CoMoDA : Continuous Monocular Depth Adaptation Using Past Experiences Supplementary Materials

Yevhen Kuznietsov[1] Marc Proesmans[1] Luc Van Gool[1,2]
[1]KU Leuven/ESAT-PSI [2]ETH Zurich/CVL
{yevhen.kuznietsov, marc.proesmans, luc.vangool}@esat.kuleuven.be

Contents

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[2] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8001–8008, 2019.

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[6] Clement Godard, Mac Aodha Oisin, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[8] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[12] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.

[13] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.

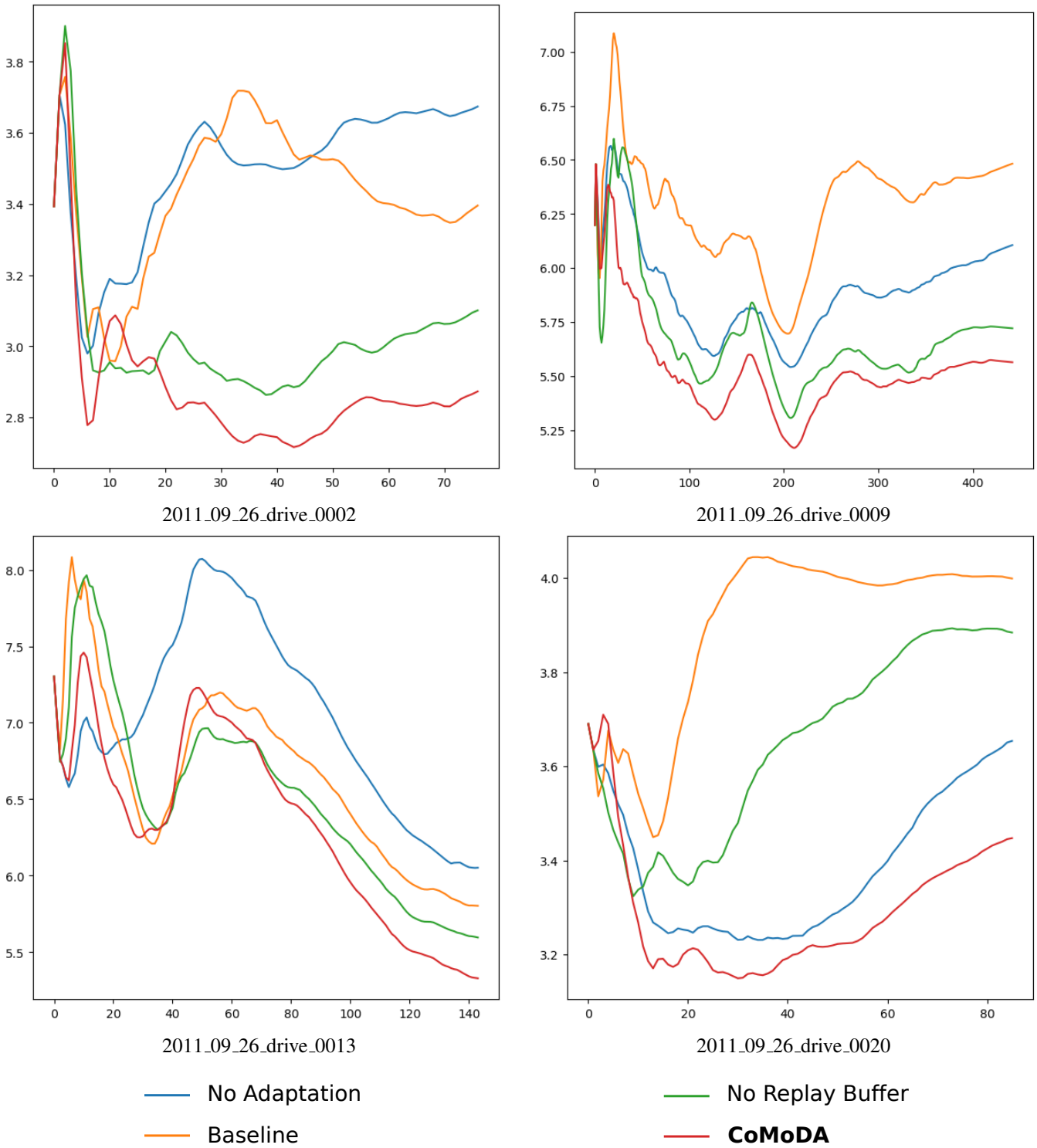
[14] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

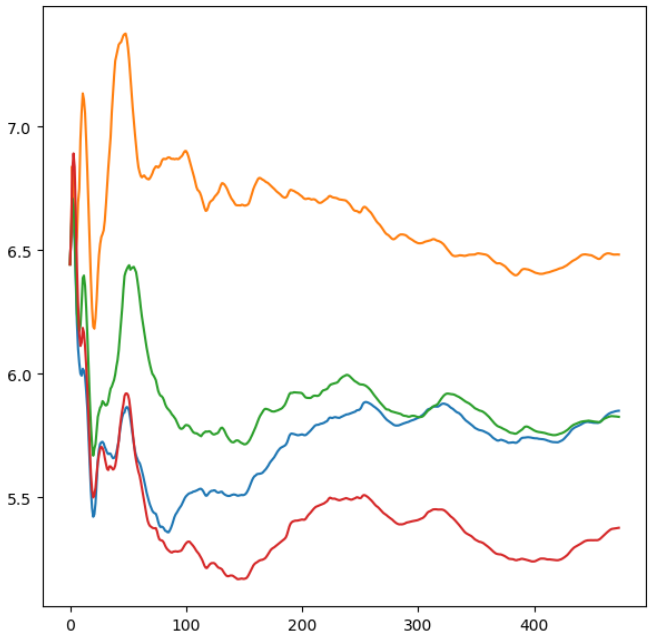
[15] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

1. Additional Results for Intra-Dataset Adaptation

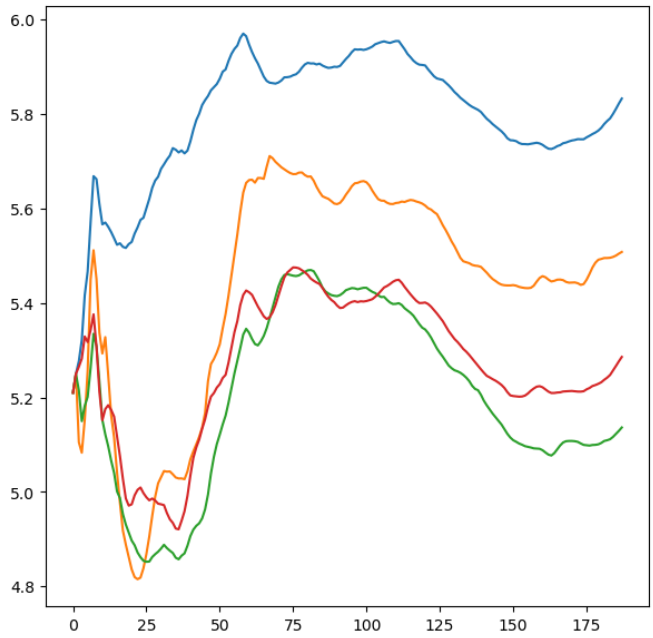
1.1. Error Evolution

We show running mean RMSE (vertical axis) for various design choices on all KITTI [4] test videos. As a reminder, our non-adapted model is Monodepth2 [6] trained with velocity supervision [8]. Horizontal axis represents video frames.

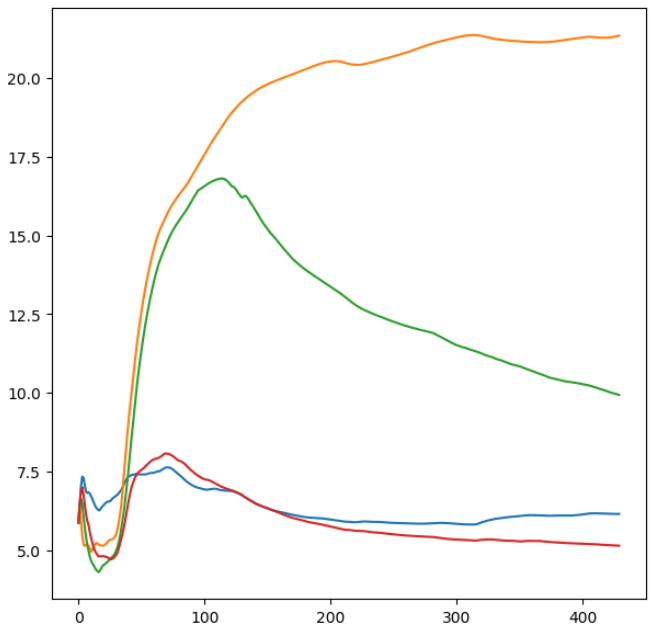




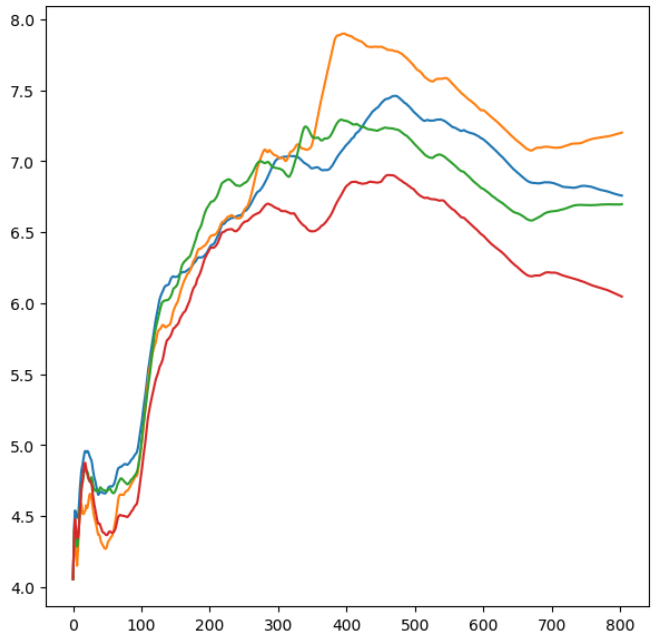
2011_09_26_drive_0023



2011_09_26_drive_0027



2011_09_26_drive_0029



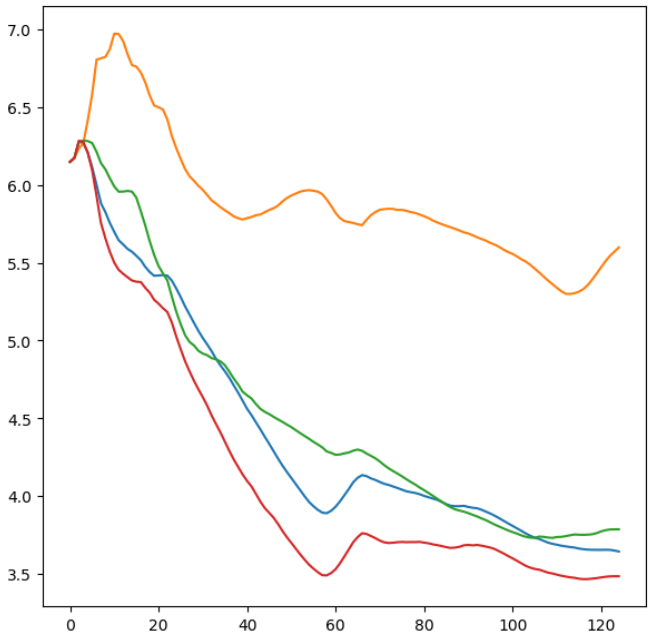
2011_09_26_drive_0036

No Adaptation

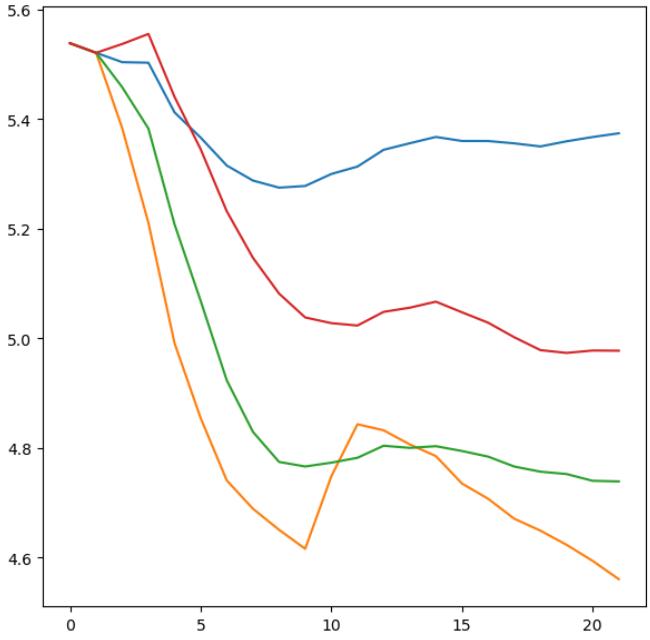
Baseline

No Replay Buffer

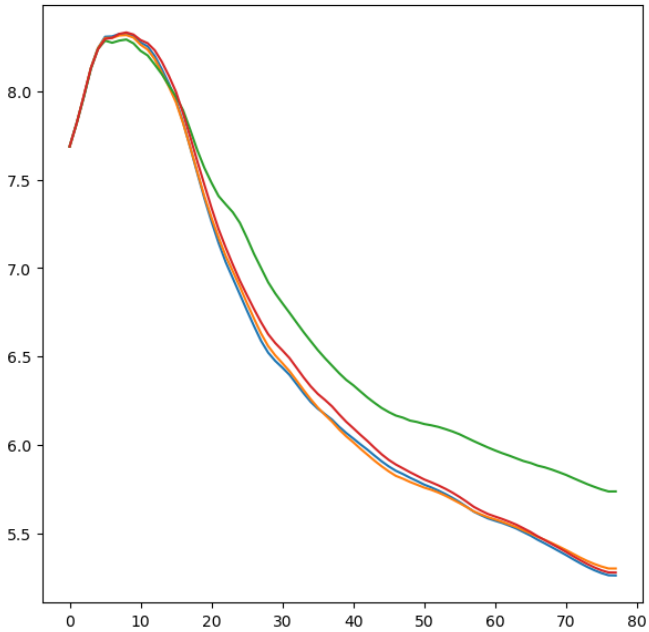
CoMoDA



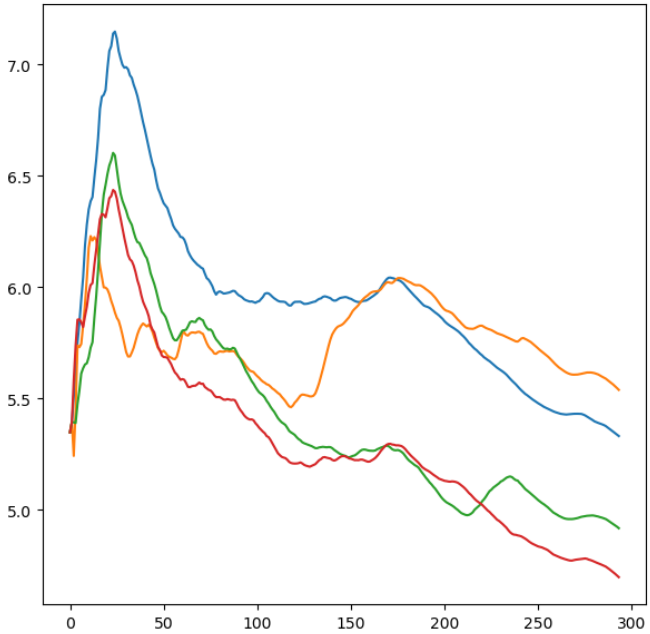
2011_09_26_drive_0046



2011_09_26_drive_0048



2011_09_26_drive_0052



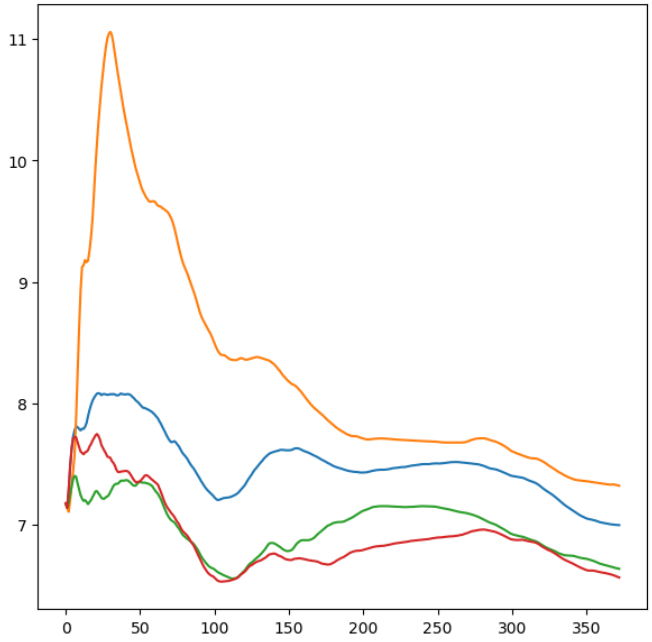
2011_09_26_drive_0056

No Adaptation

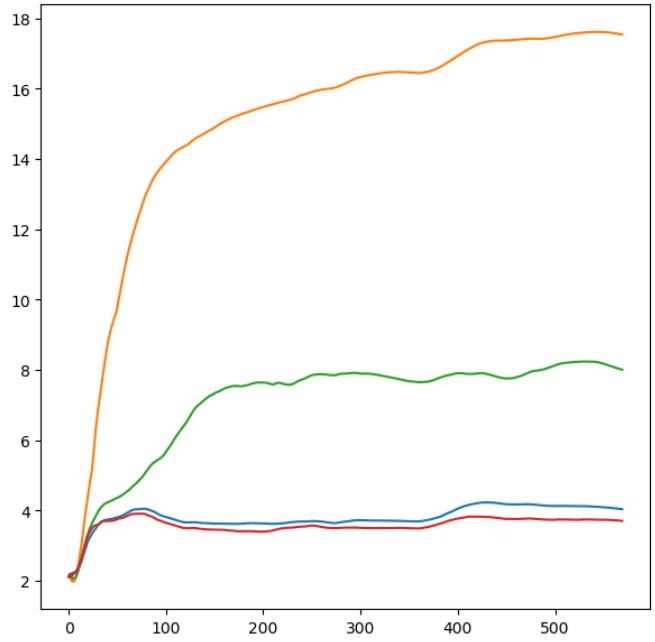
Baseline

No Replay Buffer

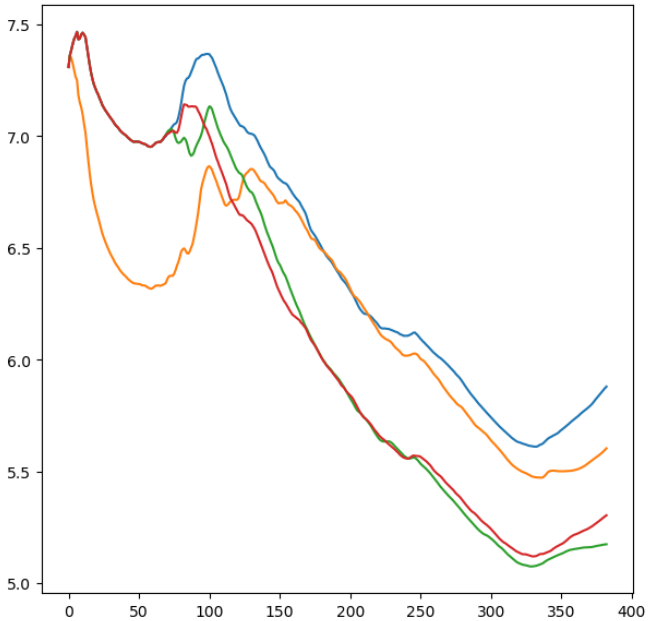
CoMoDA



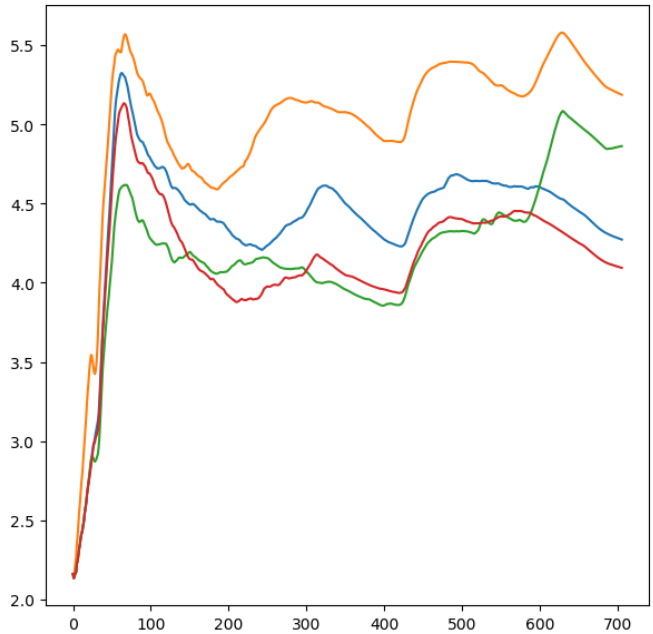
2011_09_26_drive_0059



2011_09_26_drive_0064



2011_09_26_drive_0084



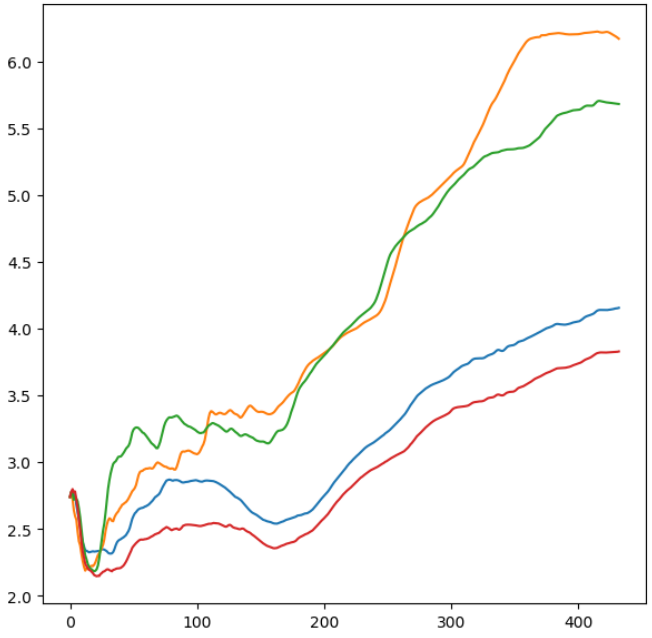
2011_09_26_drive_0086

No Adaptation

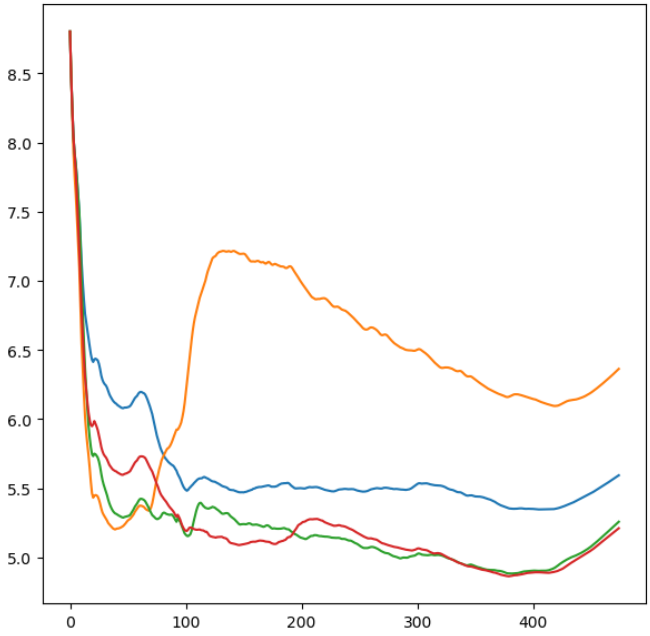
Baseline

No Replay Buffer

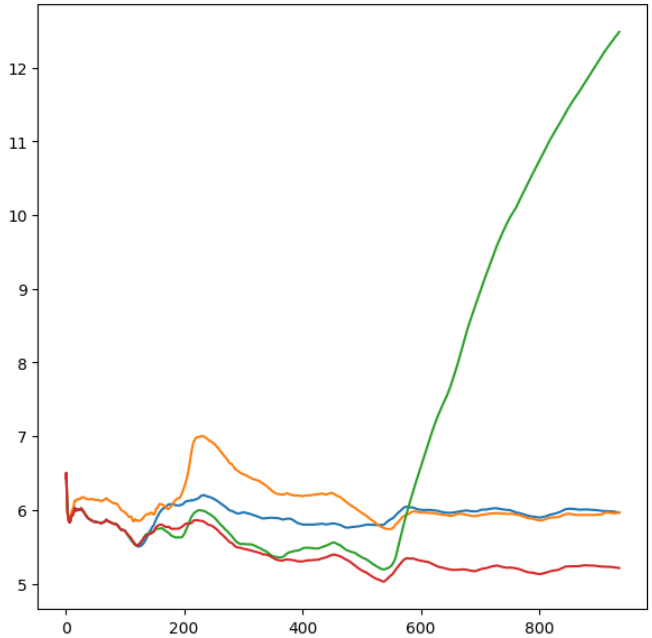
CoMoDA



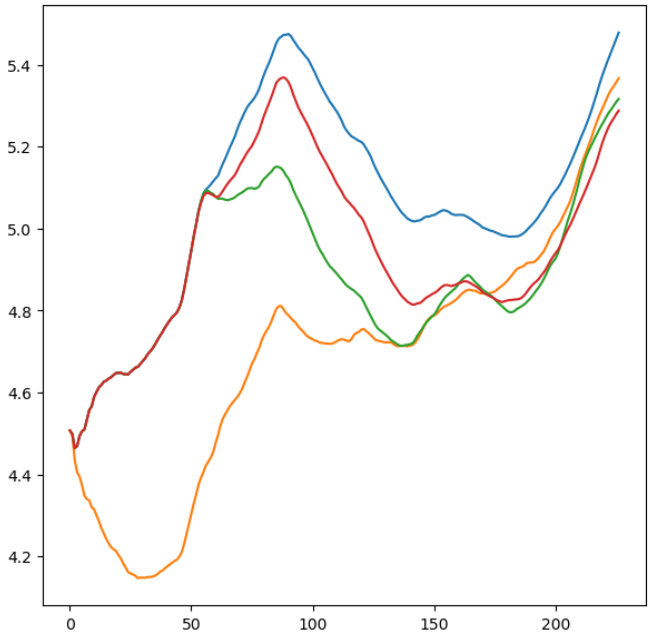
2011_09_26_drive_0093



2011_09_26_drive_0096



2011_09_26_drive_0101



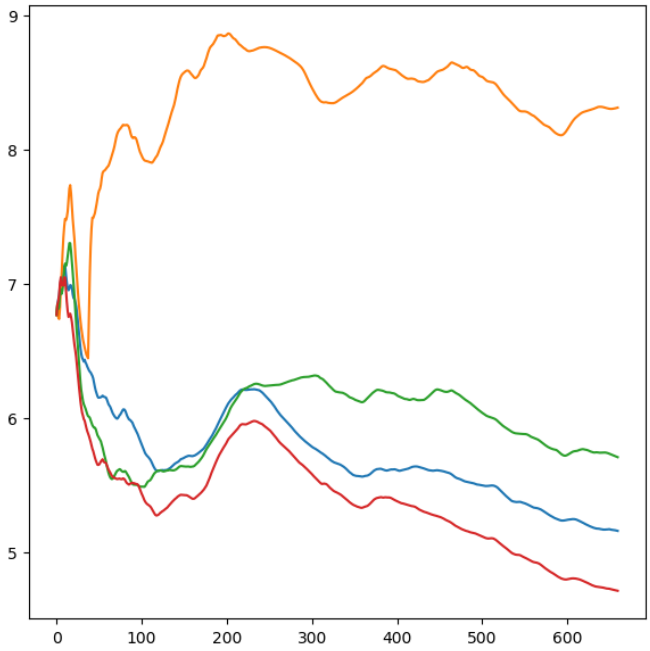
2011_09_26_drive_0106

No Adaptation

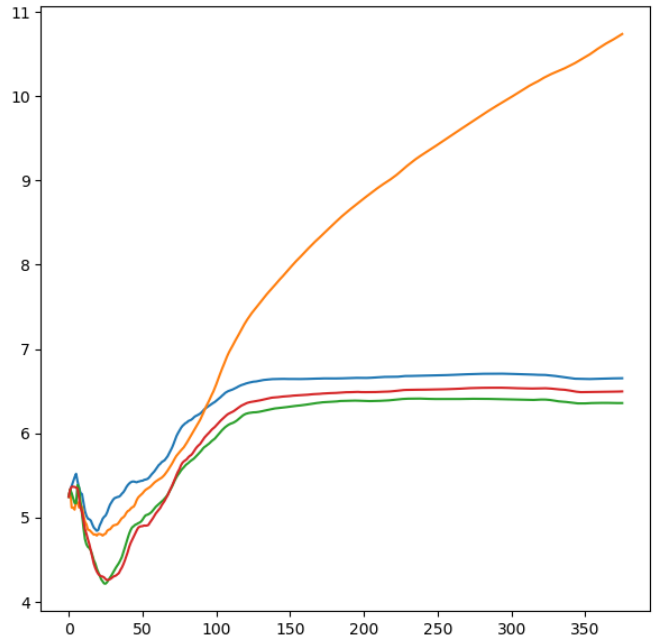
Baseline

No Replay Buffer

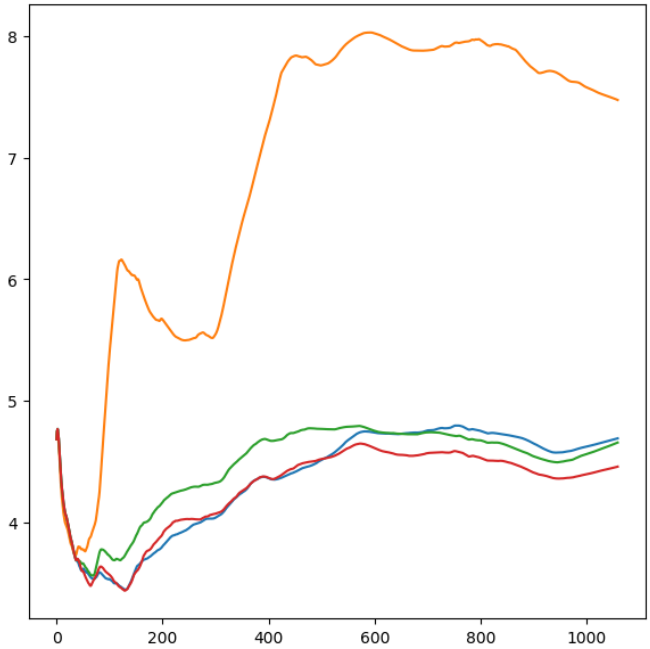
CoMoDA



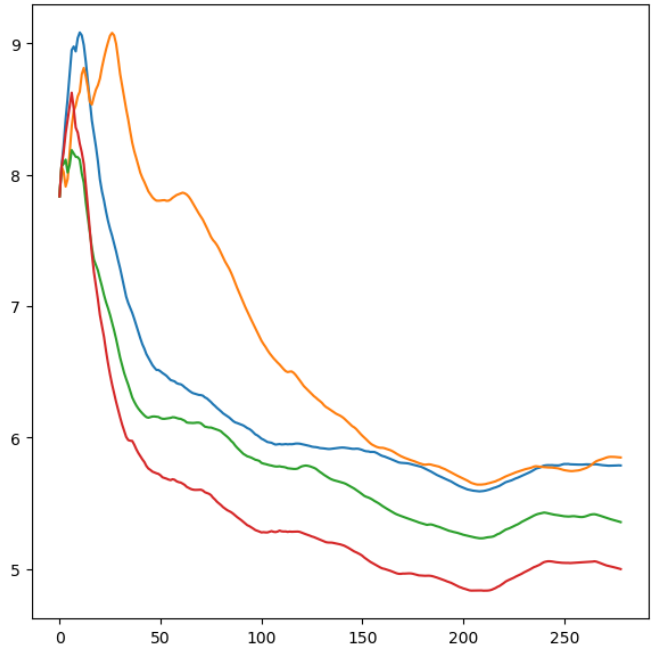
2011_09_26_drive_0117



2011_09_28_drive_0002



2011_09_29_drive_0071



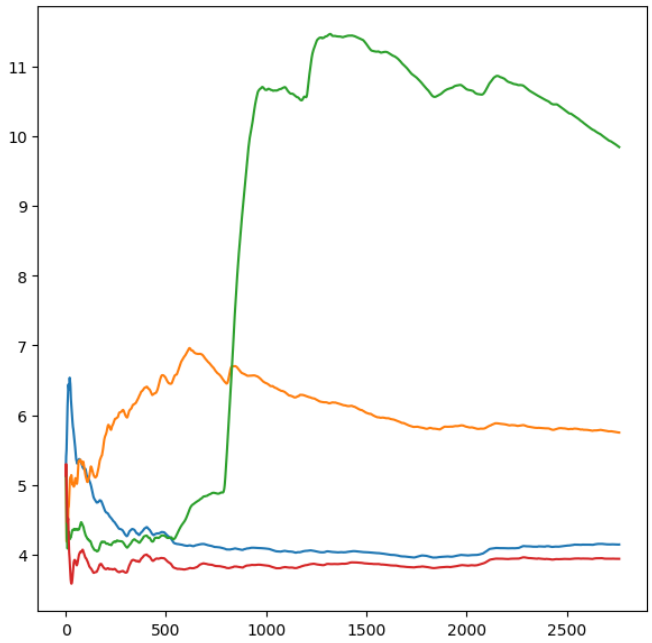
2011_09_30_drive_0016

No Adaptation

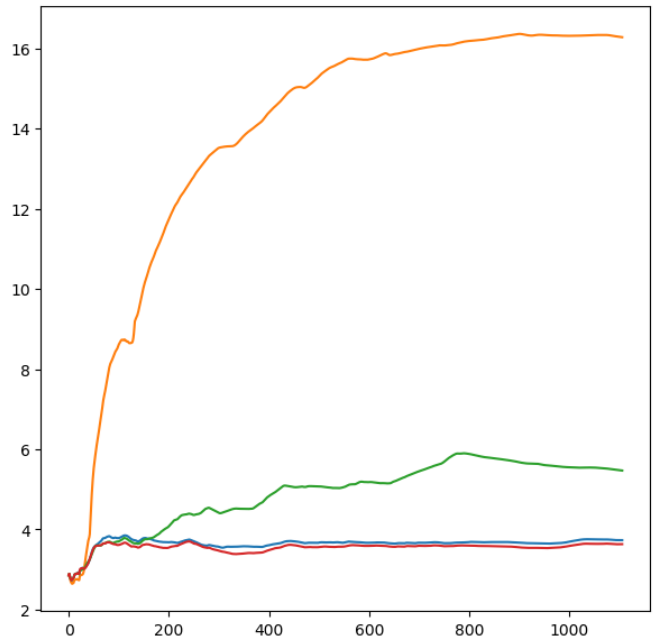
Baseline

No Replay Buffer

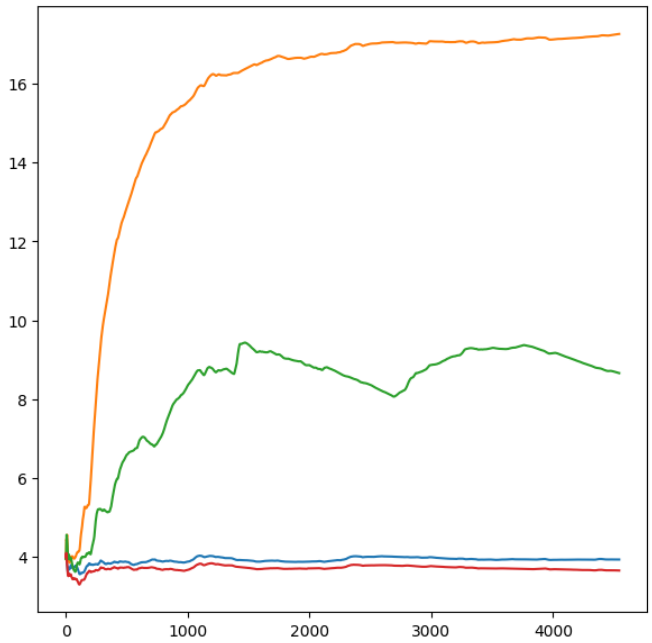
CoMoDA



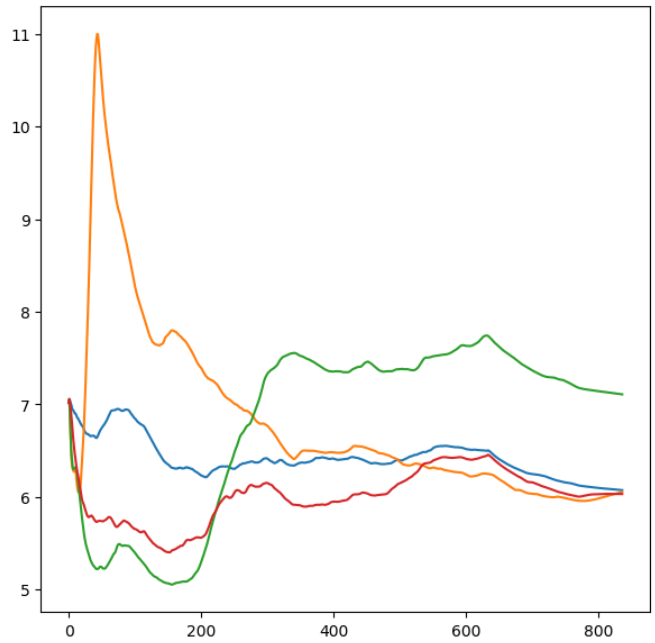
2011_09_30_drive_0018



2011_09_30_drive_0027



2011_10_03_drive_0027



2011_10_03_drive_0047

No Adaptation

Baseline

No Replay Buffer

CoMoDA

1.2. Extra Qualitative Comparison

Due to space limitations, we omitted some methods from the qualitative comparison in the paper. Figure 1 shows the predictions of more methods with code.

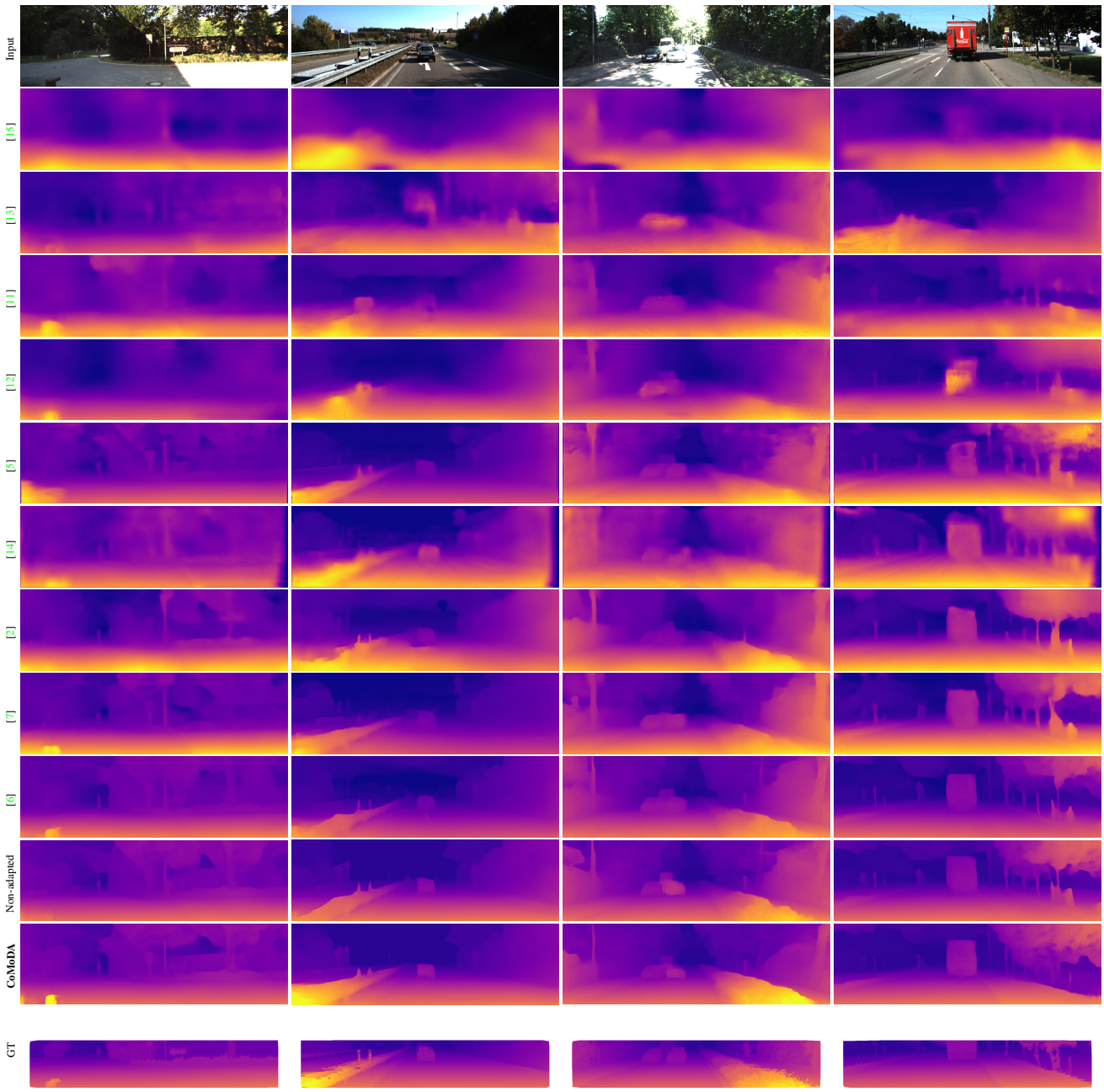
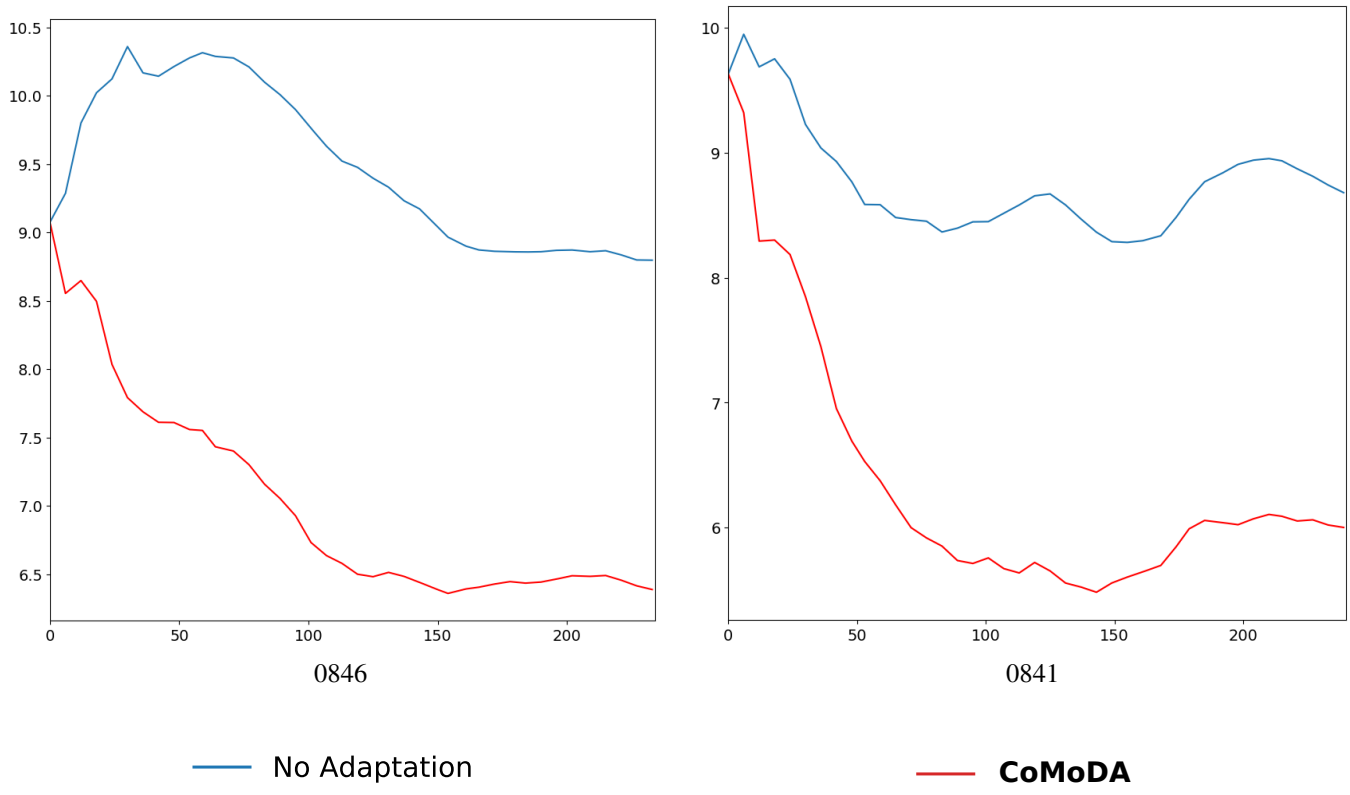


Figure 1. Qualitative results of self-supervised methods on Eigen test split [3] of KITTI [4]. GT stands for the interpolated depth ground truth. Note that our non-adapted model is Monodepth2 [6] with added velocity supervision. Just like in the paper, only CoMoDA captures the ground altitude change on the right of the right-most image (near the sidewalk).

2. Additional Results for Cross-Dataset Adaptation

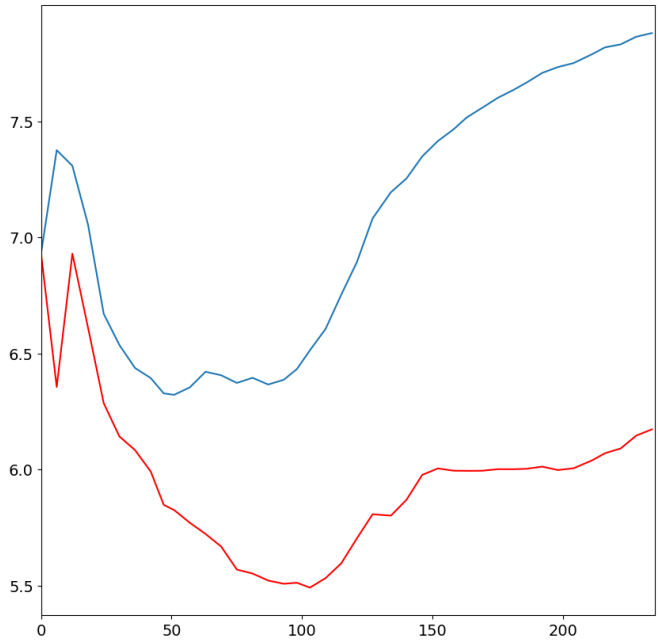
2.1. Error Evolution

As already stated in the paper, we expect further cross-dataset adaptation performance improvement, given the availability of longer videos to adapt on. To support our expectation, we provide the running mean RMSE (vertical axis) visualizations for more NuScenes [1] test videos. In most videos, we observe that the RMSE drop (from the non-adapted model trained on KITTI [4] to CoMoDA) continues to increase towards the end of those videos. Horizontal axis represents video frames.

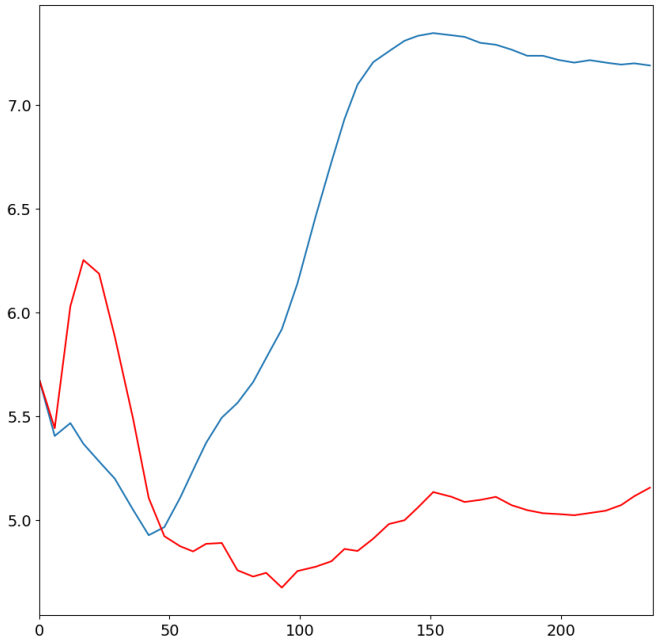


0846 0841

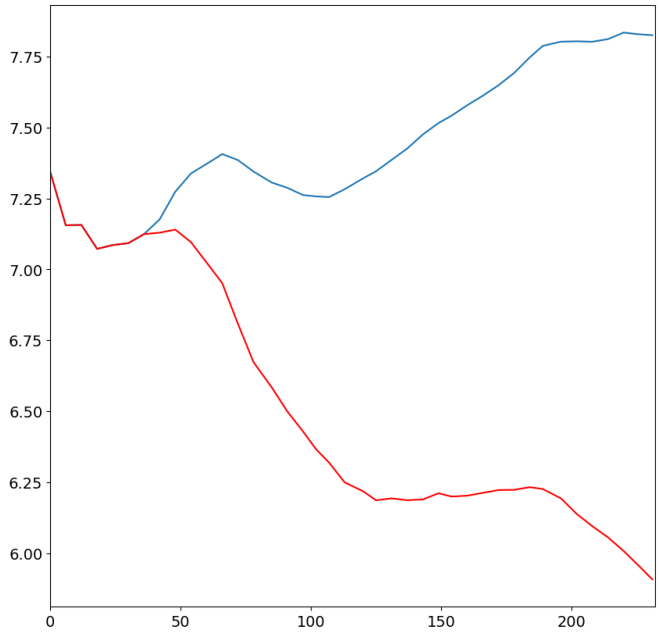
No Adaptation **CoMoDA**



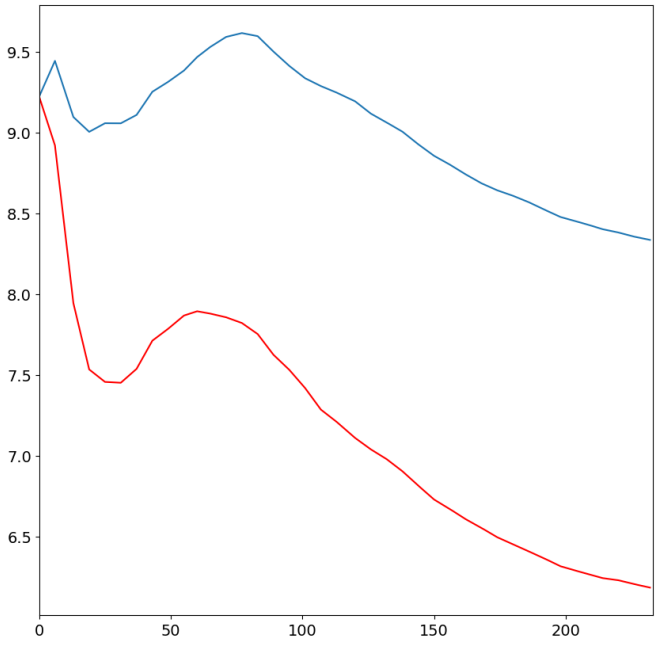
0551



0549



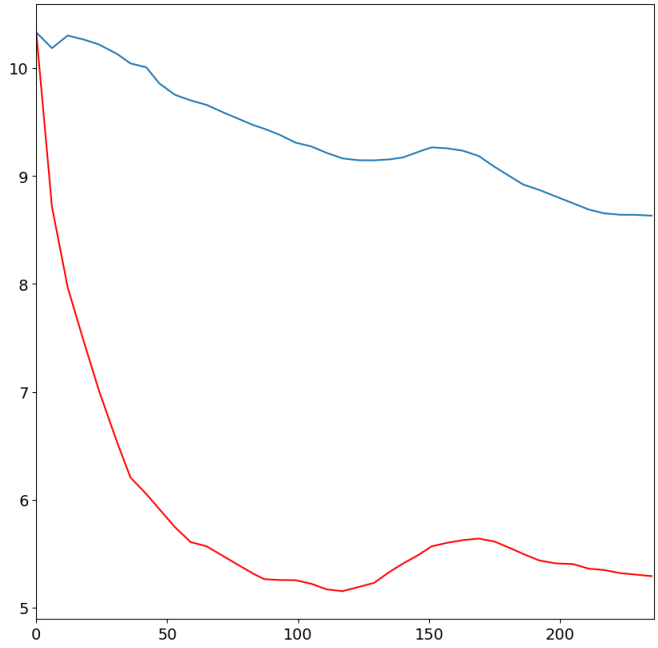
0495



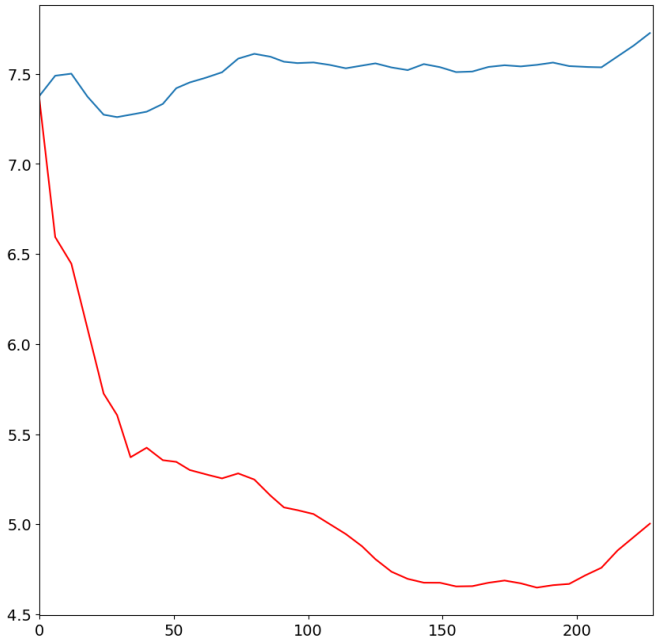
0494

No Adaptation

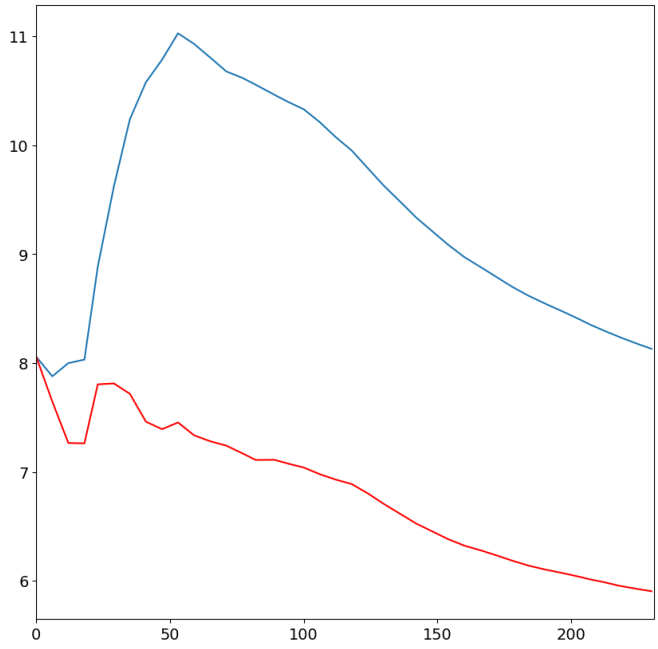
CoMoDA



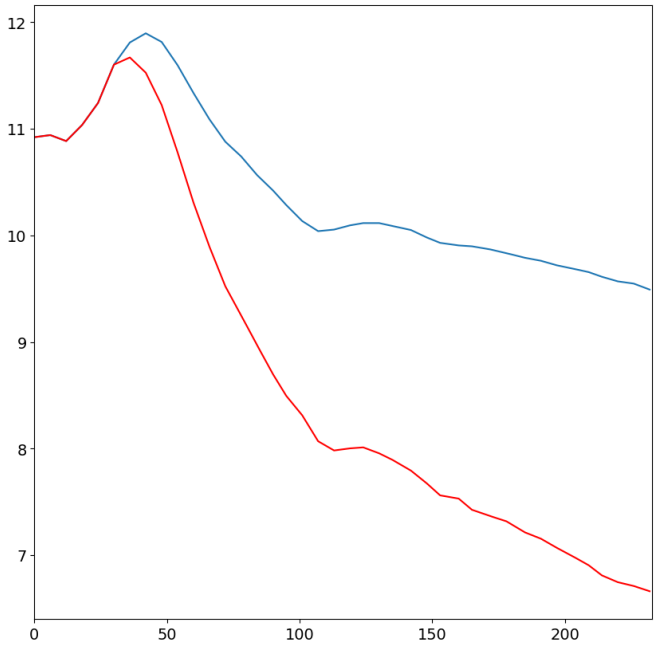
0487



0342



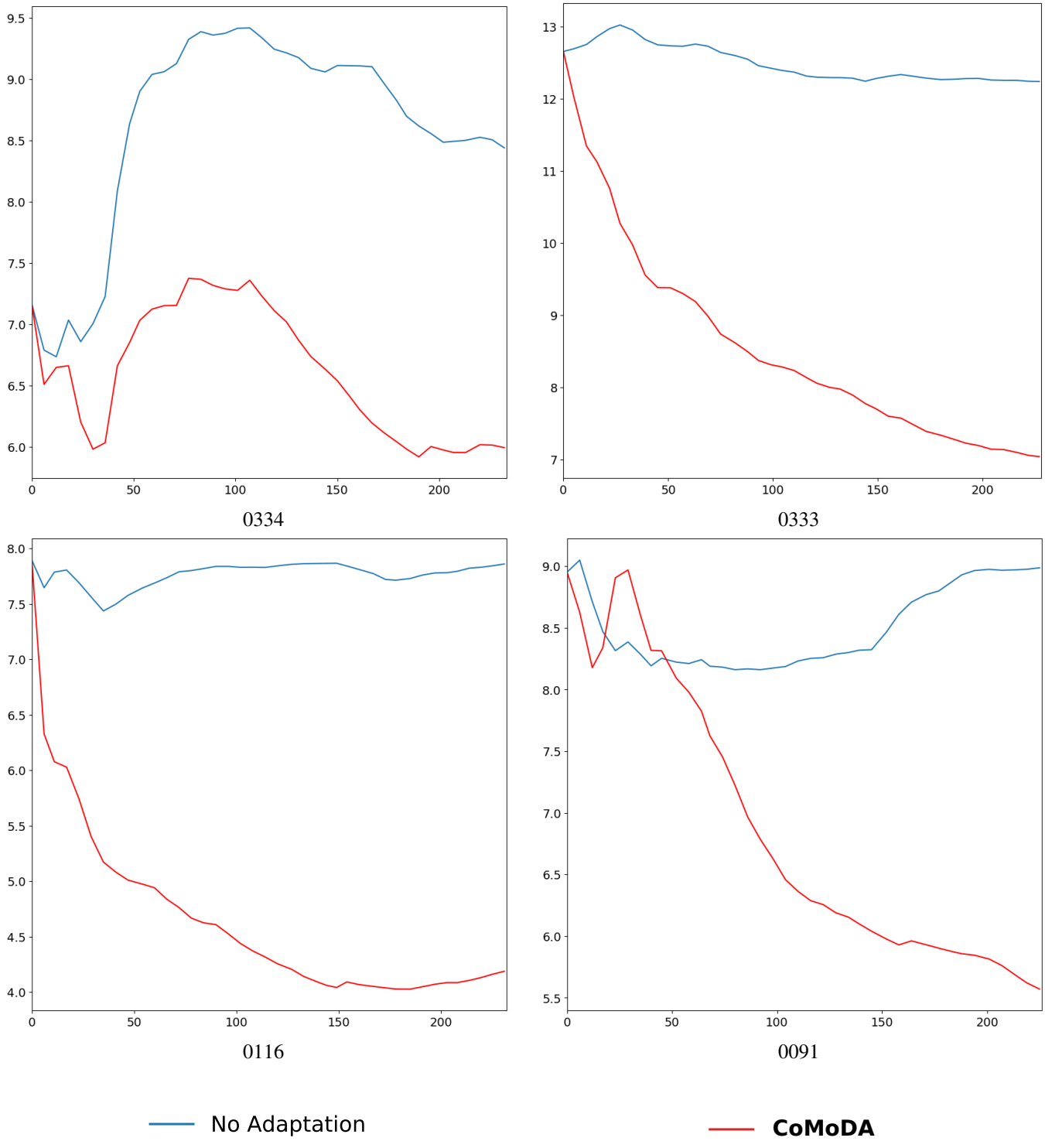
0340

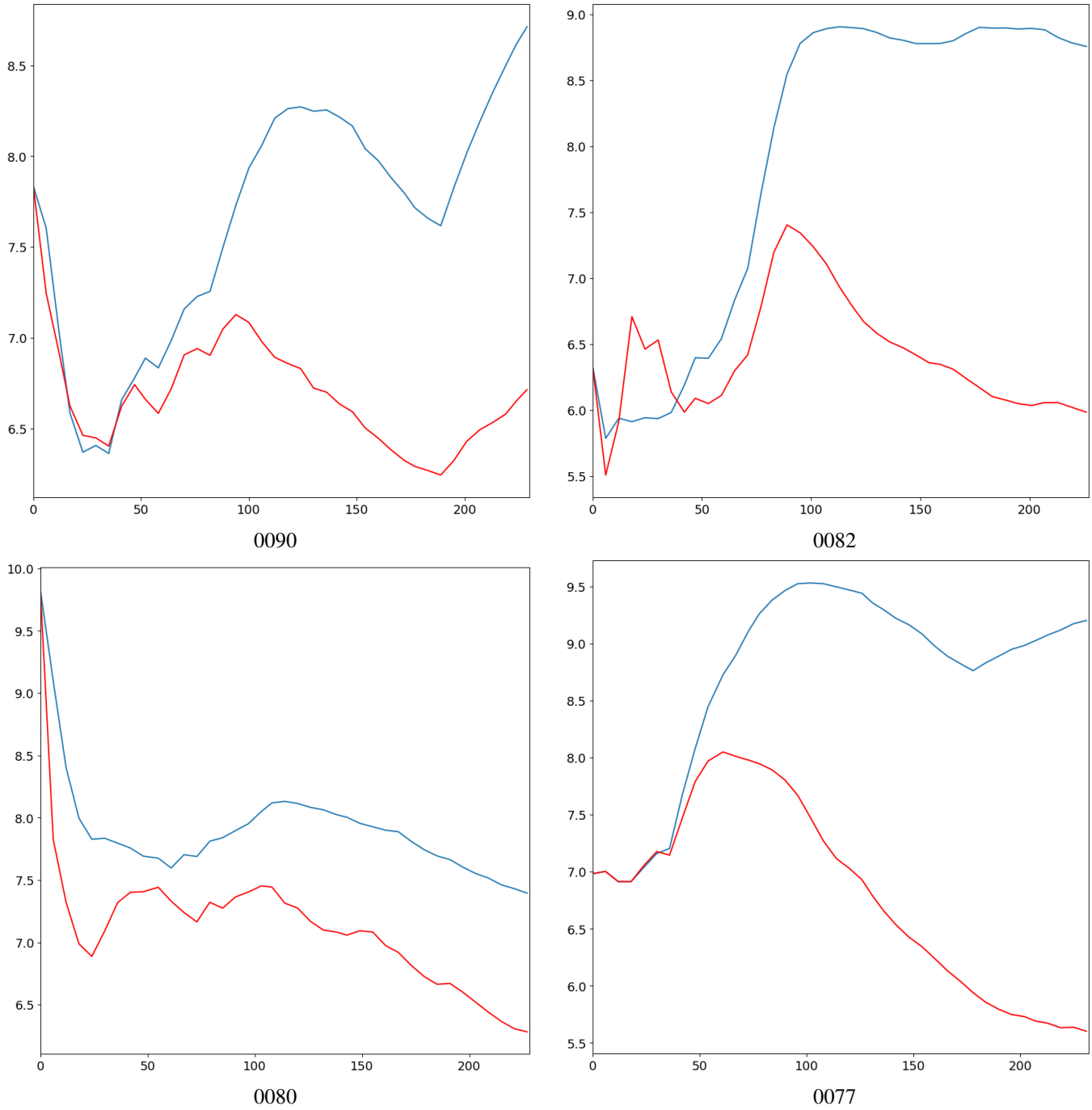


0336

No Adaptation

CoMoDA


0334
0333
0116
0091
No Adaptation
CoMoDA


0090
0082
0080
0077
No Adaptation
CoMoDA

2.2. "Failure" case

Similarly to the relative $\delta < 1.25$ improvement shown in the paper, we visualize the relative RMSE improvement for NuScenes [1] test videos (Fig. 2). These metrics are formalized by Eq. 1:

$$RI_{RMSE}(s) = \left(\frac{RMSE(s)}{RMSE^*(s)} - 1\right) \cdot 100\%, \qquad RI_{\delta}(s) = \left(\frac{\delta^*(s)}{\delta(s)} - 1\right) \cdot 100\%, \tag{1}$$

where $RMSE(s)$ and $\delta(s)$ denote the mean RMSE and $\delta < 1.25$ of the non-adapted model evaluated on video s, and * indicates the use of CoMoDA.

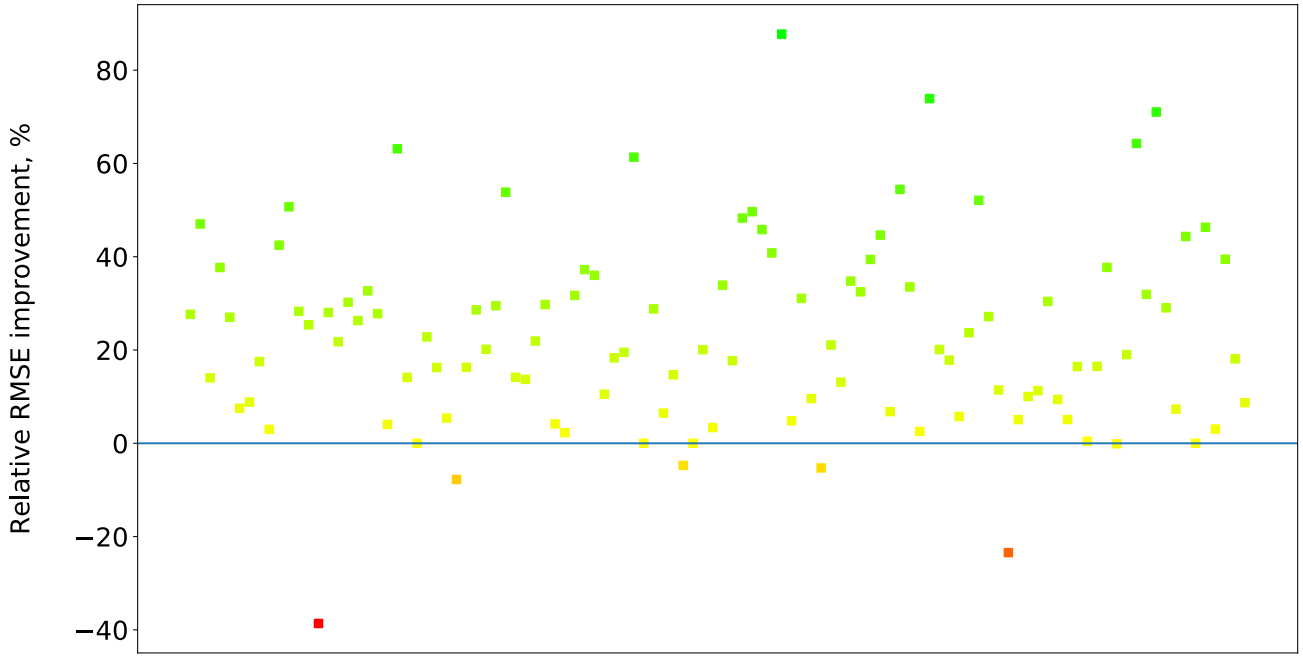


Figure 2. Relative RMSE improvements of our method compared to the non-adapted model (Monodepth2 [6] with velocity supervision [8]) trained on KITTI. Horizontal axis represents test video IDs.

While CoMoDA shows substantial RMSE improvement for most videos, our method demonstrates the noticeably worse RMSE (red dots) for videos 0112 and 0084.

Both videos 0112 and 0084 contain many frames with such objects as metal grid fences. Fig. 3 shows the example of the frame with a fence from video 0112 and the errors produced by our method. The area of interest is highlighted with the white box on top of the input image.

As shown, the LIDAR manages to capture some fine elements of the fence. On the other hand, these elements are not visible in the images, even in full resolution. This explains the difference between the predictions of our appearance-based method and the LIDAR measurements, and at the same time makes the huge errors in this area irrelevant for the assessment of our method.

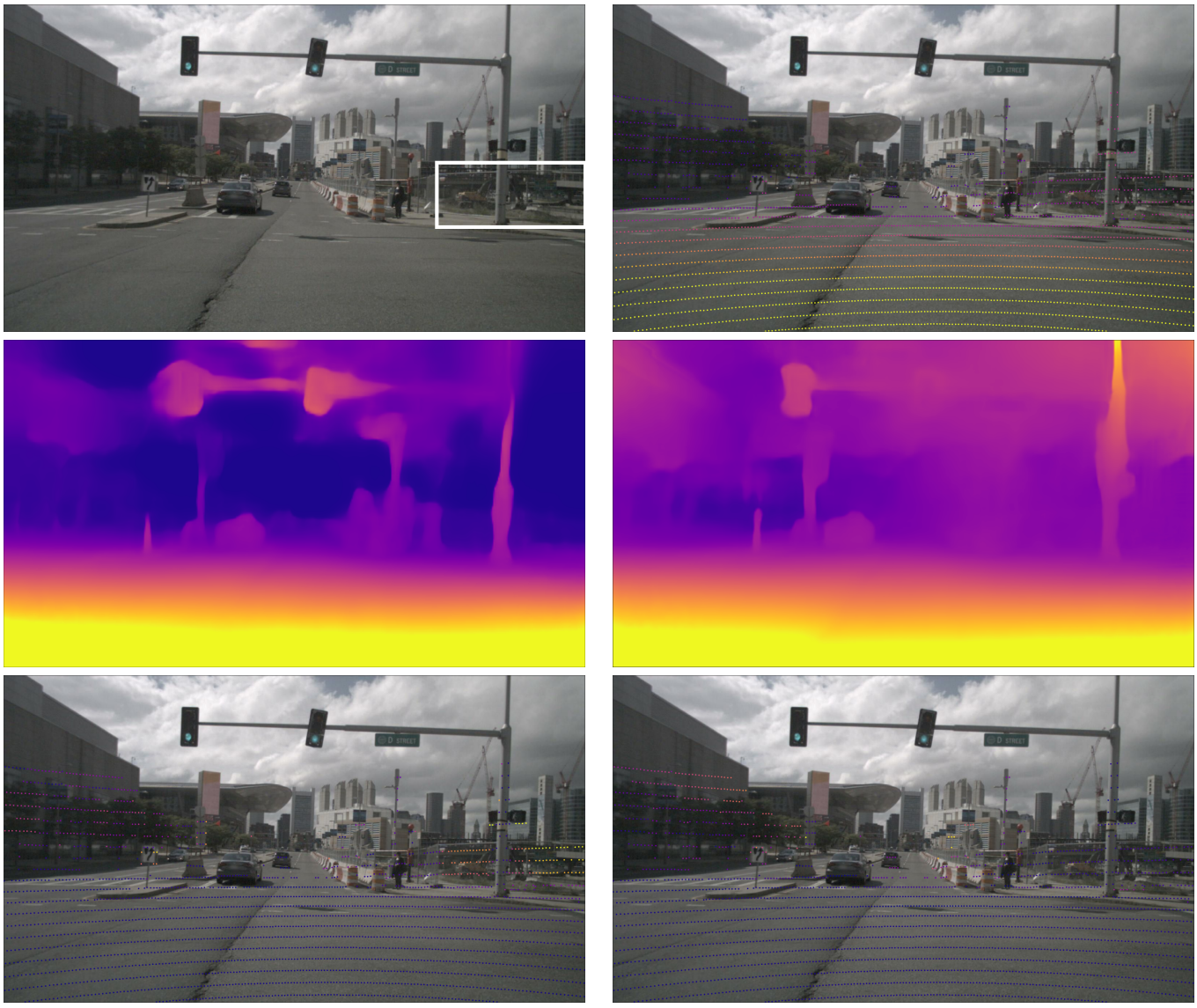


Figure 3. ”Failure” case example. Top row shows the input image and the LIDAR ground truth overlay. Middle row shows the predictions produced by CoMoDA (left) and the non-adapted model (right). Bottom row shows depth errors overlayed with the input (left – for the errors of our method, right – for the non-adapted model respectively).

3. Hyperparameters and other settings

- Depth network: ResNet18 [9] -based encoder-decoder, same as in [6]
- Pose network: ResNet18 with 3 input frames, followed by several convolutions, same as "separate ResNet" in [6]
- KITTI image size: 192 x 640
- NuScenes image size: 256 x 480
- Smoothness term weight: 1e-3
- SSIM α: 0.85
- Optimizer: Adam [10]
- Learning rate: 1e-4
- Number of samples drawn from the replay buffer: 3
- Velocity supervision term weight: 0.005
- Minimum translation to adapt the model: 0.2m
- Number of scales for the image reconstruction loss: 4 (1, 0.5, 0.25, 0.125)
- Data augmentation during pretraining: brightness (0.8, 1.2), contrast (0.8, 1.2), saturation (0.8, 1.2), hue (-0.1, 0.1), horizontal flips with probability 0.5
- Batch size during pretraining: 12
- Number of epochs for pretraining: 50