Conflict-Alleviated Gradient Descent for Adaptive Object Detection

Wenxu Shi[1] , **Bochuan Zheng**[2*]
[1]School of Microelectronics and Communication Engineering, Chongqing University, China
[2]School of Computer Science, China West Normal University, China
wxshi@cqu.edu.cn, zhengbc@vip.163.com

Abstract

Unsupervised domain adaptive object detection (DAOD) aims to adapt detectors from a labeled source domain to an unlabeled target domain. Existing DAOD works learn feature representations that are both class-discriminative and domain-invariant by jointly minimizing the loss across domain alignment and detection tasks. However, jointly resolving different tasks may lead to conflicts, one contributing factor being gradient conflicts during optimization. If left unaddressed, such disagreement may degrade adaptation performance. In this work, we propose an efficient optimization strategy named Conflict-Alleviated Gradient Descent (CAGrad), which aims to alleviate the conflict between two tasks (i.e., alignment and detection). Specifically, we alter the gradients by projecting each onto the normal plane of the other. The projection operation changes conflicting gradients from obtuse to acute angles, thus alleviating the conflict and achieving gradient harmonization. We further validate our theoretical analysis and methods on several DAOD tasks, including cross-camera, weather, scene, and synthetic-to-real-world adaptation. Extensive experiments on multiple DAOD benchmarks demonstrate the effectiveness and superiority of our CAGrad approach.

1 Introduction

The rapid expansion of image data has driven significant advancements in machine learning, particularly in the field of computer vision [Redmon and Farhadi, 2018; Ren *et al.*, 2015; Cai and Vasconcelos, 2018; Zhu *et al.*, 2020; Carion *et al.*, 2020; Liu *et al.*, 2023]. However, a major limitation of these advancements is the assumption that the distribution of the training dataset perfectly matches that of practical application scenarios, an alignment rarely achieved in real-world settings. To tackle the challenges arising from this mismatch in domain distributions, Unsupervised Domain Adaptation (UDA) [Pan and Yang, 2009; Cui *et al.*, 2020; Ganin and Lempitsky, 2015; Kang *et al.*, 2019] has emerged as a practical solution. UDA aims to reduce the performance decline caused by the disparity in domain distributions, removing the need for annotated data in the target domain.

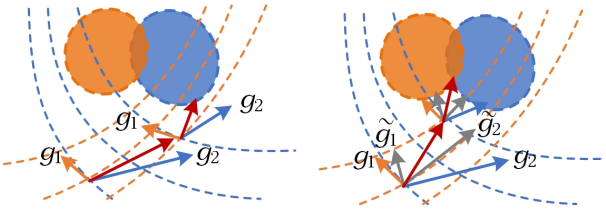

Figure 1: The combined update vector g (in red) of a two-task learning problem with gradient descent (GD) and our approach CAGrad. The two task-specific gradients are labeled g_1 (in blue) and g_2 (in orange). (a) The update vector g is the sum gradient $g_1 + g_2$. Due to the conflict between g_1 and g_2, the update vector is dominated by g_2. (b) Our CAGrad projects each gradient onto the normal plane of the other one and uses the projected gradients $\tilde{g_1}$ and $\tilde{g_2}$ (in gray arrows) as the update vector $g = \tilde{g_1} + \tilde{g_2}$.

In recent years, UDA algorithms have rapidly developed and achieved significant performance in object detection, known as domain adaptive object detection (DAOD) [Chen *et al.*, 2018; He and Zhang, 2019; Xu *et al.*, 2020; Wang *et al.*, 2021a; Wu *et al.*, 2022]. Specifically, [Chen *et al.*, 2018] is the first method to achieve domain alignment through the game between feature detector and domain classifier. [Wang *et al.*, 2021a] adjusts adversarial domain adaptation models on detection transformers to reduce distribution shift between different domains through aligning sequence features. These approaches are characteristically designed to concurrently optimize both domain alignment and detection tasks throughout the training phase. However, due to the functional differences between various tasks, the optimal gradient descent directions for these tasks may be uncoordinated or imbalanced. Directly sharing network parameters, such as the feature generator, can lead to optimization conflicts and negatively impact the learning of domain-invariant features. Figure 1 (a) illustrates a scenario of suboptimal direction updates. The orange arrow represents the optimal gradient direction for domain alignment, and the blue arrow for detection. In joint optimization, the update gradient (red arrow) is mostly dominated by detection task, owing to existing con-

*Corresponding author

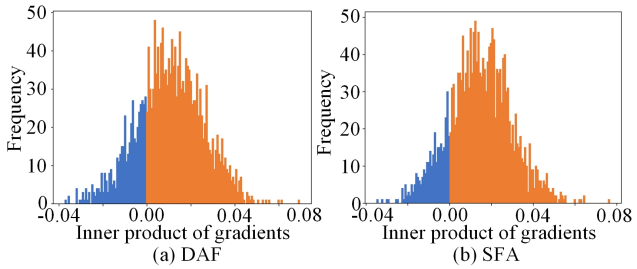

Figure 2: The histograms show gradient inner product frequencies for DAF and SFA during training, with the x-axis for inner product values and the y-axis for frequency. Blue marks obtuse angles, and orange marks acute angles, indicating conflict and non-conflict.

flicts, which diminishes the effectiveness of both tasks. In this work, we define two gradients to be conflicting if they point away from one another, i.e., have a negative cosine similarity.

To validate our observations, we conducted experiments in cross-domain object detection scenarios. Figure 2 shows the gradient inner product distribution between two tasks (alignment and detection) during training, using DAF [Chen *et al.*, 2018] and SFA [Wang *et al.*, 2021a] on tasks transitioning from Cityscapes [Cordts *et al.*, 2016a] to Foggy Cityscapes [Sakaridis *et al.*, 2018a]. The results revealed a mix of acute and obtuse angles in gradient distributions, with obtuse angles being common. Conflicting gradients cause the optimizer to favor certain tasks, leading to a poor update direction for a subset or even all tasks. If left untouched, this will degrade adaptation performance. Addressing these conflicts while preserving task functionality is a crucial yet underexplored challenge.

In this paper, we introduce a novel and effective module called Conflict-Alleviated Gradient descent (CAGrad), designed to address the optimization conflicts between domain alignment and detection tasks. When gradients from these two tasks conflict, we modify them by projecting each onto the normal plane of the other, ensuring coordinated development of both tasks during joint training while preserving their specific functions. As depicted in Figure 1 (b), the strategic integration of gradient magnitude and direction-homogenization plays a pivotal role in maintaining the stability of the overall learning process. Moreover, CAGrad is model-agnostic and requires only a single modification in gradient application, making it easy to integrate with existing state-of-the-art DAOD methods for enhanced performance. We theoretically prove the local conditions under which CAGrad improves upon standard multi-task gradient descent and empirically evaluate it on various challenging DAOD tasks, including cross-camera, weather, scene, and synthetic-to-real-world adaptation. Overall, CAGrad proves to be a powerful tool, capable of not only alleviating gradient conflicts but also enhancing detection performance. The main contributions of this paper are summarized as follows:

- We discover the gradient conflict exists in the DAOD methods,which can cause the optimizer to prioritize certain tasks over others, leading to difficulties in converging to a desirable solution.
- We propose an efficient optimization strategy named Conflict-Alleviated Gradient descent (CAGrad), which alters the gradients by projecting each onto the normal plane of the other. Figure 3 is a representative and illustrative example that depicts the usage of CAGrad.
- Extensive experiments validate the effectiveness and universality of CAGrad on various benchmark databases on several DAOD baselines. More insights and analyses of our model are also provided to justify the reasonability of CAGrad and demonstrate its superiority.

CAGrad is a simple and effective tool that easily integrates with existing adversarial learning methods for Domain Adaptive Object Detection (DAOD), without requiring changes to the network architecture. This attribute renders CAGrad highly compatible and user-friendly with various adversarial domain adaptation approaches.

2 Related Work

Object detection. Object detection, a key area in computer vision, has seen significant progress over the years. Models like Faster-RCNN [Ren *et al.*, 2015] and Mask-RCNN [He *et al.*, 2017] represent successful two-stage approaches, while one-stage models such as YOLO [Redmon *et al.*, 2016], SSD [Liu *et al.*, 2016], and FCOS [Tian *et al.*, 2019] are known for their real-time capabilities. Unlike these CNN-based detectors, DETR [Carion *et al.*, 2020] introduces transformers to object detection and shows promising results. Deformable DETR [Zhu *et al.*, 2020] further improves this by speeding up convergence and efficiently handling multi-scale features. These developments provide a strong foundation for domain adaptive object detection research.

Domain Adaptive Object Detection. To circumvent performance degradation caused by distribution shifts, the research of domain adaptive object detection has drawn attention recently. As a pioneering work, [Chen *et al.*, 2018] propose the domain adaptive Faster-RCNN method (DAF),which achieves image-level and instance-level feature alignment. Inspired by it, MAF [He and Zhang, 2019] introduces a hierarchical adversarial feature alignment strategy that reduces domain disparity at different scales. HTCN [Chen *et al.*, 2020] employ CycleGAN for data augmentation, generating intermediate domain images to facilitate model alignment between source and target domains. VDD [Wu *et al.*, 2021] tackles the problem by disentangling domain-invariant and domain-specific representations using vector decomposition, while also exploring the extraction of instance-invariant features [Wu *et al.*, 2022]. IDF [Lang *et al.*, 2023] propose a non-adversarial domain discriminator to extract domain-specific features. Additionally, PTMAF[He *et al.*, 2023] and PAATF[He *et al.*, 2022] introduce additional constraints during the adversarial learning stage. Recently, regarding the Transformer object detector, existing adaptation techniques for DETR predominantly rely on model-based approaches SFA [Wang *et al.*, 2021b], aiming to reduce the distribution shift between different domains through sequence feature alignment. AQT [Huang *et al.*, 2022] employs a novel adversarial token and a stack of cross-attention layers as the discriminator.

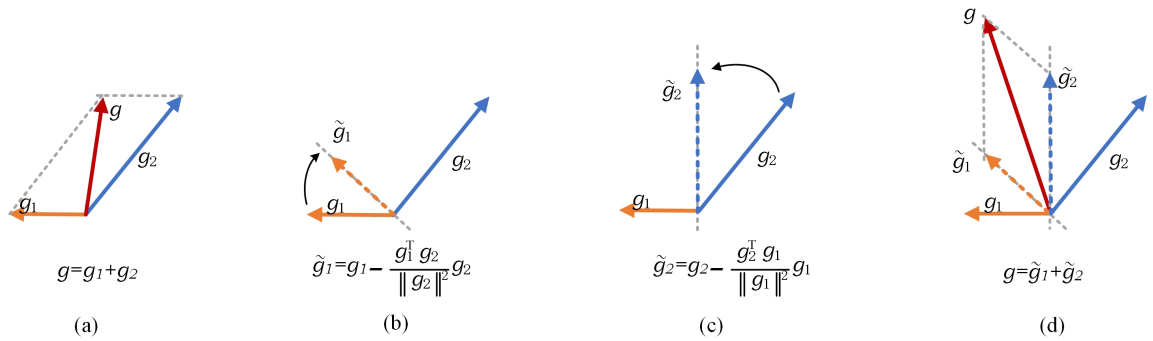

Figure 3: Conflicting gradients and CAGrad. In (a), the gradients g_1 and g_2, pertaining to two distinct tasks, exhibit opposing directional tendencies. This condition potentially results in adverse interference effects. In (b) and (c), we illustrate the CAGrad algorithm in the case where gradients are conflicting. CAGrad projects task 1's gradient onto the normal vector of task 2's gradient, and vice versa. Non-conflicting task gradients (d) are not altered under CAGrad, allowing for constructive interaction.

Previous methodologies have focused on acquiring domain-invariant and class-discriminative feature representations through the concurrent optimization of domain alignment and object detection tasks. However, these approaches often overlook the issue of imbalance and uncoordinated optimization between tasks, a phenomenon we identify as 'gradient conflict' in this study. To address this issue, we introduce the Conflict-Alleviated Gradient descent (CAGrad), a strategy designed to independently manage the gradients of domain alignment and detection tasks. This approach aims to harmonize the two tasks throughout the training process, thereby mitigating the gradient conflict.

3 Proposed Approach

3.1 Problem Definition

In domain adaptive object detection, the labeled source domain $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ comprises n_s samples across C classes. In contrast, the unlabeled target domain $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$ contains n_t samples of the same classes. Both $\mathcal{D}_s$ and $\mathcal{D}_t$ share the same feature and category spaces but differ in data distributions. Our goal is to use the labeled data from $\mathcal{D}_s$ and the unlabeled data from $\mathcal{D}_t$ to train a deep model that accurately predicts class labels and identifies bounding boxes in the target domain.

3.2 A General Framework of DAOD

Adversarial learning, pioneered by the Domain Adversarial Neural Network (DANN) [Ganin *et al.*, 2016], aligns domains effectively. It uses a feature generator $\mathcal{G}$ to deceive the domain discriminator $\mathcal{D}$, which predicts the origin domain of the generated features. Training involves an adversarial game between $\mathcal{G}$ and $\mathcal{D}$, optimizing parameters θ_g and θ_d using a domain alignment objective function as follows:

$$\begin{aligned}\mathcal{L}_{adv}(\theta_g, \theta_d) =& \mathbb{E}_{x_i^s \sim \mathcal{D}_s} \log\left[\mathcal{D}(\mathcal{G}(x_i^s))\right] \\ &+ \mathbb{E}_{x_i^t \sim \mathcal{D}_t} \log\left[1 - \mathcal{D}(\mathcal{G}(x_i^s))\right]\end{aligned} \tag{1}$$

To enhance detection performance in the target domain, it is crucial to first ensure that the detector $\mathcal{C}$ accurately identifies samples from the source domain. Consequently, we describe the supervised detection loss as follows:

$$\mathcal{L}_{det}(\theta_g, \theta_c) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}_{ce}(\mathcal{C}(\mathcal{G}(x_i^s; \theta_g); \theta_c), y_i^s) \tag{2}$$

where $\mathcal{L}ce$ represents the standard cross-entropy loss function. During the training phase, conventional methods typically optimize both the adversarial ($\mathcal{L}adv$) and detection ($\mathcal{L}_{det}$) objective functions concurrently to achieve domain-invariant and class-discriminant feature representations. The overall minimax objective function is given by:

$$\min_{\theta_g, \theta_c} \max_{\theta_d} \mathcal{L}_{det} + \mathcal{L}_{adv} \tag{3}$$

where θ_g, θ_d, and θ_c represent the parameters of the feature generator, domain discriminator, and category detector, respectively.

3.3 Conflict-Alleviated Gradient Descent

In the realm of DAOD, domain alignment and detection stand as distinct tasks, each possessing its own set of optimal gradient descent directions, which may exhibit inconsistency. This inconsistency can engender conflicts in the optimization of two loss functions during training, thereby exerting an impact on the ultimate performance of domain adaptation. To enable seamless coordination in optimizing these two tasks, we introduce the CAGrad technique.

To streamline our approach, we define the overall loss function as $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, where $\mathcal{L}_1 = \mathcal{L}_{adv}$ pertains to domain alignment and $\mathcal{L}_2 = \mathcal{L}_{det}$ relates to detection. Here, $g_1 = \nabla\mathcal{L}_1(\theta)$ and $g_2 = \nabla\mathcal{L}_2(\theta)$ signify the gradients for each task, while g denotes the gradient of the total loss $\mathcal{L}$. Our analysis delineates two scenarios based on the correlation between g_1 and g_2:

In Scenario I, a **positive** correlation between g_1 and g_2 (i.e., $cos(g_1, g_2) > 0$ or $g_1^T g_2 > 0$) suggests an **acute** angle between the gradients, indicating that they are effectively aligned. Consequently, no further modification of g_1 and g_2 is required.

In Scenario II, a **negative** correlation between g_1 and g_2 (i.e., $cos(g_1, g_2) < 0$ or $g_1^T g_2 < 0$) indicates an **obtuse** angle between the gradients, suggesting a conflict in their optimization directions. To address this effectively, projecting each gradient onto the normal plane of the other's gradient can minimize this conflict and enhance the optimization efficiency. This is achieved by:

$$\tilde{g_1} = g_1 - \frac{g_1^T g_2}{||g_2||^2} g_2 \tag{4}$$

$$\tilde{g_2} = g_2 - \frac{g_2^T g_1}{||g_1||^2} g_1 \tag{5}$$

The Conflict-Alleviated Gradient Descent process is illustrated in Figure 3, leading to a combined gradient expression:

$$g = \tilde{g_1} + \tilde{g_2} = g_1 + g_2 - \frac{g_1^T g_2}{||g_2||^2} g_2 - \frac{g_2^T g_1}{||g_1||^2} g_1 \tag{6}$$

Considering both positive and negative correlations ($g_1^T g_2 > 0$ and $g_1^T g_2 < 0$), we derive the following theorem.

Theorem 1. *In the context of DAOD, conflict resolution in optimization can be facilitated by directly altering the gradients to prevent potential conflicts. The combined gradient expression, g, post-modification, is given by:*

$$\begin{aligned} g &= \tilde{g_1} + \tilde{g_2} \\ &= g_1 + g_2 - \delta(g_1^T g_2 < 0) \frac{g_1^T g_2}{||g_2||^2} g_2 - \delta(g_1^T g_2 < 0) \frac{g_2^T g_1}{||g_1||^2} g_1 \end{aligned} \tag{7}$$

Here, $\delta(\cdot)$ denotes an indicator function, which returns a value of 1 if the condition is true and 0 if false, defined as:

$$\delta(A) = \begin{cases} 1, & if A \; is \; true, \\ 0, & if A \; is \; false \end{cases} \tag{8}$$

The proposed CAGrad methodology effectively coordinates two distinct tasks during optimization.

3.4 Efficient Equivalent Model of CAGrad+DAOD

Using the generalized DAOD model from Eq. 3 in Section 3.2, along with the CAGrad method from Eq. 7, improves training and helps achieve equilibrium in the DAOD framework. We introduce a more proficient loss function that integrates CAGrad principles into the DAOD model, improving convergence and performance. In alignment with CAGrad principles, we reformulate Eq. 7 as follows:

$$g = (1 - \delta(g_1^T g_2 < 0) \frac{g_2^T g_1}{||g_1||^2}) g_1 + (1 - \delta(g_1^T g_2 < 0) \frac{g_1^T g_2}{||g_2||^2}) g_2 \tag{9}$$

To streamline the presentation, let us define τ_1 and τ_2 as:

$$\begin{aligned} \tau_1 &= 1 - \delta(g_1^T g_2 < 0) \frac{g_2^T g_1}{||g_1||^2} \\ \tau_2 &= 1 - \delta(g_1^T g_2 < 0) \frac{g_1^T g_2}{||g_2||^2} \end{aligned} \tag{10}$$

These constants τ_1 and τ_2 are derived from the gradients of the original loss functions $\mathcal{L}_1(\theta)$ and $\mathcal{L}_2(\theta)$. With τ_1 and τ_2 established, we can rewrite Eq. 9 as follows:

$$g = \tau_1 g_1 + \tau_2 g_2 \tag{11}$$

Performing an integral operation on Eq. 11 yields a balanced DAOD model:

$$\tilde{\mathcal{L}} = \int (\tau_1 g_1 + \tau_2 g_2) dg_1 dg_2 = \tau_1 \mathcal{L}_1 + \tau_2 \mathcal{L}_2 \tag{12}$$

The training objective then becomes:

$$\min_{\theta_g, \theta_c} \max_{\theta_d} \tilde{\mathcal{L}} \tag{13}$$

The CAGrad module, as proposed, essentially recalibrates the original loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ through the computation of weights τ_1 and τ_2, which rely on the gradients of the original loss functions. This methodology culminates in a balanced and efficient DAOD model. In scenarios where $g_1^T g_2 > 0$, we note $\tau_1 = 1$ and $\tau_2 = 1$, leading to a simplification of the overall loss function to $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, corresponding to the general DAOD model. Figure 4 serves as an exemplary illustration of the CAGrad+DAOD application, with the complete update process delineated in Algorithm 1.

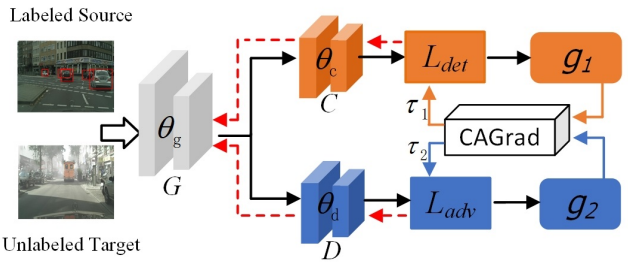

Figure 4: The usage illustration of our CAGrad+DAOD. Optimization objectives include domain alignment loss and detection loss. CAGrad module is responsible for harmonization process for the gradients of the two losses.

Algorithm 1 CAGrad+DAOD Optimization Algorithm

Input: Source data $\{x_i^s, y_i^s\}_{i=1}^{i=n_s}$, Target data $\{x_i^t\}_{i=1}^{i=n_t}$, Optimal parameters $\phi = \{\theta_g, \theta_c, \theta_d\}$, learning rate η, max_iteration;
Output: Optimal parameters $\theta_g, \theta_c, \theta_d$;
1: Initialization Optimal parameters $\theta_g, \theta_c, \theta_d$;
2: **repeat**
3: Compute domain alignment loss $\mathcal{L}_{adv}$ (Eq. 1), i.e., $\mathcal{L}_1$ and detection loss $\mathcal{L}_{det}$ (Eq. 2), i.e., $\mathcal{L}_2$;
4: Compute original gradients g_1 and g_2 and judge the inner product of two gradients $g_1^T g_2$;
5: Compute coefficients τ_1 and τ_2 by Eq. 10;
6: Compute updated total loss $\tilde{\mathcal{L}}$ (Eq. 12);
7: Update model parameters: $\phi^{t+1} \leftarrow \phi^t - \eta \nabla_{\phi^t} \tilde{\mathcal{L}}$
8: **until** max_iteration is reached;

4 Experiments and Results

4.1 Experimental Setup

Datasets and Settings. Our methodology is rigorously evaluated on a variety of datasets, including KITTI [Geiger *et al.*,

2012], Cityscapes [Cordts *et al.*, 2016b], Foggy Cityscapes [Sakaridis *et al.*, 2018b], Sim10k [Johnson-Roberson *et al.*, 2016], and BDD100k [Yu *et al.*, 2020], which present a broad spectrum of challenges for domain adaptive object detection:

- **Cross Camera Adaptation.** We evaluate the domain drift between the real domain and the similar domain caused by different cameras, which is an important factor leading to domain differences in the real-world. We employ the training set of KITTI and the Cityscapes as the source and target domains, respectively.
- **Weather Adaptation.** In this scenario, we use Cityscapes as the source dataset, consisting of 2,975 training images and 500 evaluation images. The counterpart, known as Foggy Cityscapes, is derived from Cityscapes through a fog synthesis algorithm. These datasets enable us to assess the effectiveness of our method in adapting object detection models from clear weather to foggy conditions.
- **Scene Adaptation.** Cityscapes is again the source dataset in this instance, with the daytime subset of BDD100k acting as the target. The BDD100k subset, encompassing 36,728 training and 5,258 validation images, is rich in diverse daylight scenes, each annotated with bounding boxes.
- **Synthetic to Real Adaptation.** In this scenario, the source domain is Sim10k, generated through the Grand Theft Auto game engine, containing 10,000 training images with 58,701 bounding box annotations. Cityscapes is employed as the target domain, specifically focusing on car instances for both training and evaluation.

Implementation Details. We benchmark our approach against cutting-edge domain adaptation methods across two categories: (1) The Faster RCNN series, including DAF [Chen *et al.*, 2018], MAF [He and Zhang, 2019], ATF [He and Zhang, 2020], HTCN [Chen *et al.*, 2020], UMT [Deng *et al.*, 2021], PAATF [He *et al.*, 2022], PTMAF [He *et al.*, 2023], and IDF [Lang *et al.*, 2023]. (2) The Deformable DETR series, encompassing SFA [Wang *et al.*, 2021b].

We validate the effectiveness and versatility of our method against various baselines, including **DAF**, **MAF**, **IDF**, and **SFA**. By default, we employ ImageNet pre-trained ResNet-50 [He *et al.*, 2016] and VGG-16 [Simonyan and Zisserman, 2014] as CNN backbones in all experiments. Aligning with the Faster RCNN series, we train the network using the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate is set to 1×10^{-3} and is reduced to 1×10^{-4} after 5 epochs. A total of 15 epochs are conducted, with a batch size of 2 maintained throughout. In line with the Deformable DETR series, we utilize the Adam optimizer [Kingma and Ba, 2015] for training over 50 epochs. The learning rate is initialized at 2×10^{-4} and reduced by a factor of 0.1 after 40 epochs. A batch size of 4 is employed consistently in all experiments. All these experiments are conducted using RTX3090 NVIDIA GPUs.

4.2 Comparisons with SOTA Methods

Cross Camera Adaptation. In our investigation into cross-camera adaptation, we focused on the transfer between KITTI (K) and Cityscapes (C), with detailed results outlined in Table 1. Notably, CAGrad facilitated enhancements in detection accuracy across various models DAF, MAF, IDF, and SFA. Specifically, IDF combined with CAGrad surpassed the baseline IDF by significant margins of 1.4% and 1.3% in the K→C and C→K tasks, respectively. SFA augmented with CAGrad exhibited an increment of 1.4% in K→C and an impressive 8.6% in comparison to the source-only Deformable DETR model. This underscores CAGrad's capacity to seamlessly enhance existing alignment-based DAOD methods, thereby elevating their efficacy in image detection tasks.

Methods	Detector	K → C	C → K
FRCNN	FRCNN	30.2	53.8
DAF	FRCNN	38.5	64.1
MAF	FRCNN	41.0	72.1
ATF	FRCNN	42.1	73.5
PAATF	FRCNN	42.9	74.1
IDF	FRCNN	42.1	74.0
DAF+**CAGrad**	FRCNN	40.7	68.3
MAF+**VFDD**	FRCNN	43.1	74.6
IDF+**CAGrad**	FRCNN	43.5	75.3
DefDETR	DefDETR	39.5	-
SFA	DefDETR	46.7	-
SFA+**CAGrad**	DefDETR	48.1	-

Table 1: Results of different methods for cross camera adaptation,*i.e.,* KITTI to Cityscapes and Cityscapes to KITTI.

Weather Adaptation. In assessing the robustness of our target detector under varied weather conditions, we executed a cross-domain model transition from Cityscapes to Foggy Cityscapes. The outcomes, as tabulated in Table 2, reveal that upon integrating CAGrad, models such as DAF, MAF, IDF, and SFA reported mAP values of 33.6%, 37.0%, 43.7%, and 43.5% respectively. Remarkably, the implementation of CAGrad significantly bolstered the cross-domain efficacy of the Deformable DETR, achieving a noteworthy absolute mAP increment of 2.2% (from 41.3% to 43.5%). This improvement attests to CAGrad's ability to rectify the discord and imbalance between domain alignment and detection tasks in the optimization phase, thereby effectively enhancing detection accuracy.

Scene Adaptation. The adaptability of models to varying scenes is crucial. Our proposed method, CAGrad, demonstrates its effectiveness in scene adaptation, as shown in Table 3. It can be observed that after applying CAGrad, the mAP of DAF, SWAD and SFA reaches 26.5%, 27.5%, and 31.6%, respectively. Notably, performance improvements are observed across all categories within the target dataset.

Synthetic to Real Adaptation. Within the synthetic-to-real adaptation domain, the efficacy of CAGrad was rigorously evaluated, as delineated in Table 4. Impressively, CAGrad surpassed the baseline models DAF, MAF, IDF, and SFA by margins of 2.1%, 2.4%, 1.6%, and 2.0%, respectively.

4.3 Analysis and Discussions

Convergence Analysis. Figure 5 illustrates the test accu-

Methods	Detector	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
FRCNN	FRCNN	24.1	33.1	34.3	4.1	22.3	3.0	15.3	26.5	20.3
DAF	FRCNN	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
MAF	FRCNN	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
ATF	FRCNN	34.6	48.0	50.0	23.7	43.3	38.7	33.4	38.8	38.7
HTCN	FRCNN	33.2	47.5	47.9	31.6	47.4	40.9	32.3	37.1	39.8
UMT	FRCNN	33.0	45.9	48.6	34.1	56.5	46.8	30.4	37.3	41.7
PAATF	FRCNN	37.9	49.6	52.8	27.0	46.6	48.7	33.6	39.5	42.0
PTMAF	FRCNN	37.3	49.4	52.2	26.7	49.5	34.5	34.9	41.2	40.7
IDF	FRCNN	37.4	50.1	52.8	31.3	50.6	42.0	33.7	41.7	42.4
DAF+**CAGrad**	FRCNN	30.6	40.6	43.8	25.8	38.4	32.6	24.3	32.4	33.6
MAF+**CAGrad**	FRCNN	31.2	41.8	46.0	25.7	42.3	42.1	30.5	36.1	37.0
IDF+**CAGrad**	FRCNN	38.4	51.4	54.7	33.5	51.0	43.1	35.0	42.5	43,7
DefDETR	DefDETR	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
SFA	DefDETR	46.5	48.6	62.6	25.1	46.2	29.4	28.3	44.0	41.3
SFA+**CAGrad**	DefDETR	48.0	50.2	65.7	26.2	47.8	32.2	32.0	46.1	43.5

Table 2: Results of different methods for weather adaptation, *i.e.,* Cityscapes to Foggy Cityscapes. FRCNN and DefDETR are abbreviations for Faster RCNN based on the VGG-16 and Deformable DETR based on the ResNet-50, respectively.

Methods	Detector	person	rider	car	truck	bus	mcycle	bicycle	mAP
FRCNN	FRCNN	29.3	28.2	45.7	15.5	16.6	16.0	22.1	24.8
DAF	FRCNN	26.9	22.1	44.7	17.4	16.7	17.1	18.8	23.4
SWDA	FRCNN	30.2	29.5	45.7	15.2	18.4	17.1	21.2	25.3
DAF+**CAGrad**	FRCNN	30.6	28.5	46.5	19.5	19.1	18.4	22.8	26.5
SWDA+**CAGrad**	FRCNN	33.1	31.7	46.9	19.3	19.7	18.7	23.0	27.5
DefDETR	DefDETR	38.9	26.7	55.2	15.7	19.7	10.8	16.2	26.2
SFA	DefDETR	40.2	27.6	57.5	19.1	23.4	15.4	19.2	28.9
SFA+**CAGrad**	DefDETR	43.0	30.3	59.1	23.0	25.4	16.8	23.5	31.6

Table 3: Results of different methods for scene adaptation,*i.e.,* Cityscapes to BDD100k daytime subset.

Methods	Detector	car AP
FRCNN	FRCNN	34.6
DAF	FRCNN	38.9
MAF	FRCNN	41.1
HTCN	FRCNN	42.5
PAATF	FRCNN	43.7
PTMAF	FRCNN	43.2
IDF	FRCNN	43.9
DAF+**CAGrad**	FRCNN	41.0
MAF+**CAGrad**	FRCNN	43.5
IDF+**CAGrad**	FRCNN	45.5
DefDETR	DefDETR	47.4
SFA	DefDETR	52.6
SFA+**CAGrad**	DefDETR	54.6

Table 4: Results of different methods for synthetic to real adaptation,*i.e.,* Sim10k to Cityscapes.

racy's convergence trajectories relative to iteration counts in tasks transitioning from Cityscape to Foggy Cityscape. Here, the test accuracy of the baseline models is depicted by the blue curve, while the red curve shows the test error for baseline models enhanced with CAGrad. It is evident that, when compared to the baseline alone, the combination of baseline and CAGrad achieves quicker convergence and a reduced test

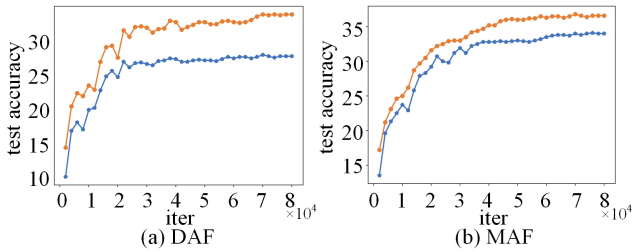

Figure 5: Convergence curves of various baselines and baseline+CAGrad on the test error (%). Clearly, the test performance is improved with CAGrad.

error. This effectively demonstrates CAGrad's significant role in facilitating the optimization process, steering both domain alignment and detection tasks towards more favorable outcomes.

Detection Results. In Figure 6 (a), we show some visual results by SFA (baseline) and our SFA+CAGrad, accompanied with the ground-truth. As can be seen, in all three scenarios, SFA+CAGrad improves the detection performance. It successfully mitigates the false positives generated by SFA and detects challenging objects overlooked by SFA.

Visualization of Feature Distribution. Figure 6 (b) presents

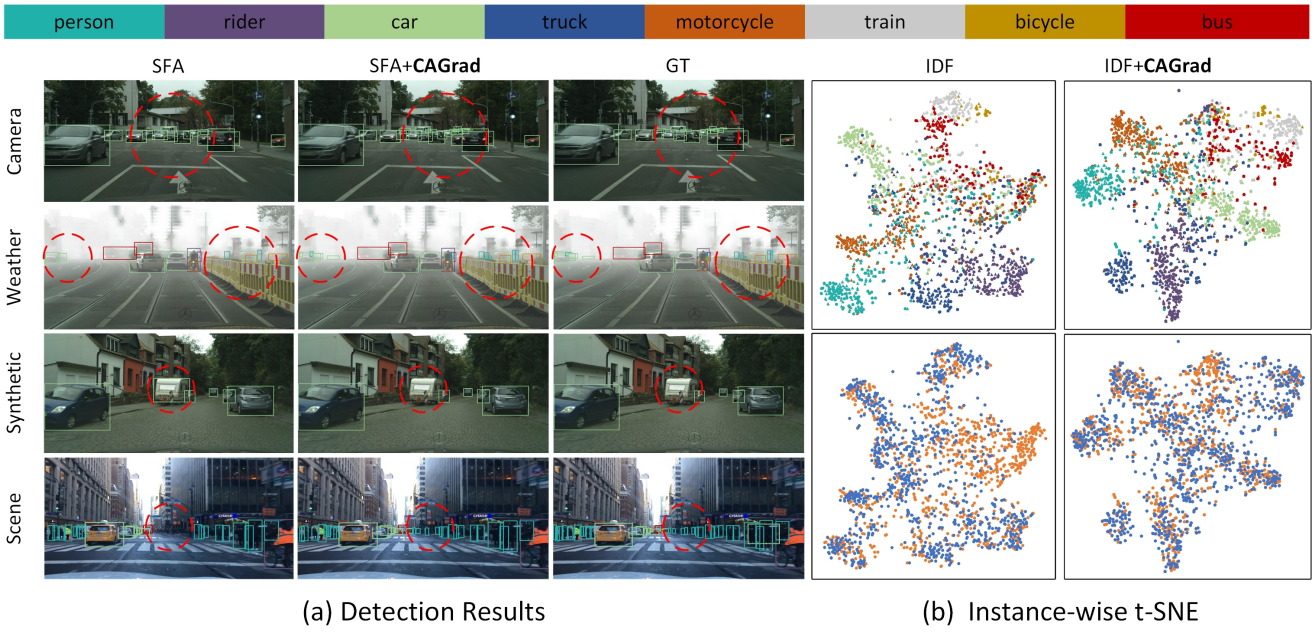

Figure 6: (a) A qualitative assessment juxtaposes SFA+CAGrad against the preceding SOTA technique and Ground Truth (GT) across four distinct scenarios. The regions highlighted in red underscore the enhanced performance achieved by our approach. (b) In Cityscapes to Foggy Cityscapes, instance-level feature t-SNE results. Colors in the first row represent classes, while orange signifies the source domain and blue signifies the target domain in second row.

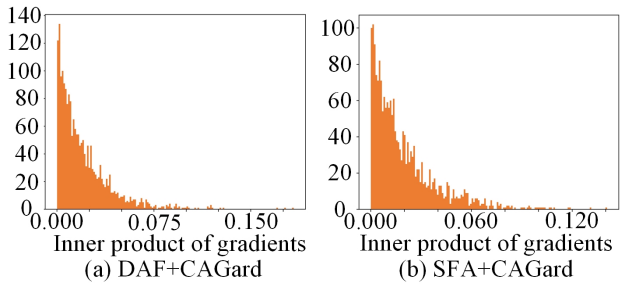

Figure 7: Inner product distributions (histogram) of two gradients before and after CAGrad on Cityscape to Foggy Cityscape. Clearly, the between-task gradient conflict is eliminated after harmonization.

the t-SNE visualizations illustrating the feature distributions generated by IDF (baseline) and IDF+CAGrad in the context of weather adaptation. The features processed using CAGrad demonstrate a more pronounced clustering effect, with a diminished presence of samples straddling class boundaries. This enhances the discriminative capacity of the features. Furthermore, these visualizations corroborate the efficacy of the CAGrad mechanism in learning enhancement.

Effectiveness Analysis. Figure 7 presents the inner product distributions of gradients between domain alignment and detection tasks after applying CAGrad to DAF and SFA during the entire training and updating process. Compared to Figure 2, after applying CAGrad, as shown in Figure 7, the inner products of both gradients are positive. This means that the gradient angles of the two tasks have been coordinated into acute angles. The proposed CAGrad avoids optimization conflicts by separately adjusting the gradients of the two tasks to achieve optimal coordination. Experimental results fully illustrate the effectiveness of the proposed CAGrad.

5 Conclusion

This research addresses the challenge of optimization conflict in unsupervised domain adaptive object detection models, specifically between the alignment and detection tasks. To address this challenge, we introduce an innovative yet straightforward technique, the Conflict-Alleviated Gradient descent (CAGrad), which effectively resolves gradient conflicts for each task. Furthermore, we propose a rapidly optimizable equivalent model, CAGrad+DAOD, implementing swift integration of CAGrad. This approach ensures that both the alignment and detection tasks are harmoniously balanced during DAOD optimization. A comprehensive array of experimental evaluations and model analyses substantiate that CAGrad markedly enhances the performance of existing alignment-based DAOD models, contributing to state-of-the-art outcomes. Additionally, CAGrad demonstrates scalability and is underpinned by a solid theoretical framework.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No.62176217).

References

[Cai and Vasconcelos, 2018] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[Chen *et al.*, 2018] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018.

[Chen *et al.*, 2020] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8869–8878, 2020.

[Cordts *et al.*, 2016a] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[Cordts *et al.*, 2016b] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[Cui *et al.*, 2020] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *CVPR*, pages 3941–3950, 2020.

[Deng *et al.*, 2021] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[He and Zhang, 2019] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019.

[He and Zhang, 2020] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *ECCV*, pages 309–324, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *CVPR*, pages 2961–2969, 2017.

[He *et al.*, 2022] Zhenwei He, Lei Zhang, Yi Yang, and Xinbo Gao. Partial alignment for object detection in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5238–5251, 2022.

[He *et al.*, 2023] Zhenwei He, Lei Zhang, Xinbo Gao, and David Zhang. Multi-adversarial faster-rcnn with paradigm teacher for unrestricted object detection. *International Journal of Computer Vision*, 131(3):680–700, 2023.

[Huang *et al.*, 2022] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *IJCAI*, pages 972–979. International Joint Conferences on Artificial Intelligence, 2022.

[Johnson-Roberson *et al.*, 2016] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016.

[Kang *et al.*, 2019] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015.

[Lang *et al.*, 2023] Qinghai Lang, Lei Zhang, Wenxu Shi, Weijie Chen, and Shiliang Pu. Exploring implicit domain-invariant features for domain adaptive object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1816–1826, 2023.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[Liu *et al.*, 2023] Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, and Christian Rupprecht. Continual detection transformer for incremental object detection. In *CVPR*, pages 23799–23808, 2023.

[Pan and Yang, 2009] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[Sakaridis *et al.*, 2018a] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[Sakaridis *et al.*, 2018b] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Tian *et al.*, 2019] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *iccv*, pages 9627–9636, 2019.

[Wang *et al.*, 2021a] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *ACM MM*, pages 1730–1738, 2021.

[Wang *et al.*, 2021b] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *ACM MM*, pages 1730–1738, 2021.

[Wu *et al.*, 2021] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, pages 9342–9351, 2021.

[Wu *et al.*, 2022] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE TPAMI*, 44(8):4178–4193, 2022.

[Xu *et al.*, 2020] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020.

[Yu *et al.*, 2020] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.