instance segmentation and disparity/depth estimation. Training and testing are done on 713 × 713 crops. Semantic segmentation is to differentiate among the commonly used 19 classes. Instance segmentation is taken as offset regression, where each pixel \( p_i = (x_i, y_i) \) approximates the relative offset \( o_i = (dx_i, dy_i) \) with respect to the centroid \( c_{id(p_i)} \) of its belonging instance \( id(p_i) \).

To conduct inference, we abandon the time-consuming and complicated clustering methods adopted by the previous method (Kendall et al., 2018). Instead, we directly use the offset vectors \( \{o_i\} \) predicted by the model to find the centroids of instances. By definition, the norm of a centroid’s offset vector should be 0, so we can transform the offset vector norm \( ||o_i|| \) to the probability \( q_i \) of being a centroid with the exponential function \( q_i = e^{-||o_i||} \). Next a 7 × 7 edge filter is applied on the centroid probability map to filter out the spurious centroids on object edges resulting from the regression target ambiguity. The locations with centroid probability \( q_i < 0.1 \) are also manually suppressed. Then 7 × 7 max-pooling on the filtered probability map is used to produce candidate centroids and filter out duplicate ones. With the predicted centroids \( \{c_i\} \), we can then assign each pixel \( p_i \) to its belonging instance \( id(p_i) \) by the distance between its approximated centroids \( p_i + o_i \) and the candidate centroids \( \{c_j\} \): \( id(p_i) = \arg\min_j ||p_i + o_i - c_j|| \). Depth is measured in pixels by the disparity between the left and right images. Fig. 5 shows the whole process. Note that we need to carefully deal with label transformation during data augmentation. For example, disparity ground truth needs to be up-scaled by \( s \) times if the image is up-sampled by \( s \) times. Also, the predicted offset vectors of the flipped input should be mirrored to comply with the normal one.

On the NYUv2 dataset, the batch size is 48 (6 × 8 GPUs) with the initial learning rate 0.03. We use the 795 training images for training and the 654 validation images for testing with 480 × 640 full resolution. 401 × 401 crops are used for training and testing. 13 coarse-grain classes are considered in semantic segmentation. The surface normal is represented by the unit normal vector of the corresponding surface. When doing data augmentation, surface normal ground truth \( n = (x, y, z) \) should be processed accordingly. If we resize the image by \( s \) times, the \( z \) coordinate of the normal vector should be scaled by \( s \) and renormalized: \( n' = (x, y, sz) / \| (x, y, sz) \| \). If the image is rotated by the rotation matrix \( R \), the normal vector should also be in-plane rotated \( (x', y') = (x, y) R^T \) with \( z \) unchanged. Moreover, the left-right flip should be applied on the normal vector \( n' = (−x, y, z) \) when mirroring the image horizontally. During testing, the normal vectors in the overlapped region of crops are averaged and renormalized to produce the aggregated results. Depth is the absolute distance to the camera and measured by meters, which is inverse-proportional to the disparity measurement adopted by Cityscapes. So the depth in meters needs to be scaled by \( 1/s \) when the image is scaled by \( s \) times, which is the reciprocal of disparity transformation.

CelebA contains 202,599 face images from 10,177 identities, where each image has 40 binary attribute annotations. We train on the 162,770 training images and test on the 19,867 validation images.
images. Most of the implementation details are the same as those on the Cityscapes dataset, except that: 1) we employ the ResNet-18 as the backbone and linear classifiers as the task-specific heads, so totally 40 heads are attached on the backbone; 2) the binary-cross entropy is used as the classification loss for each attribute; 3) the batch size is 256 (32 × 8 GPUs) and the model is trained from scratch for 100 epochs; 4) the input image has been aligned with the annotated 5 landmarks and cropped to 218 × 178.

E Qualitative Results
Figure 6: Qualitative comparisons between our IMTL and previous methods on Cityscapes.
Figure 7: Qualitative results of our IMTL on Cityscapes. Semantic segmentation, instance segmentation and disparity estimation predictions are produced by a single network. The task-shared backbone is ResNet-50 and the task-specific heads are PSPNet. The image resolution is $1024 \times 2048$. 
Figure 8: Qualitative results of our IMTL on NYUv2. Semantic segmentation, surface normal estimation and depth estimation predictions are produced by a single network. The task-shared backbone is ResNet-50 and the task-specific heads are PSPNet. The image resolution is 480 × 640.