KMAN PUBLISHING INSTITUTE

Article history:
Received 13 April 20246
Revised 18 June 202
Accepted 26 June 2026
Initial Publish 27 July 2026
Published online 01 March 2027

Journal of Resource Management and Decision Engineering

Volume 6, Issue 2, pp 1-17

Development of an Intelligent Framework for Road Semantic Segmentation Using Depth and Camera Motion Constraints: An Approach for Resource Management and Engineering Decision-Making in Transportation Systems

Mohammad. Mohammadi[1*]

[1] Graduated from Master of Science in Computer Engineering - Artificial Intelligence and Robotics, K.N. Toosi University of Technology

*** Corresponding author email address**: m.mohamadi2@email.kntu.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Mohammadi, M., (2027). Development of an Intelligent Framework for Road Semantic Segmentation Using Depth and Camera Motion Constraints: An Approach for Resource Management and Engineering Decision-Making in Transportation Systems. *Journal of Resource Management and Decision Engineering*, 6(2), 1-17.
https://doi.org/10.61838/kman.jrmde.378

A B S T R A C T

One of the major challenges in the semantic segmentation of road videos is the effective use of the three-dimensional structure of the scene and camera motion, while only two-dimensional pixel-level annotations are available. This article introduces a semi-supervised learning framework for semantic segmentation that leverages geometric constraints derived from depth and camera motion in video. In this approach, for each 15-frame sequence, only the first frame has manual pixel-level labels, and for the subsequent frames, pseudo-labels are generated through the "geometric warping" of the mask from the labeled frame to the other frames. To estimate monocular depth and the relative camera motion between frames, the Monodepth2 model is used, which performs self-supervised learning of depth and camera motion through novel-view synthesis and photometric reprojection error. Based on the estimated depth and motion, each pixel from the mask of the reference frame is mapped into three-dimensional space and then projected onto the target frame. By comparing the warped depth with the estimated depth in the target frame, a validity mask is constructed to refine the pseudo-labels. These pseudo-labels are then used for semi-supervised training of the same LR-ASPP segmentation network. Evaluation on the KITTI-STEP dataset shows that the semi-supervised model based on geometric warping (SSL-Warp) achieved an mIoU of 25.73% and a pixel accuracy of 71.42%, producing a modest performance improvement compared with the first-frame-only baseline model. The analysis of the results indicates that depth and camera motion constraints are highly beneficial for static classes such as road and vegetation; however, for independently moving objects, they need to be combined with explicit motion information.

Keywords: *video semantic segmentation, semi-supervised learning, monocular depth estimation, camera motion, geometric constraints, pseudo-labeling*

1. Introduction

The digital transformation of transportation systems has shifted road-infrastructure management from a predominantly reactive and inspection-based model toward an intelligent, data-driven, and predictive paradigm. In contemporary intelligent transportation systems, cameras installed on vehicles, road corridors, mobile robots, and autonomous platforms continuously generate large-scale visual data that can support engineering decision-making, infrastructure monitoring, safety assessment, traffic control, asset management, and operational planning. However, the managerial value of these data depends on the ability of computational systems to transform raw visual streams into reliable semantic information. Among different computer-vision tasks, semantic segmentation is particularly important because it assigns a semantic category to each pixel and therefore provides fine-grained scene understanding. In road environments, semantic segmentation enables the recognition of drivable areas, sidewalks, vegetation, buildings, traffic signs, pedestrians, vehicles, and other relevant components of the mobility environment. Such information is not only technically relevant for perception modules in autonomous systems but also strategically important for transportation management, because accurate scene decomposition can improve resource allocation, risk identification, route planning, infrastructure maintenance prioritization, and engineering decisions under uncertainty.

Despite the rapid development of deep learning, robust semantic interpretation of road videos remains a challenging problem. Road scenes are characterized by perspective changes, lighting variation, occlusion, dynamic objects, camera ego-motion, and complex spatial structures. More importantly, pixel-level annotation is expensive, time-consuming, and difficult to scale across long video sequences. From a management perspective, this creates a resource-efficiency problem: transportation authorities, engineering teams, and technology developers require large amounts of annotated visual data to train reliable models, but the production of dense labels requires extensive human labor and financial cost. Therefore, methods that can exploit unlabeled or weakly labeled video frames are highly valuable because they reduce annotation dependence while increasing the operational feasibility of vision-based transportation intelligence. This concern is consistent with the broader evolution of visual localization, mapping, and perception systems, in which recent studies increasingly seek to combine deep learning, semantic reasoning, geometric modeling, and motion-aware constraints to improve performance in complex and dynamic environments (Chen et al., 2024; Sahili et al., 2023).

The problem becomes more complex when a system must process road videos rather than isolated images. In video, consecutive frames are not independent; they are related through camera motion, scene geometry, object motion, and temporal continuity. This temporal structure represents an important source of information that can be exploited to transfer labels from annotated frames to unannotated frames. However, naive temporal transfer may produce inaccurate pseudo-labels when objects move independently, when occlusions occur, or when camera motion changes rapidly. Therefore, an effective semi-supervised segmentation framework should not only propagate semantic information across frames but should also incorporate constraints that determine whether the propagated labels are geometrically reliable. In this regard, depth estimation and camera-motion estimation provide a principled basis for label propagation because they connect two-dimensional image pixels to the three-dimensional structure of the scene. The use of geometric constraints in dynamic visual systems has become a prominent research direction, particularly in visual simultaneous localization and mapping (VSLAM), where semantic and geometric information are frequently combined to improve robustness under dynamic conditions (Li et al., 2025; Liao et al., 2025; Zhang et al., 2023).

The importance of geometry-aware perception is also evident in the literature on semantic SLAM and dynamic-environment mapping. Traditional visual SLAM systems often assume that the observed environment is mostly static; however, this assumption is frequently violated in road and urban settings where vehicles, pedestrians, cyclists, and other movable agents continuously change their positions. To address this problem, recent studies have introduced semantic constraints, object detection, instance segmentation, moving-object tracking, graph optimization, point–line feature fusion, and dynamic rejection mechanisms. For instance, approaches based on semantic segmentation and object-level reasoning have been developed to remove or down-weight dynamic regions, improve localization accuracy, and enhance map consistency in dynamic environments (Chang et al., 2023; Wei et al., 2023; Yao et al., 2023; You et al., 2022). Although these studies mainly focus on localization and mapping, their underlying insight is highly relevant to semantic segmentation: reliable visual understanding in dynamic scenes requires the integration of appearance,

semantics, motion, and geometry rather than reliance on two-dimensional image features alone.

In the context of transportation and road-scene analysis, semantic information can also support localization and map matching. Semantic map matching methods for autonomous vehicles demonstrate that road objects and scene categories can provide stable cues for positioning and environmental interpretation, especially when traditional localization signals are noisy, unavailable, or insufficient (Huang et al., 2023). Similarly, visual–inertial SLAM systems use inertial priors, semantic constraints, and adaptive mechanisms to improve stability in dynamic scenes (Sun et al., 2023). These developments indicate that semantic scene understanding is not an isolated perception problem but part of a broader decision-support architecture. In practical transportation systems, the ability to segment road scenes can directly influence engineering decisions, including which areas require maintenance, where safety risks are concentrated, how traffic participants interact with infrastructure, and how autonomous or semi-autonomous systems should allocate computational attention. Accordingly, semantic segmentation can be viewed as a technical foundation for managerial and operational intelligence in transportation systems.

Recent SLAM studies have increasingly emphasized dynamic-object handling because moving objects are one of the main sources of visual inconsistency. Systems such as DMOT-SLAM, DOT-SLAM, and Strong-SLAM employ object tracking and dynamic-object reasoning to improve robustness in changing environments (Huang et al., 2024; F. Wang et al., 2024; Zhu et al., 2024). Other approaches, such as ADS-SLAM and SFD-SLAM, use adaptive motion compensation, semantic information, and saliency-region detection to identify reliable regions and reduce the disruptive effects of dynamic components (Dai et al., 2024; Gong et al., 2024). These methods suggest that dynamic regions should not be treated in the same way as static regions. For road semantic segmentation, this distinction is equally important. Static classes such as road, building, vegetation, terrain, and sky are often more consistent under camera motion and are therefore more suitable for geometric propagation. In contrast, dynamic classes such as cars, pedestrians, riders, motorcycles, and bicycles may violate the assumptions of camera-motion-based warping because their displacement is not explained solely by camera ego-motion.

The recent literature also shows that lightweight and computationally efficient methods are increasingly necessary. Intelligent transportation systems often operate under real-time or near-real-time constraints, particularly when deployed on edge devices, mobile robots, autonomous vehicles, or infrastructure-side cameras. For this reason, several studies have focused on lightweight segmentation, low computational cost, and efficient dynamic filtering. For example, LDVI-SLAM introduced a lightweight monocular visual–inertial SLAM approach based on motion constraints for dynamic environments, while DPLS-SLAM combined point–line feature fusion with a lightweight improved YOLOv8seg network (Jiang et al., 2025; K. Wang et al., 2024). Similarly, recent work has proposed low-computational-cost semantic VSLAM using frame skipping, dual filtering, and adaptive motion estimation (Gong et al., 2026). These directions are particularly relevant for management-oriented engineering applications because the feasibility of deployment depends not only on model accuracy but also on computational cost, annotation cost, scalability, and maintainability. Therefore, a segmentation framework that reduces annotation requirements while using available geometric cues can contribute to both technical performance and resource-efficient decision-making.

Another major trend in the literature is the use of object detection and segmentation to distinguish dynamic and static regions. Systems based on YOLO variants, YOLACT++, mask propagation, and deep mask segmentation have been proposed to improve perception robustness in dynamic scenes (Gao et al., 2025; Li et al., 2024; Shen & Zhang, 2025; Zhang et al., 2025). For example, YS-SLAM uses YOLACT++-based semantic visual SLAM for mobile robots in dynamic environments, while DMS-SLAM relies on deep mask segmentation to support semantic visual SLAM under dynamic conditions (Gao et al., 2025; Li et al., 2024). These works show that segmentation masks are not merely output products; they can function as intermediate decision tools that guide localization, mapping, dynamic-object rejection, and optimization. In the present study, segmentation masks play a similar strategic role, but the focus is reversed: instead of using segmentation to improve SLAM, estimated depth and camera motion are used to improve semi-supervised semantic segmentation through geometrically informed pseudo-label generation.

Geometric constraints have been especially important in recent dynamic SLAM research because they provide an interpretable mechanism for assessing whether visual correspondences are physically plausible. DEG-SLAM, for example, uses object detection and geometric constraints to address degenerate motion, while methods based on direct

geometric constraints and depth image segmentation attempt to improve robustness when dynamic objects and camera motion complicate visual estimation (Cao, 2025; Liao et al., 2025). Similarly, semantic and geometric constraints have been tightly coupled in complex dynamic environments to improve the discrimination between reliable static information and unstable dynamic information (Li et al., 2025; Zhang et al., 2023). These ideas are central to the present research because road-video pseudo-labels generated only through appearance similarity or frame-level propagation may be unreliable. By contrast, pseudo-labels generated through depth- and motion-based geometric warping can be evaluated according to depth consistency, allowing the model to ignore pixels whose transferred labels are likely to be geometrically invalid.

At the same time, the literature indicates that no single strategy fully resolves the complexity of dynamic scenes. Some methods rely on moving-object tracking, some use semantic segmentation, some emphasize feature fusion, and others incorporate graph optimization, mask restoration, region growing, or large visual models (Liu et al., 2025; Wang et al., 2025; Wu et al., 2025; Zheng et al., 2025). For example, motion segmentation has been proposed for indoor GNSS-denied environments, while instance segmentation and point–line feature fusion have been used to move toward biologically inspired visual SLAM in dynamic environments (Liu et al., 2025; Wu et al., 2025). In another direction, semantic RGB-D systems with coarse-to-fine dynamic rejection and static weighted optimization have been developed to enhance system stability (Wang et al., 2025). The emergence of large visual models in semantic SLAM further indicates that the field is moving toward richer semantic representations and more adaptive scene interpretation (Zheng et al., 2025). These developments collectively demonstrate that dynamic-scene perception requires the integration of multiple sources of information, while also highlighting the need for methods that remain practical under limited annotation and computational resources.

For road semantic segmentation, the central methodological challenge is therefore how to use the geometric information embedded in video without requiring dense annotations for every frame. A semi-supervised setting in which only the first frame of each sequence is manually labeled reflects a realistic resource-management scenario. In such a setting, the model must generate pseudo-labels for subsequent frames by exploiting the temporal and geometric relationships between frames. Depth estimation allows the system to lift two-dimensional pixels into three-dimensional space, while camera-motion estimation allows the projected position of those pixels to be calculated in target frames. When the warped depth is consistent with the depth directly estimated in the target frame, the transferred label is more likely to be reliable; when inconsistency is high, the pixel can be ignored to avoid contaminating training with erroneous pseudo-labels. This logic aligns with the broader emphasis on adaptive filtering, dynamic rejection, and geometric verification found in contemporary semantic and dynamic SLAM systems (Cao, 2025; Gong et al., 2026; Wang et al., 2025; Zhang & Shen, 2025).

The managerial relevance of this approach lies in its potential to reduce the human and computational resources required for developing reliable transportation perception systems. Manual annotation of road videos is expensive, especially when dense pixel-level labels are required across long sequences. By using a single annotated frame to generate pseudo-labels for subsequent frames, the proposed framework addresses a practical bottleneck in the deployment of intelligent transportation analytics. Moreover, the use of depth consistency as a validity criterion introduces a quality-control mechanism that can reduce the risk of training on incorrect labels. In engineering-management terms, this contributes to more efficient data utilization, improved model-development workflows, and better decision-support systems for transportation applications. Related studies on accurate RGB-D visual SLAM, region-growing-based dynamic SLAM, and non-blocking semantic detection with mask propagation indicate that robustness, efficiency, and reliability are increasingly treated as integrated design requirements rather than separate technical objectives (Li & Luo, 2024; Zhang et al., 2025; Zhang & Shen, 2025).

Nevertheless, the use of depth and camera-motion constraints also has inherent limitations. Since the geometric transformation is based primarily on camera ego-motion and estimated scene depth, it works best when the observed scene is static or quasi-static. In road environments, however, many important classes are dynamic and may move independently of the camera. Vehicles, pedestrians, cyclists, and riders may change position between frames in ways that cannot be explained by the camera-motion matrix alone. As a result, geometrically warped labels may be inaccurate or may need to be removed by the depth-consistency filter. This limitation is consistent with recent research showing that robust performance in dynamic environments often requires explicit modeling of object

motion, semantic awareness, tracking, and dynamic filtering (Huang et al., 2024; Liu et al., 2025; F. Wang et al., 2024; Zhu et al., 2024). Therefore, a depth- and camera-motion-based semi-supervised framework should be understood as an important step toward resource-efficient segmentation, particularly for static and structurally stable classes, while future improvements may require integration with explicit optical flow, object tracking, or instance-level motion estimation.

In summary, the growing body of research on semantic SLAM, dynamic visual localization, motion-aware mapping, and geometry-based filtering shows that intelligent perception in complex environments increasingly depends on the integration of semantic and geometric information. However, most existing work emphasizes the use of semantic segmentation to improve SLAM or localization, whereas fewer studies examine how depth and camera-motion constraints can be used to generate reliable pseudo-labels for semi-supervised video semantic segmentation. This gap is particularly important in road-scene analysis, where dense annotation is costly and where transportation managers and engineering teams require scalable methods for converting video data into actionable semantic information. By developing a semi-supervised segmentation framework based on geometric warping and depth-consistency refinement, the present study responds to both a technical need in computer vision and a managerial need for efficient resource utilization in intelligent transportation systems.

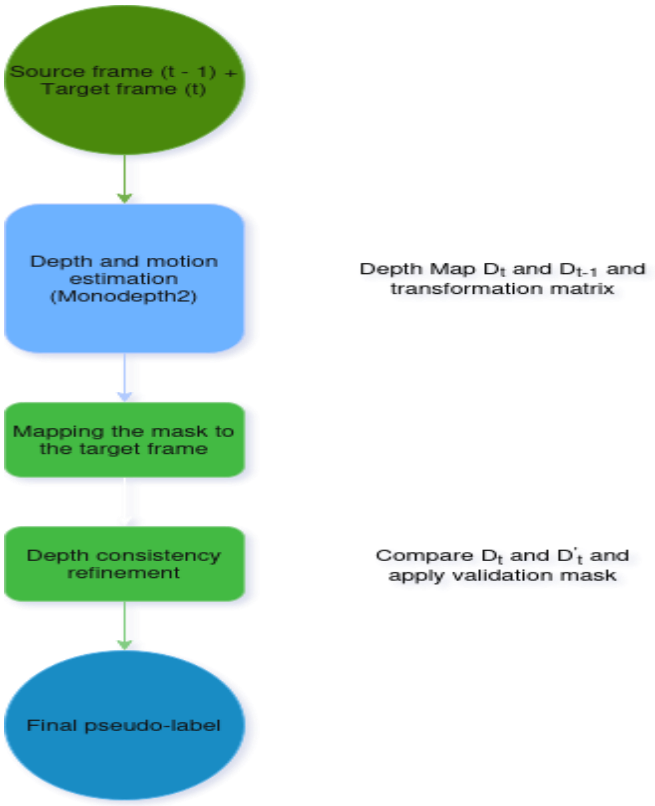
The aim of this study is to develop and evaluate a semi-supervised road semantic segmentation framework that uses monocular depth estimation and camera-motion constraints to generate and refine pseudo-labels for unlabeled video frames, thereby improving resource-efficient engineering decision-making in intelligent transportation systems.

2. Methods and Materials

The proposed method consists of three main steps: estimating depth and camera motion for consecutive frames, geometrically warping the mask of the reference frame to the target frames using this information, and refining pseudo-labels based on depth consistency. The general framework proposed for generating pseudo-labels based on depth and camera motion is shown in Figure 1.

Figure 1

Pseudo-label generation process using depth and camera motion estimation



Each of these steps is explained in this section.

2.1. Depth and Motion Estimation Using Monodepth2

For depth and motion estimation, the Monodepth2 implementation was used. This model is trained in a self-supervised manner using only monocular videos. In this model, for each image frame I_t, the depth network generates a corresponding depth map $D_t$with matching spatial dimensions. In addition, the motion network receives the pair of frames $I_t$and $I_{t-1}$and estimates the camera-motion transformation matrix T, which includes rotation and translation in three-dimensional space.

Depth and motion are jointly learned using a photometric reprojection loss function. Briefly, for each pixel in the target frame, the corresponding coordinates in the reference frame are computed using the estimated depth and camera motion, and the pixel intensity is sampled from the reference frame. The difference between the reconstructed intensity and the actual intensity in the target frame, along with a structural similarity term (SSIM), is considered the reprojection error for that pixel. To reduce the effect of occlusions, Monodepth2 uses minimum reprojection loss across multiple reference frames and applies auto-masking to remove pixels whose motion is inconsistent with the camera-motion model, such as independently moving objects.

In this study, a Monodepth2 model pre-trained on the KITTI driving dataset was used (Geiger, 2012), and no retraining was performed due to resource constraints. For each 15-frame sequence, this model provides depth maps for different frames and estimates motion between consecutive frames.

2.2. Geometric Warping of the Reference-Frame Mask to Target Frames

Assume that frame I_{t-1}, as the reference frame, has a semantic mask M_{t-1}, and the objective is to generate the corresponding mask $M_t$for frame I_t, which has no manual annotation. For each pixel $p_{t-1} = (u, v, 1)$in the reference frame whose semantic class is known in M_{t-1}, the corresponding three-dimensional coordinates in the reference-camera coordinate system are first computed using the estimated depth $D_{t-1}(p_{t-1})$and the camera intrinsic matrix K:

$$P_{t-1} = D_{t-1}(p_{t-1})K^{-1}p_{t-1} \qquad (1)$$

Then, using the camera-motion transformation matrix T, this point is reprojected into the target-camera coordinate system:

$$P_t = TP_{t-1} \qquad (2)$$

Finally, the three-dimensional point $P_t$is mapped onto the image plane of the target frame using the intrinsic matrix K:

$$p_t = KP_t$$

After normalization, the pixel coordinates $(u', v')$are obtained, and the mask class at pixel $(u, v)$is transferred to pixel (u', v')in the target frame. This process is repeated for all pixels of the reference mask and results in an initial warped mask M'_t.

2.3. Pseudo-Label Refinement Based on Depth Consistency

Geometric warping alone may lead to incorrect labels in the presence of depth-estimation errors, motion-estimation errors, occlusions, and independently moving objects. To reduce these errors, a refinement step is applied based on depth consistency between the depth map warped from the reference frame to the target frame and the directly estimated depth map for the target frame.

Specifically, in the same way that the mask $M_{t-1}$is warped into M'_t, the depth map $D_{t-1}$is also transformed into a warped depth map D'_tusing the same geometric warping procedure. Then, for each pixel, the warped depth $D'_t$is compared with the directly estimated depth D_t, and the relative error is computed as follows:

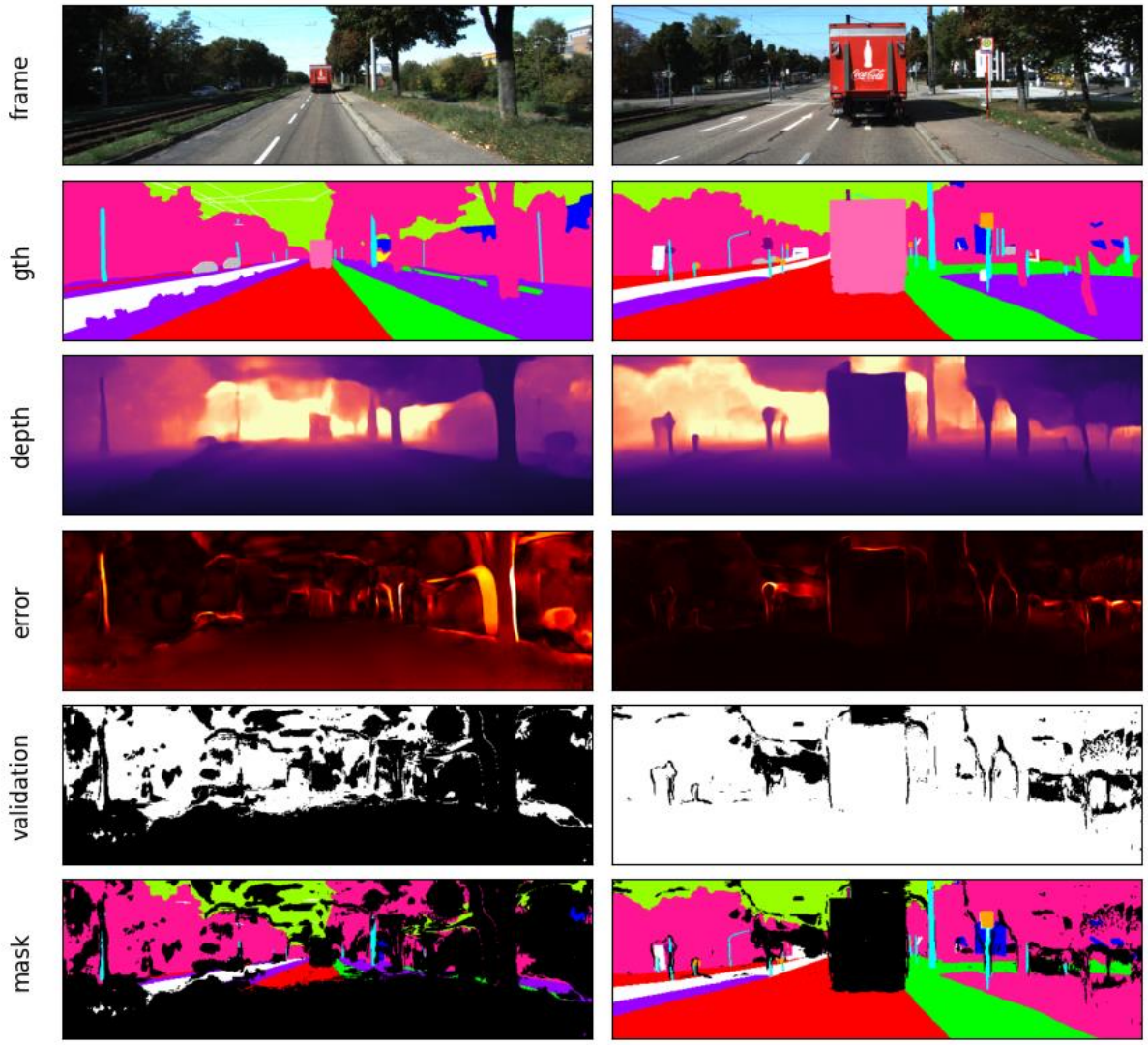
$$\text{Error} = \frac{| D'_t - D_t |}{D_t} \qquad (3)$$

If this error for a pixel is greater than a predefined threshold, for example 0.05, the pixel is considered invalid, and its label is marked as “ignore.” Thus, a validity mask is constructed that treats only those pixels as valid for which the warped depth and the estimated depth are consistent.

The final pseudo-label $M_t$is obtained by applying this validity mask to the initial mask M'_t. This refinement process removes a considerable portion of incorrect labels, particularly in regions affected by occlusion or independent motion, and improves the overall quality of the pseudo-labels. An example of the predicted depth map, the relative depth-error map, and the resulting validity mask is shown in Figure 2.

Figure 2

Illustration of the predicted depth map, consistency error, and final mask.



The LR-ASPP segmentation network with a MobileNetV3 backbone is used to learn the mapping from the image to the semantic mask. The main difference lies in the type of pseudo-labels used for unlabeled frames; here, the pseudo-labels are obtained through geometric warping based on depth and camera motion.

At each step, a batch containing frames with ground-truth labels and frames with pseudo-labels is selected. The overall loss function is the same combination of supervised and pseudo-supervised loss, controlled by the coefficient λ. In this article, $\lambda = 1$also provided the best balance between the two components.

Training was performed on the training set of the KITTI-STEP dataset, and evaluation was conducted on the validation set of the same dataset. Optimization was carried out using the SGD algorithm (Robbins, 1951) with an initial learning rate of 0.0001, and the learning rate was reduced during training using the MultiStepLR scheduler. The number of training epochs was set to 100.

3. Findings and Results

In this section, the performance of the SSL-Warp model on KITTI-STEP is presented and compared with the supervised baseline model, SSL-First.

Table 1 presents the overall results based on the mIoU and pixel accuracy metrics. The comparative mIoU chart for the baseline model and the SSL-Warp model is shown in Figure 3.

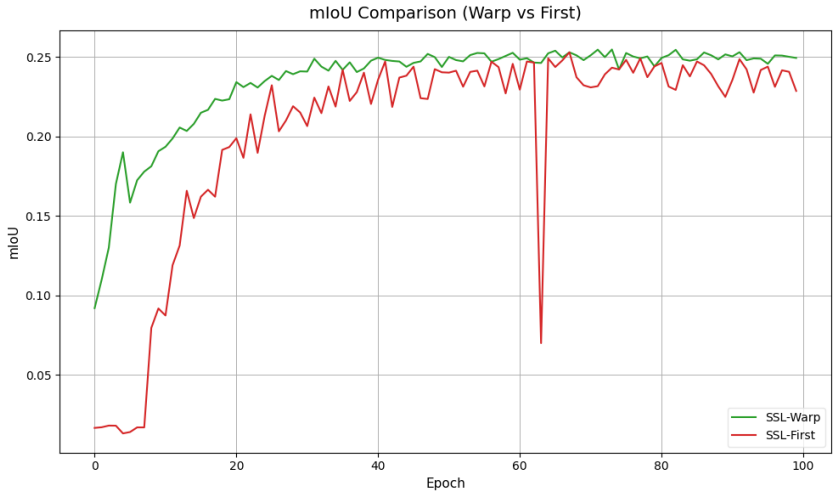
Table 1

Performance Comparison of the Methods

Method	mIoU	Accuracy
Baseline method	25.13	70.47
Proposed method	25.73	71.42

Figure 3

Comparative mIoU chart



The results indicate that the SSL-Warp model achieved approximately a 0.6%improvement in mIoU and approximately a 1%improvement in pixel accuracy compared with the supervised baseline.

Class-wise IoU analysis shows that the SSL-Warp model performs well for static classes such as “building” and “wall,” but performs more weakly for dynamic classes such as “car” and “bicycle” because these classes have independent motion relative to the scene.

This behavioral difference can be interpreted as follows: geometric warping based on depth and camera motion naturally models movements resulting from camera displacement and is effective for static scenes. However, for objects that move independently relative to the camera, the assumptions of the model are violated, and geometric warping cannot correctly predict their new locations. Although refinement based on depth consistency can remove part of these errors, because there is no explicit model for object motion, less semantic information related to such objects is transferred to the target frames. The class-wise IoU values are presented in Table 2.

Table 2

Class-Wise IoU Values

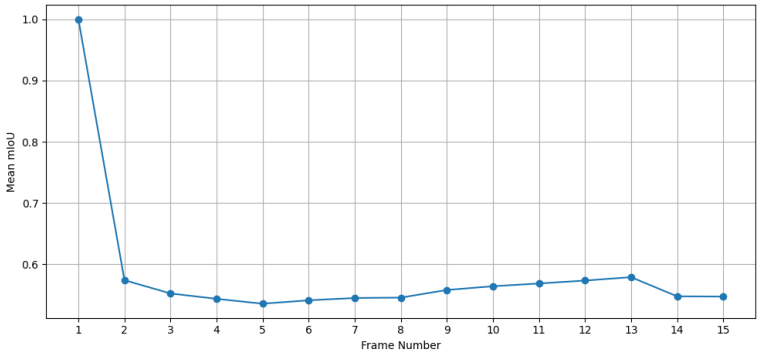
Class	SSL-Tracker
Road	63.16
Sidewalk	8.18
Building	17.04
Wall	0.04
Fence	2.95
Pole	8.11
Traffic light	1.14
Traffic sign	5.49
Vegetation	73.48
Terrain	38.79
Sky	71.33

Pedestrian	0.95
Rider	0
Car	23.97
Truck	0
Bus	0.03
Train	0
Motorcycle	0
Bicycle	0.20

In analyzing the stability of pseudo-labels across frames, it is observed that the mIoU of pseudo-labels obtained through geometric warping also decreases as the distance from the reference frame increases; however, the rate of decline varies depending on the scene structure. Changes in pseudo-label mIoU with respect to frame number are shown in Figure 4.

Figure 4

mIoU per frame for SSL-Warp

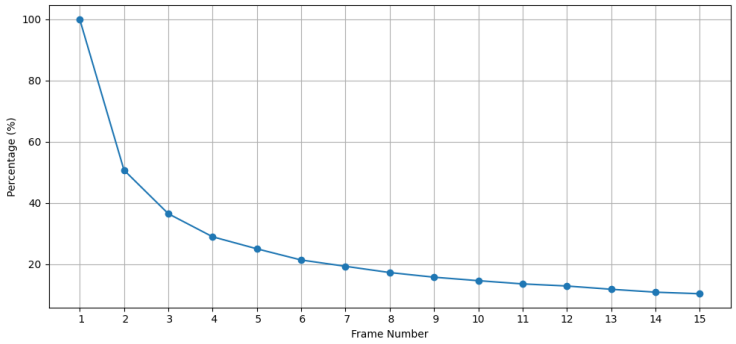


In sequences where camera motion is mild and the scene is relatively static, geometric warping can generate acceptable labels even for more distant frames. In contrast, in sequences involving intense camera maneuvers or the presence of diverse moving objects, pseudo-label quality declines more rapidly.

The role of refinement based on depth consistency was highly important in improving the results. Without this step, many pixels in the warped mask, particularly in occluded regions, would have incorrect labels. By applying the relative depth-error threshold, these pixels are removed; although some parts of the mask are converted into the "ignore" label, the remaining portion is geometrically consistent with the depth of the target frame and can be used as a valid pseudo-label in semi-supervised training. The percentage of valid pixels remaining after depth refinement across different frames is presented in Figure 5.

Figure 5

Percentage of pixels remaining after refinement



The qualitative results show that in relatively static scenes, the SSL-Warp model can effectively preserve the boundaries of the road, sidewalk, and vegetation. In many samples, the precise shape of the road is acceptably maintained across frames due to the consistent warping of the reference mask to the target frames. Examples of the generated pseudo-label masks in different parts of the sequence are shown in Figures 6 to 9.

Figure 6

Predicted masks using the proposed method, Part 1

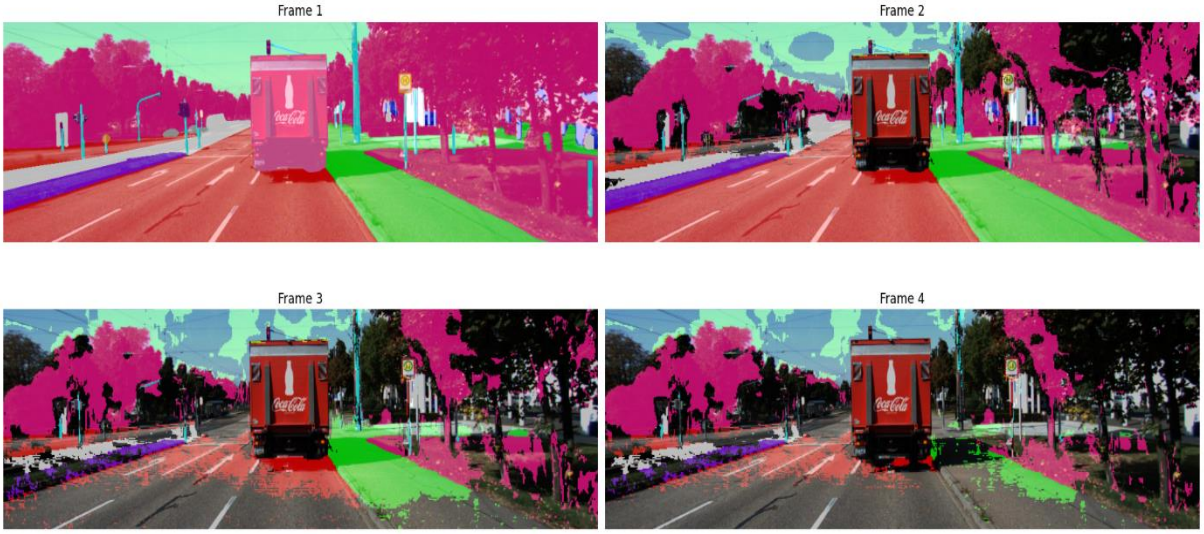


Figure 7

Predicted masks using the proposed method, Part 2

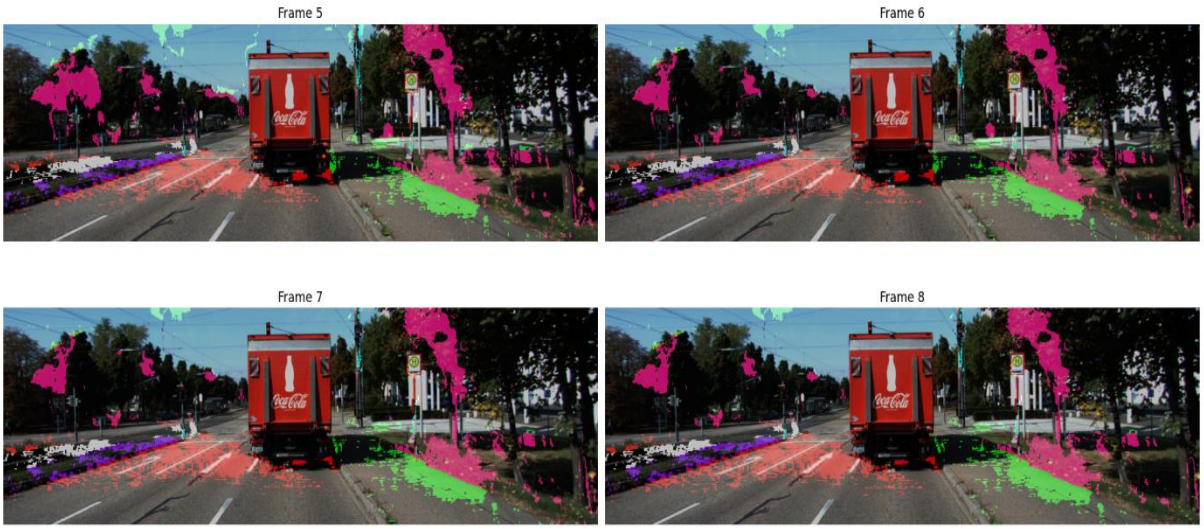


Figure 8

Predicted masks using the proposed method, Part 3

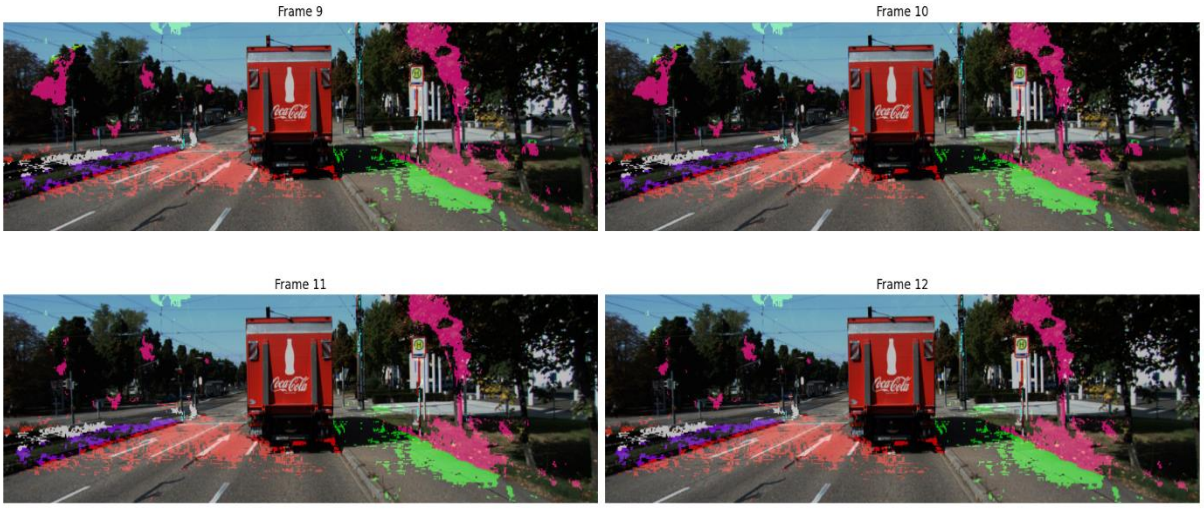
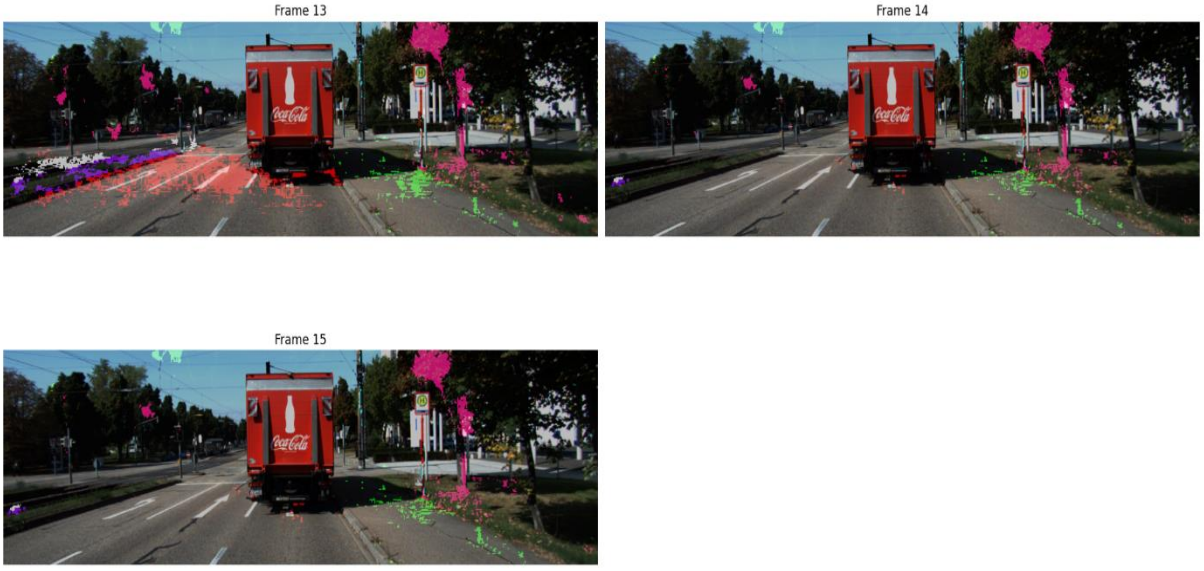


Figure 9

Predicted masks using the proposed method, Part 4



However, in the presence of overtaking vehicles or crossing pedestrians, pseudo-labels in these regions are often either removed as invalid or contain inaccurate labels, because the Monodepth2 model does not explicitly model the independent motion of these objects. A visual comparison of the outputs of the baseline model and the SSL-Warp model can be seen in Figures 10 and 11.

Figure 10

Visual comparison of the segmentation masks of the SSL-Warp and SSL-First models, Part 1

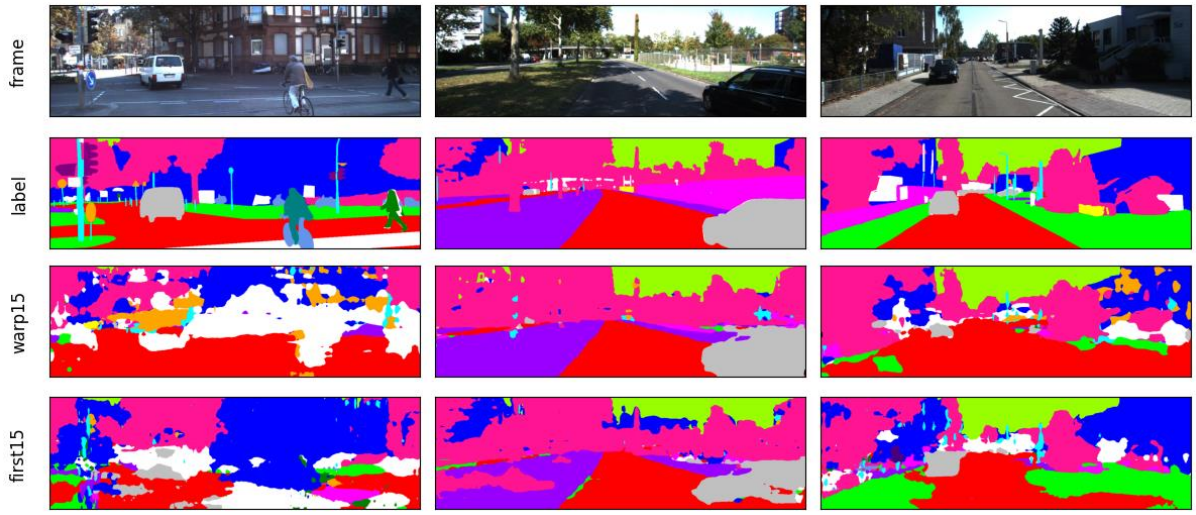
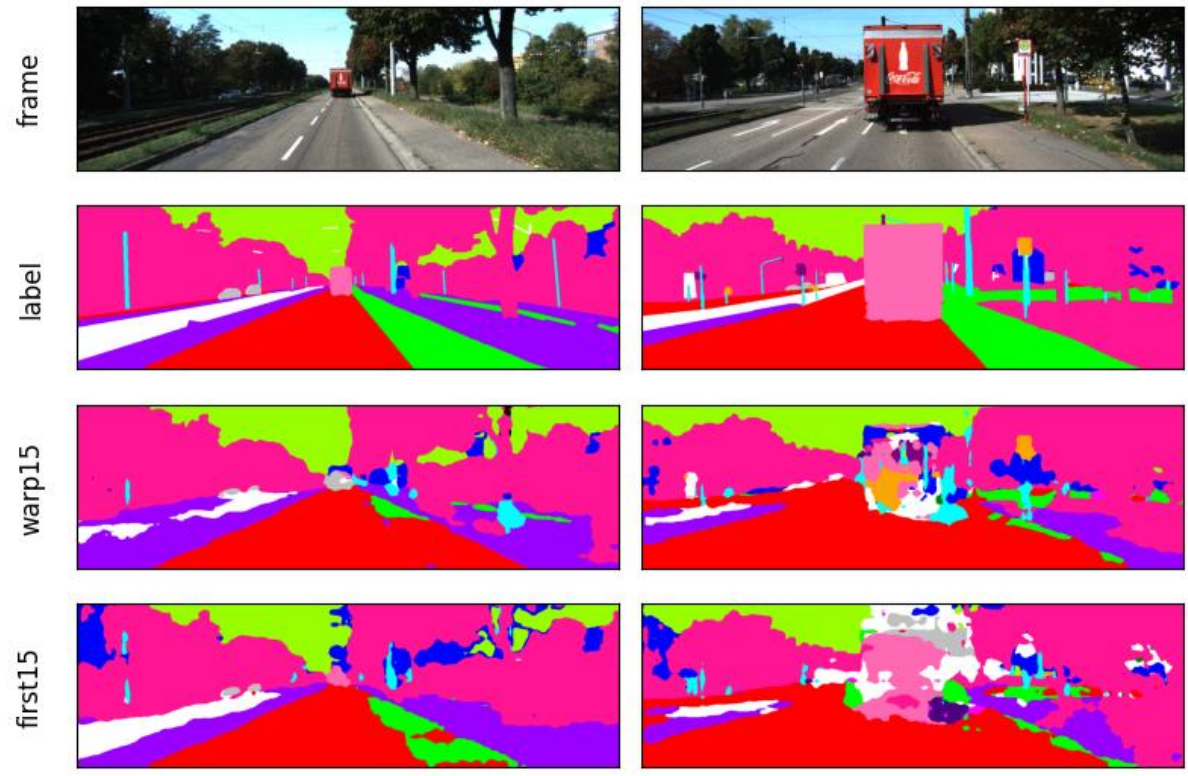


Figure 11

Visual comparison of the segmentation masks of the SSL-Warp and SSL-First models, Part 2



The effect of changing the value of the coefficient λon model performance is reported in Table 3.

Table 3

mIoU Values for Different λValues

λvalue	SSL-Warp
0.5	24.68
1	25.73
2	25.34

4. Discussion and Conclusion

The findings of this study indicate that incorporating depth and camera-motion constraints into the semi-supervised training process produced a measurable but moderate improvement in road-scene semantic segmentation. The proposed SSL-Warp model achieved an mIoU of 25.73% and a pixel accuracy of 71.42%, compared with 25.13% mIoU and 70.47% pixel accuracy for the baseline SSL-First model. This improvement of approximately 0.60 percentage points in mIoU and 0.95 percentage points in pixel accuracy suggests that geometric warping can transfer useful semantic information from a manually labeled reference frame to subsequent unlabeled frames. Although the magnitude of improvement is limited, it is meaningful in the context of a highly resource-constrained semi-supervised setting in which only the first frame of each 15-frame sequence is manually annotated. This result supports the broader argument that semantic perception systems can benefit from integrating geometric and motion-based constraints rather than relying exclusively on two-dimensional pixel-level supervision. Similar tendencies have been observed in recent visual localization and mapping studies, where the fusion of semantic information, geometric consistency, and motion estimation has been shown to improve robustness in dynamic and complex environments (Chen et al., 2024; Sahili et al., 2023; Zhang et al., 2023).

The superiority of SSL-Warp over the first-frame-only baseline can be explained by the fact that road videos contain strong temporal and geometric continuity. When camera motion and scene depth are estimated with sufficient reliability, the semantic mask of the reference frame can be projected into subsequent frames in a geometrically meaningful way. This mechanism increases the amount of training supervision without requiring additional manual annotation. In this regard, the results align with studies emphasizing that geometric constraints can improve visual understanding by filtering unreliable correspondences and preserving spatial consistency across frames (Cao, 2025; Li et al., 2025; Liao et al., 2025). The depth-consistency refinement step was particularly important because it prevented all warped pixels from being accepted uncritically. By comparing the warped depth with the directly estimated target-frame depth, the method removed pixels whose transferred labels were likely to be geometrically inconsistent. This mechanism resembles the dynamic rejection and static-weighted optimization strategies used in semantic SLAM systems, where potentially unreliable dynamic regions are suppressed to protect the stability of localization, mapping, and optimization (Gao et al., 2025; Shen & Zhang, 2025; Wang et al., 2025).

The class-wise results further clarify the nature of the obtained improvement. SSL-Warp produced stronger performance for structurally stable and relatively static classes, including vegetation, sky, road, and terrain. The high IoU values for vegetation, sky, and road indicate that the proposed depth- and camera-motion-based warping procedure is particularly effective when the spatial position of scene elements can be explained mainly by camera ego-motion. This finding is consistent with the theoretical basis of geometric warping: if the scene is static and the depth and pose estimates are sufficiently accurate, the projected position of a pixel in the target frame should remain geometrically consistent. Similar conclusions can be derived from semantic SLAM research, where static background structures and persistent geometric features are commonly treated as more reliable sources for localization and map optimization than dynamic foreground objects (Dai et al., 2024; Gong et al., 2024; Li & Luo, 2024). Therefore, the present findings confirm that depth and camera-motion constraints are especially useful for transferring labels associated with stable road-scene structures.

In contrast, the weak IoU values for dynamic or small object classes, such as pedestrian, rider, motorcycle, bicycle, truck, bus, and train, show that geometric warping based only on camera motion is insufficient for objects with independent motion or limited pixel representation. This result is expected because the transformation matrix used in the proposed framework models camera movement, not

object-specific movement. When a car overtakes, a pedestrian crosses the road, or a cyclist moves independently of the camera, the assumption that pixel displacement is explained by scene depth and camera ego-motion is violated. Consequently, the warped label may be displaced, removed by the validity mask, or assigned incorrectly. This finding is strongly aligned with studies showing that dynamic environments require explicit mechanisms for moving-object detection, object tracking, mask propagation, region growing, and motion segmentation (Huang et al., 2024; F. Wang et al., 2024; Wu et al., 2025; Zhu et al., 2024). The results also support the argument that segmentation frameworks for road videos should distinguish between camera-induced motion and independently moving objects if they aim to improve performance on dynamic traffic participants.

The relatively modest overall improvement can also be attributed to the limitations of pseudo-label quality. In semi-supervised segmentation, pseudo-labels are useful only when they provide reliable additional supervision. If pseudo-labels are noisy, incomplete, or biased toward static classes, they may improve some categories while providing little benefit for others. In the present study, the validity mask improved the reliability of pseudo-labels by removing inconsistent pixels; however, this also reduced the amount of usable supervision in difficult regions. This trade-off explains why the model improved overall pixel accuracy and mIoU but did not achieve a large performance gain. The result is comparable to dynamic SLAM studies in which aggressive filtering of dynamic regions improves stability but may also reduce the amount of information available for optimization (Chang et al., 2023; Wei et al., 2023; Yao et al., 2023; You et al., 2022). Thus, the proposed method appears to favor precision over coverage: it produces more geometrically reliable pseudo-labels, but the ignored regions limit its capacity to learn difficult dynamic classes.

The ablation analysis on the coefficient λprovides further insight into the balance between supervised and pseudo-supervised learning. The best performance was obtained when $\lambda = 1$, with an mIoU of 25.73%. When $\lambda = 0.5$, performance decreased to 24.68%, suggesting that the pseudo-label component was underweighted and therefore did not contribute sufficiently to learning from unlabeled frames. Conversely, when $\lambda = 2$, mIoU declined to 25.34%, indicating that excessive reliance on pseudo-labels may amplify residual noise and reduce generalization. This pattern demonstrates that pseudo-supervision should be integrated carefully, particularly when pseudo-labels are generated through estimated depth and camera motion rather than ground-truth annotations. The finding is consistent with broader work on dynamic semantic perception, where adaptive filtering, frame skipping, and non-blocking detection are used to regulate the influence of uncertain visual information (Gong et al., 2026; Jiang et al., 2025; Zhang & Shen, 2025). In other words, the effectiveness of geometric pseudo-labeling depends not only on the warping procedure itself but also on the training strategy that determines how strongly those pseudo-labels influence model optimization.

From a management and engineering decision-making perspective, the results are important because they show that additional annotation efficiency can be achieved through the use of geometric constraints. The proposed framework reduces dependence on dense manual labeling by using only the first annotated frame of each sequence and generating pseudo-labels for subsequent frames. Even a moderate gain in segmentation performance is valuable when it is achieved under reduced annotation demand. In intelligent transportation systems, this has direct implications for resource management, because large-scale road-video analysis requires scalable, cost-efficient, and operationally feasible model-training pipelines. Similar priorities are reflected in lightweight and adaptive SLAM systems that seek to improve computational efficiency while maintaining acceptable robustness in dynamic environments (Li et al., 2024; K. Wang et al., 2024; Zheng et al., 2025). Therefore, the proposed method contributes not only to technical segmentation performance but also to the practical problem of reducing the human labor and computational resources needed for deploying vision-based transportation analytics.

The qualitative findings reinforce the quantitative results. In relatively static scenes, the SSL-Warp model preserved the contours of road, sidewalk, and vegetation more effectively, suggesting that geometric transfer maintained semantic continuity across frames. This is especially useful for transportation applications in which stable background elements are crucial for route understanding, lane-level interpretation, infrastructure assessment, and environmental monitoring. However, the visual results also confirmed that independently moving objects remain problematic. In regions containing overtaking vehicles or crossing pedestrians, pseudo-labels were frequently invalidated or distorted. This behavior is consistent with the limitations reported in recent dynamic visual SLAM and semantic mapping research, where object-level modeling, tracking, and semantic segmentation are repeatedly introduced to

address the instability caused by moving agents (Huang et al., 2023; Liu et al., 2025; Sun et al., 2023; Zhang et al., 2025). Therefore, the present study confirms that depth and camera-motion constraints are valuable but incomplete: they are effective for static geometry but should be complemented by explicit motion information to address dynamic objects.

Overall, the findings demonstrate that semi-supervised semantic segmentation can benefit from the integration of monocular depth estimation, camera-motion estimation, geometric warping, and depth-consistency refinement. The proposed SSL-Warp approach improved performance compared with the SSL-First baseline and provided a resource-efficient mechanism for exploiting unlabeled frames in road-video sequences. The strongest gains were conceptually associated with static scene categories, while dynamic and small object classes remained difficult because their motion cannot be fully modeled by camera ego-motion. This outcome is aligned with the direction of recent research in semantic SLAM and dynamic-scene perception, which increasingly emphasizes the combination of semantic, geometric, and motion-aware modules to improve robustness under real-world conditions (Gao et al., 2025; Liao et al., 2025; Wang et al., 2025; Zhang et al., 2025). The study therefore contributes to the field by showing that depth- and motion-based pseudo-label generation can enhance road semantic segmentation in a low-annotation regime, while also identifying the need for stronger dynamic-object modeling in future intelligent transportation systems.

The main limitation of this study is that the proposed framework relies on estimated monocular depth and estimated camera motion rather than ground-truth geometric information. Any error in depth estimation, pose estimation, camera calibration, or reprojection directly affects the quality of the generated pseudo-labels. In addition, the method assumes that most pixel displacement between frames can be explained by camera motion, which limits its effectiveness for independently moving objects such as vehicles, pedestrians, riders, and bicycles. Another limitation is that the improvement over the baseline, although positive, remained modest, indicating that geometric pseudo-labeling alone is not sufficient to solve the full complexity of road-video semantic segmentation. The use of a pre-trained depth-estimation model without task-specific retraining may also have constrained performance, especially in scenes with challenging lighting, occlusion, scale variation, and dense traffic. Finally, evaluation was conducted on one dataset, which limits the generalizability of the results to other road environments, weather conditions, camera configurations, and transportation contexts.

Future research should extend this framework by integrating explicit object-motion modeling, optical flow, instance tracking, and dynamic-object segmentation into the pseudo-label generation process. Combining depth- and camera-motion constraints with object-level motion estimation may improve label propagation for vehicles, pedestrians, cyclists, and other independently moving agents. Future studies should also evaluate the framework across multiple datasets with different driving conditions, camera viewpoints, weather patterns, and urban structures to determine its robustness and transferability. Another important direction is to fine-tune the monocular depth-estimation model on target-domain road videos and compare its effect with the use of a generic pre-trained model. Future work may also examine adaptive thresholds for depth-consistency filtering, uncertainty-aware pseudo-label weighting, temporal consistency losses, and class-balanced training strategies to prevent the model from benefiting mainly from large static classes while neglecting smaller dynamic categories.

From a practical perspective, transportation organizations and engineering teams can use the proposed framework as a cost-reduction strategy for developing road-scene perception models when dense pixel-level annotation is not feasible. The method is particularly useful for applications focused on stable environmental components, such as road boundaries, vegetation, terrain, sky, and infrastructure-related classes. Practitioners should treat the generated pseudo-labels as quality-controlled but incomplete supervisory signals rather than as replacements for manual annotation. For deployment in operational transportation systems, the framework should be combined with validation procedures, periodic human review, and additional modules for moving-object detection. Engineering teams should also monitor class-wise performance rather than relying only on aggregate mIoU or pixel accuracy, because high performance on large static regions may hide weak performance on safety-critical dynamic classes.

Authors' Contributions

Authors contributed equally to this article.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethics Considerations

In this research, ethical standards including obtaining informed consent, ensuring privacy and confidentiality were considered.

References

Cao, S. (2025). DEG-SLAM: A Dynamic Visual RGB-D SLAM Based on Object Detection and Geometric Constraints for Degenerate Motion. *Measurement Science and Technology*, *36*(2), 026302. https://doi.org/10.1088/1361-6501/ada39c

Chang, Y., Hu, J., & Xu, S. (2023). OTE-SLAM: An Object Tracking Enhanced Visual SLAM System for Dynamic Environments. *Sensors*, *23*(18), 7921. https://doi.org/10.3390/s23187921

Chen, C., Wang, B., Lu, C. X., Trigoni, N., & Markham, A. (2024). Deep Learning for Visual Localization and Mapping: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(12), 17000-17020. https://doi.org/10.1109/tnnls.2023.3309809

Dai, J., Yang, M., Li, Y., Zhao, J., & Hanajima, N. (2024). ADS–SLAM: A Semantic SLAM Based on Adaptive Motion Compensation and Semantic Information for Dynamic Environments. *Measurement Science and Technology*, *36*(1), 016304. https://doi.org/10.1088/1361-6501/ad824b

Gao, S., Gao, X., & Zhang, D. (2025). DMS-SLAM: Semantic Visual SLAM Based on Deep Mask Segmentation in Dynamic Environments. *Measurement Science and Technology*, *36*(4), 046311. https://doi.org/10.1088/1361-6501/adc1f1

Gong, C., Sun, Y., Zou, C., Jiang, D., Huang, L., & Tao, B. (2024). SFD-SLAM: A Novel Dynamic RGB-D SLAM Based on Saliency Region Detection. *Measurement Science and Technology*, *35*(10), 106304. https://doi.org/10.1088/1361-6501/ad5b0e

Gong, X., Chen, H., Zhang, H., Liao, K., & Liu, X. (2026). Low Computational Cost and Misclassification Rate Semantic VSLAM Realization Based on Frame Skipping Dual Filtering and Adaptive Motion Estimation. *Measurement Science and Technology*, *37*(12), 126201. https://doi.org/10.1088/1361-6501/ae531e

Huang, H., Dong, Y., Yang, J., & Liu, X. (2023). Autonomous Vehicles Localisation Based on Semantic Map Matching Method. *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, *XLVIII-1/W2-2023*, 901-908. https://doi.org/10.5194/isprs-archives-xlviii-1-w2-2023-901-2023

Huang, W., Zou, C., Yun, J., Jiang, D., Huang, L., Liu, Y., Jiang, G. Z., & Xie, Y. (2024). Strong-Slam: Real-Time RGB-D Visual SLAM in Dynamic Environments Based on StrongSORT. *Measurement Science and Technology*, *35*(12), 126309. https://doi.org/10.1088/1361-6501/ad7a11

Jiang, D., Yun, J., Huang, L., Xie, Y., & Sun, Y. (2025). DPLS-SLAM: A Visual SLAM System Based on Point–line Feature Fusion and Lightweight Improved YOLOv8seg Network in Dynamic Environment. *Measurement Science and Technology*, *36*(7), 076304. https://doi.org/10.1088/1361-6501/ade55c

Li, J., & Luo, J. (2024). YS-SLAM: YOLACT++ Based Semantic Visual SLAM for Autonomous Adaptation to Dynamic Environments of Mobile Robots. *Complex & Intelligent Systems*, *10*(4), 5771-5792. https://doi.org/10.1007/s40747-024-01443-x

Li, Z., Zhang, X., Fan, L., & Li, J. (2024). An Accurate and Robust RGB-D Visual SLAM Method in Dynamic Environments. https://doi.org/10.21203/rs.3.rs-5311384/v1

Li, Z., Zhang, X., Fan, L., & Li, J. (2025). Visual SLAM With Semantic and Geometric Constraints in Dynamic Environments. *Journal of Electronic Imaging*, *34*(02). https://doi.org/10.1117/1.jei.34.2.023050

Liao, P., Chen, L., Tang, J., & Feng, Z. (2025). YGDD-SLAM: Direct Geometric Constraint SLAM Based on Object Detection and Depth Image Segmentation. *Journal of Field Robotics*. https://doi.org/10.1002/rob.70024

Liu, J., Liu, N., & Yuan, Y. (2025). Towards Biologically-Inspired Visual SLAM in Dynamic Environments: IPL-SLAM With Instance Segmentation and Point-Line Feature Fusion. *Biomimetics*, *10*(9), 558. https://doi.org/10.3390/biomimetics10090558

Sahili, A. R., Hassan, S., Sakhrieh, S., Mounsef, J., Maalouf, N., Arain, B., & Taha, T. (2023). A Survey of Visual SLAM Methods. *IEEE Access*, *11*, 139643-139677. https://doi.org/10.1109/access.2023.3341489

Shen, Y., & Zhang, X. (2025). A Dynamic SLAM System With YOLOv7 Segmentation and Geometric Constraints for Indoor Environments. *Robotica*, *43*(7), 2527-2545. https://doi.org/10.1017/s0263574725101823

Sun, Y., Wang, Q., Yan, C., Feng, Y., Tan, R., Shi, X., & Wang, X. (2023). D-Vins: Dynamic Adaptive Visual–Inertial SLAM With IMU Prior and Semantic Constraints in Dynamic Scenes. *Remote Sensing*, *15*(15), 3881. https://doi.org/10.3390/rs15153881

Wang, F., Zhao, L., Xu, Z., Liang, H., & Zhang, Q. (2024). LDVI-SLAM: A Lightweight Monocular Visual-Inertial SLAM System for Dynamic Environments Based on Motion Constraints. *Measurement Science and Technology*, *35*(12), 126301. https://doi.org/10.1088/1361-6501/ad71e7

Wang, K., Yao, X., Ma, N., Ran, G., & Liu, M. (2024). DMOT-SLAM: Visual SLAM in Dynamic Environments With Moving Object Tracking. *Measurement Science and Technology*, *35*(9), 096302. https://doi.org/10.1088/1361-6501/ad4dc7

Wang, S., Chen, N., Li, W., Yuan, J., Zheng, E., Wang, G., & Chen, W. (2025). SGDO-SLAM: A Semantic RGB-D SLAM System With Coarse-to-Fine Dynamic Rejection and Static Weighted Optimization. *Sensors*, *25*(12), 3734. https://doi.org/10.3390/s25123734

Wei, S., Wang, S., Li, H., Liu, G., Yang, T., & Liu, C. (2023). A Semantic Information-Based Optimized vSLAM in Indoor Dynamic Environments. *Applied Sciences*, *13*(15), 8790. https://doi.org/10.3390/app13158790

Wu, Y., Zhang, Z., Chen, H., & Li, J. (2025). A Motion Segmentation Dynamic SLAM for Indoor GNSS-Denied Environments. *Sensors*, *25*(16), 4952. https://doi.org/10.3390/s25164952

Yao, C., Ding, L., & Lan, Y. H. (2023). MOR-SLAM: A New Visual SLAM System for Indoor Dynamic Environments Based on Mask Restoration. https://doi.org/10.20944/preprints202308.1419.v1

You, Y., Peng, W., Cai, J., Huang, W., Kang, R., & Liu, H. (2022). MISD-SLAM: Multimodal Semantic SLAM for Dynamic Environments. *Wireless Communications and Mobile Computing*, *2022*(1). https://doi.org/10.1155/2022/7600669

Zhang, W., Chen, H., & Song, F. (2025). NSDM-SLAM: Non-Blocking Semantic Detection and Mask Propagation for Robust Visual SLAM in Dynamic Environments. *Measurement Science and Technology*, *36*(9), 096302. https://doi.org/10.1088/1361-6501/adfe04

Zhang, X., & Shen, Y. (2025). YER-SLAM: A Dynamic Visual SLAM Based on Object Detection With Region Growing Algorithm. *Journal of Intelligent & Fuzzy Systems Applications in Engineering and Technology*, *49*(3), 720-734. https://doi.org/10.1177/18758967251353441

Zhang, Y., Li, Y., & Chen, P. (2023). TSG-SLAM: SLAM Employing Tight Coupling of Instance Segmentation and Geometric Constraints in Complex Dynamic Environments. *Sensors*, *23*(24), 9807. https://doi.org/10.3390/s23249807

Zheng, C., Zhang, P., & Li, Y. (2025). Semantic SLAM System for Mobile Robots Based on Large Visual Model in Complex Environments. *Scientific reports*, *15*(1). https://doi.org/10.1038/s41598-025-90340-5

Zhu, Y., An, H., Wang, H., Xu, R., Sun, Z., & Lu, K. (2024). DOT-SLAM: A Stereo Visual Simultaneous Localization and Mapping (SLAM) System With Dynamic Object Tracking Based on Graph Optimization. *Sensors*, *24*(14), 4676. https://doi.org/10.3390/s24144676