Rewis3d: Reconstruction Improves Weakly-Supervised Semantic Segmentation

Jonas Ernst[*1,2], Wolfgang Boettcher[*2], Lukas Hoyer[3], Jan Eric Lenssen[2], Bernt Schiele[2]
[1]Saarland University [2]Max Planck Institute for Informatics, SIC [3]ETH Zurich
{jernst, wolfgang.boettcher, jlenssen, schiele}@mpi-inf.mpg.de
lhoyer@vision.ee.ethz.ch

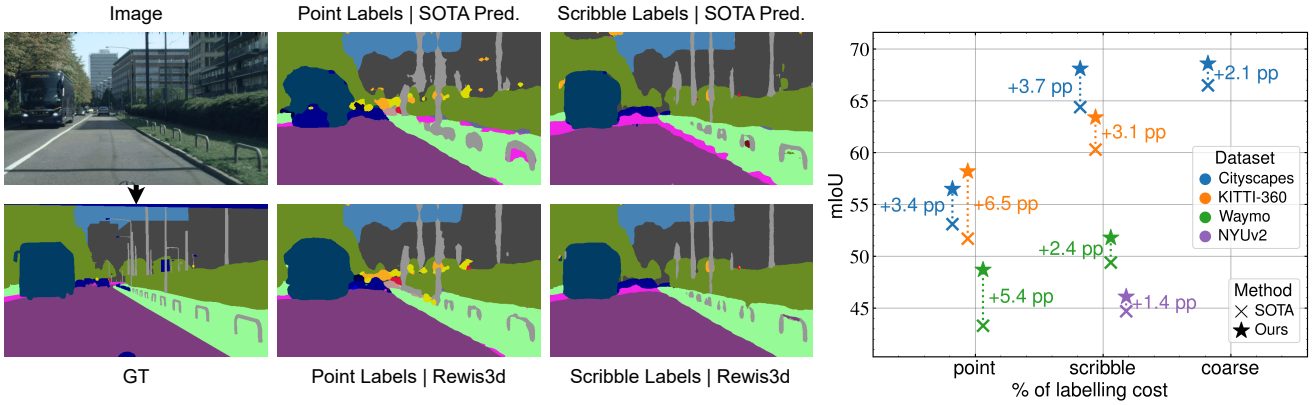

Figure 1. **Rewis3d** – *Left:* Our method (Rewis3d) greatly improves performance for weakly supervised segmentation, trained with point and scribble labels. Notably, we improve robustness to scale changes in objects and more precise class boundaries. *Right:* We consistently outperform previous SOTA methods on a range of datasets and a variety of sparse annotations by significant margins.

Abstract

We present Rewis3d, a framework that leverages recent advances in feed-forward 3D reconstruction to significantly improve weakly supervised semantic segmentation on 2D images. Obtaining dense, pixel-level annotations remains a costly bottleneck for training segmentation models. Alleviating this issue, sparse annotations offer an efficient weakly-supervised alternative. However, they still incur a performance gap. To address this, we introduce a novel approach that leverages 3D scene reconstruction as an auxiliary supervisory signal. Our key insight is that 3D geometric structure recovered from 2D videos provides strong cues that can propagate sparse annotations across entire scenes. Specifically, a dual student–teacher architecture enforces semantic consistency between 2D images and reconstructed 3D point clouds, using state-of-the-art feed-forward reconstruction to generate reliable geometric supervision. Extensive experiments demonstrate that Rewis3d achieves state-of-the-art performance in sparse supervision, outperforming existing approaches by 2-7% without requiring additional labels or inference overhead. Project page: https://rewis3d-mpi.github.io/Rewis3d/

* Equal contribution.

1. Introduction

Semantic segmentation has achieved remarkable progress through deep learning, enabling critical applications in autonomous driving, robotics, and medical imaging [2, 21, 44, 45]. However, these advances heavily rely on large-scale datasets with dense, pixel-accurate annotations, resources that are prohibitively expensive and time-consuming to obtain [3, 6, 23]. A potential mitigation strategy is weakly-supervised semantic segmentation (WSSS), as it allows leveraging incomplete or imprecise annotations. Sparse spatial annotations such as points [1], scribbles [3], and coarse labels [6] offer a compelling trade-off, providing explicit spatial localization while requiring only a fraction of the labeling effort needed for dense masks. In this work, we present an approach to significantly improve segmentation quality for all WSSS settings (see Fig. 1 for multiple label types and datasets) without requiring more labels or additional computation during inference.

Current approaches to bridging the weak-to-full supervision gap [1, 3, 26, 52], such as SASFormer [22] or TreeEnergy [15], introduce specialized architectures and loss functions to propagate information from labeled to unlabeled pixels. While effective, these techniques struggle to fully

compensate for the limited supervisory signal, especially in geometrically complex outdoor scenes. We propose a fundamentally different strategy: leveraging reconstructed 3D geometric structure as an auxiliary supervisory signal to enhance 2D weakly-supervised segmentation. Recent advances in feed-forward 3D reconstruction [12, 35, 36] enable the recovery of high-fidelity 3D point clouds directly from casual 2D video sequences, without requiring specialized sensors like LiDAR. This breakthrough allows us to inject geometric constraints into the learning process while maintaining a purely 2D inference pipeline. The core principle of our method is that 3D geometry provides complementary cross-view consistency constraints that propagate sparse supervision across entire scenes: when an object is annotated with a scribble, point, or coarse label in one view, its 3D structure enables knowledge transfer to all other views in which it appears.

To implement this concept, we introduce Rewis3d, a framework featuring a novel dual student–teacher architecture that enforces bidirectional consistency between 2D image-based and 3D geometry-based segmentation. Within Rewis3d, we define a cross-modal consistency (CMC) loss that aligns the predictions across modalities, effectively bridging the gap between reconstructed geometry and weak 2D supervision (c.f. Fig. 2). Addressing the inevitable noise in both reconstructed geometry and weak annotations, we propose dual confidence filtering and view-aware sampling strategies that prioritize reliable 2D–3D correspondences and suppress erroneous pseudo-labels. We extend the typical evaluation scenario of WSSS methods to additional large-scale, scene-centric datasets, such as Waymo [23] and KITTI-360 [16], and show that our approach improves on the state of the art in mIoU by 2-7%, achieving consistent and significant gains across a variety of datasets and supervision types.

In summary, our main contributions are as follows:

- We present Rewis3d, the first weakly-supervised framework to integrate sparse 2D annotations with 3D geometry reconstructed solely from 2D images, proving geometry as a powerful supervisory signal.
- We introduce a novel dual student-teacher mechanism with confidence-guided filtering and view-aware sampling to ensure robust 2D–3D alignment and knowledge transfer.
- Using only sparse supervision, our method achieves state-of-the-art results on several datasets, outperforming existing WSSS methods.

2. Related Work

Weakly-Supervised Segmentation. Weakly supervised semantic segmentation (WSSS) aims to reduce the annotation burden of dense pixel-level labels by leveraging weaker or incomplete forms of supervision. Among these, sparse

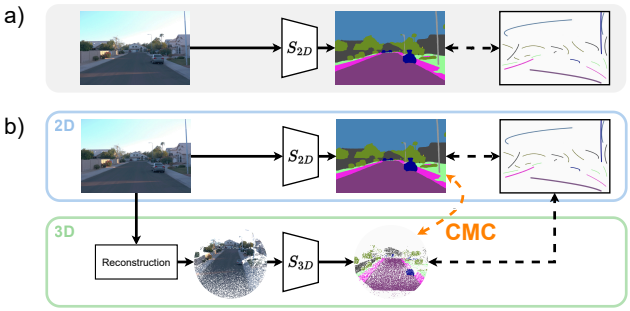

Figure 2. **Conceptual overview of weakly-supervised segmentation approaches.** (a) Traditional methods rely solely on sparse 2D annotations, limiting supervision propagation. (b) Our proposed method Rewis3d introduces a 3D branch, enforcing cross-modal consistency (CMC) between 2D predictions and 3D predictions from reconstructed geometry.

annotations, such as points [1, 30, 50], scribbles [3, 17], and coarse polygons [9, 37], offer a practical compromise between annotation efficiency and spatial precision. Points provide minimal but explicit localization, scribbles capture regional structure, and coarse polygons represent a denser yet still lightweight annotation form. In contrast, image-level labels [5, 38, 51], which indicate only the presence of object categories without spatial cues, are less effective for scene-centric datasets where multiple object instances and overlapping categories demand precise localization to achieve meaningful segmentation.

Early sparse-label approaches, such as ScribbleSup [17], used expectation–maximization to iteratively refine pseudo-labels. Later methods incorporated graph-based regularization [27, 28], self-supervised constraints [34], and uncertainty-aware refinement [41]. Recent transformer-based models like SASFormer [22] leverage self-attention to propagate sparse supervision across the image, while Tree Energy Loss (TEL) [15] captures hierarchical relationships via minimum spanning trees to generate coarse-to-fine pseudo-labels. More recently, Boettcher et al. [3] showed that mean teacher frameworks [29] remain competitive for sparse supervision, effectively propagating information from limited annotations to unlabeled regions. Our work builds on this foundation by incorporating 3D geometric supervision to further improve consistency and segmentation quality.

Weakly Supervised 3D Segmentation. The high cost of annotating 3D point clouds has spurred research into weakly supervised 3D segmentation [10]. Common strategies project 3D points onto 2D images to exploit abundant 2D labels [4, 19] or adapt sparse annotations such as scribbles directly in 3D [31, 33, 49]. Recent work further explores cross-modal guidance, transferring knowledge from unlabeled 2D images via association learning [24] or aligning 3D points with textual semantics through vision-

language supervision [42]. Many methods extend established 2D frameworks, such as the Mean Teacher consistency model, to the 3D domain [31, 33]. In contrast, our approach employs 3D reconstruction purely as an auxiliary supervisory signal to enhance 2D segmentation during training, while inference remains entirely in 2D.

Learning-based Multi-View Stereo. Recent advances in multi-view reconstruction enable robust joint prediction of depth, camera parameters, and point maps in a single forward pass. MVSNet [46] pioneered the use of deep features for depth estimation, inspiring a new generation of methods that reconstruct dense point clouds from uncalibrated images. More recent approaches, such as DUSt3R [36] and its successors [14, 48], estimate point clouds in a canonical space, though global alignment remains necessary beyond image pairs. State-of-the-art models like VGGT [35] and MapAnything [12] generalize this concept into unified transformer-based frameworks for metric-scale 3D reconstruction, with the latter supporting additional inputs such as intrinsics and poses. These feed-forward MVS methods highlight the increasing ability of learning-based models to infer accurate 3D geometry without classical optimization or explicit 3D sensors, which we utilize in this work.

Cross-Modal Consistency Learning. Cross-modal learning leverages complementary information from different modalities to improve task performance. In fully supervised settings, depth information has been shown to enhance 2D segmentation [11]. Under weak supervision, some methods use 3D data to refine 2D predictions [25, 47], but often require 3D data at inference. Methods like 2DPASS [43] and Unal et al. [32] distill 2D knowledge into 3D networks for LiDAR segmentation. In contrast, our proposed approach transfers geometric knowledge from reconstructed 3D to 2D during training only, enabling robust and efficient 2D-only inference without expensive 3D sensors. This bidirectional consistency between modalities, weighted by both reconstruction and prediction confidence, distinguishes our work from prior cross-modal approaches.

3. Rewis3d

Our work improves 2D semantic segmentation from sparse annotations by leveraging 3D geometric information reconstructed from video sequences. We propose Rewis3d, illustrated in Fig. 3, that enforces bidirectional consistency between the semantic predictions for 2D images and 3D point clouds. Crucially, the 3D geometry is generated as a pre-processing step from 2D image sequences, and the final inference is performed using only a 2D image, making our approach broadly applicable without specialized 3D sensors or requiring specific segmentation architectures.

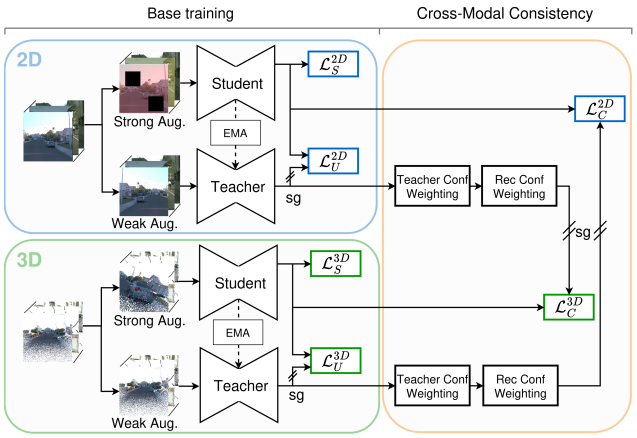

Figure 3. **Overview of the training pipeline.** Our framework operates in two stages. *Base Training* (blue and green) establishes independent student-teacher setups for each modality using sparse supervision. *Cross-Modal Consistency* (orange) introduces our core contribution: bidirectional knowledge transfer where the teacher of one modality supervises the student of the other, weighted by our dual confidence mechanism leveraging prediction certainty and reconstruction quality.

3.1. Framework Overview

Our framework consists of three key components that work in synergy: (1) a 2D segmentation branch, (2) a 3D segmentation branch, and (3) Cross-Modal Consistency (CMC) that enables bidirectional knowledge transfer between modalities. Each branch employs a Mean Teacher [29] architecture with student-teacher structures. The CMC component introduces a bidirectional consistency loss, where the teacher model from each modality generates confidence-weighted pseudo-labels to supervise the student model of the complementary modality, enabling mutual knowledge transfer between the 2D and 3D domains.

3.2. 3D Scene Reconstruction and Preprocessing

To generate the 3D data used for supervision, we employ MapAnything [12], a state-of-the-art multi-view stereo model that reconstructs dense, metric point clouds $P = \{p_i\}$ and per-point reconstruction confidences c_i^{rec} from 2D video sequences in a single forward pass. Unlike methods requiring post-processing optimization [36], MapAnything directly outputs camera parameters, depth maps, and point clouds, making it highly suitable for our pipeline.

View-Aware Point Cloud Sampling. Processing fully reconstructed point clouds derived from long captured sequences (often 60M+ points from 200+ images) is computationally prohibitive. Furthermore, our cross-modal consistency (CMC) loss operates on a per-image basis, requiring a dense set of 2D-3D correspondences between a target image and the points visible within its field-of-view. A simple

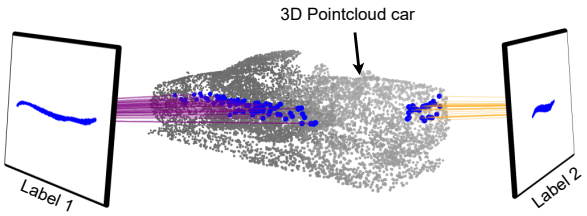

Figure 4. **Sparse label accumulation.** Firstly, an image sequence is unprojected to a 3D point cloud via a multi-view reconstruction model. Subsequently, we establish correspondences between the 3D points and the 2D pixels in the source images. This allows for label accumulation in the 3D space, and by projection, also in the 2D images.

random sampling of the entire 60M+ point scene (e.g., to 120K points) cannot satisfy this. It would yield an average of only ∼140 corresponding points per image, which is far too sparse to effectively train the consistency loss.

To solve this, we propose a *view-aware sampling strategy* that generates a dedicated 120K-point subsample for each target image, balancing dense correspondence with global context. For a given image, its subsample is created by: (1) sampling 60% (72K points) exclusively from the subset of points derived from that specific view, ensuring a rich set of 2D-3D correspondences for the CMC loss, and (2) sampling the remaining 40% (48K points) from the surrounding scene within a spatial radius, providing crucial context for robust 3D segmentation. This strategy guarantees ∼72K correspondences for the cross-modal loss while maintaining holistic scene understanding for the 3D branch.

3D Label Generation. As illustrated in Fig. 4, we generate sparse 3D labels by directly transferring annotations during the point unprojection. Since each 3D point in our cloud originates from a single 2D pixel, we assign its label in a direct one-to-one mapping. If a source 2D pixel contains an annotation, its corresponding unprojected 3D point is assigned that semantic class. Conversely, if the pixel is unlabeled, the resulting 3D point is also marked as unlabeled. By applying this process across all source images, we effectively aggregate the sparse 2D annotations from every view into a single, unified, and sparsely-labeled 3D point cloud. This sparse set of 3D labels is used for the supervised loss $\mathcal{L}_S^{3D}$, while the vast majority of unlabeled points, originating from unlabeled pixels, are supervised through teacher pseudo-labels, analogous to the 2D case.

3.3. Dual Student-Teacher Architecture

Our framework consists of a 2D and 3D branch, each with identical student-teacher structures. Both branches adopt a Mean Teacher [29] setup. This architecture is particularly effective for two reasons: (1) it enhances robustness in weakly-supervised settings by using teacher predictions as reliable pseudo-labels, and (2) the teacher's weights, updated as an Exponential Moving Average (EMA) of the student's weights, provide stable supervision for our cross-modal loss. Formally, teacher weights

$$\boldsymbol{\theta}_t^{\text{teacher}} \leftarrow \alpha \boldsymbol{\theta}_{t-1}^{\text{teacher}} + (1-\alpha)\boldsymbol{\theta}_t^{\text{student}} \tag{1}$$

are updated at each step t, where we set $\alpha = 0.99$ in our experiments.

Within each branch, the student is trained with a supervised cross-entropy loss $\mathcal{L}_S$ on labeled regions and an unsupervised consistency loss

$$\mathcal{L}_U = D_{\text{KL}}\left(\sigma(z_{\mathcal{U}}^t) \,\|\, \sigma(z_{\mathcal{U}}^s)\right) \tag{2}$$

on unlabeled regions using teacher pseudo-labels, where $z_{\mathcal{U}}^s$ and $z_{\mathcal{U}}^t$ denote student and teacher logits on unlabeled pixels $\mathcal{U}$, D_{KL} the Kullback–Leibler divergence, and $\sigma(\cdot)$ is the softmax function. To ensure high-quality pseudo-labels, we compute a confidence weight

$$w_t = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \mathbf{1}\left[\max_c \sigma(z_i^t)_c \geq \tau\right] \tag{3}$$

representing the fraction of pixels where the teacher's maximum class probability exceeds threshold τ. The training objective for base training

$$\mathcal{L} = (1-\beta)\mathcal{L}_S + \beta\, w_t\, \mathcal{L}_U \tag{4}$$

combines both supervised and unsupervised components, where β balances the contribution of labeled and unlabeled regions, and w_t adaptively scales the consistency loss based on pseudo-label quality.

3.4. Weighted Cross-Modal Consistency

A core component of our method is a bidirectional Cross-Modal Consistency (CMC) loss that uses the teacher of one modality to supervise the student of the other. To mitigate noise from uncertain predictions and unreliable geometry, we formulate a *dual confidence weighting mechanism*.

For supervising the 2D student with the 3D teacher, we compute a weighted cross-entropy loss

$$\mathcal{L}_C^{2D} = -\sum_j w_i \cdot \log(S_{2D}^{y_i}(I_j)), \tag{5}$$

for each pixel j corresponding to a valid and visible 3D point p_i, where $y_i = \arg\max(T_{3D}(p_i))$ is the hard pseudo-label from the 3D teacher, $S_{2D}^{y_i}(I_j)$ is the 2D student's probability output for class y_i at pixel j, and the weight

$$w_i = \underbrace{\max(\text{softmax}(T_{3D}(p_i)))}_{\text{prediction confidence}} \cdot \underbrace{c_i^{\text{rec}}}_{\text{reconstruction confidence}} \tag{6}$$

combines two confidence scores, *prediction confidence* and *reconstruction confidence*. Here, $T_{3D}(p_i)$ is the 3D

Table 1. **Semantic segmentation results with scribble supervision.** We report mean Intersection-over-Union (mIoU, %) and the percentage of the supervision gap closed between scribble-supervised and fully-supervised baselines (SS/FS). Our proposed Rewis3d framework, which leverages reconstructed 3D geometry for cross-modal consistency, achieves state-of-the-art performance and consistently outperforms existing weakly supervised semantic segmentation (WSSS) methods across all datasets. **Bold** indicates the best scribble-supervised result, and underline the second best.

Method	Ours	Backbone	3D Supervision	Waymo		KITTI-360		NYUv2	
				mIoU	SS/FS (%)	mIoU	SS/FS (%)	mIoU	SS/FS (%)
Fully Supervised		Segformer-B4	–	59.0	–	68.4	–	51.1	–
EMA		Segformer-B4	–	49.4	83.7	60.3	88.2	42.9	84.0
SASFormer [22]		Segformer-B4	–	37.8	64.1	46.4	67.8	<u>44.7</u>	87.5
TEL [15]		DeepLabV3+	–	42.4	71.9	59.2	86.6	38.3	75.0
Ours (Real 3D)	✓	Segformer-B4	LiDAR/Depth	<u>51.8</u>	<u>87.8</u>	<u>61.7</u>	<u>90.2</u>	<u>44.7</u>	<u>87.6</u>
Ours (Recon)	✓	Segformer-B4	Recon.	**53.3**	**90.3**	**63.4**	**93.4**	**46.1**	**90.2**

teacher's logit output, and c_i^{rec} is MapAnything's per-point reconstruction confidence. This dual filtering ensures that supervision is dominated by reliable predictions on well-reconstructed geometry. A symmetric loss $\mathcal{L}_C^{3D}$ supervises the 3D student using the 2D teacher.

To prevent overconfident wrong predictions from dominating, we apply stronger augmentations (RandomCrop, Cutout, AugMix for 2D; RandomRotation, RandomScale, RandomJitter for 3D) to student inputs compared to teachers, encouraging the student to learn robust features while the teacher provides stable targets.

3.5. Training Objective

We combine our proposed cross-modal consistency loss with standard intra-modal supervised ($\mathcal{L}_S^m$) and unsupervised ($\mathcal{L}_U^m$) losses for each modality $m \in \{2D, 3D\}$. The final training objective $\mathcal{L}_{\text{Total}}$ is a weighted sum of all components:

$$\mathcal{L}_{\text{Total}} = \sum_{m \in \{2D,3D\}} (\mathcal{L}_S^m + \mathcal{L}_U^m) \quad + \quad \lambda_{2D}\mathcal{L}_C^{2D} + \lambda_{3D}\mathcal{L}_C^{3D} \quad (7)$$

where λ_{2D} and λ_{3D} are hyperparameters that balance the contribution of the cross-modal consistency terms.

4. Experiments

We evaluate our method on four datasets spanning outdoor and indoor scenes: KITTI-360 [16], Waymo Open Dataset [23], Cityscapes [6] and NYUv2 [7]. Our experiments demonstrate that cross-modal consistency from reconstructed 3D geometry substantially outperforms traditional sparsely annotated segmentation methods.

4.1. Experimental Setup

Datasets. *KITTI-360* is a large-scale outdoor dataset with RGB images from the city of Karlsruhe, covering 19 semantic classes. The dataset provides accumulated LiDAR scans, which we use for baseline comparisons with ground-truth 3D geometry. *Waymo Open Dataset* provides images from diverse driving environments across multiple US cities, with 25 semantic classes. Unlike KITTI-360, it provides non-accumulated, single-scan LiDAR point clouds. *Cityscapes* is a widely-used urban driving dataset captured in 50 European cities, featuring 19 semantic classes. *NYUv2* is an indoor RGB-D dataset with images across 40 semantic classes. We use the provided depth maps to generate 3D point clouds for baseline comparisons. For all datasets, we generate scribble labels using Scribbles4All [3], covering approx. 2-4% of pixels while maintaining class distribution similar to full annotations. For Cityscapes, we also utilize the provided coarse labels and derive point annotations for a comprehensive comparison of sparse annotations. Specifically, we obtain point labels by randomly sampling one pixel per class in an image, following [8]. We treat Cityscapes as a collection of single images, as the training split consists of irregular drives, and thus perform reconstruction on a per-image basis.

Implementation Details. We use SegFormer-B4 [40] for 2D and Point Transformer V3 [39] for 3D segmentation respectively. Training runs for 50 epochs (250 for NYUv2) with batch size 12 on two H100 GPUs. We use AdamW [18] optimizer with learning rates 5×10^{-5} (2D) and 1×10^{-3} (3D). We train our framework in two stages. Base Training (15 epochs for KITTI-360/Waymo, 150 for NYUv2) establishes independent student-teacher setups for each modality using sparse supervision. We then introduce the CMC loss and ramp it up linearly over 5 epochs to maximum weight $\lambda = 0.1$. For data augmentation, student's images receive more severe augmentations (Cutout, Blur, AugMix for 2D; RandomRotation, RandomScale, RandomJitter for 3D) while teacher images receive weaker ones, encouraging robust feature learning. We reconstruct scenes using MapAnything [12] with up to 200 images per batch, then apply our view-aware sampling to create 120K-point clouds (60% from the current view, 40% from surrounding context).

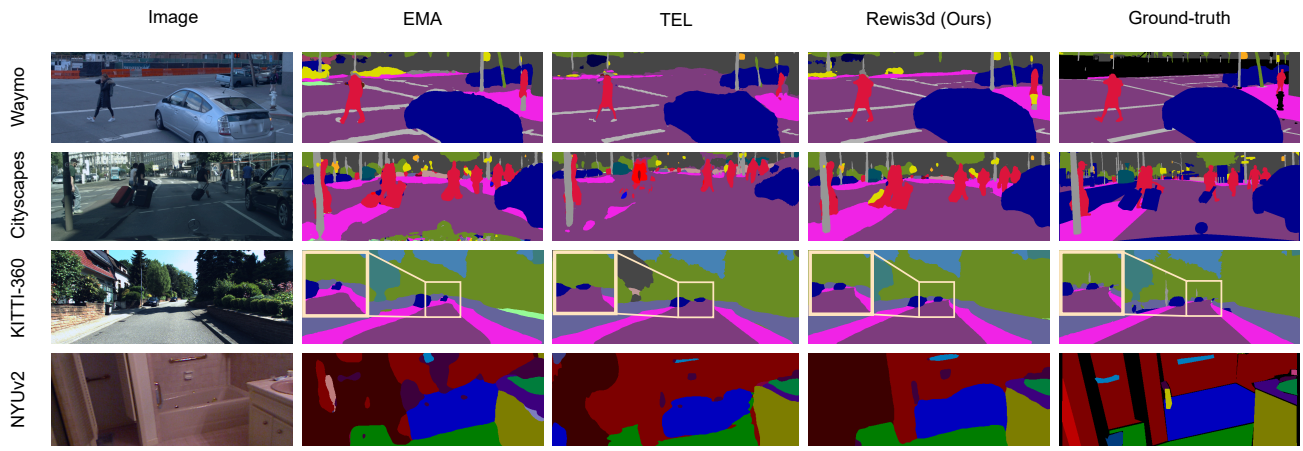

Figure 5. **Qualitative comparison across outdoor and indoor datasets.** Rewis3d produces sharper boundaries, more accurate fine-grained predictions, and better long-range segmentation compared to the Mean Teacher baseline (EMA) and TEL, even in regions where 3D reconstruction is uncertain. Colormaps are provided in the appendix.

4.2. Baselines

We compare against state-of-the-art methods specifically designed for sparsely annotated semantic segmentation, Tree Energy Loss (TEL) [15] and SASFormer [22]. These methods represent current best practices for learning directly from sparse supervision without external priors. As a foundational baseline, we train SegFormer-B4 with a standard Mean Teacher setup on scribbles only (denoted as EMA), which serves as the basis for our CMC framework and demonstrates weakly-supervised consistency regularization without geometric guidance.

To isolate the contribution of 3D geometry, we evaluate two variants of Rewis3d. Ours (Recon) represents our main contribution, using point clouds reconstructed from multi-view stereo. Ours (Real 3D) uses ground-truth LiDAR data, where available, to serve as a reference for 3D geometry.

Finally, we include a fully supervised baseline with dense annotations to quantify how much of the supervision gap our method closes. All methods are trained according to their official implementations.

4.3. Main Results

Tab. 1 presents our main 2D semantic segmentation results. Our method achieves substantial improvements across all datasets, demonstrating the effectiveness of cross-modal geometric supervision.

Improvements over Specialized WSSS Methods. Rewis3d substantially outperforms existing weakly-supervised methods specifically designed for scribble supervision, with particularly pronounced gains on scene-centric outdoor datasets. This performance gap highlights the crucial role of geometric context in propagating weak labels. On Waymo, Ours (Recon) achieves 53.3% mIoU compared to 49.4% for EMA, 42.4% for TEL, and 37.8% for SASFormer, representing improvements of 3.9,10.9, and 15.5 points, respectively. The trend is consistent on KITTI-360, where we reach 63.9% mIoU versus 60.3% (EMA), 59.2% (TEL), and 46.4% (SASFormer). This geometric advantage also extends to indoor environments, which often contain objects with less distinct 3D structure. On the NYUv2 dataset, Rewis3d (Recon) achieves 46.1% mIoU, once again outperforming all scribble-supervised baselines, including the next-best SASFormer (44.7%).

These results highlight a fundamental limitation of WSSS methods that operate solely in the 2D image plane: they struggle to robustly resolve ambiguities and propagate supervision across complex, occluded regions using only appearance cues. In contrast, our approach leverages 3D geometric information (derived from either single-view or multi-view reconstruction) to project weak labels into a stable 3D space. This geometric grounding provides a powerful mechanism for consistency enforcement and long-range supervision propagation, effectively enabling the sparse scribble to govern the segmentation of an entire 3D scene volume, rather than being confined to propagation within single frames.

Reconstructed 3D Outperforms Real 3D. Interestingly, Rewis3d using reconstructed point clouds consistently outperforms the variant using ground-truth LiDAR/depth sensors. We argue that this counterintuitive result stems from two primary advantages of the reconstructed data. First, the reconstructions typically offer a denser and more complete point cloud representation compared to the sparsity inherent in real-world LiDAR or depth sensor scans, which is critical for robust geometric label propagation. Second, the reconstructed 3D allows us to leverage our Dual Confidence Filtering Approach, which assigns lower weights to uncertain points derived from the reconstruction process itself. In stark contrast, when using real LiDAR or depth sensor data (Real 3D), we do not have an estimation of the recon-

struction confidence, forcing us to treat all real 3D points equally. This lack of a confidence measure prevents the robust filtering of noise and imprecise measurements inherent in raw sensor data, leading to less effective supervision propagation compared to the confidence-weighted reconstructed point clouds. Our ablation studies (Table 3) quantitatively confirm both the value of dense multi-view reconstruction and the effectiveness of dual confidence filtering.

4.4. Generalization to Diverse Annotation Types

To validate the general applicability of our framework beyond scribble-based supervision, we conduct experiments on the Cityscapes dataset using three distinct types of sparse annotations: points, scribbles, and coarse labels. Point labels represent the sparsest form of supervision, where each object instance is marked with a single pixel. Scribble labels are generated as on other datasets, and the provided coarse labels serve as an upper bound for sparse, yet inexact, supervision. Importantly, as Cityscapes is treated as a collection of single images, our 3D reconstruction is performed on a per-frame basis, demonstrating our method's utility even without multi-view video context and label accumulation.

Tab. 2 presents the results of this analysis. Rewis3d demonstrates significant performance gains over the EMA baseline across all three annotation types. With point supervision, our method improves the mIoU by 6.0% (from 50.5% to 56.5%), demonstrating that even from minimal spatial cues, our 3D-to-2D consistency mechanism can effectively propagate semantic information. The performance gain is even larger for scribbles at 6.9% (from 61.2% to 68.1%). For coarse annotations, which already provide a strong baseline, our method still achieves a consistent improvement of 2.1% (from 66.5% to 68.6%). These findings generalize to KITTI-360 and Waymo in the case of points and scribbles as illustrated in Fig. 1.

These insights underscore the robustness and versatility of our approach. The consistent improvements, regardless of the specific form of sparse annotation, confirm that leveraging reconstructed 3D geometry is a powerful and generalizable strategy for weakly-supervised semantic segmentation. This demonstrates that our method is not tailored to a single annotation type but rather provides a general improvement for a wide range of sparse supervision scenarios.

4.5. Qualitative Results

Figure 5 reveals dataset-specific qualitative strengths of our approach. On Waymo, Rewis3d significantly improves nearby object predictions, producing sharper class boundaries and more accurate classification. Fine-grained classes (e.g., road and lane markers) that challenge the Mean Teacher (EMA) baseline are predicted with substantially higher detail, as our geometric consistency constraints effectively resolve ambiguities in these complex regions.

Table 2. **Generalization across annotation types on Cityscapes (mIoU %).** Rewis3d consistently outperforms all competing methods, including the EMA baseline, across three sparse annotation types—points, scribbles, and coarse masks—demonstrating strong versatility and broad applicability. Reported improvements (+6.0, +6.9, +2.1) are relative to the EMA baseline, with the largest gains observed under minimal supervision (points and scribbles), confirming the effectiveness of geometric consistency in diverse weak supervision scenarios.

Method	**Annotation Type**		
	Points	*Scribbles*	*Coarse*
Fully Supervised	77.6	77.6	77.6
TEL [15]	53.1	64.4	64.9
SASFormer [22]	42.7	55.6	42.8
EMA (Baseline)	50.5	61.2	66.5
Ours	**56.5**	**68.1**	**68.6**
Improvement	+6.0	+6.9	+2.1

For KITTI-360, Rewis3d demonstrates remarkable improvements in distant scene regions. Despite inherent 3D reconstruction uncertainty at long ranges, our framework successfully propagates semantic information, producing cleaner road-sidewalk boundaries and more coherent distant vehicles. This suggests the bidirectional consistency mechanism effectively leverages even noisy geometric cues to enhance segmentation.

On NYUv2, improvements reflect the dataset's unique challenges and 40 classes. Many classes lack strong 3D structure (e.g., towels), limiting the distinctiveness of our geometric signal compared to outdoor scenes. Nevertheless, Rewis3d still clearly improves over baselines: boundaries are more precisely delineated, and predictions show better spatial coherence, particularly for structured objects like furniture. This demonstrates that while most effective with pronounced 3D structure, our geometric supervision still provides measurable benefits and state-of-the-art performance in ambiguous indoor environments.

4.6. Ablation Studies

We conduct comprehensive ablation studies to validate our key design choices. All experiments are performed on Waymo except for scribble length ablations, for which we use KITTI-360. Results are summarized in Tab. 3.

Key Findings from Ablations. (1) Dual confidence filtering improves reliability: Both prediction confidence (+0.8 mIoU) and reconstruction confidence (+0.2 mIoU) contribute to filtering noisy pseudo-labels. Without any filtering, performance reaches 51.9 mIoU. Their combination achieves the best result (53.3 mIoU), showing they capture complementary aspects of reliability.

(2) View-aware sampling enhances correspondence quality: Random sampling yields only $\sim$140 correspon-

Table 3. **Ablation study on Waymo (mIoU %).** We systematically validate each component of our framework. Base configuration uses EMA baseline (49.4 mIoU). Each row progressively adds components, demonstrating their individual and cumulative contributions.

Configuration	Confidence Filtering			Sampling		3D Source	mIoU
	None	Pred.	Recon.	Random	View-Aware		
EMA Baseline (2D only)	—	—	—	—	—	—	49.4
Effect of Confidence Filtering (with View-Aware sampling, Multi-view Recon.):							
No filtering	✓				✓	Multi-view	51.9
+ Prediction conf.		✓			✓	Multi-view	52.7
+ Reconstruction conf.			✓		✓	Multi-view	52.1
+ Both (Ours)		✓	✓		✓	Multi-view	**53.3**
Effect of Sampling Strategy (with dual confidence, Multi-view Recon.):							
Random sampling		✓	✓	✓		Multi-view	51.9
View-Aware		✓	✓		✓	Multi-view	**53.3**
Effect of 3D Reconstruction Quality (with dual confidence, View-Aware):							
Single frame		✓	✓		✓	Single	52.1
Multi-view (Ours)		✓	✓		✓	Multi-view	**53.3**

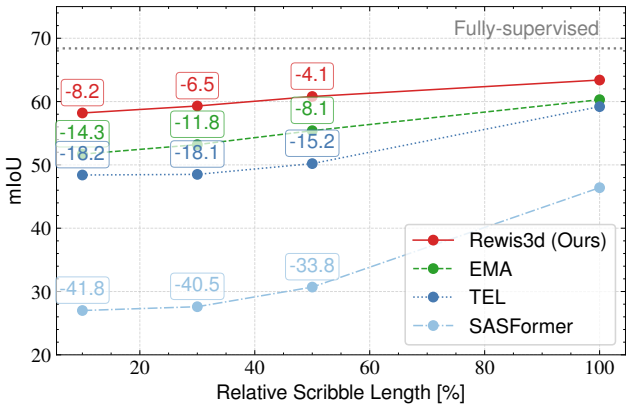

Figure 6. **Scribble length ablation on KITTI-360.** Rewis3d (red) maintains strong performance across varying scribble lengths compared to baselines, with largest gains in sparse settings, demonstrating the value of geometric supervision when annotations are scarce.

dences per image, limiting cross-modal learning (51.9 mIoU). Our view-aware strategy improves performance by +1.4 mIoU, achieving 53.3 mIoU.

(3) Multi-view reconstruction provides richer geometry: Dense point clouds from multi-view video sequences (53.3 mIoU) outperform single-frame reconstruction (52.1 mIoU) by +1.2 mIoU, as they provide richer geometric context and more reliable depth estimates through multi-view consistency. However, single-frame reconstruction still improves over the baseline by +2.7 mIoU, enabling application even in datasets without video data.

Scribble Length Robustness. Fig. 6 demonstrates that our framework maintains strong performance even with extremely sparse scribbles. While the EMA baseline and other methods degrade significantly as annotation density decreases, our method remains robust, with the performance gap widening in sparser regimes—precisely where annotation efficiency matters most.

Orthogonality to Backbone Architecture. Our proposed framework is architecture-agnostic and functions independently of the underlying segmentation backbone. To validate this, we replaced our standard network with the Encoder-only Mask Transformer (EoMT) [13], which relies on semantic features distilled from DINOv2 [20]. As shown in App. C, integrating Rewis3d consistently elevates performance over the vanilla EoMT baseline across all scenarios, demonstrating that our geometric priors successfully complement semantic foundational features.

5. Conclusion

We introduced Rewis3d, which successfully leverages 3D geometry from reconstructed point clouds to significantly improve sparsely-supervised semantic segmentation across multiple supervision types, closing a substantial portion of the performance gap to fully supervised models. Our work demonstrates that for complex, scene-centric data, targeted geometric consistency provides an effective supervisory signal. By generating this 3D supervision from standard 2D videos, our framework obviates the need for specialized 3D sensors, making it broadly applicable.

Limitations and Future Work. Our framework achieves state-of-the-art results by leveraging a 3D reconstruction model that is not explicitly optimized for dynamic content. On sequential driving data, this can introduce geometric noise and depth uncertainties from unmodeled dynamic objects or distant regions. The robustness of Rewis3d, which succeeds despite these geometric artifacts, underscores the strength of our core cross-modal consistency (CMC) principle. This also defines a clear and compelling direction for future research: integrating reconstruction models that explicitly handle dynamic scenes. We hypothesize that providing the CMC loss with a temporally consistent and geometrically cleaner 3D signal would further strengthen supervision, unlocking substantial new performance gains.

Acknowledgements

Jan Eric Lenssen is supported by the German Research Foundation (DFG) - 556415750 (Emmy Noether Programme, project: Spatial Modeling and Reasoning).

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. *What's the Point: Semantic Segmentation with Point Supervision*, page 549–565. Springer International Publishing, 2016. 1, 2

[2] Wolfgang Boettcher, Lukas Hoyer, Ozan Unal, Ke Li, and Dengxin Dai. LiDAR meta depth completion. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 7750–7756. IEEE, 2023. 1

[3] Wolfgang Boettcher, Lukas Hoyer, Ozan Unal, Jan Eric Lenssen, and Bernt Schiele. Scribbles for All: Benchmarking Scribble Supervised Segmentation Across Datasets. *Advances in Neural Information Processing Systems*, 37: 46002–46024, 2024. 1, 2, 5, 19

[4] Yilong Chen, Zongyi Xu, Xiaoshui Huang, Shanshan Zhao, Xinqi Jiang, Xinyu Gao, and Xinbo Gao. Weakly supervised lidar semantic segmentation via scatter image annotation. *IEEE Transactions on Multimedia*, 2025. 2

[5] Zhaozheng Chen and Qianru Sun. Weakly-supervised semantic segmentation with image-level labels: From traditional models to foundation models. *ACM Computing Surveys*, 57(5):1–29, 2025. 2

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1, 5

[7] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information: 1st international conference on learning representations, iclr 2013. In *1st International Conference on Learning Representations, ICLR 2013*, 2013. 5

[8] Anurag Das, Yongqin Xian, Dengxin Dai, and Bernt Schiele. Weakly-supervised domain adaptive semantic segmentation with prototypical contrastive learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15434–15443. IEEE, 2023. 5

[9] Anurag Das, Yongqin Xian, Yang He, Zeynep Akata, and Bernt Schiele. Urban scene semantic segmentation with low-cost coarse annotation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 5967–5976. IEEE, 2023. 2

[10] Yuan Gao, Shaobo Xia, Pu Wang, Xiaohuan Xi, Sheng Nie, and Cheng Wang. Lidar remote sensing meets weak supervision: Concepts, methods, and perspectives. *arXiv preprint arXiv:2503.18384*, 2025. 2

[11] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016. 3

[12] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. MapAnything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2, 3, 5

[13] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan De Geus. Your vit is secretly an image segmentation model. In *Proceedings of the computer vision and pattern recognition conference*, pages 25303–25313, 2025. 8

[14] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding Image Matching in 3D with MASt3R. In *Computer Vision – ECCV 2024*, pages 71–91, Cham, 2025. Springer Nature Switzerland. 3

[15] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree Energy Loss: Towards Sparsely Annotated Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16907–16916, 2022. 1, 2, 5, 6, 7

[16] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023. 2, 5

[17] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 2

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[19] Yongwei Miao, Guoxiang Ren, Jinrong Wang, and Fuchang Liu. Weakly supervised semantic segmentation for point cloud based on view-based adversarial training and self-attention fusion. *Computers & Graphics*, 116:46–54, 2023. 2

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2024. arXiv:2304.07193 [cs]. 8

[21] Giulia Rizzoli, Francesco Barbato, and Pietro Zanuttigh. Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives. *Technologies*, 10(4):90, 2022. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. 1

[22] Hui Su, Yue Ye, Wei Hua, Lechao Cheng, and Mingli Song. SASFormer: Transformers for Sparsely Annotated Semantic

Segmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 390–395, 2023. ISSN: 1945-788X. 1, 2, 5, 6, 7

[23] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1, 2, 5

[24] Tianfang Sun, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Image understands point cloud: Weakly supervised 3d semantic segmentation via association learning. *IEEE Transactions on Image Processing*, 33:1838–1852, 2024. 2

[25] Weixuan Sun, Jing Zhang, and Nick Barnes. 3D Guided Weakly Supervised Semantic Segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[26] Saksham Suri, Saketh Rambhatla, Rama Chellappa, and Abhinav Shrivastava. SparseDet: Improving sparsely annotated object detection with pseudo-positive mining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 6747–6758. IEEE, 2023. 1

[27] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized Cut Loss for Weakly-Supervised CNN Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2018. 2

[28] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On Regularized Losses for Weakly-supervised CNN Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 2

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4

[30] Kuan Tian, Jun Zhang, Haocheng Shen, Kezhou Yan, Pei Dong, Jianhua Yao, Shannon Che, Pifu Luo, and Xiao Han. *Weakly-Supervised Nucleus Segmentation Based on Point Annotations: A Coarse-to-Fine Self-Stimulated Learning Strategy*, page 299–308. Springer International Publishing, 2020. 2

[31] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-Supervised LiDAR Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2697–2707, 2022. 2, 3

[32] Ozan Unal, Dengxin Dai, Lukas Hoyer, Yigit Baran Can, and Luc Van Gool. 2D Feature Distillation for Weakly- and Semi-Supervised 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7336–7345, 2024. 3

[33] Ozan Unal, Christos Sakaridis, and Luc Van Gool. Bayesian Self-training for Semi-supervised 3D Segmentation. In *Computer Vision – ECCV 2024*, pages 89–107, Cham, 2025. Springer Nature Switzerland. 2, 3

[34] Gabriele Valvano, Andrea Leo, and Sotirios A. Tsaftaris. Self-supervised Multi-scale Consistency for Weakly Supervised Segmentation Learning. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 14–24, Cham, 2021. Springer International Publishing. 2

[35] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5294–5306, 2025. 2, 3

[36] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 3

[37] Zhenzhen Wang, Carla Saoud, Sintawat Wangsiricharoen, Aaron W. James, Aleksander S. Popel, and Jeremias Sulam. Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. *IEEE Transactions on Medical Imaging*, 41(12):3952–3968, 2022. 2

[38] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6488–6496. IEEE, 2017. 2

[39] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024. 5

[40] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 5

[41] Jingshan Xu, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, and Jian Yang. Scribble-Supervised Semantic Segmentation Inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15354–15363, 2021. 2

[42] Xiaoxu Xu, Yitian Yuan, Jinlong Li, Qiudan Zhang, Zequn Jie, Lin Ma, Hao Tang, Nicu Sebe, and Xu Wang. 3d weakly supervised semantic segmentation with 2d vision-language guidance. In *European Conference on Computer Vision*, pages 87–104. Springer, 2024. 3

[43] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds. In *Computer Vision – ECCV 2022*, pages 677–695, Cham, 2022. Springer Nature Switzerland. 3

[44] Jingwei Yang, Sicen Guo, Mohammud Junaid Bocus, Qijun Chen, and Rui Fan. Semantic Segmentation for Autonomous Driving. In *Autonomous Driving Perception: Fundamentals*

and Applications, pages 101–137. Springer Nature, Singapore, 2023. 1

[45] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From CNN to Transformer: A Review of Medical Image Segmentation Models. *Journal of Imaging Informatics in Medicine*, 37(4):1529–1547, 2024. 1

[46] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 3

[47] Xinyi Yu, Ling Yan, Pengtao Jiang, Hao Chen, Bo Li, Lin Yuanbo Wu, and Linlin Ou. Boosting Box-supervised Instance Segmentation with Pseudo Depth, 2024. arXiv:2403.01214 [cs]. 3

[48] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[49] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. GrowSP: Unsupervised Semantic Segmentation of 3D Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17619–17629, 2023. 2

[50] Tianyi Zhao and Zhaozheng Yin. Weakly supervised cell segmentation by point annotation. *IEEE Transactions on Medical Imaging*, 40(10):2736–2747, 2020. 2

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2

[52] Kaiyin Zhu, Neal N. Xiong, and Mingming Lu. A survey of weakly-supervised semantic segmentation. In *2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 10–15, 2023. 1