



Monocular Depth in the Real World



Toyota Research Institute · Follow

Published in Toyota Research Institute

11 min read · May 19, 2022



Listen



Share

By Vitor Guizilini et al.

Understanding scenes in 3D is both a fundamental challenge in computer vision and a breakthrough capability in practice. Learning how to infer depth only from cameras can indeed help to save lives, increase mobility, reduce costs, and improve manufacturing processes, among many other applications. In our previous blog posts ([part1](#) / [part2](#)), we presented some of our research on *self-supervised learning*, showing that deep networks can learn from raw videos how to predict *accurate* depth and motion information

In this follow-up post, we go one big step further: how to bring monodepth to the real world. Our core insight lies in jointly learning end-to-end multiple geometrically related prediction tasks, such as learning camera models, depth, and per-pixel motion in 2D (a.k.a. optical flow) and 3D (a.k.a. scene flow). Furthermore, we go beyond relaxing geometric constraints and show how breakthroughs in neural architectures, including the famous transformers [7], enable generalization to multi-camera and multi-frame setups that are the norm in practice.

Learning Arbitrary Camera Geometries

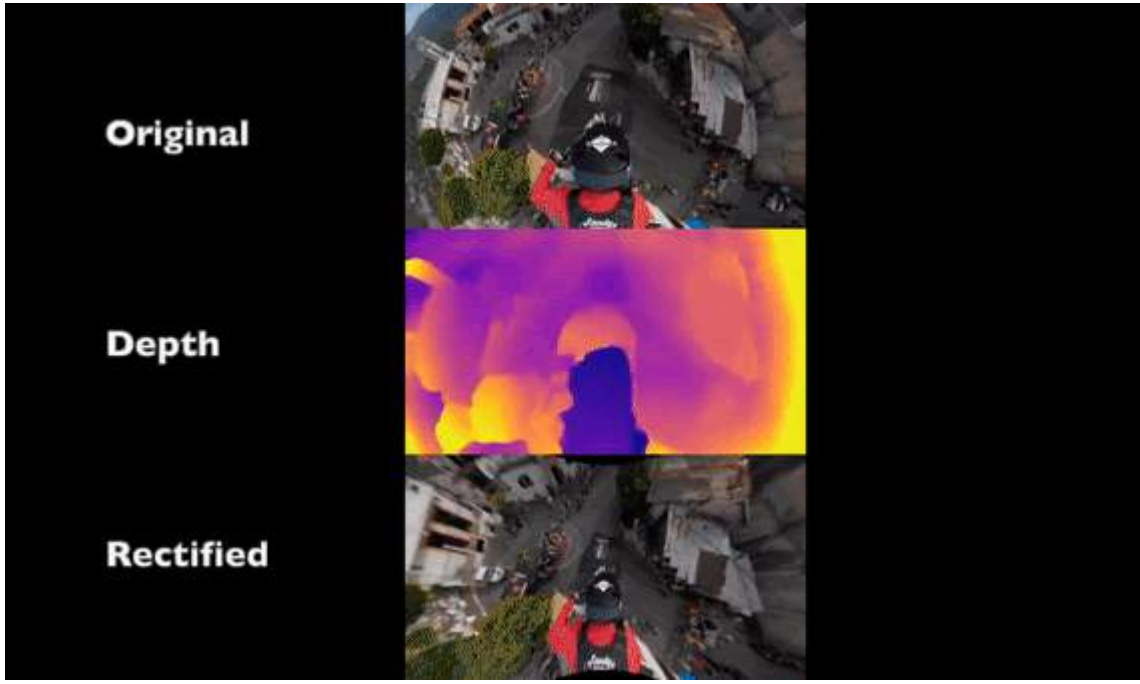
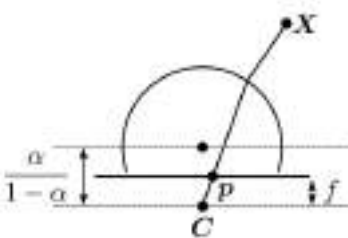


Figure 1. Our self-calibration framework learns depth and ego-motion from unlabeled videos collected with arbitrary camera models.

Most works in self-supervised monocular depth estimation focus on only two of the three components required to use geometry as inductive biases for training: **depth** and **ego-motion** (a.k.a., pose). But there is another component that is crucial to perform these lifting and projection operations: the **camera model**. This information is usually assumed to be known, which is a strong assumption that violates the ultimate goal of self-supervised learning: to extract information from unlabeled data. Even worse, most methods are restricted to the standard rectified pinhole model, composed of a 3 x 3 intrinsic matrix containing focal lengths and principal points. While this pinhole model enables the efficient projection from 2D to 3D and vice-versa, it severely limits the applicability of such methods, as it is only an approximation to real-world cameras. It leaves out many other widely used cameras, such as perspective cameras with radial distortion, fisheye cameras and catadioptric cameras.



$$\pi(P, i) = \begin{bmatrix} f_x \frac{x}{\alpha d + (1-\alpha)z} \\ f_y \frac{y}{\alpha d + (1-\alpha)z} \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix}$$

Figure 2. The Unified Camera Model (UCM) approximates a wide variety of cameras. It introduces an alpha parameter ranging from [0,1] that offsets a pinhole camera model.

In our recent paper, *Self-Supervised Camera Self-Calibration from Video* (ICRA 2022), we propose a self-supervised monocular depth estimation framework capable of adapting to a wide variety of cameras, using the **Unified Camera Model (UCM)** [5,6]. With 5 parameters, only one more than the traditional pinhole model, we can approximate cameras ranging from pinhole to fisheye, and even catadioptric, with the same model. Unlike some other parametric models, UCM possesses closed-form projection and unprojection operations, which makes it a better fit for auto-diff based pipelines (deep learning). In practice, instead of predicting per-image camera parameters, as is common in the literature, we learn **per-dataset** parameters, using shared estimates. We show that this approach leads to more stable results, with the only added mild assumption that the same camera is used in each video sequence.

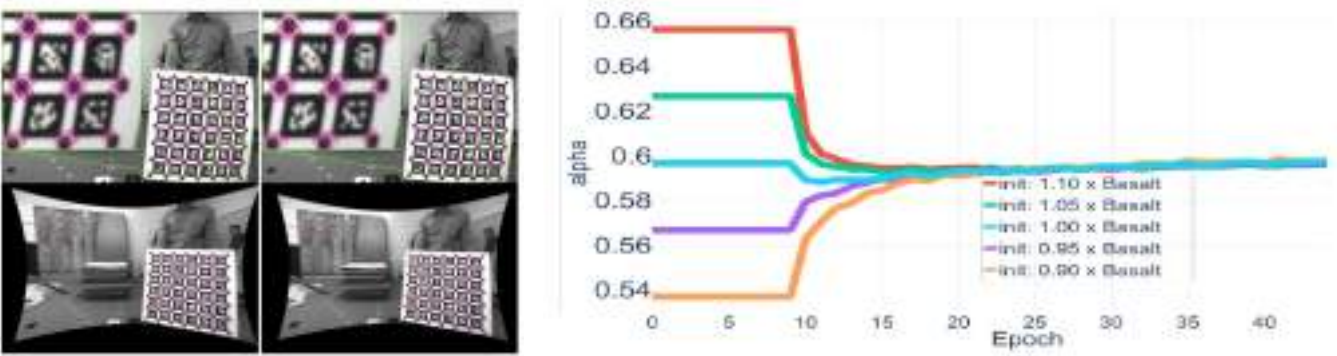


Figure 3. Our learned calibration parameters achieve sub-pixel reprojection error, and can be used to (left) rectify images, as well as (right) online recover from calibration errors

These learned parameters can then be used as calibration parameters for that camera, under the UCM geometry. We show that our approach achieves **sub-pixel reprojection error**, relative to standard target-based calibration methods. However, because we operate in a fully self-supervised setting, calibration can still be performed online after deployment, which allows us to recover from deviations to the originally predicted parameters. These deviations are common in deployed systems, due to temperature changes, vibrations, and other unforeseen circumstances, and methods that use hard-coded parameters will severely suffer in these cases. Our proposed approach, however, is capable of **self-calibration**, and therefore can adapt to these changes and quickly recover from such failures.

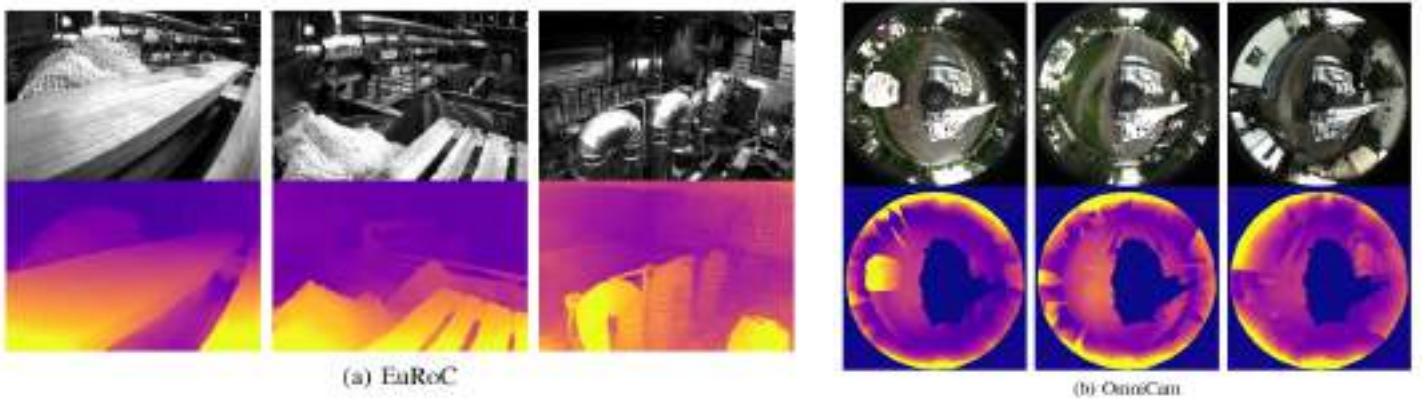


Figure 4. Examples of depth maps generated from fisheye (left) and catadioptric (right) cameras.

Explicit Modeling of Dynamic Objects

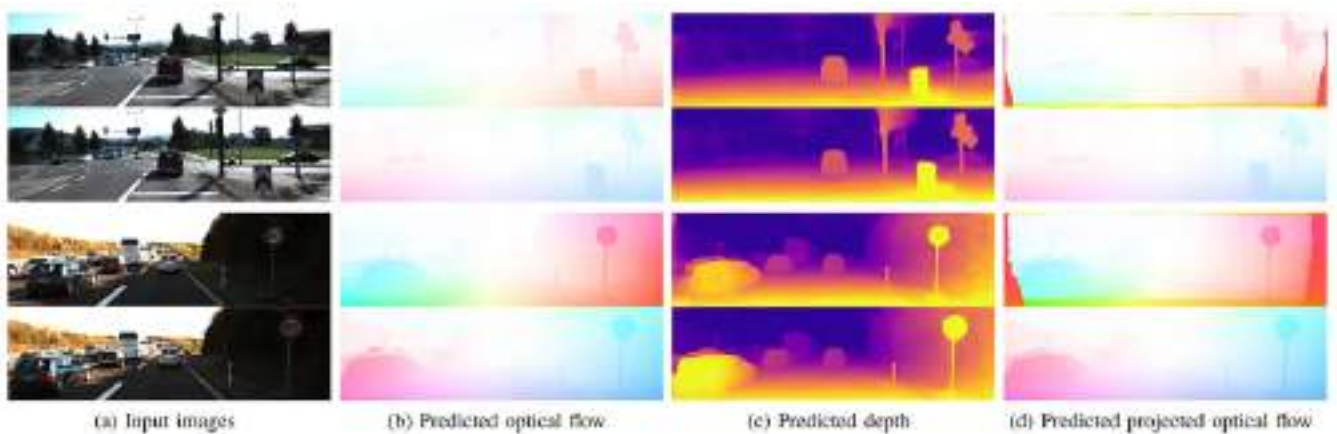


Figure 5. Our multi-task DRAFT architecture generates state of the art depth, optical flow, and scene flow estimates. It is trained using a combination of unlabeled real-world data and synthetic labeled data.

Another key assumption in the standard framework for self-supervised monocular depth estimation is the **static world assumption**. Because we are warping information between frames, it is necessary to know the transformation between viewpoints. However, because we are dealing with temporal information, there is always the possibility that objects have moved between timesteps. The usual approach to mitigate these problems is via dataset filtering, either at a pixel level [9] or at an image level (something we have already explored [here](#)). However, it is clear that the best solution would be to explicitly model it as an additional task. Luckily, such a task already exists, and it is called **scene flow**: per-pixel vectors estimating 3D motion. Many scene flow methods have been proposed over the years, achieving very promising results as a complement to depth estimation. There is one caveat, though: none of these methods are self-supervised from a monocular camera. More specifically, depth is always either supervised with ground-truth or self-supervised

from stereo cameras. The reason for that is simple: jointly estimating depth and scene flow in a monocular depth estimation setting is an **ill-posed problem**, meaning that there are infinite solutions to the same initial state.

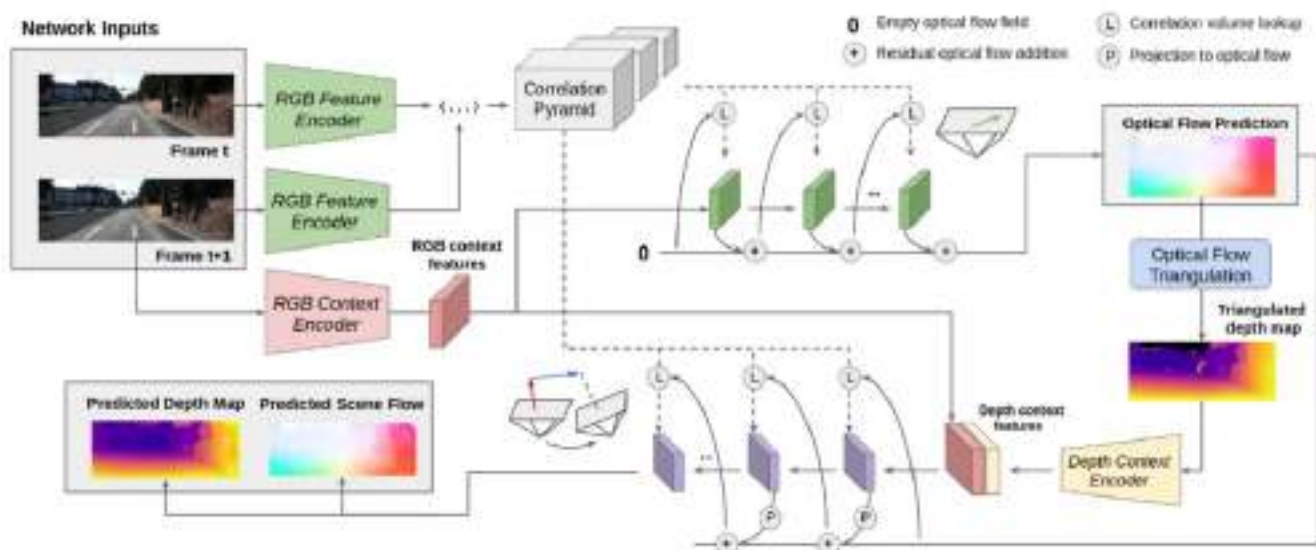


Figure 6. DRAFT diagram. Two input images are used to generate optical flow estimates. These estimates are triangulated to generate initial depth maps, and refined with scene flow to produce final estimates for these tasks.

Even so, in our recent paper *Learning Optical Flow, Depth, and Scene Flow without Real-World Labels* (RA-L + ICRA 2022), we propose DRAFT, a method capable of doing just that. We extend a very recent optical flow architecture [8] to (a) also estimate depth and scene flow, and (b) to operate in the self-supervised monocular setting. The first key component is **multi-task consistency**, to decrease the state space of possible solutions to this ill-posed problem. The second key component is to learn **strong geometric priors from synthetic data**, where exact supervision is always available, and transfer this information to the real-world using mixed-batch training. To achieve the former, we jointly learn optical flow, depth, and scene flow, using the same network, and enforce a series of cross-task regularizations by producing the same output in multiple ways. To achieve the latter, we use a novel procedurally generated dataset from Parallel Domain, including photo-realistic multi-camera images with corresponding depth, optical flow, and scene flow labels.

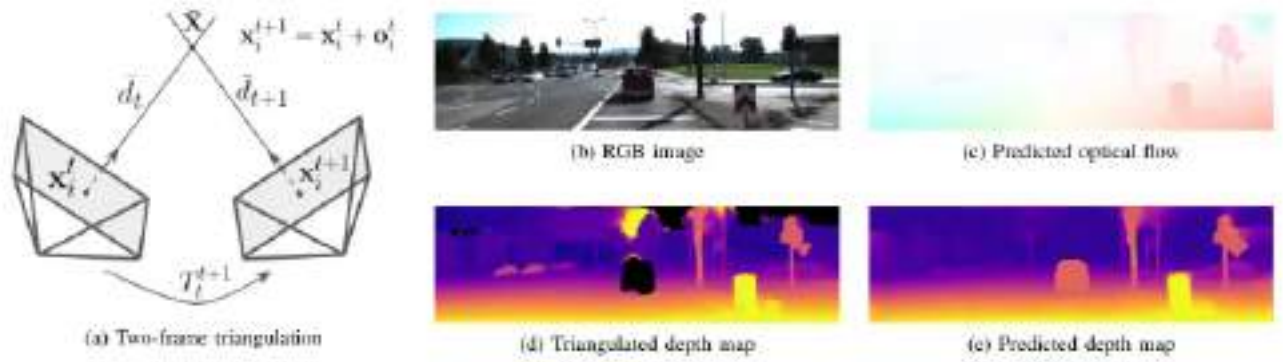
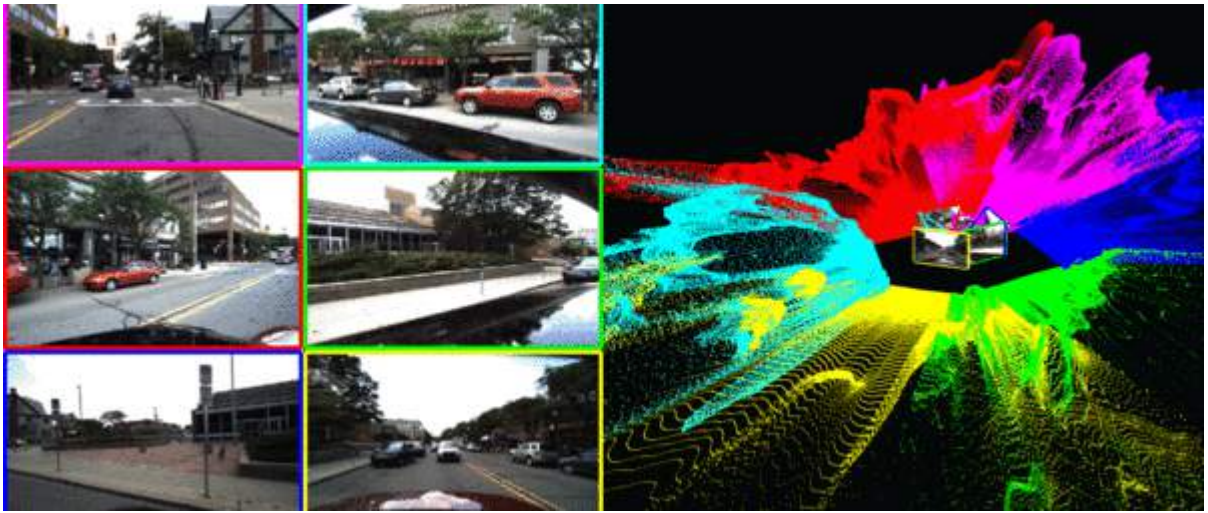


Figure 7. Example of triangulation for depth estimation.

We also explore another way of leveraging the multi-task aspect of our DRAFT architecture, by using estimated optical flow as initialization for depth and scene flow estimation. We achieve that in two ways: (a) by triangulation, we can produce depth from optical flow, generating highly accurate predictions for static portions of the environment, and (b) using optical flow estimates as initialization for depth and scene flow refinement, taking advantage of our recursive architecture. The combination of all these techniques enables DRAFT to achieve **state of the art results in all three considered tasks using the same model**.

Multi-Camera Consistent Pointclouds



One of the primary goals of monocular depth estimation is to complement or substitute LiDAR sensors. There are several benefits in doing so: cameras are cheaper, and provide dense and textured information. However, LiDAR sensors naturally provide 360o measurements around the vehicle, while cameras are limited to a much smaller field of view, and often have to deal with issues such as self-occlusion. To increase coverage, we can always use more cameras pointing in different directions. However, because each frame is processed independently,

there is no guaranteed consistency across pointclouds, which is made even worse due to the scale ambiguity inherent to self-supervised monocular depth estimation.



Figure 9. Example of pointcloud aggregation from multiple cameras with (right) and without (left) our proposed consistency objective. Note that the baseline is highly misaligned.

Stereo methods are capable of enforcing multi-view consistency, at the expense of requiring large overlaps in field of view and fronto-parallel rectification. However, the configurations we investigate include very small overlaps between frames, and the cameras can be placed at arbitrary locations. This is a common scenario in autonomous driving, and the most popular datasets (nuScenes, DDAD, Waymo) have this type of camera arrangement. In this setting, stereo methods cannot be used, and that is why in our recent paper *Full Surround Monodepth from Multiple Cameras* (RA-L + ICRA 2022) we propose FSM, a generalization of self-supervised depth estimation that can be applied both in the monocular (single camera in different time steps) and generalized stereo (different cameras in the same timestep). In fact, we go further and show that spatio-temporal constraints (different cameras in different timesteps) also lead to improvements, by increasing the amount of cross-camera overlap.

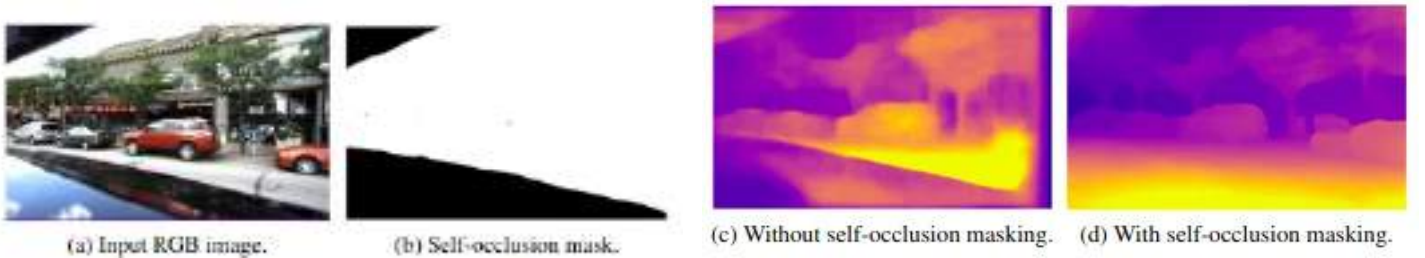


Figure 10. Example of self-occlusion mask and how it affects self-supervised depth estimation results.

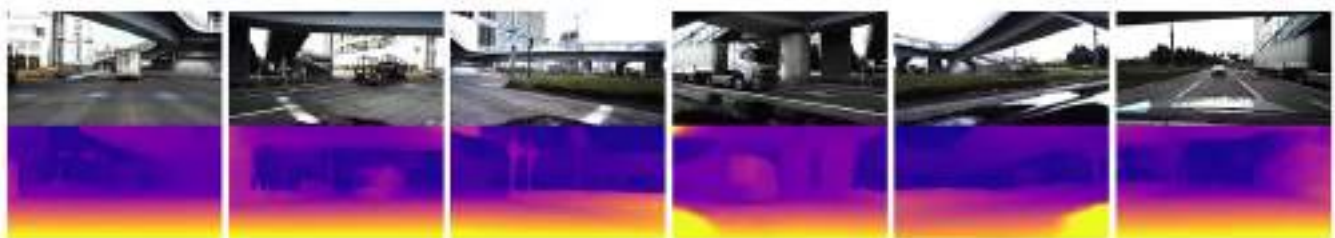
Some other key technical insights are made and used to further boost performance and provide a better analysis of such improvements. The first of them is **self-occlusion masks**. When placing cameras on a vehicle, sometimes compromises have to be made, resulting in some parts of the image being covered by the vehicle itself. These areas do not contain any useful information, and to avoid using them

we manually generate a mask for each camera, and use it to remove these pixels during the photometric loss calculation. Additionally, we propose a novel regularization term that enforces the **estimated ego-motion from each camera to be the same**, following the assumption that they are attached to a rigid body. Finally, we propose **multi-camera median scaling** as a way to better capture the inconsistencies in depth map generation when using multiple cameras.

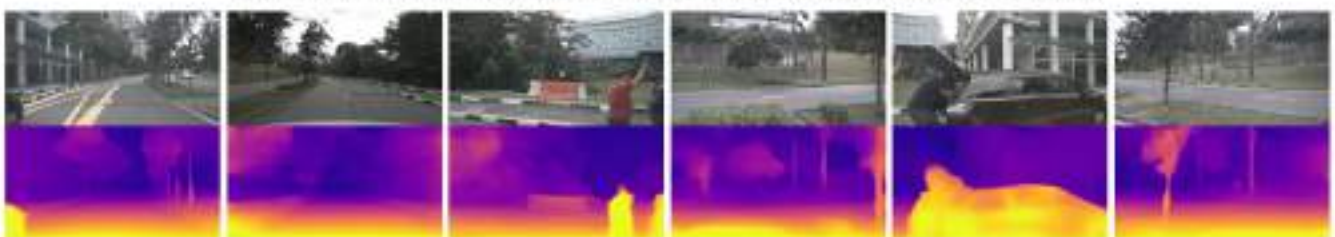


Figure 11. Overlap between cameras and point cloud predictions for DDAD (left) and nuScenes (right) datasets.

The combination of all these techniques enable FSM to achieve state-of-the-art results on the DDAD and nuScenes datasets, two widely used multi-camera driving datasets. Furthermore, our estimates improve upon other methods in several other ways: (a) even though we do not assume known ego-motion, our estimates are **scale-aware**, due to the use of known cross-camera extrinsics; (b) in fact, our metric scales are **better than median-scaling**, and can be used in real-world applications where ground-truth information is not available at test time; © even though our estimates are still single-frame, the resulting pointclouds are **much more spatially consistent**, due to the cross-camera photometric losses enforced at training time, and evaluated with our novel multi-camera metrics.



a) Multi-camera depth estimation results on DDAD.



b) Multi-camera depth estimation results on NuScenes.

Figure 12. Examples of FSM depth maps on various multi-camera datasets.

Multi-Frame Monocular Depth Estimation

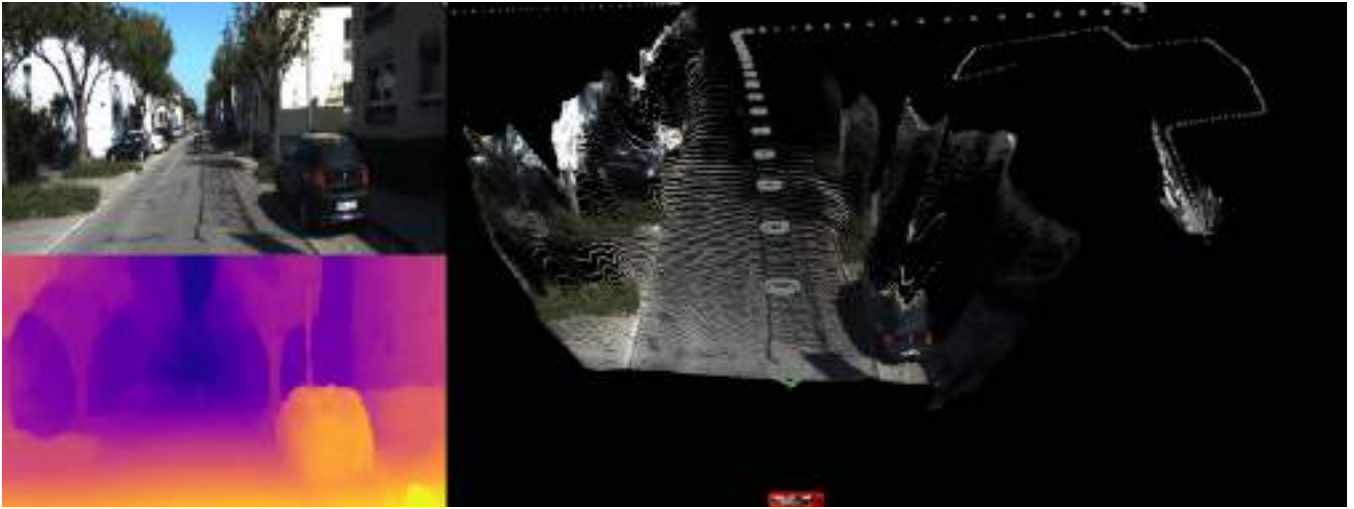


Figure 13. We can use predicted depth maps and ego-motion to generate consistent pointclouds for entire sequences, as shown here on the KITTI dataset.

The final paradigm of self-supervised monocular depth estimation we revisit in this series is **multi-frame inference**. As we discussed before, multi-view consistency is key in self-supervision for depth estimation. However, the depth network itself operates on a single frame, and therefore, even though geometry is used at training time, the learned features are purely **appearance-based**. This severely limits the expressivity of such features, leading to suboptimal performance relative to multi-frame methods. Even so, while multi-frame stereo methods have found great success in the literature, multi-frame monocular methods have so far remained largely overlooked.

The reason for that is simple, and goes back to the core challenges we discussed earlier: external motion and static frames. When these issues are present only at training time, there are tools to mitigate their influence. When multiple frames are used during inference as well, however, these issues can severely degrade performance. This is because the learned features will also be **geometry-based**, which leads to much better results, but also catastrophic failures when hard-coded geometric assumptions do not hold (e.g. a mis-calibrated camera, or dynamic objects).

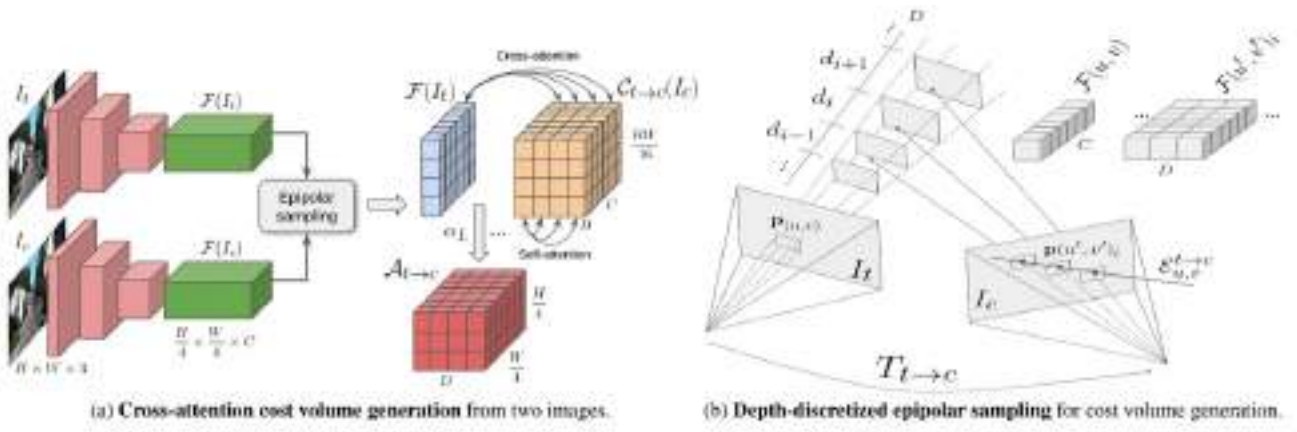


Figure 14. Our DepthFormer architecture uses depth-discretized epipolar sampling to generate a multi-view cost volume, and attention values as the similarity metric to perform feature matching.

However, at the core of these challenges lies an even deeper aspect of multi-view geometry: **feature matching**, and that’s what we address in our recent paper *Multi-Frame Self-Supervised Depth with Transformers* (CVPR 2022). Our core contribution is DepthFormer, a novel **attention-based** feature matching module that improves upon traditional similarity metrics for feature matching. Because this is a highly ambiguous setting, due to inherent data limitations (e.g., low texture areas, luminosity changes, dynamic objects, etc.), we propose to instead **learn a better similarity metric**. We use depth-discretized epipolar sampling to select match candidates across frames, and a series of cross- and self-attention layers to generate a matching distribution.

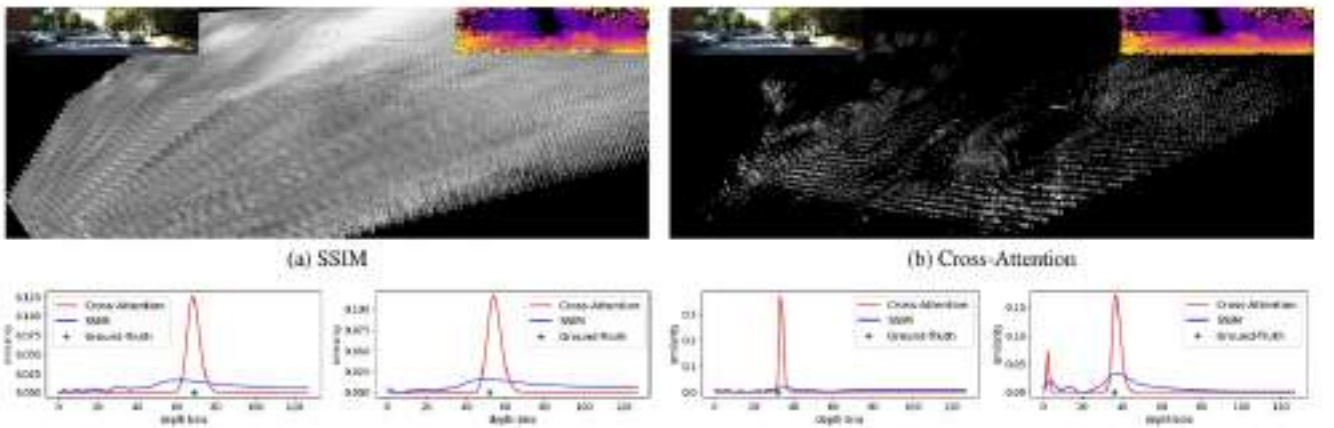


Figure 15. Our learned similarity metric leads to much sharper distributions for feature matching.

The resulting cost volume is then decoded into depth estimates in two ways: (a) directly, using the most probable candidate (i.e., highest similarity) as the predicted depth, and (b) in combination with single-frame features from a traditional depth network [9], thus creating a “safety net” for the settings where multi-frame estimation fails. These depth maps are optimized jointly using the self-supervised

photometric loss, alongside a learned pose network to estimate motion between frames. As we show [10], this feature refinement module leads to state-of-the-art results on the KITTI and DDAD datasets, outperforming other multi-frame depth estimation methods, and is even competitive with highly specialized single-frame supervised architectures. Finally, we show that our learned feature matching module can be directly repurposed across datasets, which greatly decreases training time and cost.

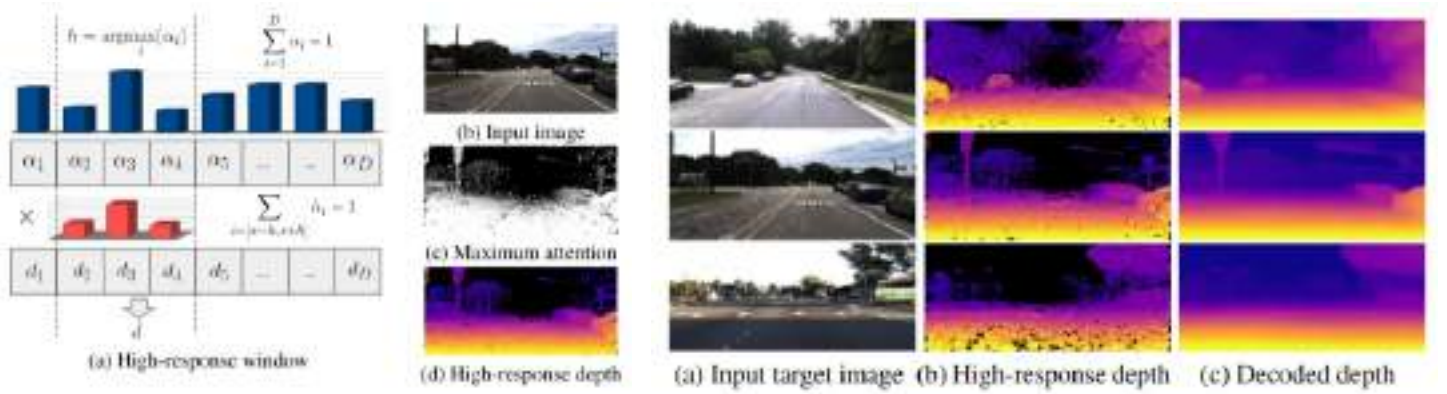


Figure 16. Estimated high-response and multi-level decoded depth maps from DepthFormer.

Conclusion

In this post we discussed how recent research breakthroughs in self-supervised monocular depth estimation have enabled it to move beyond academic settings, towards becoming useful tools for real-world applications. With a strong foundation in geometric principles, we can eliminate the need for *hard* labels in favor of *soft* constraints, imposed either during the loss calculation or in the network architecture itself. As examples of these breakthroughs, we have shown how to (a) explicitly model dynamic objects via geometric multi-task learning, (b) generate scale-aware and consistent point clouds from multiple cameras, (c) calibrate arbitrary cameras with sub-pixel reprojection error from raw videos, and (d) leverage multi-frame information from a single camera during inference. We will be presenting these papers and a few more in the upcoming [ICRA](#) and [CVPR](#) conferences, so if you have the chance please come and talk to us!

References

[1] Gordon, Ariel, Hanhan Li, Rico Jonschkowski, Anelia Angelova. [Depth from Videos in the Wild: Unsupervised Monocular Depth Learning from Unknown Cameras.](#)

CVPR 2019.

[2] Vasiljevic, Igor, Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Wolfram Burgard, Greg Shakhnarovich, and Adrien Gaidon. **Neural Ray Surfaces for Self-Supervised Learning of Depth and Ego-Motion.** 3DV 2020.

[3] Fang, Jiading, Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Greg Shakhnarovich, Adrien Gaidon, and Matthew R. Walter. **Self-Supervised Camera Self-Calibration from Video.** ICRA 2022.

[4] Guizilini, Vitor, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, Adrien Gaidon. **Full Surround Monodepth from Multiple Cameras.** RA-L 2022.

[5] C. Geyer, K. Daniilidis. **A Unifying Theory for Central Panoramic Systems and Practical Implications.** ECCV 2000.

[6] Usenko, Vladyslav C., N. Demmel, Daniel Cremers. **The Double Sphere Camera Model.** 3DV 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. **Attention is All You Need.** NeurIPS 2017.

[8] Zachary Teed, Jia Deng. **RAFT: Recurrent All Pairs Field Transforms for Optical Flow.** ECCV 2020.

[9] Clément Godard, Oisin Mac Aodha, Michael Firman and Gabriel J. Brostow. **Digging into Self-Supervised Monocular Depth Prediction.** ICCV 2019.

[10] Vitor Guizilini, Rares Ambrus, Dian Chen, Sergey Zakharov, Adrien Gaidon. **Multi-Frame Self-Supervised Depth with Transformers.** CVPR 2022.

Computer Vision

Deep Learning

Machine Learning

Self Supervised Learning

Artificial Intelligence
