Towards Sharper Object Boundaries in Self-Supervised Depth Estimation

Aurélien Cecille[1,2]
Stefan Duffner[2]
Franck Davoine[2]
Rémi Agier[1]
Thibault Neveu[1]

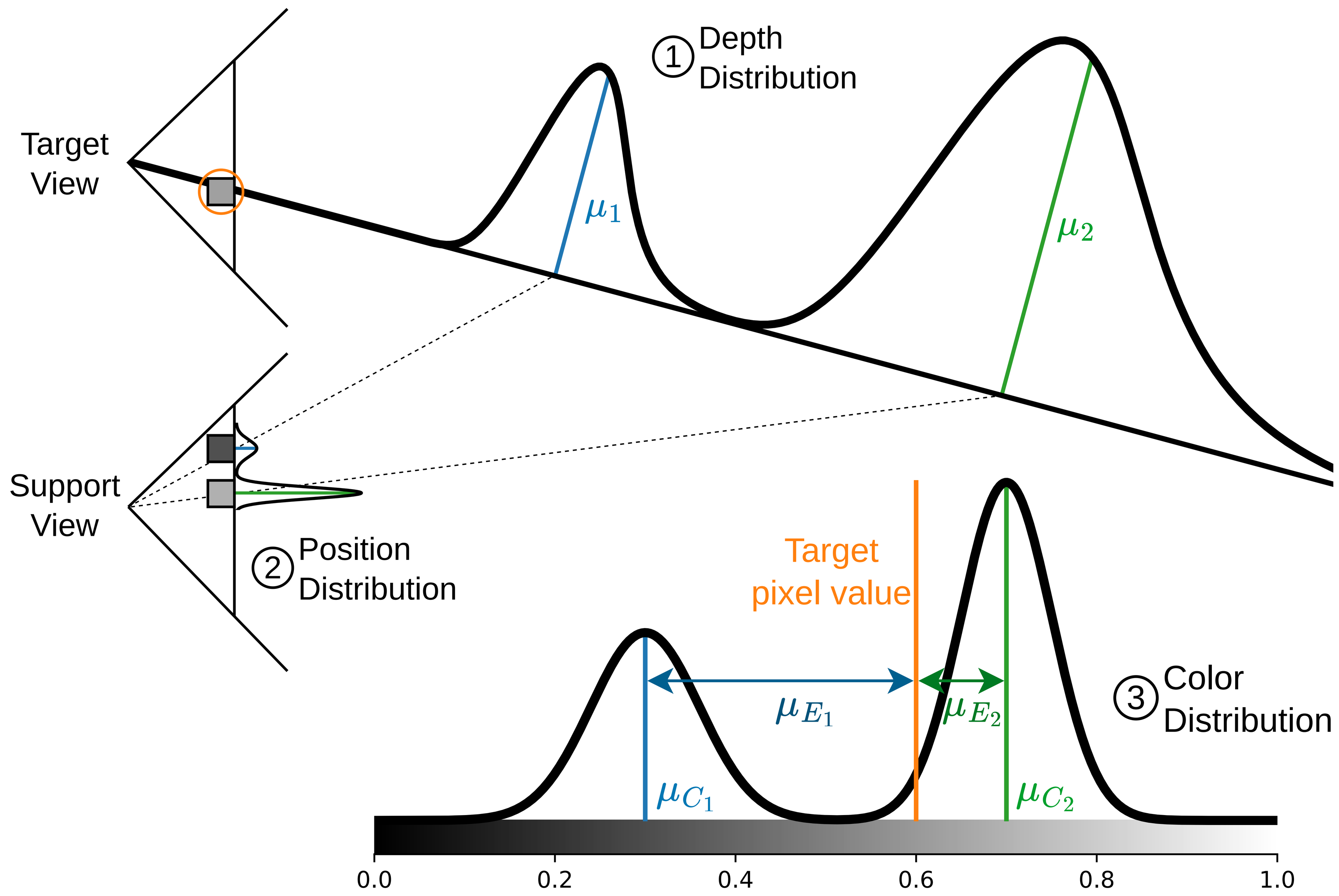
BMVC 2025

[1]Visual Behavior, Lyon, France
[2]LIRIS, Villeurbanne, France
INSA Lyon, CNRS, École Centrale de Lyon, Université Lumière Lyon 2, Universite Claude Bernard Lyon 1

Abstract

Accurate monocular depth estimation is crucial for 3D scene understanding, but existing methods often blur depth at object boundaries, introducing spurious intermediate 3D points. While achieving sharp edges usually requires very fine-grained supervision, our method produces crisp depth discontinuities using only self-supervision. Specifically, we model per-pixel depth as a mixture distribution, capturing multiple plausible depths and shifting uncertainty from direct regression to the mixture weights. This formulation integrates seamlessly into existing pipelines via variance-aware loss functions and uncertainty propagation. Extensive evaluations on KITTI and VKITTIv2 show that our method achieves up to 35% higher boundary sharpness and improves point cloud quality compared to state-of-the-art baselines.

Mixture Reconstruction



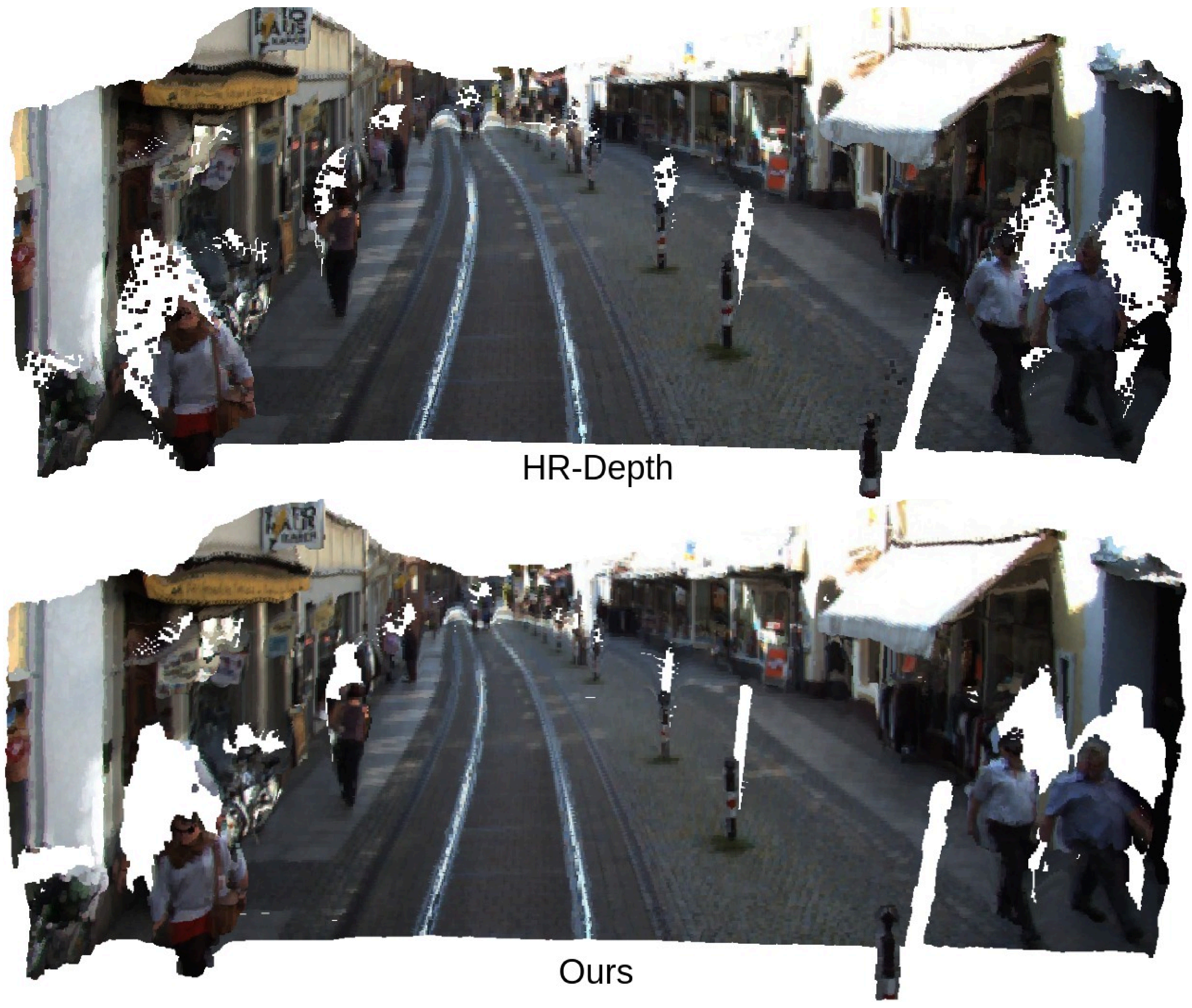
Our method builds on standard self-supervised depth estimation: predict depth and camera pose between frames, then use these to warp the support image to match the target. The reconstruction quality provides supervision without ground truth depth labels.

To handle ambiguity at object boundaries, we use a two-component disparity mixture per pixel. This captures foreground/background alternatives and prevents averaging conflicting estimates.
We propagate the uncertainty of these distributions through the reprojection pipeline using linear approximation, translating disparity uncertainties to color uncertainties.

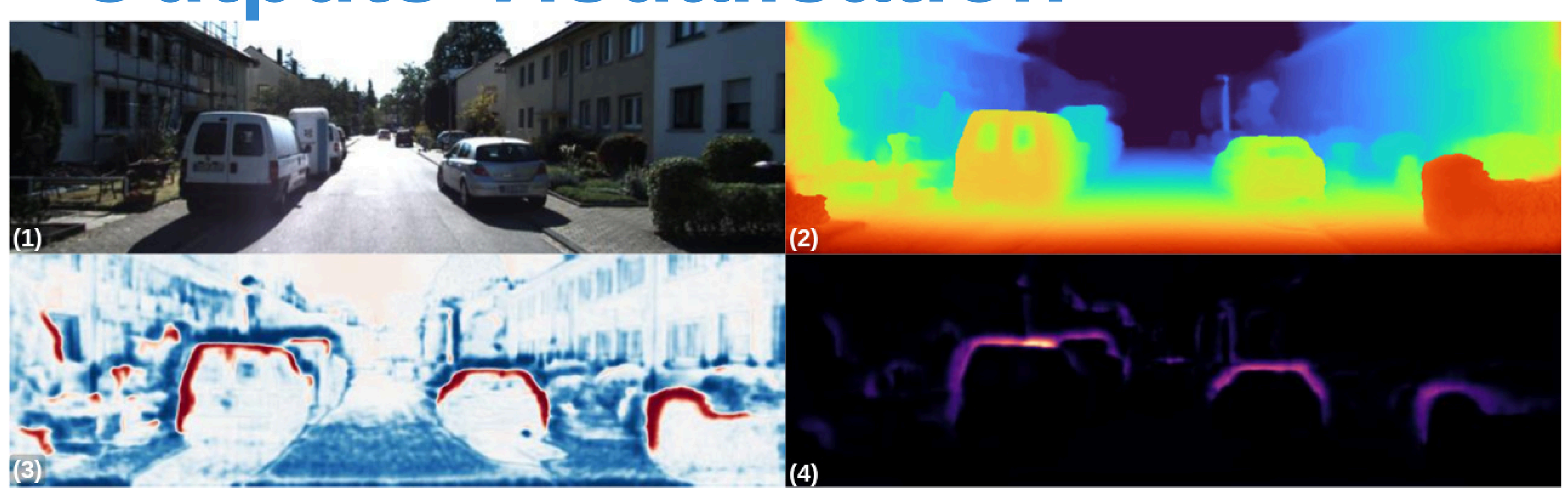
Components compete by comparing their errors, focusing training on the predicted winner. Three loss terms encourage: minimizing winner's error, estimating uncertainty and learning mixture weighting.

At inference, we select the most confident component rather than averaging, preserving crisp depth estimates.

Removes Pointcloud Artefacts



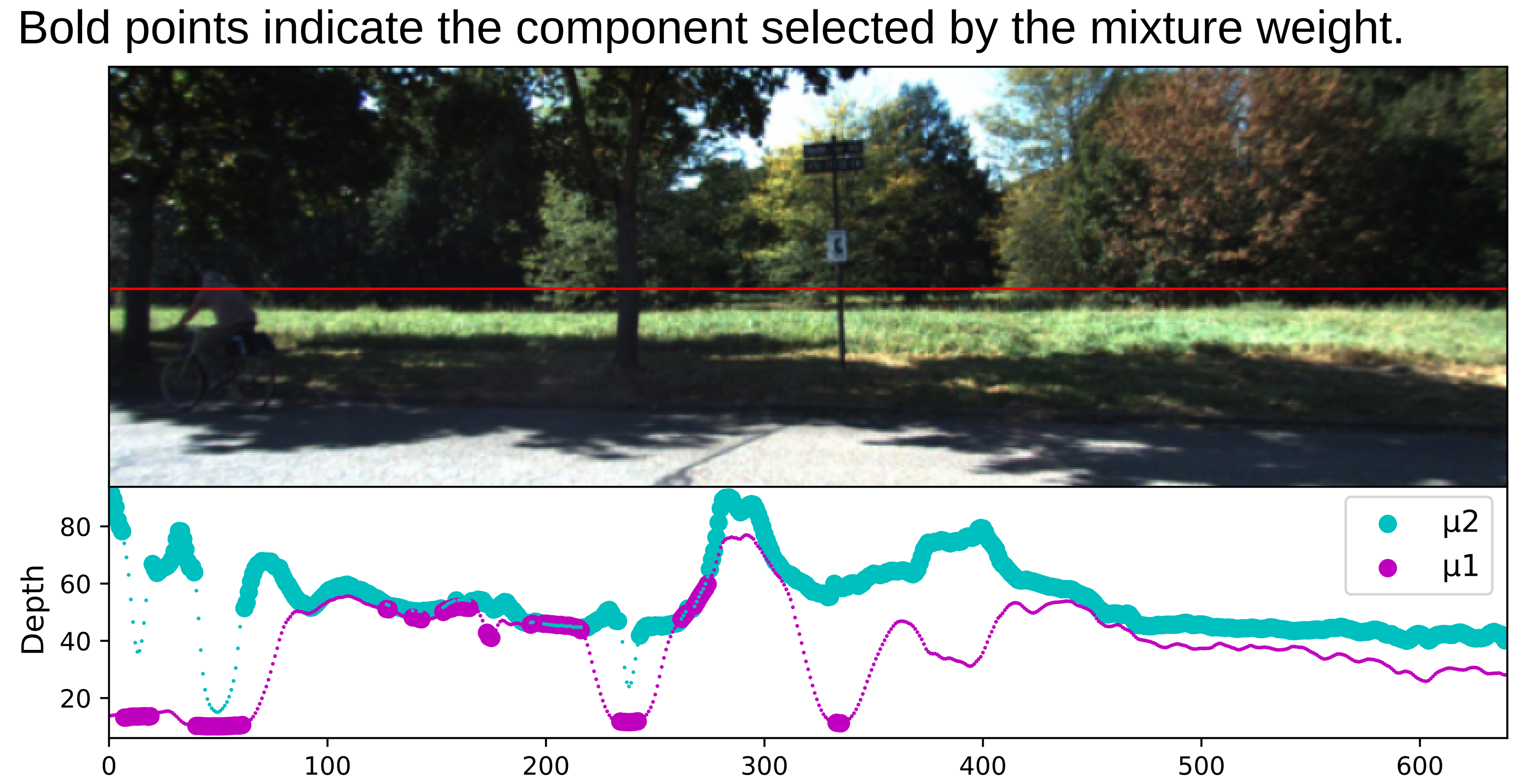
Outputs Visualisation



(1) Input Image (2) Most likely depth (3) Mixture weights α
(4) Difference between the two components means

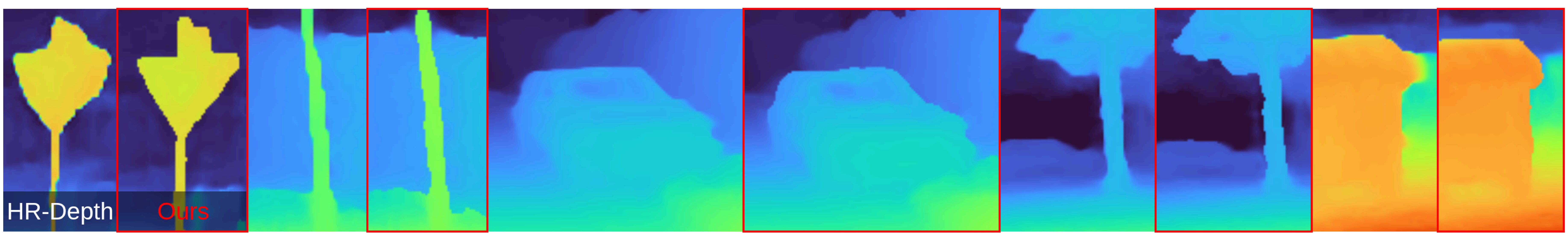
Components Specialization

Looking at the depth components on a single row of pixels, we see that they complementarily over/underestimate obstacle width smoothly, and that it's the selection process that creates the discontinuity.
Bold points indicate the component selected by the mixture weight.



Sharp Boundaries

Our method predicts sharp objects borders, removing blurry edges



VISUAL BEHAVIOR LIRIS cnrs INSA Lyon 1 université Lumière Lyon 2 ÉCOLE CENTRALE LYON