Foundation Model-Driven Mask Consistency for Weakly Supervised Domain Adaptation in Monocular 3D Detection

Benjamin Lorenz* Joshua Niemeijer†

Abstract: Monocular 3D object detection offers a scalable perception solution for autonomous driving, yet models trained on perfectly annotated synthetic data suffer performance drops in real-world environments due to the severe sim-to-real domain gap. We study a weakly supervised domain adaptation setting in which the synthetic source domain provides full 3D annotations, while the real target domain provides only auxiliary 2D bounding-box annotations and no target-domain 3D labels for detector supervision. Mean Teacher self-training attempts to bridge this gap but frequently falls into a "confidence trap." Confronted with unfamiliar real-world data, the teacher model generates noisy pseudo-labels, causing standard confidence-thresholding methods to aggressively discard predictions and ultimately deprive the student model of a viable supervisory signal. To overcome this, we propose a novel Weakly Supervised Domain Adaptation (WSDA) framework that integrates off-the-shelf foundation models into the self-training loop to provide external object-level guidance. By prompting Segment Anything v2 (SAM2) with target-domain 2D box centers to generate car-candidate object masks, we independently validate the teacher's 3D bounding box predictions against mask consistency criteria, filtering out mask-inconsistent False Positives. Furthermore, we introduce a False Negative (FN)-Whitening strategy that explicitly erases the visual features of missed objects in the input image. This creates a don't-care-like input masking region that preserves surrounding scene context while reducing destructive background associations. Our mask-based filtering mechanism stabilizes the teacher-student dynamic in our experiments and mitigates the confirmation bias cycle. Extensive evaluations on our custom KITTI split show that the proposed framework improves over a standard self-training baseline, achieving a relative performance gain of over 34% in Moderate AP_{3D}.

Keywords: Weakly Supervised Domain Adaptation, Monocular 3D Object Detection, Sim-to-Real, Foundation Models, Self-Training, 2D Box Supervision

*Research Assistant, German Aerospace Center (DLR), Institute of Transportation Systems (e-mail: Benjamin.Lorenz@dlr.de).

†Research Assistant, German Aerospace Center (DLR), Institute of Transportation Systems (e-mail: Joshua.Niemeijer@dlr.de)

1 Introduction

Monocular 3D object detection provides a scalable, inexpensive perception solution for autonomous driving, extracting spatial information from a single RGB image instead of costly LiDAR point clouds. However, training robust models requires large datasets. Manual annotation of real-world 3D scenes is prohibitively expensive, time-consuming, and prone to error. To bypass this bottleneck, one can use synthetic datasets from simulation engines (e.g., Virtual KITTI 2) that provide almost limitless, perfectly annotated data. Unfortunately, a severe sim-to-real domain gap persists. Due to complex prior, covariate, and concept shifts between rendered graphics and physical sensors, models trained exclusively on simulated data fail in the real world [1] (Figure 1).

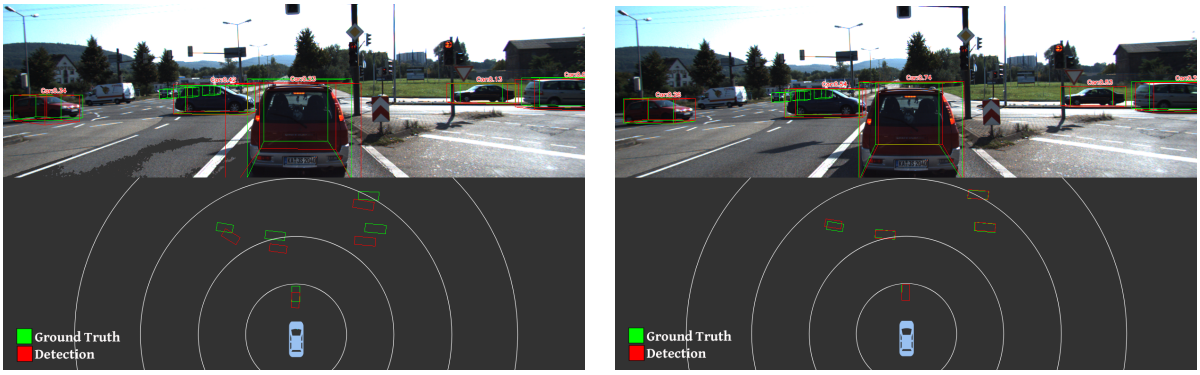

Figure 1: The Sim-to-Real Domain Gap. A model trained exclusively on the synthetic source domain completely fails to generalize to real-world target data (left), compared to a model trained on real target data (right).

To bridge this gap while avoiding expensive target-domain 3D annotation, we consider a weakly supervised domain adaptation setting. The synthetic source domain provides full 3D annotations, while the real target domain provides RGB images with auxiliary 2D bounding-box annotations but no target-domain 3D labels for detector supervision. For output-space adaptation, the Mean Teacher paradigm [2, 3] is highly effective. As illustrated in Figure 2, a teacher model, stabilized via an exponential moving average of the student's weights, generates pseudo-labels on real-world target data to supervise the actively learning student.
However, visual differences between simulation and reality make this dynamic volatile. Unfiltered, a synthetic-trained teacher generates noisy predictions on real-world images, trapping the student into a cycle of confirmation bias [4]. Standard self-training methods apply strict confidence thresholds to prevent error accumulation by discarding uncertain predictions. We reveal that this strict filtering introduces a bottleneck: the "confidence trap" (Figure 4). When confronted with unfamiliar target domain data, the teacher model yields systematically low-confidence predictions. Because these predictions fail to exceed the required thresholds, valid pseudo-label generation effectively collapses, as demonstrated in Figure 4. This state of severe *pseudo-label scarcity* deprives the student model of the supervisory signal required to continue adaptation.

To mitigate this cycle, we propose a novel foundation model-guided WSDA framework. Instead of relying on the teacher's fragile internal confidence scores alone, we deploy an

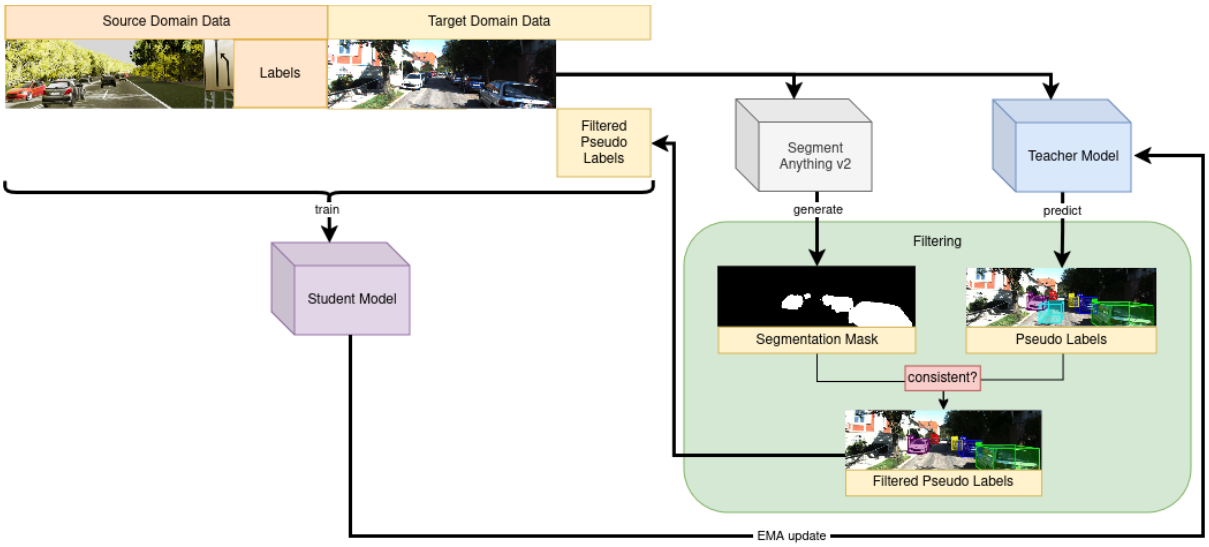

Figure 2: Overview of our proposed WSDA framework. The standard Mean Teacher loop is stabilized by using SAM2 to generate external object masks, filtering out noisy pseudo-labels before the student model is updated.

off-the-shelf foundation model, SAM2 [5], as an independent validator. SAM2 generates target-domain object masks from auxiliary 2D box-center prompts, enforcing mask consistency on the teacher's predictions. We discard any pseudo-label misaligned with the mask, neutralizing False Positive (FP)s before they corrupt the student. Additionally, we introduce a FN-Whitening strategy. When the teacher misses an object identified by the mask, we erase (whiten) its pixels in the input image. This creates a don't-care-like input region, preserving surrounding context while reducing destructive background associations.

By stabilizing self-training with external object-level guidance, we improve adaptation under severe sim-to-real shift. Our main contributions are:

- A weakly supervised domain adaptation framework for monocular 3D object detection that uses synthetic 3D source labels and auxiliary target-domain 2D bounding boxes, without relying on target-domain 3D labels for detector supervision.
- A foundation model-guided mask-consistency strategy that prompts SAM2 with target-domain 2D box centers to generate car-candidate masks and validate teacher-generated 3D pseudo-labels.
- The FN-Whitening strategy, an augmentation technique addressing FNs by masking missed objects. This creates don't-care-like input regions while preserving visual context important for depth estimation.
- Extensive evaluations on our custom KITTI split show that our framework mitigates the confidence trap, yielding over 34% relative performance improvement compared to a standard self-training baseline.

2 Related Work

2.1 Monocular 3D Object Detection

Monocular 3D object detection connects 2D images with 3D spatial understanding. Early anchor-based frameworks [6] relied on predefined bounding boxes, but the field shifted toward anchor-free architectures [7] that directly predict attributes from deep feature maps. We adopt MonoLSS [8] as our 3D detection backbone. MonoLSS reformulates 3D property regression as a sample selection problem, using a Learnable Sample Selection (LSS) module to target optimal features and ignore occlusions. However, MonoLSS struggles with large domain shifts. We address this limitation by wrapping MonoLSS in a mask-guided self-training loop, enabling its feature selection to adapt to real-world target domains.

2.2 Domain Adaptation and Self-Training

While Unsupervised Domain Adaptation (UDA) is prevalent in LiDAR detection, research addressing the extreme sim-to-real shift for monocular models remains scarce [9]. The Mean Teacher paradigm [2] remains a popular output-space adaptation strategy, generating pseudo-labels via an exponential moving average of student weights. However, the sim-to-real gap renders this self-training fragile. Unfiltered teachers generate inaccurate predictions, trapping students in a confirmation bias cycle [4]. Standard self-training mitigates this with strict confidence thresholds. We demonstrate that this creates a "confidence trap": the teacher model's prediction scores degrade, halting pseudo-label generation and depriving the student model of supervisory data. Our proposed weakly supervised framework mitigates this cycle by using auxiliary 2D target annotations to guide foundation-model mask generation. Instead of relying only on internal thresholds, we apply external mask-based validation. Filtering pseudo-labels against independent masks stabilizes self-training, ensuring a continuous flow of targets.

2.3 Foundation Models as Priors

Foundation models exhibit powerful zero-shot generalization without task-specific retraining [10]. We leverage SAM2 [5] for segmentation. While foundation models aid various tasks, their use as external validators in monocular 3D domain adaptation remains underutilized. Standard self-training relies on the teacher's noisy internal features. We instead position SAM2 as an independent validator. Cross-referencing target-domain masks against the teacher's 3D bounding box projections enforces strict mask consistency. When validation fails due to a missed object (FN), our FN-Whitening strategy erases the corresponding input pixels. This reduces destructive background associations caused by teacher omissions and preserves surrounding visual context, unlike naive filtering. Ultimately, our SAM2-guided framework leverages foundation models to mitigate the confidence trap and stabilize sim-to-real adaptation.

3 Methodology

We propose a mask-guided WSDA framework to bridge the sim-to-real domain gap without using target-domain 3D labels for detector supervision. Building upon the Mean

Teacher paradigm, we deploy a foundation model to independently validate pseudo-labels. This mitigates the "confidence trap" (Figure 4) that typically halts the adaptation process, ensuring a more continuous and stable supervisory signal.

3.1 Preliminaries

Problem Formulation. For weakly supervised domain adaptation in monocular 3D object detection, we define a labeled synthetic source domain $D^S = \{(x_S^i, y_S^i, K_S^i)\}_{i=1}^{N_S}$, where each of the N_S samples includes a 2D image x_S^i, camera intrinsics K_S^i, and a ground-truth 3D label $y_S^i = (k, l, s, \psi)$ denoting object class, location, size, and orientation. The real-world target domain is $D^T = \{(x_T^i, K_T^i, b_T^i)\}_{i=1}^{N_T}$, where b_T^i denotes auxiliary 2D bounding-box annotations used only for mask prompting. No target-domain 3D labels are used for detector supervision during adaptation. We train on $\{D^S, D^T\}$ and evaluate on an unseen target test set, D_{test}^T.

Baseline Framework. We use MonoLSS [8] as our 3D detection backbone. It uses a Learnable Sample Selection (LSS) module to sample optimal features for 3D regression. We adapt this using a Mean Teacher framework [2]. A student model (Θ^{stud}) learns via gradient descent, while a teacher (Θ^{teach}) provides target-domain pseudo-labels. To maintain stability, the teacher updates via an Exponential Moving Average (EMA) of student weights at step t:

$$\Theta_t^{teach} = \alpha\Theta_{t-1}^{teach} + (1-\alpha)\Theta_t^{stud}$$

where α is a smoothing coefficient. The student minimizes a combined loss [11]:

$$L_{total} = L_{sup} + \lambda_{wsda} \cdot L_{wsda}$$

Here, L_{sup} is the supervised loss on D^S, and L_{wsda} measures the discrepancy between student predictions and teacher pseudo-labels $\hat{y}_T$ on D^T. The hyperparameter λ_{wsda} balances adaptation.

3.2 Foundation Model-Guided Filtering

Standard self-training degrades when noisy pseudo-labels trap the student in confirmation bias. Rather than relying on naive confidence thresholds that cause pseudo-label scarcity, we introduce an external validation mechanism driven by mask consistency.

Auxiliary Mask Generation. For external guidance, we leverage the SAM2 foundation model [5]. To generate car-candidate masks, we extract the geometric center (x, y) of each target-domain 2D bounding box annotation and provide these single coordinate points as prompts to SAM2. These 2D annotations constitute weak target-domain supervision and are used only to generate object masks, not as 3D detector labels. Rather than dynamically prompting SAM2 during the adaptation phase, we directly evaluate these pre-computed binary masks (white for object, black for background) to validate the teacher's proposals.

Mask-Consistency Filtering. As shown in Figures 2 and 3, we evaluate pseudo-label quality by checking alignment with target image masks. A consistent 3D prediction must project into a 2D bounding box that overlaps the object's segmentation mask. We quantify this using a Mask Overlap Score ($\mathcal{S}_{mask}$). Let B be the pixel coordinates within a

predicted 2D bounding box, and M_{car} be the pixels belonging to the selected car-candidate mask. The score is:

$$\mathcal{S}_{mask} = \frac{|B \cap M_{car}|}{|B|}$$

If $\mathcal{S}_{mask}$ falls below a threshold (empirically 0.4), the detection lacks sufficient mask support and is flagged as a FP. We discard these mask-inconsistent pseudo-labels, preventing the student from incorporating FPs into its training signal.

FN-Whitening Strategy. The teacher model also suffers from missed objects (FNs), failing to generate pseudo-labels for valid objects present in the image. We identify these missed vehicles using the foundation model's object mask. Ignoring them inadvertently trains the student to classify real objects as background. To combat this, we introduce FN-Whitening. When the mask reveals a missing pseudo-label for an object, we assign a white color to its corresponding pixels in the target image, effectively erasing it to create a don't-care-like input region. This reduces the risk of penalizing the student for teacher omissions and limits destructive background associations. Crucially, we restrict whitening exclusively to FNs (missed objects). Because FPs typically occur on background regions such as empty roads, whitening them introduces large, artificial occlusions on the ground plane. Monocular 3D object detection heavily relies on this spatial context and road geometry to infer the depth of valid objects. In contrast, whitening FNs simply masks out an unlabeled vehicle without disrupting the surrounding structural cues.

4 Experiments

4.1 Implementation Details

Datasets. To evaluate our WSDA framework across the sim-to-real domain shift, we use Virtual KITTI 2.0.3 [12] as our labeled synthetic source domain. For the weakly annotated real-world target domain, we use the KITTI 3D Object Detection Benchmark [13]. Since official KITTI test labels are withheld, we divide the 7,481 training frames into a custom split: 70% training, 20% validation, and 10% testing. All experiments focus exclusively on the "Car" category, the most frequent class in both domains. During adaptation, target-domain 2D bounding boxes are used only to prompt SAM2; target-domain 3D labels are not used for detector supervision.

Baselines. We compare our method against three baselines:

- **Lower Bound (Source-Only):** MonoLSS trained exclusively on Virtual KITTI 2 and evaluated on the real-world KITTI split, establishing the performance "floor" of the unmitigated sim-to-real gap.
- **Upper Bound (Target-Only):** A fully supervised model trained using ground-truth KITTI target 3D labels, establishing the theoretical performance "ceiling".
- **Standard Self-Training:** A baseline Mean Teacher self-training implementation adapting from source to target without our mask-based filtering.

Implementation Details. During adaptation, we train the student model with a batch size of 56 (28 source, 28 target images). We use the Adam optimizer [14] with a

learning rate of 0.00005 and a 5-epoch warmup over 100 total epochs. The teacher generates new pseudo-labels every 10 epochs with an EMA decay of 0.999. The unsupervised consistency loss weight (λ_{wsda}) is 0.5, slightly favoring pristine source labels over noisier target pseudo-labels. The Mask Overlap Score ($\mathcal{S}_{mask}$) threshold is set to 0.4.

4.2 Main Results

Performance Analysis. We evaluate 3D detection performance using Average Precision (AP_{3D}) at an Intersection over Union (IoU) threshold of 0.7 for the Car category, summarizing quantitative results in Table 1.

Table 1: Sim-to-real adaptation results (AP_{3D} Car, IoU=0.7). Evaluated on the custom KITTI test split. The best adaptation result is highlighted in bold.

Method	**Easy (%)**	**Moderate (%)**	**Hard (%)**
Baselines			
Lower Bound (Source Only)	3.98	2.60	2.57
Upper Bound (Target Only)	91.12	63.79	56.42
Ablation of Filtering Strategies			
Standard Self-Training Baseline	8.57	5.54	4.90
Filtering-Only	10.41	6.53	5.98
FP-FN-Whitening	10.72	6.89	6.06
Final Approach			
Proposed (FN-Whitening)	**10.99**	**7.44**	**6.22**

The Lower Bound (2.60% Moderate) illustrates the failure of purely synthetic-trained models in real-world deployments. While Standard Self-Training doubles this floor (5.54% Moderate), it remains bottlenecked by noisy pseudo-labels.

In contrast, our foundation model-guided framework with FN-Whitening achieves 7.44% Moderate AP_{3D}. This represents a 34% relative improvement over the Standard Self-Training baseline.

Difficulty Breakdown. These gains are consistent across all KITTI difficulty levels. On the Easy split, mask-based filtering reaches 10.99% (vs. the baseline's 8.57%). On the challenging Hard split, our method maintains 6.22% AP_{3D}. This demonstrates that enforcing mask consistency enhances resilience to complex, cluttered scenes. By filtering FPs and neutralizing FNs via image whitening, our framework extracts a stabilizing supervisory signal from target data.

4.3 Ablation Studies

We analyze our WSDA framework through targeted ablation studies to understand its driving mechanisms. We evaluate pseudo-label filtering strategies, analyze the stability of teacher-student learning dynamics, and study the influence of foundation model guidance quality.

4.3.1 Impact of Filtering Strategies

Managing teacher-generated noise on the target domain is a primary challenge in Mean Teacher self-training. We compared three mask-based filtering strategies against our Standard Self-Training baseline (5.54% Moderate AP_{3D}, see Table 1).

Filtering-Only: The simplest approach discards mask-inconsistent predictions. Pseudo-labels failing the Mask Overlap Score threshold are flagged as FP and removed. Preventing the student from learning incorrect background associations yielded an immediate performance jump to 6.53% (Table 1).

FN-Whitening (Proposed Method): Filtering-Only ignores FNs, i.e., missed detections that are nevertheless clearly identified by the object mask. Standard self-training ignores FNs, inadvertently teaching the student to classify valid objects as background. To counteract this, FN-Whitening erases the corresponding pixel area in the input image. This creates a don't-care-like input region, reducing the risk of penalizing the student for teacher omissions. This strategy yielded our peak performance of 7.44% Moderate AP_{3D} (see Table 1), providing a more stable training signal.

FP-FN-Whitening: Applying the same whitening augmentation to FPs, erasing features that caused teacher hallucinations, resulted in a performance degradation to 6.89% (Table 1). We attribute this decline to *context destruction*. Erasing the FP image patch destroys critical contextual information, such as road surface planes or visual boundaries of adjacent vehicles, which are vital for depth estimation. Thus, the optimal strategy is discarding incorrect FP labels without altering the image, while selectively masking FN mistakes.

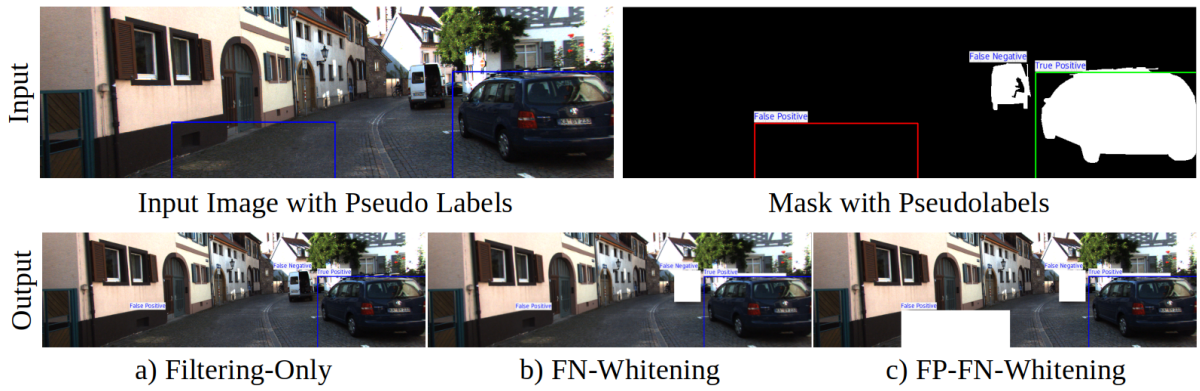

Figure 3: Comparison of mask-based filtering strategies. Top row: Input pseudo-labels and their overlap with the SAM2 mask, identifying FPs and FNs. Bottom row: **a) Filtering-Only** discards FPs but leaves FNs as background data, penalizing the student model. **b) FN-Whitening** (our approach) masks the FN to create a don't-care-like input region while preserving scene context. **c) FP-FN-Whitening** masks both FNs and FPs, destroying the surrounding visual context required to localize nearby valid objects.

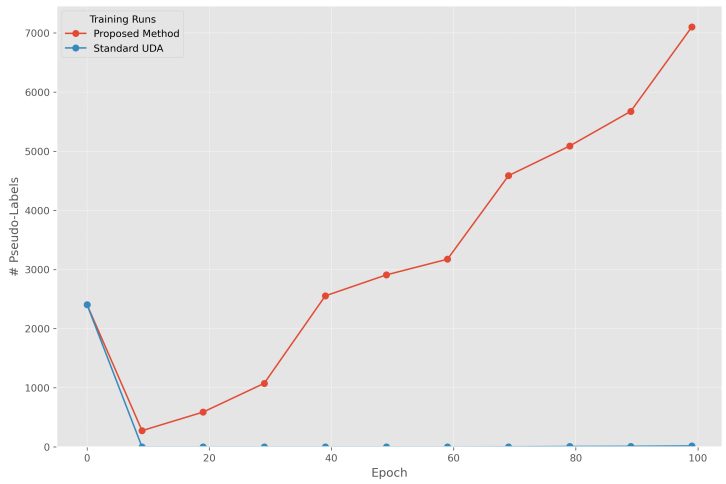

Figure 4: Total number of accepted target-domain pseudo-labels over training. Pseudo-labels are regenerated every 10 epochs. Standard Self-Training suffers a total label collapse after epoch 10 due to the confidence trap. Our proposed mask-based filtering stabilizes the learning loop, enabling a steady growth to over 7,000 valid pseudo-labels.

4.3.2 Overcoming the Confidence Trap

A fundamental precondition of the Mean Teacher paradigm is balancing pseudo-label quality and quantity. To prevent confirmation bias [4], standard self-training frameworks typically employ high confidence thresholds. However, under extreme sim-to-real domain shifts, this strict filtering triggers a degenerative cycle of pseudo-label scarcity.

We tracked pseudo-label generation across training epochs. In Standard Self-Training, the teacher encounters unfamiliar target data, causing confidence scores to collapse. Even with a decaying threshold, teacher generation collapses effectively to zero after 10 epochs. This drops the student into total pseudo-label scarcity, an unrecoverable loop where it needs target data to adapt but cannot clear the threshold to acquire it.

Conversely, our mask-based filtering stabilizes this dynamic. Despite an initial drop in label quantity, external mask-based validation guarantees a quality baseline, anchoring the framework on roughly 275 highly consistent pseudo-labels. By training strictly on validated targets, the student avoids noise accumulation, ensuring stable and accurate EMA updates for the teacher. From this stable foundation, accepted pseudo-labels grow steadily, exceeding 7,000 by the final epoch. External mask-based filtering mitigates the confidence trap in our experiments, transforming a collapsed pseudo-labeling process into a more stable adaptation loop.

4.4 Impact of Guidance Quality

Our framework depends on external object masks to validate pseudo-labels and to identify missed target-domain objects. Since SAM2 is used in a zero-shot setting, its masks are not always reliable. Qualitative inspection of Mask Set A revealed several failure cases, including incomplete masks under heavy occlusion and masks that occasionally focus on distinctive vehicle sub-parts rather than the full object.

To estimate how much these segmentation errors affect adaptation, we compare the

SAM2-derived Mask Set A with an oracle Mask Set B constructed from ground-truth target annotations. This oracle setting is used only for analysis and is not part of the weakly supervised adaptation protocol.

Using oracle masks with FN-Whitening increases Moderate AP_{3D} from 7.44% to 7.75%. Similarly, FP-FN-Whitening reaches 7.60% when supplied with oracle masks. These results indicate that imperfect foundation-model guidance does reduce performance. However, the relatively small gain obtained with oracle masks also shows that segmentation quality is not the only limiting factor. Even with perfect object masks, the framework remains constrained by the quality of the teacher's 3D pseudo-labels, the source-to-target domain gap, and the current filtering and training design.

Thus, the oracle-mask experiment should be interpreted as an upper-bound analysis for the guidance component under the current implementation, rather than as a theoretical ceiling for the full WSDA framework. While improved foundation models may further benefit the proposed approach, additional advances in pseudo-label generation, matching criteria, and adaptation losses are likely required to close the remaining gap to fully supervised target-domain training.

5 Conclusion

Monocular 3D object detection is a highly scalable perception solution for autonomous driving, yet the severe sim-to-real domain gap prevents synthetic-trained models from safely generalizing to physical environments. Weakly supervised domain adaptation offers a practical compromise between fully synthetic training and expensive target-domain 3D annotation by using cheaper auxiliary target-domain annotations. In our setting, target-domain 2D bounding boxes are used only to prompt SAM2, while target-domain 3D labels are never used for detector supervision. Building on Mean Teacher self-training, we proposed a foundation model-guided WSDA framework that validates target-domain pseudo-labels through object-mask consistency.

Cross-referencing 3D bounding box projections against SAM2 masks allows our method to systematically discard mask-inconsistent FPs. Additionally, our FN-Whitening strategy erases the visual features of objects the teacher missed. This creates a don't-care-like input region that preserves surrounding scene context while reducing destructive background associations. This mask-based filtering stabilizes the teacher-student learning loop in our experiments, mitigating the confidence trap and maintaining a substantially larger set of validated target-domain pseudo-labels than the standard self-training baseline. Consequently, our framework achieved 7.44% Moderate AP_{3D} on our custom KITTI split, corresponding to a 34% relative improvement over the Standard Self-Training.

Ablation studies using ground-truth masks indicate that the quality of external guidance influences adaptation performance, but they also show that segmentation quality alone does not explain the remaining gap to fully supervised target-domain training. Under the current implementation, oracle masks improve Moderate AP_{3D} only modestly over the SAM2-based setting, suggesting that other factors, including teacher pseudo-label quality, the remaining sim-to-real gap, and the current filtering and training design, also limit performance. This analysis highlights external object-level guidance as a useful stabilizing signal, while pointing to improved pseudo-label generation, matching criteria, and adaptation losses as important directions for future work.

References

[1] Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, Z. He, K. Kitani, and P. Vajda, "Cross-domain adaptive teacher for object detection," 2022. [Online]. Available: https://arxiv.org/abs/2111.13216

[2] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2018. [Online]. Available: https://arxiv.org/abs/1703.01780

[3] M. Schwonberg, J. Niemeijer, J.-A. Termöhlen, J. P. schäfer, N. M. Schmidt, H. Gottschalk, and T. Fingscheidt, "Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving," *IEEE Access*, vol. 11, pp. 54 296–54 336, 2023.

[4] A. Radhakrishnan, J. Davis, Z. Rabin, B. Lewis, M. Scherreik, and R. Ilin, "Design choices for enhancing noisy student self-training," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 1915–1924.

[5] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[6] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," 2019. [Online]. Available: https://arxiv.org/abs/1907.06038

[7] Z. Liu, Z. Wu, and R. Tóth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation," 2020. [Online]. Available: https://arxiv.org/abs/2002.10111

[8] Z. Li, J. Jia, and Y. Shi, "Monolss: Learnable sample selection for monocular 3d detection," 2024. [Online]. Available: https://arxiv.org/abs/2312.14474

[9] Z. Li, Z. Chen, A. Li, L. Fang, Q. Jiang, X. Liu, and J. Jiang, "Unsupervised domain adaptation for monocular 3d object detection via self-training," in *European Conference on Computer Vision*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248377732

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[11] M.-R. Amini, V. Feofanov, L. Pauletto, L. Hadjadj, E. Devijver, and Y. Maximov, "Self-training: A survey," *Neurocomputing*, vol. 616, p. 128904, Feb. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2024.128904

[12] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," 2020.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. [Online]. Available: https://doi.org/10.1177/0278364913491297

[14] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017. [Online]. Available: https://arxiv.org/abs/1412.6980