This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FULLER: Unified Multi-modality Multi-task 3D Perception via Multi-level Gradient Calibration

Zhijian Huang[1*] Sihao Lin[2*] Guiyu Liu[3*] Mukun Luo[4] Chaoqiang Ye[3] Hang Xu[3]
Xiaojun Chang[5,6] Xiaodan Liang[1,6†]
[1]Shenzhen Campus of Sun Yat-sen University [2]RMIT University [3]Huawei Noah's Ark Lab
[4]Shanghai Jiao Tong University [5]University of Technology Sydney [6]MBZUAI
huangzhj56@mail2.sysu.edu.cn, {linsihao6,guiyuliou,chromexbjxh,cxj273,xdliang328}@gmail.com,
luomukun@sjtu.edu.cn, yechaoqiang@huawei.com

Abstract

Multi-modality fusion and multi-task learning are becoming trendy in 3D autonomous driving scenario, considering robust prediction and computation budget. However, naively extending the existing framework to the domain of multi-modality multi-task learning remains ineffective and even poisonous due to the notorious modality bias and task conflict. Previous works manually coordinate the learning framework with empirical knowledge, which may lead to sub-optima. To mitigate the issue, we propose a novel yet simple multi-level gradient calibration learning framework across tasks and modalities during optimization. Specifically, the gradients, produced by the task heads and used to update the shared backbone, will be calibrated at the backbone's last layer to alleviate the task conflict. Before the calibrated gradients are further propagated to the modality branches of the backbone, their magnitudes will be calibrated again to the same level, ensuring the downstream tasks pay balanced attention to different modalities. Experiments on large-scale benchmark nuScenes demonstrate the effectiveness of the proposed method, e.g.*, an absolute 14.4% mIoU improvement on map segmentation and 1.4% mAP improvement on 3D detection, advancing the application of 3D autonomous driving in the domain of multi-modality fusion and multi-task learning. We also discuss the links between modalities and tasks.*

1. Introduction

3D perception task plays an important role in autonomous driving. Previous works are mainly developed on single modality [44, 20, 16, 43, 7, 35, 21, 41, 27, 28] and different perception tasks are separated into individual models [2, 19, 18, 45, 5]. Often we wish to leverage complementary modalities to produce robust prediction and integrate multiple tasks within a model for the sake of computation budget. For instance, with the development of hardware, it is affordable to deploy both LiDAR and camera on a car, which are responsible to provide spatial information and semantic information. Integrating semantic-complementary vision tasks within a framework would greatly facilitate the deployment of real-world application [3].

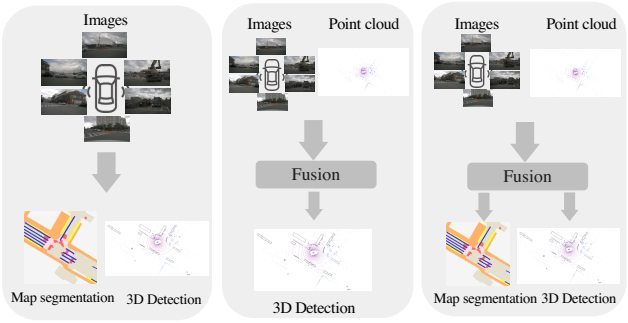


Figure 1. **Comparison of paradigms on 3D perception.** (a) BEVFormer [21] focuses on multi-task learning, which could save the computation burden and thus facilitate the depolyment of real-world application. (b) Transfusion [2] is proposed for multi-modality fusion for robust prediction since point cloud and images are complementary. (c) Our proposed Fuller is a unified framework that integrates these ingredients organically by solving the notorious problems of modality bias and task conflict.

Recent advances have stayed tuned for multi-modality fusion [30, 23] and multi-task learning [47, 21] in the applications of 3D autonomous driving scenario. Meanwhile, it is of great interest to unify multi-modality fusion and multi-task learning within a framework. In fact, it is unlikely to expect that dumping all the individual components into one framework and they would function organically. We build up a competitive baseline based on BEVFusion [30], which

*Equal contribution.
†Corresponding author.

takes as input both the point cloud and image, and serves for two complementary vision tasks: 3D detection (foreground) and map segmentation (background). However, we observe the severe issues of modality bias and task conflict: a) different tasks prefer specific modality, *e.g.*, 3D detection relies on spatial information provided by LiDAR sensor while segmentation task relies more on image inputs. b) adding a new task will degrade both tasks: -3.0 % mAP for detection and -18.3% mIoU for map segmentation.

From the perspective of optimization, we investigate the potential gradient imbalance that occurs during end-to-end training in a hierarchical view. First, we study the gradients which are produced by different task heads and are applied to update the parameters of the shared backbone. We observe that simply summing up these raw gradients to update the shared backbone would damage the performance of both tasks, suggesting an imbalance between them. Empirical findings prove that there is a great discrepancy between the gradient magnitudes w.r.t. the task objectives. Second, we inspect the gradients produced in the intra-gradient layer, which is to be separated into successive modality branches. Given a trained baseline, we visualize the gradient distributions of different modality branches and find their magnitudes imbalanced greatly. We further calculate the task accuracy by dropping one of the modalities to measure the modality bias. Our findings catch up with the theoretical analysis of [40], which suggests that the point cloud and image branches are suffering from the imbalanced convergence rate w.r.t. the downstream tasks.

We motivate our method by noting the findings discussed above, which is proposed to uni**f**y m**u**lti-modality mu**l**ti-task 3D perception via mu**l**ti-level gradi**e**nt calib**r**ation, dubbed as ***Fuller***. Specifically, we devise the multi-level gradient calibration, comprised of inter-gradient and intra-gradient calibration, to address the associated issues. In terms of the task conflict, we find that the task with lower gradient magnitude would be overwhelmed by another task with higher gradient magnitude. Thus, we propose to calibrate the gradients of different task losses at the backbone. Since the gradient would be manipulated at the layer level, this technique is referred to as **inter-gradient** calibration. Regarding modality bias, we expect the different modalities can update and converge at the same pace. Hence, before the gradients are separated into the modality branches, we calibrate their magnitudes to the same level, which is performed in the intra-gradient layer internally and thus called **intra-gradient** calibration.

On top of the gradient calibration, we introduce two lightweight heads for our tasks. These two heads are both transformer-based. With our specially designed initialization methods, they can generate fine-grained results with just a one-layer decoder, allowing to save much more parameters than dense heads.

We thoroughly evaluate the Fuller on the popular benchmark nuScenes [3]. Regarding the sensory input, we adopt the point cloud to provide accurate spatial information and use the image to compensate for the lack of visual semantics. In terms of perception tasks, we select two representative and complementary tasks: 3D detection and map segmentation, which are responsible for dynamic foreground objects and static road elements understanding. Note that BEVFusion [30] only organizes these ingredients empirically without mentioning the problems discussed above. To summarize, our contributions are:

- We propose the Fuller which organically integrates multi-modality fusion and multi-task learning for 3D perception via multi-level gradient calibration during end-to-end optimization.
- We introduce the new architecture design for task heads, which outperforms or is comparable with the previous head design while saving $\sim$40% parameters.
- Extensive experiments demonstrate that Fuller can prevent the notorious problems of modality bias and task conflict, *e.g.*, an absolute 14.4% mIoU improvement on map segmentation and 1.4% mAP improvement on 3D detection.

2. Related Work

3D perception tasks in autonomous driving. Lidar and image are the two most powerful and widely used modalities in the area of autonomous driving. Multimodal fusion has been well-studied to boost the performance of 3D object detection task[38, 45, 5, 2, 23]. Multi-task networks of 3D perception also arouse significant interest in autonomous driving community. These multi-task studies are limited on uni-modal network architectures, either with a Lidar backbone[16, 43, 7] or an image backbone[35, 21, 41, 27, 28]. MMF[22] works on depth completion and object detection with both camera and LiDAR inputs, but depth estimation only works as an auxiliary head and only object detection was evaluated. BEVFusion[30] is the first multimodal network to perform object detection and map segmentation simultaneously. However, BEVFusion[30] focuses on single task and network acceleration, and only provides two pieces of joint training results. Our proposed method is the first multimodal multitask network, and we evaluate each task and analyze them from the perspectives of multimodal and multitask.

Multimodal learning. Multimodal learning is increasingly used to improve the performance of certain tasks, such as action recognition[8, 14, 15], visual question answering [1, 13] and perception tasks in autonomous driving[30, 2, 23]. Most multi-modality research focuses on the network structure, such as concatenation, convolution or gated fusion in the middle or later part of the network[17, 33, 12].

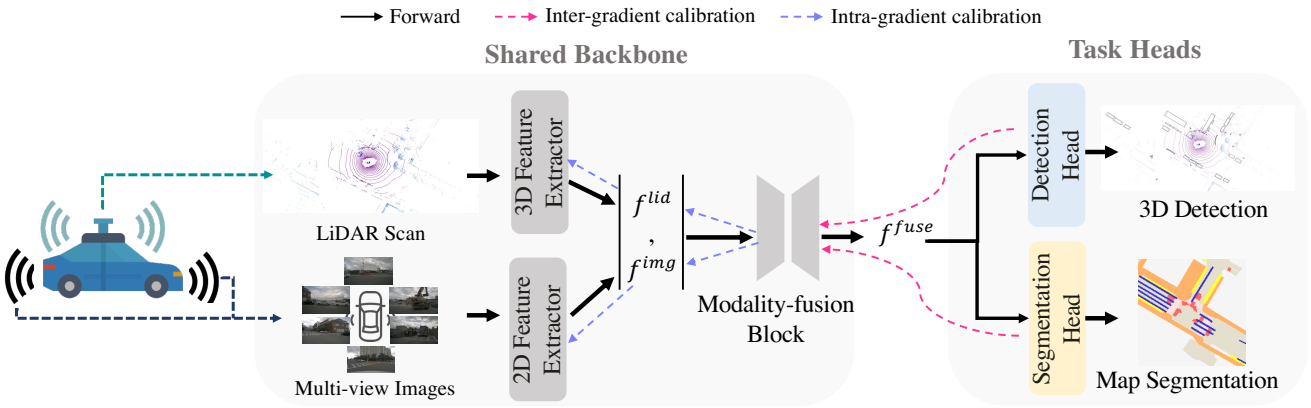


Figure 2. **Framework of the Fuller.** Generally, Fuller takes as input the LiDAR scan and multi-view images and predicts two tasks: 3D detection and map segmentation. We propose multi-level gradient calibration to deal with the problems of task conflict and modality bias during optimization: i) The gradients, produced by the task heads and applied on the shared backbone, will be calibrated on the last layer of the backbone, namely, inter-gradient calibration (pink dashed line). ii) When it comes to the subsequent modality branches of the shared backbone, the gradient magnitudes will be calibrated again to the same level within the intra-gradient layer, referred to as intra-gradient calibration (blue dashed line). We also introduce a lightweight design for the task heads, which saves ∼40% parameters.

Few studies[39, 34] concentrate on multimodal optimization methods during the training process. [39] proposes a metric OGR to quantize the significance of overfitting and try to solve it with Gradient Blending. It designs modal heads for each task thus difficult to expand to multi-task network. OGM-GE[34] try to solve the optimization imbalance problem by dynamically adjusting the gradients of different modalities. Since it separates parameters of different modalities in the linear classification head, it is hard to generalize to other complicated task heads. Differently, our method can be used in networks with any task head as long as the network has modal-specific parameters.

Multi-task optimization methods. Multi-task methods are mainly divided into two categories in [37], network architecture improvement[31, 42, 36, 10] and optimization methods[26, 6, 25, 46, 32]. Our approach focuses on the optimization methods. The goal of multi-task optimization methods is to balance the loss weights of different tasks to prevent one task from overwhelming another during training. DWA[26] adjusts the loss weights based on the rate at which the task-specific losses change, but it requires to balance the loss magnitudes beforehand. Gradnorm[6] balances the loss weights automatically by stimulating the task-specific gradients to be of similar magnitude. IMTL[25] optimizes the training process by guaranteeing the aggregated gradient has equal projections onto individual tasks. Yet they have not been studied in the domain of multi-modality multi-task learning. Our method complements these analysis.

3. Method

In this section, we introduce the Fuller, a framework that unifies the multi-modality multi-task 3D perception in autonomous driving scenarios. Fuller aims to mitigate the problem of modality bias and task conflict during the end-to-end training by gradient calibration. Regarding the network architecture, we introduce a lightweight design for the task heads, named Fuller-det and Fuller-seg.

3.1. Network architecture

As shown in Fig. 2, our proposed Fuller extracts features from both point cloud and images, then transforms them into a unified bird's-eye view (BEV) representation. It relies on VoxelNet [48] as LiDAR backbone and Swin-T [29] as image backbone. As for image features from multi-view cameras, we project them onto BEV feature using the scheme as same as LSS [35]. We adopt the modality fusion strategy where the features of two branches, f^{img} and f^{lid}, are first concatenated and then fed into the fusion block:

$$f^{fuse} = \texttt{conv}(f^{lid} \oplus f^{img}), \tag{1}$$

where $\texttt{conv}$ is the modal fusion block (*i.e.*, 2-layer FPN) and $\oplus$ is concatenation operation. f^{fuse} is then connected to task-specific heads.

The detection head Fuller-det follows a DETR-style [4] architecture with object queries. Given the fusion featrure f^{fuse}, Fuller-det initializes the queries by an auxiliary heatmap head according to TransFusion[2]. Also, Fuller-seg utilizes a query-based semantic segmentation head with segmentation queries. The BEV feature f^{fuse} is transformed into the output shape feature F. The initialized queries and F are then used to obtain mask embeddings M, processed by the transformer decoder layer. Finally, the binary mask prediction S is computed via a dot product between M and F, followed by a sigmoid activation. We refer the reader to App. B for more details.

Both Fuller-det and Fuller-seg have only one transformer

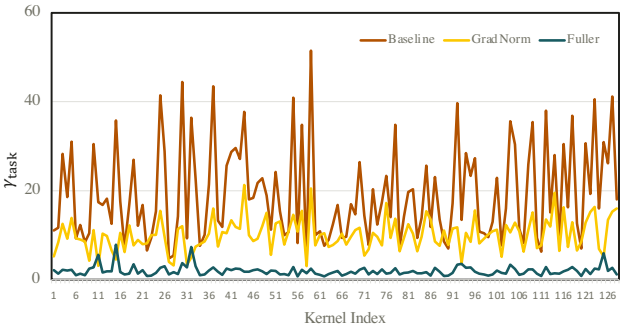


Figure 3. We visulize the γ_{task} (Eq. (2)) in the last layer of the modality-fusion block. The gradient tensors are unfolded along the first axis. It is easy to observe that the gradient magnitude of seg loss dramatically lags behind that of the det loss. We apply the proposed Fuller and compare it with GradNorm [6]. We find that our method is able to balance the gradients from two tasks. Importantly, our method yields a more stable and balanced γ_{task}.

decoder layer and could achieve competitive results compared to state-of-the-art methods, as will see in Sec. 4.4.

3.2. Multi-level Gradient Calibration

We now introduce the multi-level gradient calibration. First, it will calibrate the gradient between tasks via inter-gradient calibration. When it comes to the subsequent modality branches of the backbone, the gradient will be calibrated again by intra-gradient calibration.

3.2.1 Inter-Gradient Calibration for Task Conflict

By definition, the gradients will be propagated from the task heads to the shared backbone. Without any regularization, multi-task learning would simply sum up the individual gradients for backbone update. Since the gradients of the downstream tasks tend to exhibit great distinction, this naive manner will inevitably result in task conflict. For example, an objective with low gradient magnitude would be overwhelmed by another one with high gradient magnitude. Therefore, existing works [37, 6, 26, 25] propose to *manipulate* the gradients to interfere the optimization process.

Following this philosophy, we visualize the gradient distribution of the two tasks to inspect the inferior performance. Specifically, we compute the ratio of $L2$ norm between the gradients computed by raw individual losses:

$$\gamma_{\text{task}} = \frac{||\nabla_{\texttt{shared\_L}}\mathcal{L}_{\text{Det}}||}{||\nabla_{\texttt{shared\_L}}\mathcal{L}_{\text{Seg}}||}, \tag{2}$$

where ∇ denotes gradient computation operator, $\mathcal{L}_{\text{Det}}$ and $\mathcal{L}_{\text{Seg}}$ are the output losses of 3D detection and map segmentation, respectively. Typically, the gradients of shared backbone computed by different task losses are utilized to measure task characteristics. To save computation, we select the last layer of shared backbone, denoted as `shared_L`. Thus, γ_{task} is a metric that reflects the gradient discrepancy.

As we might notice in Fig. 3, the value of γ_{task} between the two tasks is significantly huge. Based on this finding, we consider that the emergence of task conflict is probably because the gradients of segmentation task are overwhelmed by that of detection task. Inspired by the loss weighting methods [37, 6, 26, 25], we balance the gradients of different tasks by balancing their loss weights. At each iteration, we obtain the gradients corresponding to individual loss on the last layer of the shared backbone. These gradients are utilized to derive the new loss weights. Then, the aggregated loss is applied to calibrate the gradients of the entire network. We evaluate existing literature and choose the IMTL_G [25] as the technique for this purpose given its superior performance, as discussed in App. C.2.

3.2.2 Intra-Gradient Calibration for Modality Bias

We have analyzed the impact of different task objectives on the backbone holistically. Another complicated situation arises when optimizing modality branches. During our experiments, we observe the issue of modality bias which undermines the assumption that multiple modalities can collaboratively support the downstream tasks. This phenomenon is also known as semantic inconsistency [9] and modality imbalance [40].

The first layer of the modality fusion block is referred to as the intra-gradient layer, parameterized by θ^F. It consists of two parts, θ^F_{lid} and θ^F_{img}, that represent the parameters directly connected to the LiDAR and image backbones during backpropagation. Let H denotes the modality branches, where θ^H_{lid} and θ^H_{img} represent the parameters of the LiDAR and image branch. According to the chain rule, the gradient for a certain modality branch is defined as:

$$G_{mod} = \frac{\partial \mathcal{L}}{\partial \theta^H_{mod}} = \frac{\partial \mathcal{L}}{\partial \theta^F_{mod}} \cdot \frac{\partial \theta^F_{mod}}{\partial \theta^H_{mod}}, \tag{3}$$

where $mod = \{lid, img\}$, G_{lid} and G_{img} mean the gradients of the two modality branches. According to Eq. (3), $\nabla\theta^F_{lid} = \frac{\partial \mathcal{L}}{\partial \theta^F_{lid}}$ would carry out the updating message from the task heads to the LiDAR branch, similarly for image branch.

Regarding the term $\nabla_{\theta^F}\mathcal{L}$, this gradient corresponds to the optimization process that determines how the intra-gradient layer would coordinate the fusion of the two modalities to adapt to downstream tasks. Therefore, we use $\nabla\theta^F_{lid}$ and $\nabla\theta^F_{img}$ within the intra-gradient layer to establish the connection between two modality branches. Since they will be separated into different branches, we consider their relative magnitude during end-to-end training:

$$\gamma_{\text{modal}} = \frac{||\nabla\theta^F_{lid}||}{||\nabla\theta^F_{img}||}. \tag{4}$$

The result displayed in Fig. 4 indicates that for most of the time, $||\nabla\theta^F_{lid}||$ would surpass $||\nabla\theta^F_{img}||$, which means

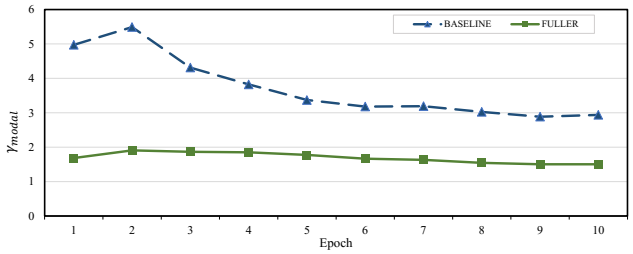


Figure 4. Compared with baseline, Fuller has a balanced γ_{modal} (Eq. (4)), meaning that two modalities can be learned in a balanced manner.

the LiDAR and image branches receive uneven attention from the downstream tasks.

To solve this problem, we propose calibrating the gradients between two branches, *i.e.*, G_{lid} and G_{img}. In practice, we gate the one with greater magnitude to slow down its pace, ensuring that the tasks pay balanced attention to both modalities. At t step, we obtain the gating factors by:

$$\begin{aligned} w_{lid}^t &= \sigma(||\nabla {\theta_{lid}^F}^t||, ||\nabla {\theta_{img}^F}^t||) \in (0,1], \\ w_{img}^t &= \sigma(||\nabla {\theta_{img}^F}^t||, ||\nabla {\theta_{lid}^F}^t||) \in (0,1], \end{aligned} \tag{5}$$

$$\sigma(x,y) = \mathbf{1}_{\frac{x}{y}>1}(1 - \texttt{tanh}(\alpha \cdot \frac{x}{y})) + \mathbf{1}_{\frac{x}{y}<=1}, \tag{6}$$

where $\sigma(\cdot,\cdot)$ is a composition function conditioned by the indicator function and used to measure a paired input. α is a weight factor. The gating factors in Eq. (5) are further smoothed by momentum update with coefficient m to stabilize the training. Then the calibrated gradient will be backpropagated to the associated branch:

$$w_{mod}^t = m \cdot w_{mod}^{t-1} + (1-m) \cdot w_{mod}^t, \tag{7}$$

$$G_{mod}^t = w_{mod}^t \cdot G_{mod}^t. \tag{8}$$

We refer this technique as intra-gradient calibration where it is performed between modalities.

3.3. Fuller: The Blueprint

We have presented a hierarchical view of the inter-gradient and intra-gradient calibration techniques, which are proposed to optimize the entire backbone and the associated modality branches, respectively. The procedure of Fuller is summarized in Algorithm 1. At each update step, we first calculate the gradients w.r.t. the two objectives in the last layer of the shared backbone. The two gradients are calibrated to alleviate the problem of task conflict, where a pair of weights are derived. After applying the weights to the raw losses, we obtain the calibrated gradient of the total loss on the intra-gradient layer, $\nabla\theta_{lid}^F$ and $\nabla\theta_{img}^F$. To mitigate the issue of modality bias, we utilize them to calibrate the gradients of corresponding branches.

Algorithm 1 Training Procedure of Fuller

Input composition function σ, modality branches' parameter θ^H, intra-gradient layer's parameter θ^F, off-the-modality's parameter θ^K, learning rate η.

Output θ^H, θ^K

for $t = 0, 1, 2, ..., T$ **do**
  Inter-gradient Calibration For Task Conflict:
  $\alpha_{\text{Seg}}, \alpha_{\text{Det}} \leftarrow \text{IMTL}(\nabla_{\texttt{shared\_L}}\mathcal{L}_{\text{Det}}, \nabla_{\texttt{shared\_L}}\mathcal{L}_{\text{Seg}})$
  $\mathcal{L} \leftarrow \alpha_{\text{Seg}} \cdot \mathcal{L}_{\text{Seg}} + \alpha_{\text{Det}} \cdot \mathcal{L}_{\text{Det}}$
  $\theta^K \leftarrow \theta^K - \eta \cdot \frac{\partial \mathcal{L}}{\partial \theta^K}$ ◁ Update task heads
  Intra-gradient Calibration for Modality Bias:
  $\nabla\theta_{lid}^F, \nabla\theta_{img}^F \leftarrow \frac{\partial \mathcal{L}}{\partial \theta_{lid}^F}, \frac{\partial \mathcal{L}}{\partial \theta_{img}^F}$
  $w_{lid}^t \leftarrow \sigma(||\nabla {\theta_{lid}^F}^t||, ||\nabla {\theta_{img}^F}^t||)$
  $w_{img}^t \leftarrow \sigma(||\nabla {\theta_{img}^F}^t||, ||\nabla {\theta_{lid}^F}^t||)$
  $w_{lid}^t \leftarrow m \cdot w_{lid}^{t-1} + (1-m) \cdot w_{lid}^t$
  $w_{img}^t \leftarrow m \cdot w_{img}^{t-1} + (1-m) \cdot w_{img}^t$
  $G_{lid}^t \leftarrow w_{lid}^t \cdot G_{lid}^t$; $G_{img}^t \leftarrow w_{img}^t \cdot G_{img}^t$
  backward G_{lid}^t and update θ_{lid}^H ◁ LiDAR branch
  backward G_{img}^t and update θ_{img}^H ◁ Image branch
end for

4. Experiments

We first introduce our baseline setting and benchmark dataset. We also investigate the potential of Fuller by evaluating it under different loss weights and dataset distribution settings. Finally, we ablate the proposed components to validate their individual effectiveness.

4.1. Experimental Settings

Implementation details. We adopt BEVFusion [30] as the strong baseline with a few modifications. The detailed design has been discussed in Sec. 3.1. The AdamW optimizer is utilized with a weight decay of 10^{-2} and momentum of 0.9. The models are trained for 10 epochs with a learning rate of 10^{-3}. We use 8 NVIDIA V100 GPUs with 2 samples per GPU, resulting in a total batch size of 16. Additionally, the value of α used in Eq. (6) is set to 0.1, while m in Eq. (7) is set to 0.2.

nuScenes dataset. nuScenes [3] is a multi-sensor dataset that provides diverse annotations for multiple tasks, including detection, tracking, and especially BEV map segmentation, which is typically absent in other datasets. The dataset comprises 28,130 training samples and 6,019 validation samples, each containing a 32-beam LiDAR scan and 6 multi-view images. The 3D detection task involves 10 foreground categories, and the performance is evaluated by mean Average Precision (mAP) and nuScenes Detection Score (NDS). For map segmentation, the model is required to segment 6 background categories in BEV view, which is measured by the mean Intersection over Union (mIoU).

Evaluation protocol. We evaluate the performance of

Table 1. Sensitivity analysis and ablation study of the proposed gradient calibration with different initial loss weights.

Intra.	Inter.	mAP(%)↑	NDS↑	mIoU(%)↑	Δ_{MTL}(%)↓
det_weight:seg_weight=1:1					
		59.1	65.0	44.0	18.3
✓		59.5	**65.4**	45.0	16.9
	✓	57.1	63.3	**59.5**	8.8
✓	✓	**60.5**	65.3	58.4	**5.4**
det_weight:seg_weight=1:5					
		59.8	65.5	55.7	8.0
✓		59.8	65.3	56.1	7.8
	✓	56.9	63.3	**59.8**	8.7
✓	✓	**60.1**	**65.6**	58.2	**5.7**
det_weight:seg_weight=1:10					
		59.3	65.0	57.9	7.0
✓		**60.1**	**65.4**	57.3	6.5
	✓	58.2	64.2	**60.1**	6.7
✓	✓	59.9	65.2	59.2	**5.3**

Table 2. Sensitivity analysis and ablation study of the proposed gradient calibration under different dataset distribution.

Intra.	Inter.	mAP(%)↑	NDS↑	mIoU(%)↑	Δ_{MTL}(%)↓
Full					
		59.1	65.0	44.0	18.3
✓		59.5	**65.4**	45.0	16.9
	✓	57.1	63.3	**59.5**	8.8
✓	✓	**60.5**	65.3	58.4	**5.4**
Disjoint-balance					
		58.7	64.2	41.5	21.3
✓		58.3	**65.0**	42.4	20.5
	✓	57.4	62.7	**57.3**	10.8
✓	✓	58.4	63.6	56.7	**9.8**
Disjoint-normal					
		59.1	64.7	43.1	19.3
✓		**59.3**	**65.3**	44.0	17.9
	✓	57.4	62.9	53.8	13.4
✓	✓	58.9	64.9	**55.0**	**9.7**

multi-task learning based on the metric in [37]:

$$\Delta_{\text{MTL}} = \frac{(-1)^l}{T}\sum_{i=1}^{T}(M_{m,i} - M_{b,i})/M_{b,i}, \qquad (9)$$

where T is the number of tasks. $M_{m,i}$ and $M_{b,i}$ are the performance of the i-th task of the evaluated model and baseline, respectively. Δ_{MTL} could be intuitively understood as the average performance drop, where we set $l = 1$, *i.e.*, lower value means better performance. Following Liang *et al*. [24], we also evaluate the Fuller in three annotation schemes.

Full setting. We leverage all available annotations by default, which serves as the upper bound for the following two settings.

Disjoint-normal. Given the limited budget, the annotation complexity determines the quantity of task labels. In a realistic practice, we split the full dataset into 3D detection and map segmentation parts using a 3:1 ratio, where each sample is labeled for one task.

Disjoint-balance. Similarly, each sample is endowed with a task label and each task can leverage half of the dataset.

4.2. Sensitivity Analysis

Initial states. To investigate the robustness of our method w.r.t. initial loss weights, we incrementally increase the weight of segmentation loss and inspect its impact on performance. As illustrated in Tab. 1, the loss weights between detection and segmentation are set to 1:1, 1:5, and 1:10. We find that increasing the loss weight of map segmentation can improve the subsequent performance of the baseline model. However, manually adjusting the loss weights can lead to sub-optimal results. For instance, when the loss weight of map segmentation is increased from 5 to 10, it benefits map segmentation (55.7%→57.9% mIoU) but damages the performance of detection task (59.8%→59.3% mAP). Nonetheless, the proposed method can facilitate model training despite variations in initial loss weights.

Dataset distributions. We verify the model under different dataset distributions [24]. In Tab. 2, the baseline performance of disjoint dataset is notably inferior compared with that of full dataset, posing a greater challenge to multi-task learning. In the full setting, Fuller was observed to improve both tasks. Given the significant Δ_{MTL} metric (42.5%) of the disjoint-balance baseline, Fuller would pay more attention to improving map segmentation while slightly degrading 3D detection to address task conflict. In the case of disjoint-normal, the baseline's Δ_{MTL} (38.5%) is relatively small and has minor effect on 3D detection. Generally, Fuller can boost Δ_{MTL} metric under these scenarios.

4.3. Ablation Study

Validating the inter- and intra-gradient calibration. We conduct thorough experiments, including three different initial states, to validate the individual effectiveness of the proposed multi-level gradient calibration in Tab. 1. Generally, the intra-gradient calibration leads to considerable improvements in the downstream tasks compared to the baseline. Regarding the inter-gradient calibration, it can largely enhance the performance of the map segmentation while deteriorating the detection task at the acceptable cost. The combination of the two techniques yields remarkable improvement in both tasks, ultimately achieving best Δ_{MTL}. We further evaluate the validity of the two proposed components across various dataset distribution, as shown in Tab. 2. Again, both components can individually improve the Δ_{MTL} in all settings.

Relation between two calibration techniques. The above experiments (Tab. 1&Tab. 2) evidence the individual effectiveness of the two proposed calibration techniques. We are curious whether the calibrations are consistently cooperative or could be adversarial in certain scenarios. To inves-

Table 3. Comparison with benchmark. The upper two sub-tables are single task results while the bottom one is multi-task result.'L' and 'C' represent LiDAR and Camera, respectively. We treat single task result as our upper bound because multi-task will generally decrease the performance. Baseline means Fuller is naively trained where detection loss and segmentation loss are set to 1:1. '-' means inapplicable.

	Modality	VoxelSize	LiDAR	Image	mAP(%)↑	NDS↑	mIoU(%)↑
3D Detection							
BEVFormer [21]	C	-	-	ResNet101 [11]	41.6	51.7	-
CenterPoint [44]	L	0.075	VoxelNet	-	59.6	66.8	-
MVP‡ [45]	C+L	0.075	VoxelNet	DLA-34	66.1	70.0	-
TransFusion [2]	C+L	0.075	VoxelNet	DLA-34	67.5	71.3	-
BEVFusion [30]	C+L	0.075	VoxelNet	Swin-T	68.5	71.4	-
Fuller-det	C+L	0.075	VoxelNet	Swin-T	**67.6**	**71.3**	-
Fuller-det (upper bound)	C+L	0.1	VoxelNet	Swin-T	**62.1**	**66.6**	-
BEV Map Segmentation							
LSS‡ [35]	C	-	-	EfficientNet-B0	-	-	44.4
CenterPoint‡ [44]	L	0.1	VoxelNet	-	-	-	48.6
BEVFusion [30]	C+L	0.1	VoxelNet	Swin-T	-	-	62.7
Fuller-seg(upper bound)	C+L	0.1	VoxelNet	Swin-T	-	-	**62.3**
3D Detection + BEV Map Segmentation							
BEVFusion† [30] (share)	C+L	0.1	VoxelNet	Swin-T	-	69.7	54.0
BEVFusion† [30] (sep)	C+L	0.1	VoxelNet	Swin-T	-	69.9	58.4
Baseline(share)	C+L	0.1	VoxelNet	Swin-T	59.1	65.0	44.0
Fuller(share)	C+L	0.1	VoxelNet	Swin-T	**60.5**	**65.3**	**58.4**

† means the multi-task result in BEVFusion[30]. ‡ means re-implementation result in BEVFusion[30]. 'share' means multi-task heads share one BEV encoder to process the fused multimodal feature. 'sep' means task heads have separate encoders.

Table 4. Relation between inter- and intra-gradient calibration.

Intra.	Inter.	$\gamma_{\text{task}}\downarrow$	$\gamma_{\text{modal}}\downarrow$	$\Delta_{\text{MTL}}\downarrow$
		19.2	3.7	18.3
✓		12.5	1.9	16.9
	✓	2.5	2.8	8.8
✓	✓	**1.9**	**1.7**	**5.4**

tigate this, we visualize the γ_{modal} and γ_{task} by ablating one of the calibration techniques, as shown in Tab. 4. Interestingly, we found that applying either calibration technique could simultaneously mitigate both issues of modality bias and task conflict, resulting in more balanced γ_{modal} and γ_{task}. The result indicates that the two proposed calibration techniques are cooperative.

4.4. More Results

Comparison with the benchmark. We compare the Fuller with current state-of-the-art methods and report the result on nuScenes validation set (Tab. 3). We list each model's modality and group them by task setting. Our baseline (*i.e.*, penultimate row) is adapted from the competitive BEVFusion [30]. Given hardware capacity of V100 GPU, the voxel size is set to 0.1m for multi-task learning. For fair comparison, the single-task model Fuller-det and Fuller-seg using voxel size of 0.1m are set as the upper bounds for 3D detection and map segmentation. As shown in Fig. 5, Fuller-det converges after 7 epochs with lr=1e-4. Fuller-seg converges after 10 epochs with lr=1e-3. We train the Fuller using the scheme as same as Fuller-seg.

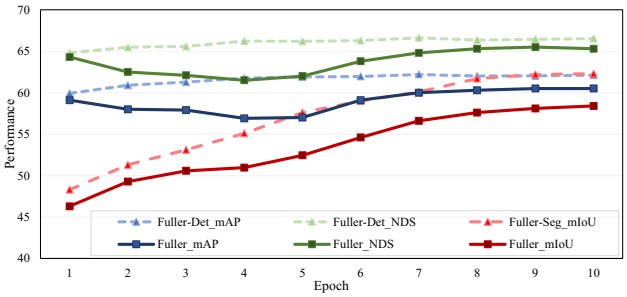
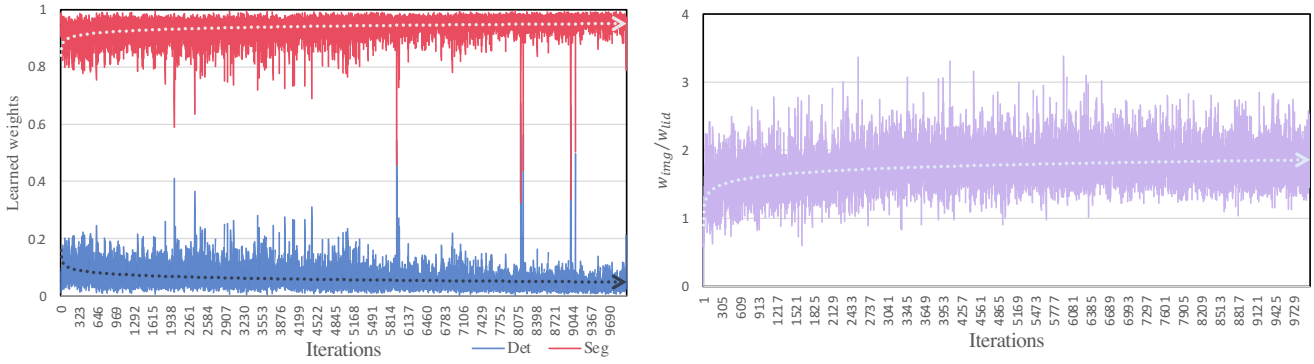


Figure 5. Model convergence of single-task and multi-task models.

As illustrated in Tab. 3, the performance of multi-task baseline suffers a significant decline compared to the upper bounds. Particularly, the mIoU of segmentation task drops drastically from 62.3% to 44.0%, which discourages the multi-task applications in autonomous driving scenarios. Rather, the proposed model Fuller demonstrates substantial improvement in bridging the gap between single-task and multi-task models, which improves mIoU from 44.0% to 58.4% in map segmentation and facilitates the mAP from 59.1% to 60.5% in 3D detection.

Visualization of multi-level gradient calibration. To better understand the mechanism of the proposed multi-level gradient calibration, we visualize how Fuller manipulates the gradients during network optimization. As shown in Fig. 6 (a), Fuller adapts its loss functions to mitigate task conflict by decreasing the detection loss and elevating the



(a) Learned weights for two tasks via inter-gradient calibration

(b) Learned weights for two modalities via intra-gradient calibration

Figure 6. Visualization of gradient calibration.

Table 5. Memory cost and inference speed. All the speeds are evaluated on an Tesla V100 GPU.

	Memory↓	Parameter↓	FPS↑
STL	6144MB	81.97M	1.20
FULLER	3103MB	44.12M	2.30

Table 6. Verifying the framework with more learning tasks.

Method	mAP(%)↑	NDS↑	mIoU_{map}(%)↑	$\text{mIoU}_{\text{fore}}$(%)↑	Δ_{MTL} ↓
Upper bound	62.1	66.6	62.3	63.8	-
MTL baseline	**59.9**	**65.8**	46.9	58.1	12.8
Fuller	58.6	64.6	**57.1**	**62.0**	**9.9**

segmentation loss. In terms of modality bias, we plot the ratio w_{img}/w_{lid} (Eq. (5)) for visualization. According to Fig. 6 (b), the weight w_{lid} is smaller than w_{img} for most of the time, indicating that Fuller would gate the gradients of LiDAR branch to take maximum advantage of both modalities, thereby improving the subsequent result.

Association between task and modality. To identify the association between modalities and tasks, we propose evaluating the *trained* model with one modality removed at a time. Specifically, we examine the performance of the Fuller and baseline models in the absence of image input (Fig. 7). The results indicate that 3D detection retains a considerable level of accuracy even without image input, thanks to the precise spatial information provided by LiDAR scans. In contrast, the absence of image input significantly impairs the performance of map segmentation. Our findings are consistent with the theoretical analysis in [34], which suggests that each modality carries out a unique mechanism and contributes distinct functionality to downstream tasks. In App. C.5, we provide additional experiments and discussion in which LiDAR scans are absent.

Generalization ability. By sharing a common backbone, Fuller can save substantial memory cost and speed up the inference as shown in Tab. 5. Additionally, we augment our framework to 3 tasks by introducing foreground segmentation, as demonstrated in Tab. 6. Foreground segmentation is a task related to 3D detection and map segmentation, which is segmentation of foreground objects under BEV. Our proposed approach still achieves performance gains.

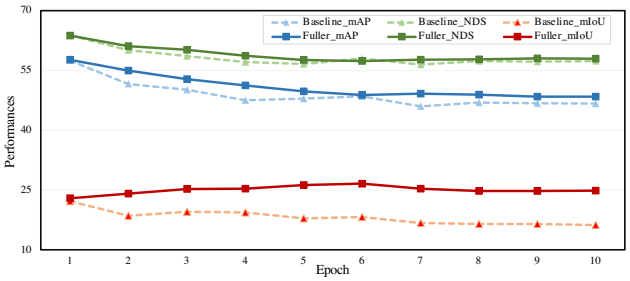


Figure 7. A model trained with both modalities is evaluated by dropping off the image input. LiDAR can still support the detection task. However, without image input, the performance on map segmentation drops largely. Fuller proves to be able to mitigate the modality bias and significantly improves map segmentation result.

5. Conclusion

We present Fuller, a framework that addresses the challenges of modality bias and task conflict in multi-modality multi-task learning for 3D perception tasks. To cope with these problems, we propose multi-level gradient calibration to guide the learning process of the model. Our approach includes inter-gradient calibration to balance the gradients w.r.t. downstream tasks on the last layer of the shared backbone. Before being separated into different branches, the magnitude of these gradients will be calibrated again within the intra-gradient layer.

6. Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0109700, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061), Shenzhen Science and Technology Program (Grant No. RCYX20200714114642083) Shenzhen Fundamental Research Program(Grant No. JCYJ20190807154211365), Nansha Key RD Program under Grant No.2022ZD014 and Sun Yat-sen University under Grant No. 22lgqb38 and 76160-12220011. We thank MindSpore for the partial support of this work, which is a new deep learning computing framwork*.

*https://www.mindspore.cn/

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022.

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of European Conference on Computer Vision*, pages 213–229, 2020.

[5] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022.

[6] 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, 2018.

[7] Di Feng, Yiyang Zhou, Chenfeng Xu, Masayoshi Tomizuka, and Wei Zhan. A simple and efficient multi-task network for 3d object detection and road understanding. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7067–7074, 2021.

[8] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.

[9] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

[10] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *International Conference on Machine Learning*, pages 3854–3863, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitions*, pages 770–778, 2016.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[13] Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. *Advances in Neural Information Processing Systems*, 30:551–562, 2017.

[14] Aya Abdelsalam Ismail, Mahmudul Hasan, and Faisal Ishtiaq. Improving multimodal accuracy through modality pretraining and attention. *arXiv preprint arXiv:2011.06102*, 2020.

[15] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[16] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[17] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5198–5204, 2018.

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[19] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online HD map construction and evaluation framework. In *2022 International Conference on Robotics and Automation*, pages 4628–4634, 2022.

[20] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022.

[21] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision*, pages 1–18, 2022.

[22] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019.

[23] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework. *arXiv preprint arXiv:2205.13790*, 2022.

[24] Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. *arXiv preprint arXiv:2209.08953*, 2022.

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

[26] 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.

[27] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022.

[28] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[30] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022.

[31] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

[32] Taylor Mordan, Matthieu Cord, Patrick Pérez, and Alexandre Alahi. Detecting 32 pedestrian attributes for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11823–11835, 2021.

[33] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, pages 631–648, 2018.

[34] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022.

[35] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of European Conference on Computer Vision*, pages 194–210, 2020.

[36] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4822–4829, 2019.

[37] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 3614–3633, 2021.

[38] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4612, 2020.

[39] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

[40] Xun Wang, Bingqing Ke, Xuanping Li, Fangyu Liu, Mingyu Zhang, Xiao Liang, and Qiushi Xiao. Modality-balanced embedding for video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2578–2582, 2022.

[41] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. Mˆ 2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022.

[42] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.

[43] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. *arXiv preprint arXiv:2209.09385*, 2022.

[44] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021.

[45] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multi-modal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.

[46] 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.

[47] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022.

[48] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.