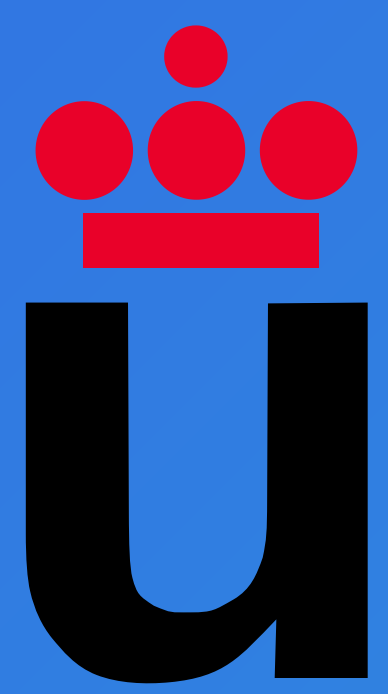




Semantic-driven Alignment for Robust Vision-LiDAR Sensor Fusion in Off-Road 3D Segmentation



David Pascual-Hernández¹, Roberto Calvo-Palomino¹, Inmaculada Mora-Jiménez¹, Jose María Cañas-Plaza¹

Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Fuenlabrada, 28943, Spain

Motivation

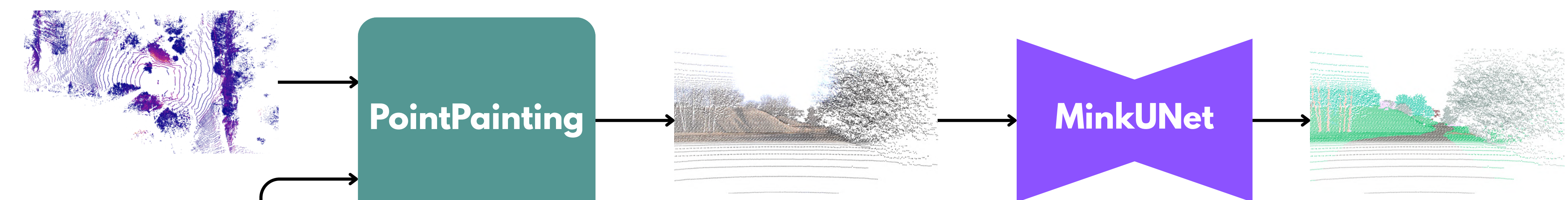
- **Semantic 3D scene understanding** is essential for safe navigation in unstructured environments, where the diversity of hazards and varying terrain traversability pose significant challenges.
- **Why fusion?** LiDAR provides rich geometric data, but off-road terrains are often geometrically ambiguous. RGB information helps disambiguate these elements.
- **Precise calibration** is fundamental to robust sensor fusion. However, while operating in off-road environments **spatiotemporal synchronization may degrade** due to rugged terrain and adverse conditions.
- There is a **gap** in existing methods, which typically require:
 - Urban geometry → Edges and planes are scarce in off-road settings.
 - Dense intensity → Limited by hardware capabilities.
 - Costly multi-frame optimization → Not feasible for online updates.

Experimental Setup

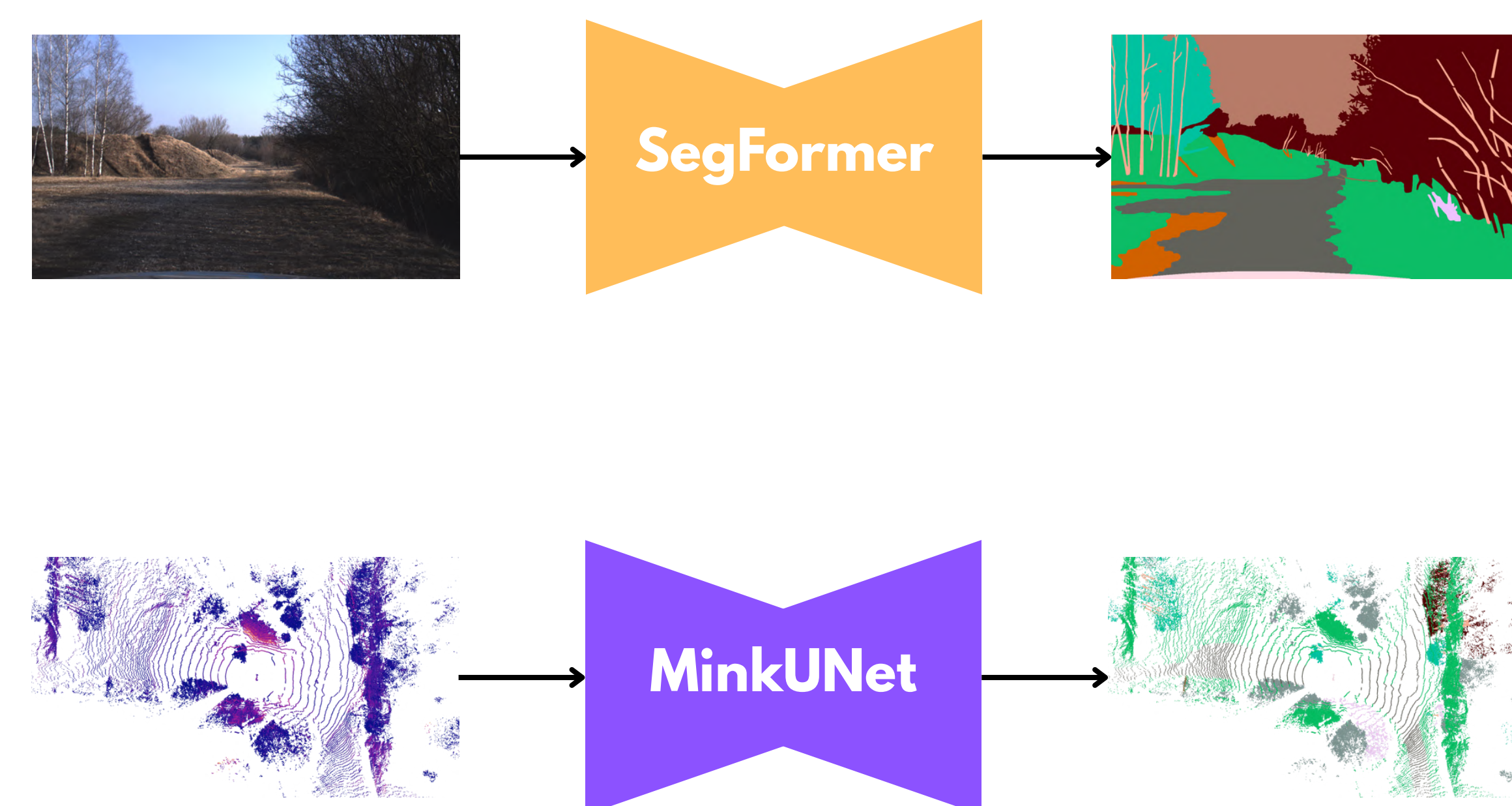
- **Off-road Perception Datasets:** RELLIS-3D [1] and GOOSE¹ [2].
- **Model Architectures:**
 - 2D → SegFormer [3] (Transformer-based).
 - 3D → MinkUNet [4] (Sparse 3D Convolutions).
- **Fusion Strategy:** PointPainting [5], i.e., projecting 2D class probabilities onto the 3D point cloud using our estimated extrinsics.
- **Metrics:** mean Intersection over Union (mIoU).

We propose a **semantic-driven alignment framework for sensor fusion** that uses **gradient descent** to minimize the **Jensen-Shannon Divergence** between **2D and 3D class probabilities**. To ensure robustness, we introduce a **dual-weighting scheme** using **class frequency** and **CLIP distance**, effectively addressing semantic inconsistency and class imbalance.

Fusion



Alignment



$$[R|t]^* = \arg \min_{[R|t]} \frac{1}{|\mathcal{X}_{\text{vis}}|} \sum_{i \in \mathcal{X}_{\text{vis}}} \omega_i \cdot D_{JS}(P_{2D}(K[R|t]\mathbf{x}_i) \parallel P_{3D}(\mathbf{x}_i))$$

$$\omega_i = \left(\frac{\sum_{j=1}^C n_j}{n_{c_i} + \epsilon} \right) \cdot \underbrace{\exp(d_{\text{CLIP}})}_{\text{CLIP Weight}}$$

Freq. Weight

CLIP Weight

Jensen-Shannon Divergence

$[R|t]^*$: Optimal extrinsic transformation matrix.

$\mathcal{X}_{\text{vis}}$: Subset of 3D points visible within the camera frustum.

$\mathbf{x}_i \in \mathbb{R}^4$: i -th LiDAR point.

P_{2D}, P_{3D} : Semantic probability distributions (logits).

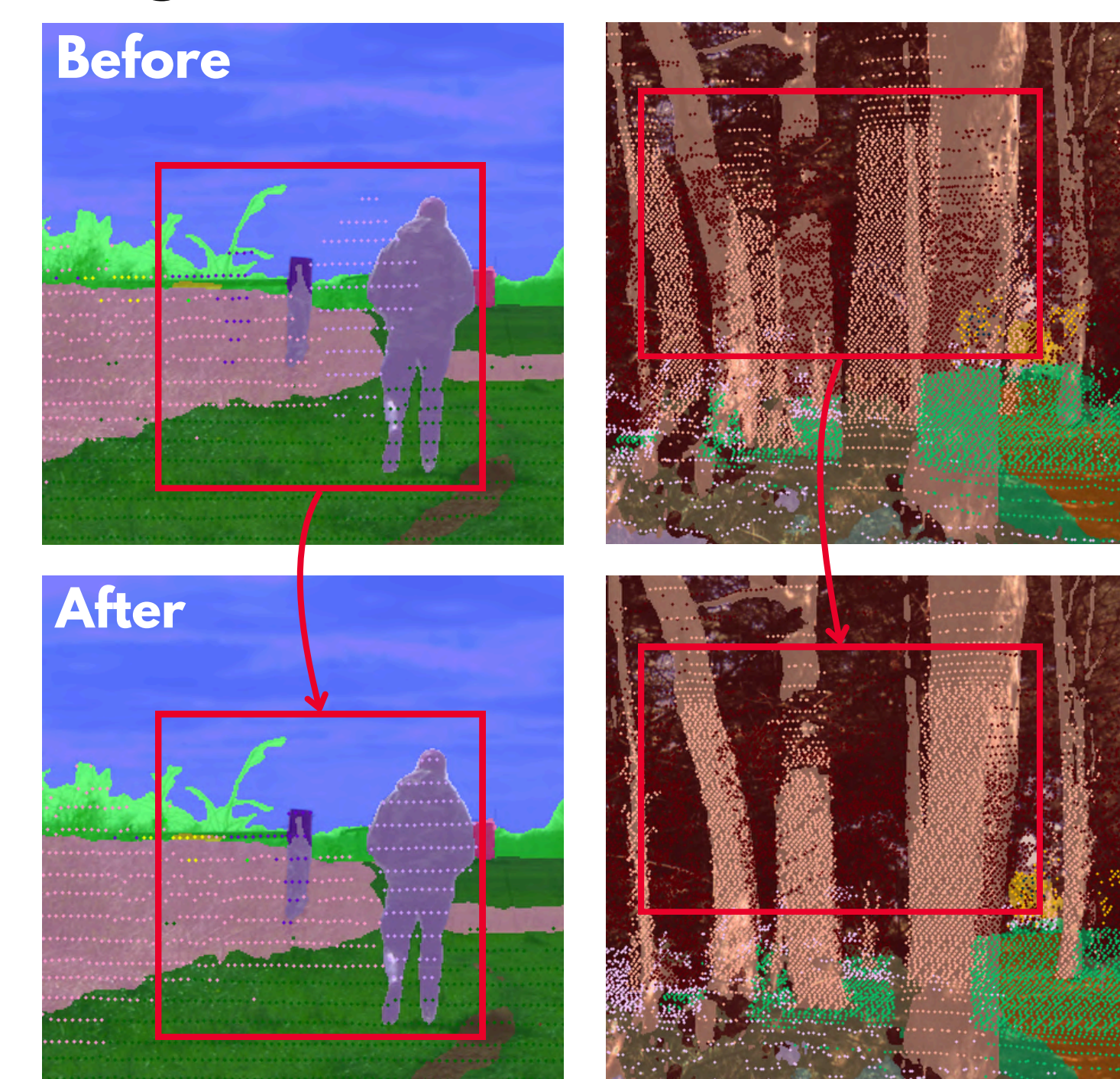
K : Camera intrinsic matrix.

n_{c_i} : Total count of points belonging to class c_i .

d_{CLIP} : Cosine distance between CLIP embeddings of 2D/3D labels.

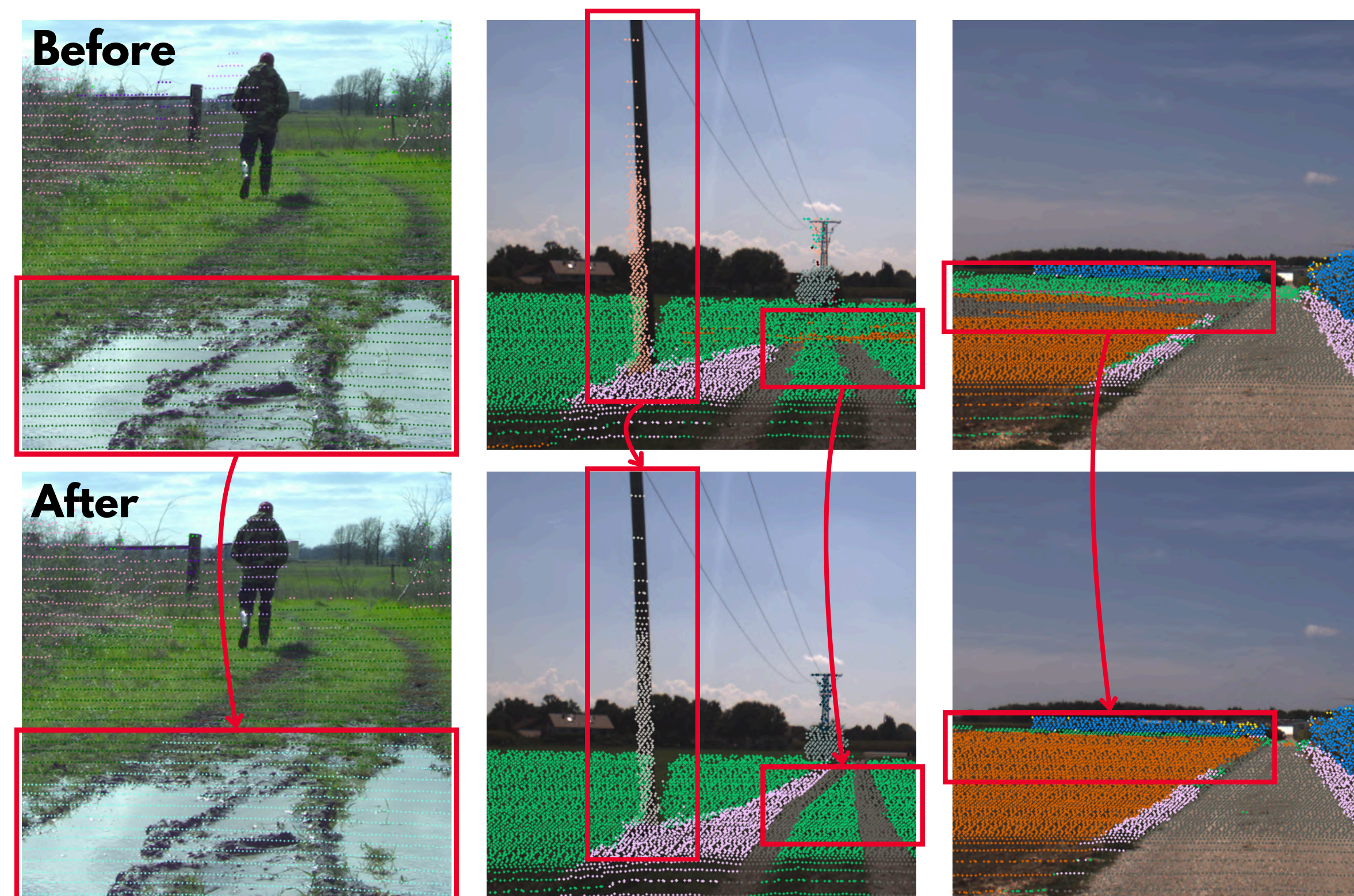
Results

Alignment



Fusion

Dataset	Fusion	Aligned	Weighted	mIoU(%)
RELLIS-3D	$\times$	$\times$	$\times$	43.54
	$\checkmark$	$\times$	$\times$	45.46
	$\checkmark$	$\checkmark$	$\times$	47.67
	$\checkmark$	$\checkmark$	$\checkmark$	47.94
GOOSE (12 categories)	$\times$	$\times$	$\times$	65.16
	$\checkmark$	$\times$	$\times$	65.92
	$\checkmark$	$\checkmark$	$\times$	65.98
	$\checkmark$	$\checkmark$	$\checkmark$	66.25
GOOSE (64 classes)	$\times$	$\times$	$\times$	30.70
	$\checkmark$	$\times$	$\times$	31.42
	$\checkmark$	$\checkmark$	$\times$	32.12
	$\checkmark$	$\checkmark$	$\checkmark$	32.29



Key Takeaways

- **Overcoming geometric scarcity** → In unstructured environments where geometric primitives are scarce, our semantic-driven alignment boosts mIoU by +2.5% mIoU on RELLIS, +0.3% on GOOSE-12, and +0.9% on GOOSE-64.
- **Semantic consistency and class imbalance** → CLIP and frequency-based weighting yield mIoU gains of +0.3% on RELLIS and GOOSE-12, and +0.7% on GOOSE-64.
- **Sensor fusion impact** → The full pipeline achieves cumulative mIoU improvements of +4.4% (RELLIS), +1.1% (GOOSE-12), and +1.6% (GOOSE-64) over LiDAR-only baseline.

Ongoing & Future Work

- **Online calibration:** Optimizing latency and computational overhead to enable continuous extrinsic correction during field operation.
- **Fusion strategies:** Exploring the impact of semantic-driven alignment across a broader range of RGB-LiDAR fusion strategies.
- **Environmental robustness:** Evaluating alignment stability under adverse conditions, such as low-light, dust, smoke, or adverse weather.

[1] Jiang, P., Osteen, P., Wigness, M., & Saripalli, S. (2021). Rellis-3d dataset: Data, benchmarks and analysis. In IEEE ICRA [2] Mortimer, P., Hagmanns, R., Granero, M., Luettel, T., Petereit, J., & Wuensche, H. J. (2024). The goose dataset for perception in unstructured environments. In IEEE ICRA; [3] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 34, 12077-12090; [4] Choy, C., Gwak, J., & Savarese, S. (2019). 4d spatio-temporal convnets: Minkowski convolutional neural networks. In IEEE/CVF CVPR; [5] Vora, S., Lang, A. H., Helou, B., & Beijbom, O. (2020). Pointpainting: Sequential fusion for 3d object detection. In IEEE/CVF CVPR.

This work was supported by Spanish Agencies AEI and CDTI as part of the Ministry of Science, Innovation and Universities with grant number PLEC2023-010303 (GAIA); by Comunidad de Madrid with grant number TEC-2024/TEC-62 (IROBOCITY 2030-CM); and by the Spanish Agency AEI with Grant number PID2022-136887NB-I00 (POLIGRAPH).

