# Adaptive Monocular Depth Estimation with Masked Image Consistency

**Damian Sójka** [1 2] **Marc Masana** [3] **Bartłomiej Twardowski** [4 5 6] **Sebastian Cygert** [7 8]

## Abstract

Current Continual Test-Time Adaptation methods for Monocular Depth Estimation rely on extra data and lack efficiency, using auxiliary source models or adjacent video frames, which increase computational demand. We propose to use masked image modeling, extending Masked Image Consistency, to address these limitations. Together with the use of scale alignment to account for varying camera setups, our proposed approach enforces consistency between masked and unmasked image predictions, which shows empirical results that highlight its effectiveness in autonomous driving scenarios, achieving performance comparable with state-of-the-art.

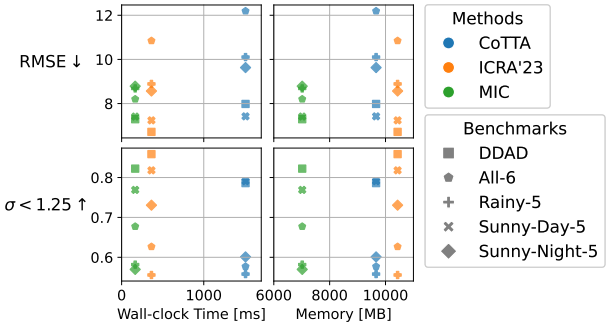


Figure 1: The simple Masked Image Consistency (MIC) technique performs on par with state-of-the-art approaches, with lower computational demand and without the need for additional source models or adjacent video frames.

## 1. Introduction

Accurate 3D perception is essential for robotics and autonomous driving applications. Monocular Depth Estimation (MDE) enables this by solving the ill-posed problem of estimating depth from a single RGB image, which is challenging due to the infinitely many 3D scenes that can project onto the same 2D image (Lee et al., 2019).

Some existing deep learning techniques tackle this issue (Godard et al., 2019; Yuan et al., 2022). However, in general, neural network-based approaches struggle in open real-world environments, where the independent and identically distributed assumption from training often fails, causing significant performance drops or model failure. Out-of-distribution data arises from factors like varying locations, weather, or camera setups. Continual Test-Time Adaptation (CTTA) (Wang et al., 2022) adapts a pre-trained model to continually shifting distributions of unlabeled data during test-time without access to the source data.

The field of CTTA primarily focuses on classification-based tasks, which differ fundamentally from depth estimation due to the latter's regression-based nature. Classic techniques (Wang et al., 2021; Niu et al., 2022; 2023; Yuan et al., 2023; Döbler et al., 2022) rely on categorical network outputs and are not directly applicable to depth estimation. Few CTTA approaches address MDE of self-supervised models, often referred to as test-time refinement. These methods typically involve continuous finetuning of two networks (for depth and ego-motion prediction) using photometric consistency across adjacent video frames (Casser et al., 2019; Shu et al., 2020; McCraith et al., 2020; Kuznietsov et al., 2021). Such models either predict only relative depth or incorporate additional data (*e.g.*, ego-motion velocity) during adaptation to achieve scale awareness (Kuznietsov et al., 2021).

To the best of our knowledge, CTTA methods adapting supervised scale-aware MDE models are scarce, with only two notable exceptions: (Li et al., 2023) (referred hereon as ICRA'23) and SVDP (Yang et al., 2024). ICRA'23 simultaneously adapts self-supervised and supervised models on adjacent video frames while preserving scale-awareness of the supervised model. However, it requires two separately trained neural networks (when counting self-supervised depth and ego-motion prediction models as one), to concurrently process multiple images, and to employ copies

[1]IDEAS NCBR [2]Poznan University of Technology [3]Graz University of Technology [4]IDEAS [5]Autonomous University of Barcelona [6]Computer Vision Center [7]NASK - National Research Institute [8]Gdańsk University of Technology. Correspondence to: Damian Sójka <damian.sojka@doctorate.put.poznan.pl>.

*2nd Workshop on Test-Time Adaptation: Putting Updates to the Test! (PUT) at $42^{nd}$ International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. 

of the supervised model for regularization. These factors make it computationally intensive and less practical compared to simpler CTTA methods for classification, such as TENT (Wang et al., 2021). In the SVDP paper, experiments include MDE, but the authors did not provide the depth estimation code, the setup description was unclear, and we could not reproduce their results as presented.

The literature highlights the significant role of global image context in depth estimation models (Zhao et al., 2022; Ibrahem et al., 2022; Rajapaksha et al., 2024). This context enables models to better understand scene layouts and capture inter-object relationships. To leverage this insight, we employ Masked Image Modeling, a powerful self-supervised learning technique that promotes the use of the entire image context to predict masked image patches. Specifically, we choose the Masked Image Consistency (MIC) approach, which enforces consistency between model predictions on masked and unmasked versions of an image, promoting an effective use of the global image context. Originally developed for Unsupervised Domain Adaptation (Hoyer et al., 2023), significantly different from CTTA, MIC has primarily been evaluated on classification tasks.

In this work, we extend MIC to CTTA for the MDE task. **We believe that we are among the first to present the simple and effective, similar to entropy minimization (Wang et al., 2021) for classification, CTTA approach for supervised MDE models.** The presented approach enhances the use of global image context, while adapting to continuously shifting domains, significantly improving performance over the source model without relying on secondary models or adjacent video frames. We experimentally show that our method achieves performance comparable to state-of-the-art approaches (ICRA'23) with lower computational overhead. Additionally, we adopt a classic CTTA method (CoTTA (Wang et al., 2022)) for MDE and show that the masked consistency used in MIC significantly outperforms the augmentation consistency technique from CoTTA.

## 2. Method

### 2.1. Preliminaries

CTTA aims to adapt the neural network model $f_\theta$ pre-trained on the labeled source data $(\mathcal{X}^S, \mathcal{Y}^S)$ to the stream of unlabeled data with continually changing data distributions $\mathcal{X}^{T_1}, \mathcal{X}^{T_2}, \ldots, \mathcal{X}^{T_n}$ at test time. There is no access to the source data during adaptation, and the test data has to be processed on-the-fly, without the possibility of revisiting it.

### 2.2. Masked Image Consistency (MIC)

We adapt Masked Image Context (MIC) (Hoyer et al., 2023) (Figure 2), a masked image modeling technique which enhances the model's focus on the entire scene by masking a portion of image patches and training the model to predict depth across the full image.

Following (Hoyer et al., 2023), we generate the random mask $\mathcal{M}$ sampled from a uniform distribution to mask a fraction $r$ of the test input image $x^T$ patches:

$$\mathcal{M}_{\substack{mb+1:(m+1)b, \\ nb+1:(n+1)b}} = [v > r] \quad \text{with } v \sim \mathcal{U}(0,1)\,, \tag{1}$$

$$x^M = \mathcal{M} \odot x^T\,, \tag{2}$$

where $[\cdot]$ indicates the Iverson bracket, $b$ the patch size, and $m, n \in [0, \ldots, W/b–1]$ the patch indices. Then, the network $f_\theta$ predicts the depth $\hat{y}^M$ for the whole masked image $x^M$, using the unmasked context for the masked regions:

$$\hat{y}^M = f_\theta(x^M)\,. \tag{3}$$

We aim to ensure consistency between the depth prediction on the masked image $\hat{y}^M$ and the depth prediction on the original image $\hat{y}^T$. In the absence of ground truth depth data, $\hat{y}^T$ serves as pseudo-labels for adaptation. However, noisy pseudo-labels can impede model improvement. Therefore, we employ an Exponential Moving Average (EMA) teacher model $(g_\phi)$ to generate more reliable pseudo-labels (Tarvainen & Valpola, 2017):

$$\hat{y}^T = g_\phi(x^T)\,. \tag{4}$$

The teacher predicts the depth using the original image $x^T$, therefore it can utilize both local and global context, making the pseudo-labels as accurate as possible. We employ the L1-norm as the consistency loss $\mathcal{L}_t$:

$$\mathcal{L}_t = \|\hat{y}^M - \hat{y}^T\|_1, \tag{5}$$

Finally, the EMA teacher $g_\phi$ is updated using the EMA of student's weights $\theta$ with $\alpha$ as a smoothing factor:

$$\phi_{t+1} = \alpha\phi_t + (1-\alpha)\theta_t\,, \tag{6}$$

where $t$ indicates the adaptation step.

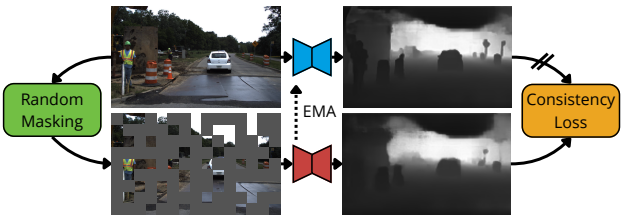


Figure 2: The Masked Image Consistency (MIC) technique.

### 2.3. Scale Alignment

To ensure a fair comparison and account for varying camera setups, we adopt the very effective Scale Alignment

(SA) from ICRA'23 (Li et al., 2023). This component is essential for accurate scale-aware depth predictions on test images, as the MDE network relies solely on object proportions in the images to estimate depth with the correct scale. Given the focal lengths of the source camera ($f^S$) and target camera ($f^T$), along with the camera heights above the ground in an autonomous driving setup for the source ($h^S$) and target ($h^T$), the objective is to resize the target image $x^T(H^T, W^T)$, with height $H^T$ and width $W^T$, based on the ratios of these parameters:

$$\hat{x}^T = x^T(H^T \cdot \frac{f^S}{f^T} \cdot \frac{h^S}{h^T}, W^T \cdot \frac{f^S}{f^T} \cdot \frac{h^S}{h^T}), \quad (7)$$

where $\hat{x}^T$ is the resized target image. **We use this operation for every experiment presented in the paper, including baselines.** It allows to obtain reasonable metric depth predictions without the need for ground truth median scaling.

## 3. Experiments

**Benchmarks.** Following (Li et al., 2023; Kuznietsov et al., 2021), we evaluate on the split of KITTI dataset (Geiger et al., 2013) introduced in (Eigen et al., 2014) as the source data. For benchmarking CTTA, as in (Li et al., 2023), we utilize datasets from an autonomous driving perspective: DDAD (Guizilini et al., 2020) and Waymo (Sun et al., 2020). The DDAD test sequence includes the entire validation split. , consisting of 50 videos with a total of 3,850 frames. For Waymo, we evaluate multiple scenarios: *Sunny-Night-5* includes the first five "sunny" night scenes from the validation set; *Rainy-5* comprises five rainy scenes from the training split (one daytime, three dawn, and one night); *All-6* includes six sequences from the training set, with one sequence for each combination of time of day ("day," "dawn," or "night") and weather ("sunny" or "rain"), selected as the first from each category; *Sunny-Day-5* includes the first five "sunny" day scenes from the evaluation split. Models adapt continuously to each video without reset in between.

**Baselines.** We compare our approach with the state-of-the-art CTTA method for depth estimation, ICRA'23 (Li et al., 2023), which requires an additional self-supervised source model and adjacent video frames for the model update. Additionally, we present the results of a well-known CTTA method, CoTTA (Wang et al., 2022), originally designed for image classification and semantic segmentation, and which we adjust for the depth estimation task.

**Implementation Details.** We conduct experiments with SwinTransformer (Liu et al., 2021) as the encoder and NewCRF as the depth decoder (Yuan et al., 2022). We use the model checkpoint provided by ICRA'23 (Li et al., 2023) as the pre-trained model. Depth predictions are evaluated using the standard metrics described in (Eigen et al., 2014). The parameters for CoTTA and our presented MIC are selected using an oracle method. The parameters for ICRA'23 are adopted directly from the paper, since they use the same benchmarks. We set the mask ratio $r$ to 0.5 and the smoothing factor $\alpha$ to 0.999. The batch size is set to 1.

### 3.1. Analysis and results

**ICRA'23 and MIC achieve comparable performance.** Results for all benchmarks are presented in Table 1. Both methods achieve significant improvement over the Source model in most cases, with ICRA'23 excelling on DDAD and *Sunny-Night-5*, and MIC on *All-6* and *Rainy-5*. MIC achieves the best overall performance across most metrics, despite ICRA'23 requiring an additional self-supervised model and adjacent video frames for photometric consistency, which highlights the effectiveness of our proposed method. However, both ICRA'23 and MIC are outperformed by the Source model on the *Sunny-Day-5* benchmark, which has the weather and time-of-day conditions most similar to the source data, and thus requires the least adaptation. This suggests that these methods might struggle to maintain performance on data with minimal domain shift.

**Masked consistency $>$ augmentation consistency.** The results in Table 1 show that CoTTA is ineffective at adapting the MDE model and is outperformed by the Source model in most cases. This suggests that augmentation consistency, the primary adaptation mechanism of CoTTA, is significantly less effective for depth estimation adaptation compared to masked image consistency (MIC).

**MIC is significantly more efficient.** Table 2 presents the average wall-clock time and memory usage for processing a single batch on an NVIDIA RTX 4080 GPU. Among the methods evaluated, MIC stands out as the most efficient due to its simpler design.

**EMA teacher is a key component.** Table 3 highlights the importance of the EMA teacher for adaptation within the proposed MIC. It shows that the improved accuracy of pseudo-labels provided by the EMA teacher significantly enhances the adaptation process.

**The potential scale drifting problem.** In theory, the scale of depth predictions of a continually adapted model within the MIC framework could drift over time, leading to inaccurate metric depth estimates. We explore this issue using the *All-6* benchmark repeated 20 times (see Fig. 3). We measure RMSE and Median Ratio (ratio of median ground truth depth to median predicted depth). Results show a slight Median Ratio increase which stabilizes over time, while RMSE rises initially but then continuously decreases. Overall, minor depth prediction deviations occur for both MIC and ICRA'23 but do not seem significant, especially given the sequence length and consistent RMSE reduction.

| Benchmark | Method | AbsRel↓ | SqRel↓ | RMSE↓ | RMSElog↓ | $\sigma < 1.25 \uparrow$ | $\sigma < 1.25^2 \uparrow$ | $\sigma < 1.25^3 \uparrow$ |
|---|---|---|---|---|---|---|---|---|
| DDAD | Source | 0.144 | 1.516 | 7.950 | 0.228 | 0.788 | 0.938 | 0.976 |
| | CoTTA | 0.144 | 1.523 | 7.983 | 0.229 | 0.786 | 0.937 | 0.976 |
| | ICRA'23 | **0.117** | **1.187** | **6.716** | **0.185** | **0.859** | **0.960** | **0.985** |
| | **MIC** | 0.133 | 1.319 | 7.290 | 0.207 | 0.822 | 0.951 | 0.983 |
| *All-6* | Source | 0.260 | 4.765 | 12.130 | 0.457 | 0.580 | 0.762 | 0.831 |
| | CoTTA | 0.262 | 4.814 | 12.190 | 0.461 | 0.577 | 0.759 | 0.828 |
| | ICRA'23 | 0.221 | 3.725 | 10.846 | 0.357 | 0.627 | 0.810 | 0.884 |
| | **MIC** | **0.196** | **2.400** | **8.202** | **0.252** | **0.677** | **0.927** | **0.970** |
| *Rainy-5* | Source | 0.244 | 3.439 | 10.021 | 0.374 | 0.601 | 0.815 | 0.890 |
| | CoTTA | 0.248 | 3.481 | 10.105 | 0.384 | 0.558 | 0.811 | 0.887 |
| | ICRA'23 | **0.233** | **2.731** | 8.889 | 0.317 | 0.556 | 0.850 | 0.938 |
| | **MIC** | 0.238 | 2.862 | **8.684** | **0.290** | **0.582** | **0.877** | **0.949** |
| *Sunny-Day-5* | Source | **0.162** | **2.076** | 7.346 | **0.211** | **0.826** | **0.949** | **0.977** |
| | CoTTA | 0.173 | 2.148 | 7.420 | 0.222 | 0.791 | 0.946 | 0.976 |
| | ICRA'23 | 0.168 | 2.179 | **7.239** | 0.212 | 0.818 | 0.947 | 0.976 |
| | **MIC** | 0.185 | 2.271 | 7.410 | 0.229 | 0.769 | 0.941 | 0.975 |
| *Sunny-Night-5* | Source | 0.199 | 2.116 | 9.558 | 0.272 | 0.607 | 0.903 | 0.972 |
| | CoTTA | 0.200 | 2.144 | 9.630 | 0.274 | 0.601 | 0.901 | 0.971 |
| | ICRA'23 | **0.165** | **1.726** | **8.568** | **0.229** | **0.731** | 0.930 | 0.984 |
| | **MIC** | 0.197 | 1.836 | 8.787 | 0.253 | 0.570 | **0.935** | **0.985** |
| **Mean** | Source | 0.202 | 2.782 | 9.401 | 0.308 | 0.680 | 0.873 | 0.929 |
| | CoTTA | 0.205 | 2.822 | 9.466 | 0.314 | 0.663 | 0.871 | 0.929 |
| | ICRA'23 | **0.181** | 2.310 | 8.452 | 0.260 | **0.718** | 0.900 | 0.953 |
| | **MIC** | 0.190 | **2.138** | **8.075** | **0.246** | 0.684 | **0.926** | **0.972** |

Table 1: Depth estimation results on CTTA benchmarks using DDAD and Waymo datasets (*All-6*, *Rainy-5*, *Sunny-Day-5*, *Sunny-Night-5*). **Mean** indicate the results averaged over all datasets. The results are averaged over 3 seeds.

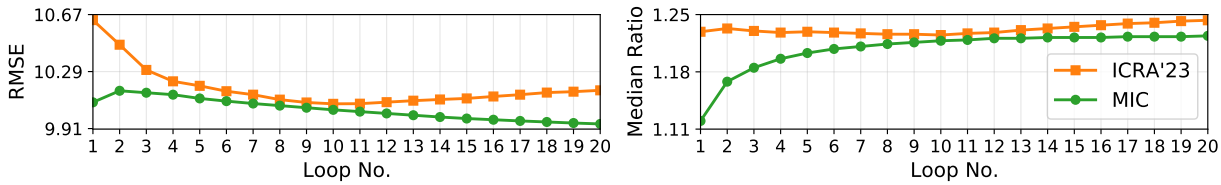


Figure 3: Mean RMSE and Median Ratio of MIC and ICRA'23 techniques for each loop of repeating the *All-6* benchmark 20 times.

| Method | Time [ms] | Memory [MB] |
|---|---|---|
| Source | 27.2 | 1215.9 |
| CoTTA | 1508.6 | 9663.2 |
| ICRA'23 | 363.1 | 10433.2 |
| **MIC** | **165.3** | **7017.7** |

Table 2: The average wall-clock time (milliseconds) and memory usage (MB) for processing a single batch on an NVIDIA RTX 4080 GPU.

| Method | RMSE↓ | $\sigma < 1.25 \uparrow$ |
|---|---|---|
| Source | 12.130 | 0.580 |
| + MIC (w/o EMA teacher) | 10.775 | 0.552 |
| + MIC (w/ EMA teacher) | **8.050** | **0.681** |

Table 3: MIC ablation study on *All-6* benchmark. The learning rate for 2nd row was selected separately considering increased plasticity without the EMA teacher model.

## 4. Conclusions

We explore the MIC approach for CTTA in a scale-aware MDE model, one of the first studies to apply CTTA to supervised MDE networks. We demonstrate that adapting the model while enhancing its focus on global image context significantly improves performance in depth estimation task. Our comprehensive evaluations reveal that MIC achieves performance comparable to state-of-the-art approach while requiring less data and computational resources. Additionally, we find that CoTTA's augmentation consistency is significantly less effective for depth estimation compared to masked consistency.

## Acknowledgments

This research was funded in whole or in part by National Science Centre, Poland, grant no 2024/53/N/ST6/03156 and 2023/51/D/ST6/02846. For the purpose of Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission. This work was supported by Horizon Europe Programme under GA no. 101120237, project "ELIAS: European Lighthouse of AI for Sustainability". This work has been supported by the Polish National Agency for Academic Exchange (NAWA) under the STER program, Towards Internationalization of Poznan University of Technology Doctoral School (2022-2024). This work has been supported by the "Bilateral AI" Cluster of Excellence.

## References

Casser, V., Pirk, S., Mahjourian, R., and Angelova, A. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8001–8008, 2019.

Döbler, M., Marsden, R. A., and Yang, B. Robust mean teacher for continual and gradual test-time adaptation. *CoRR*, abs/2211.13081, 2022.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., and Gaidon, A. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2485–2494, 2020.

Hoyer, L., Dai, D., Wang, H., and Van Gool, L. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11721–11732, 2023.

Ibrahem, H., Salem, A., and Kang, H.-S. Rt-vit: Real-time monocular depth estimation using lightweight vision transformers. *Sensors*, 22(10):3849, 2022.

Kuznietsov, Y., Proesmans, M., and Van Gool, L. Comoda: Continuous monocular depth adaptation using past experiences. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2907–2917, 2021.

Lee, J. H., Han, M.-K., Ko, D. W., and Suh, I. H. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

Li, Z., Shi, S., Schiele, B., and Dai, D. Test-time domain adaptation for monocular depth estimation. In *International Conference on Robotics and Automation (ICRA)*, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

McCraith, R., Neumann, L., Zisserman, A., and Vedaldi, A. Monocular depth estimation with self-supervised instance adaptation. *arXiv preprint arXiv:2004.05821*, 2020.

Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., and Tan, M. Efficient test-time model adaptation without forgetting. In *ICML*, volume 162, pp. 16888–16905, 2022.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., and Tan, M. Towards stable test-time adaptation in dynamic wild world. In *ICLR*, 2023.

Rajapaksha, U., Sohel, F., Laga, H., Diepeveen, D., and Bennamoun, M. Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. *ACM Computing Surveys*, 56(12):1–51, 2024.

Shu, C., Yu, K., Duan, Z., and Yang, K. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pp. 572–588. Springer, 2020.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B. A., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

Wang, Q., Fink, O., Gool, L. V., and Dai, D. Continual test-time domain adaptation. In *CVPR*, pp. 7191–7201, 2022.

Yang, S., Wu, J., Liu, J., Li, X., Zhang, Q., Pan, M., Gan, Y., Chen, Z., and Zhang, S. Exploring sparse visual prompt for domain adaptive dense prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16334–16342, 2024.

Yuan, L., Xie, B., and Li, S. Robust test-time adaptation in dynamic scenarios. In *CVPR*, pp. 15922–15932, 2023.

Yuan, W., Gu, X., Dai, Z., Zhu, S., and Tan, P. Newcrfs: Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., and Mattoccia, S. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*, pp. 668–678. IEEE, 2022.

# Appendix

This appendix provides comprehensive analysis to support the main paper's findings and outlines the details of experiments. Section A.1 includes further details regarding the implementation of tested methods A.1.1 and the description of CoTTA method adjustment for depth estimation task A.1.2. Section A.2 provides additional experimental results, including mask ratio $r$ value ablation A.2.1, example performance over time of tested methods A.2.2 and qualitative results A.2.3.

## A.1. Additional Details

### A.1.1. Methods Implementation Details

The experiments are conducted utilizing and adjusting the code repository of ICRA'23 (Li et al., 2023) method. We integrated the code of CoTTA (Wang et al., 2022) and our approach into this unified code base.

The parameters for CoTTA and our presented MIC are selected using an oracle method. The parameters for ICRA'23 (Li et al., 2023) are adopted directly from the paper, since they use the same benchmarks.

All of the tested methods utilize Adam optimizer. We set learning rate for CoTTA to 1e-7. ICRA'23 use the learning rate set to 1e-5. The presented MIC utilize the learning rate of 5e-4 for Waymo-based benchmarks and 1e-4 for DDAD dataset.

All methods utilize the Adam optimizer. The learning rate for CoTTA is set to 1e-7, while ICRA'23 uses a learning rate of 1e-5. The presented MIC method uses a learning rate of 5e-4 for Waymo-based benchmarks and 1e-4 for the DDAD dataset.

For CoTTA, as in the original implementation for semantic segmentation task, we use the multi-scaling input with flip as the augmentation method to generate augmentation-weighted pseudo-label with the scale factors of [0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0]. The restoration probability parameter $p$ is set to 0.01 and smoothing factor $\alpha$ to 0.999.

For MIC, we set the mask ratio $r$ to 0.5 and the smoothing factor $\alpha$ to 0.999.

### A.1.2. CoTTA Adjustment for Depth Estimation Task

The original CoTTA method (Wang et al., 2022), designed for image classification and semantic segmentation, leverages the categorical nature of predictions to update the network. To adapt it for depth estimation, modifications are necessary. It originally updates the student model by minimizing the cross-entropy consistency between the teacher and the student predictions. Depending on the prediction confidence, the pseudo-labels generated by the teacher model are the result of averaging predictions on multiple, differently augmented images. Since traditional depth estimation networks do not provide prediction confidence, the adapted version consistently uses averaged predictions from augmented images as pseudo-labels. Additionally, cross-entropy loss is replaced with L1 loss to suit the regression nature of depth estimation.

## A.2. Additional Experiments

### A.2.1. Mask Ratio Ablation

In Table A.1, we evaluate the impact of varying the mask ratio $r$ parameter, as defined in Eq. 1. The results show that optimal values differ across benchmarks, with no consistent pattern emerging. Nevertheless, a mask ratio of 0.5 appears to offer the most robust performance overall.

| Benchmark | Mask Ratio $r$ | AbsRel↓ | SqRel↓ | RMSE↓ | RMSElog↓ | $\sigma < 1.25 \uparrow$ | $\sigma < 1.25^2 \uparrow$ | $\sigma < 1.25^3 \uparrow$ |
|---|---|---|---|---|---|---|---|---|
| DDAD | 0.1 | 0.134 | 1.364 | 7.583 | 0.213 | 0.814 | 0.948 | 0.981 |
| | 0.3 | 0.134 | 1.339 | 7.432 | 0.210 | 0.818 | 0.950 | 0.982 |
| | 0.5 (Def.) | **0.133** | 1.319 | 7.290 | 0.207 | 0.822 | 0.951 | **0.983** |
| | 0.7 | **0.133** | **1.311** | 7.171 | 0.204 | 0.825 | **0.952** | **0.983** |
| | 0.9 | 0.135 | 1.355 | **6.938** | **0.201** | **0.827** | 0.950 | **0.983** |
| *All-6* | 0.1 | 0.230 | 3.549 | 10.231 | 0.343 | 0.622 | 0.846 | 0.911 |
| | 0.3 | 0.210 | 2.872 | 9.116 | 0.283 | 0.648 | 0.901 | 0.951 |
| | 0.5 (Def.) | 0.196 | 2.400 | 8.202 | 0.252 | 0.677 | 0.927 | 0.970 |
| | 0.7 | **0.188** | **2.178** | **7.962** | **0.245** | **0.693** | **0.932** | **0.973** |
| | 0.9 | 0.208 | 2.674 | 8.247 | 0.258 | 0.661 | 0.927 | 0.970 |
| *Rainy-5* | 0.1 | 0.250 | 3.365 | 8.689 | 0.291 | 0.592 | 0.879 | 0.945 |
| | 0.3 | **0.236** | 2.871 | **8.488** | **0.284** | **0.594** | **0.882** | **0.950** |
| | 0.5 (Def.) | 0.238 | **2.862** | 8.684 | 0.290 | 0.582 | 0.877 | 0.949 |
| | 0.7 | 0.251 | 3.331 | 8.745 | 0.293 | 0.590 | 0.872 | 0.941 |
| | 0.9 | 0.247 | 2.955 | 9.975 | 0.325 | 0.521 | 0.831 | 0.937 |
| *Sunny-Day-5* | 0.1 | 0.182 | 2.300 | 7.502 | 0.230 | **0.774** | 0.941 | **0.975** |
| | 0.3 | 0.194 | 2.497 | 7.824 | 0.236 | 0.753 | 0.942 | **0.975** |
| | 0.5 (Def.) | **0.185** | **2.271** | **7.410** | **0.229** | 0.769 | 0.941 | **0.975** |
| | 0.7 | 0.186 | 2.278 | 7.460 | 0.231 | 0.765 | **0.943** | **0.975** |
| | 0.9 | 0.228 | 3.025 | 8.944 | 0.267 | 0.677 | 0.913 | 0.970 |
| *Sunny-Night-5* | 0.1 | 0.195 | 1.882 | **8.514** | **0.245** | **0.620** | **0.942** | **0.986** |
| | 0.3 | **0.193** | 1.845 | 8.855 | 0.250 | 0.607 | 0.934 | 0.985 |
| | 0.5 (Def.) | 0.197 | **1.836** | 8.787 | 0.253 | 0.570 | 0.935 | 0.985 |
| | 0.7 | 0.203 | 1.893 | 8.946 | 0.261 | 0.522 | 0.933 | 0.985 |
| | 0.9 | 0.254 | 3.472 | 11.887 | 0.357 | 0.452 | 0.809 | 0.912 |
| **Mean** | 0.1 | 0.198 | 2.492 | 8.504 | 0.264 | **0.684** | 0.911 | 0.959 |
| | 0.3 | 0.193 | 2.285 | 8.343 | 0.253 | **0.684** | 0.922 | 0.969 |
| | 0.5 (Def.) | **0.190** | **2.138** | 8.075 | **0.246** | **0.684** | **0.926** | **0.972** |
| | 0.7 | 0.192 | 2.198 | **8.057** | 0.247 | 0.679 | **0.926** | 0.971 |
| | 0.9 | 0.214 | 2.696 | 9.198 | 0.282 | 0.627 | 0.886 | 0.954 |

Table A.1: Depth estimation results on CTTA benchmarks using DDAD and Waymo datasets (*All-6*, *Rainy-5*, *Sunny-Day-5*, *Sunny-Night-5*). **Mean** indicate the results averaged over all datasets. The results are averaged over 3 seeds.

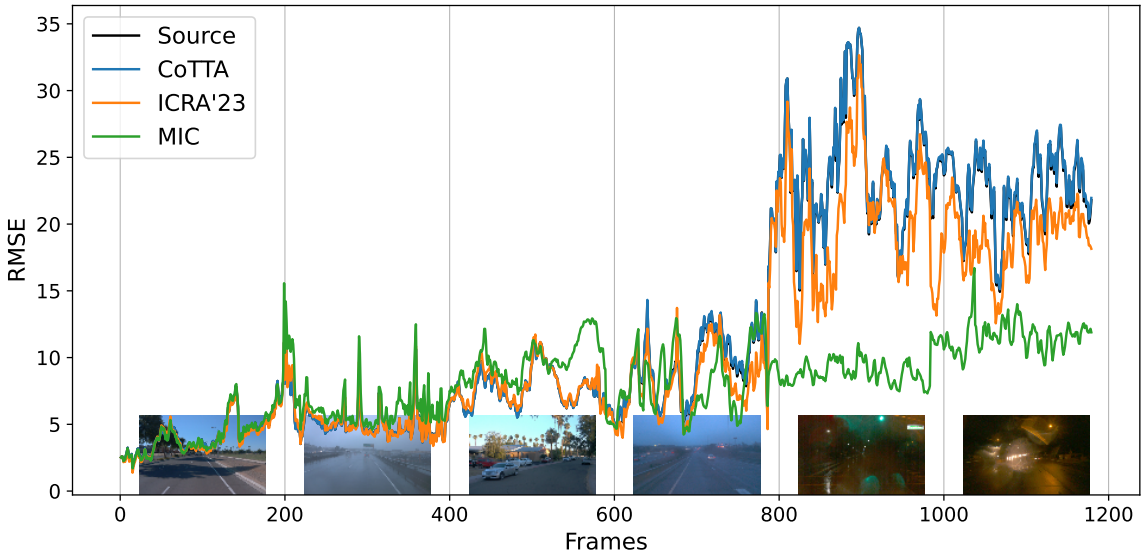


Figure A.1: The RMSE for each frame of *All-6* benchmark for each tested method (lower is better). The displayed images are examples from the video in corresponding benchmark section.

### A.2.2. Performance Over Time

Figure A.1 presents the RMSE for each frame of *All-6* benchmark. It can be seen that MIC improves the performance of the Source model even under significant domain shifts (towards the end of *All-6* benchmark).

### A.2.3. Qualitative Results

Figures A.2, A.3, A.4, A.5 and A.6 show the qualitative results for teach of the benchmarks and tested methods. Our analysis indicates that the MIC method shows greater improvement in depth estimation at higher depth values.

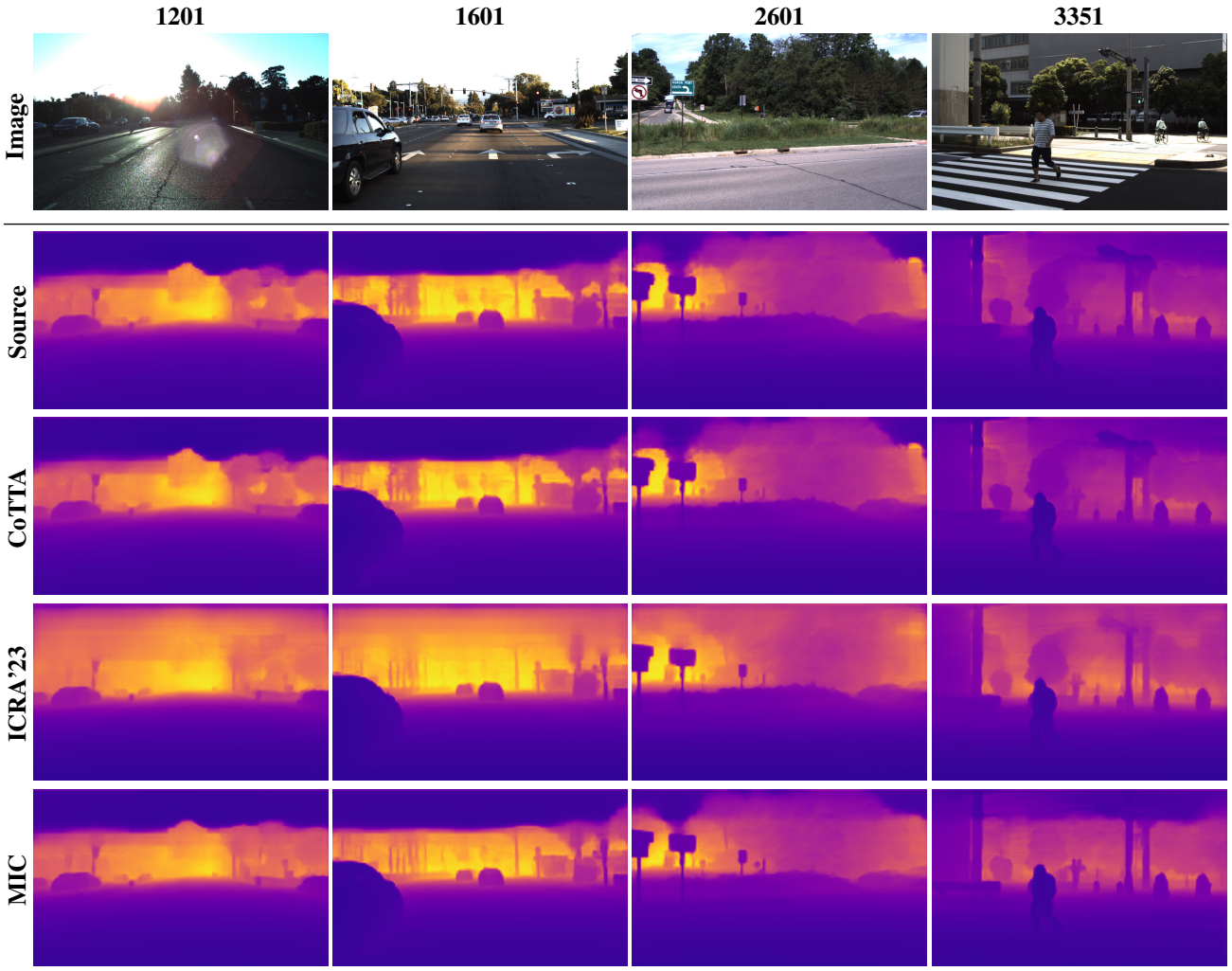


Figure A.2: Qualitative results on DDAD benchmark. Column labels indicate the frame number in the benchmark.

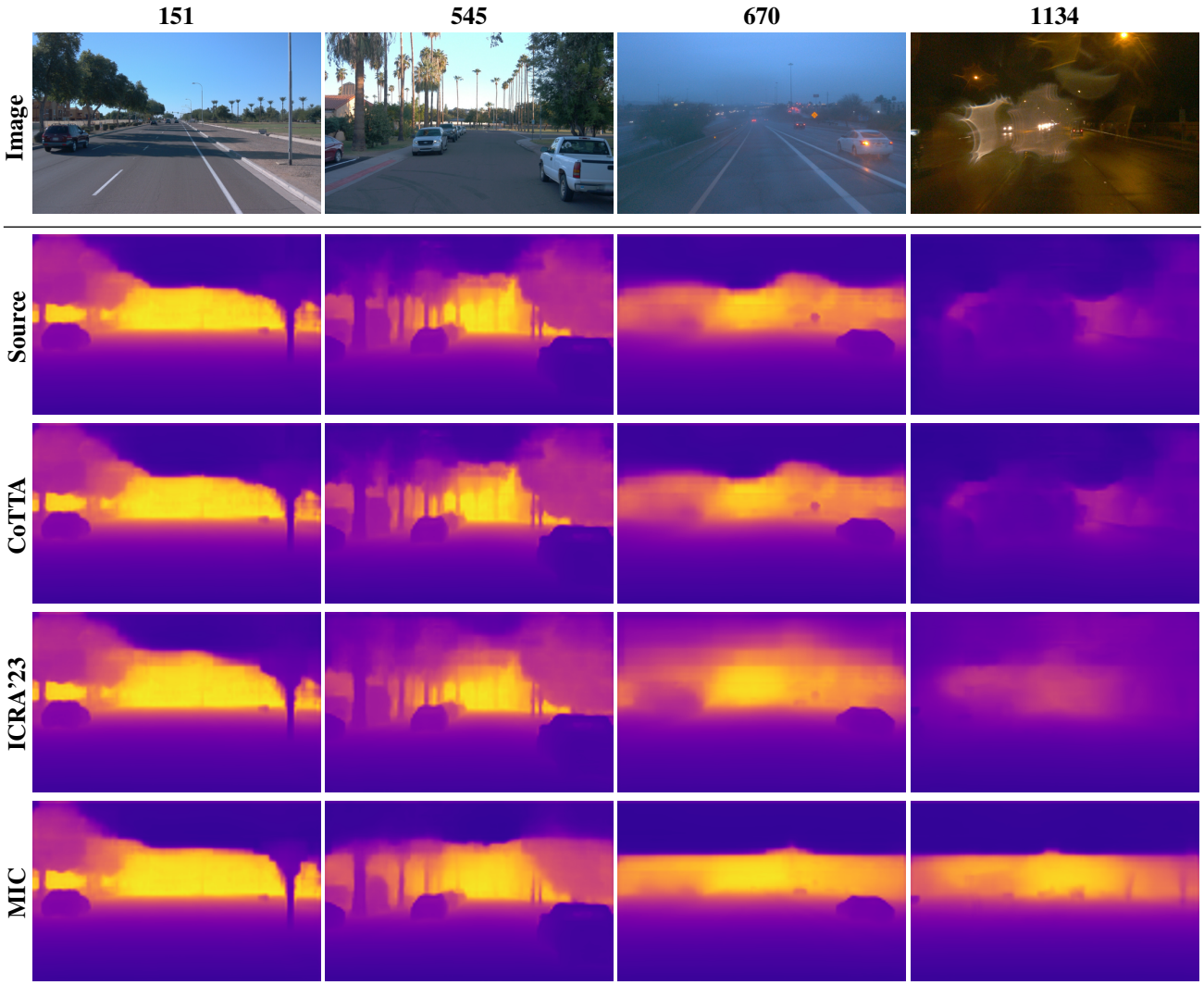


Figure A.3: Qualitative results on *All-6* benchmark. Column labels indicate the frame number in the benchmark.

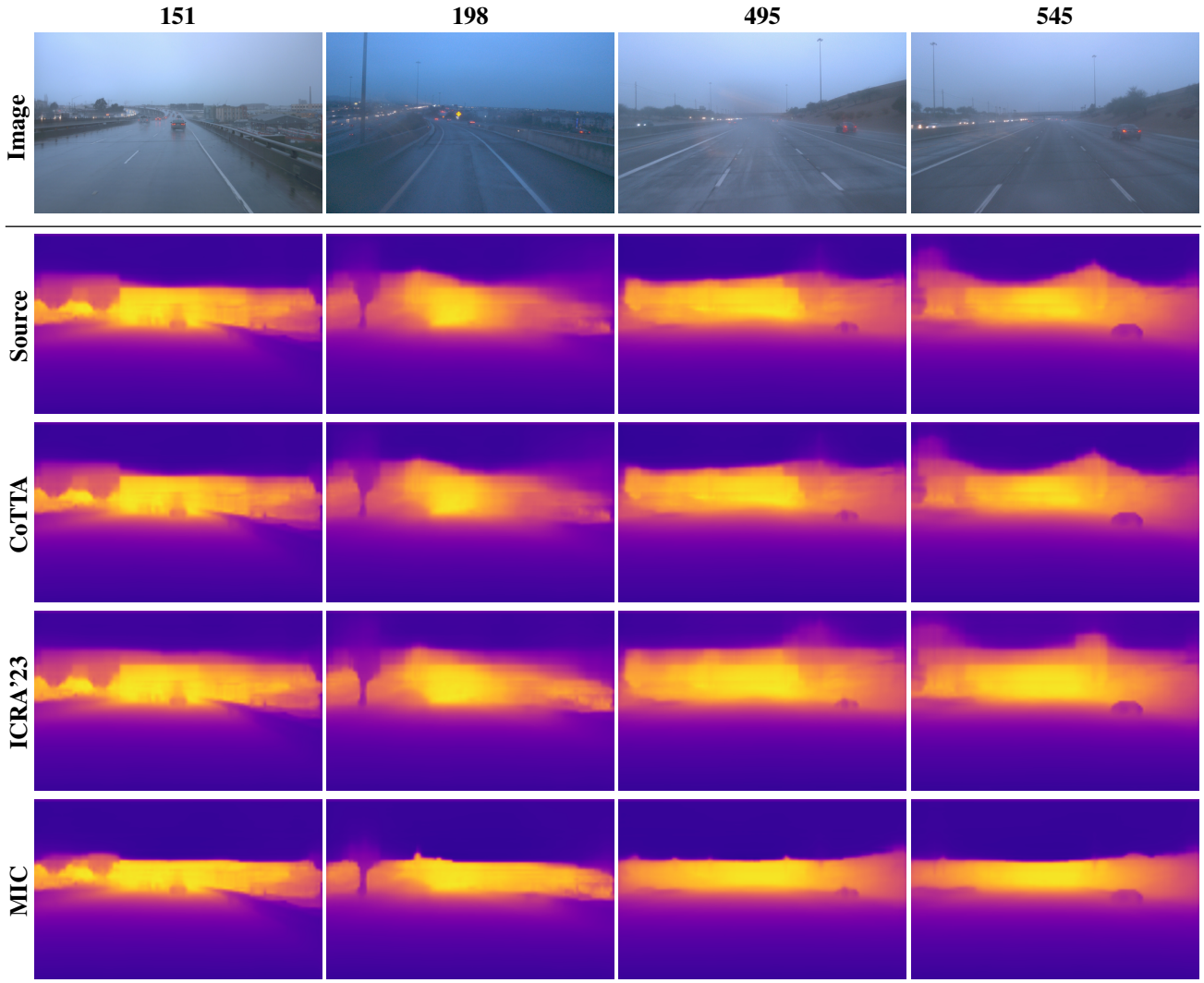


Figure A.4: Qualitative results on *Rainy-5* benchmark. Column labels indicate the frame number in the benchmark.

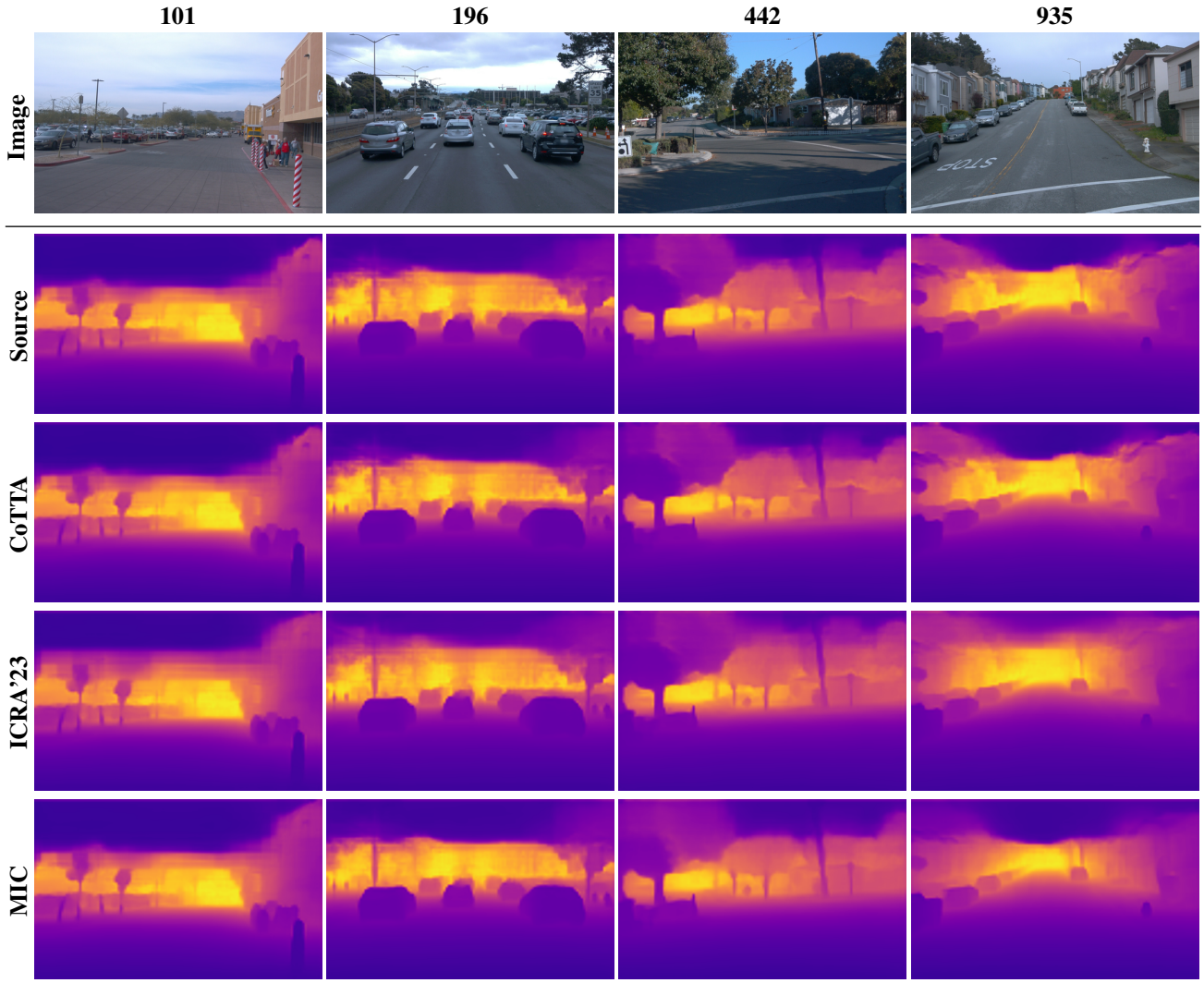


Figure A.5: Qualitative results on *Sunny-Day-5* benchmark. Column labels indicate the frame number in the benchmark.

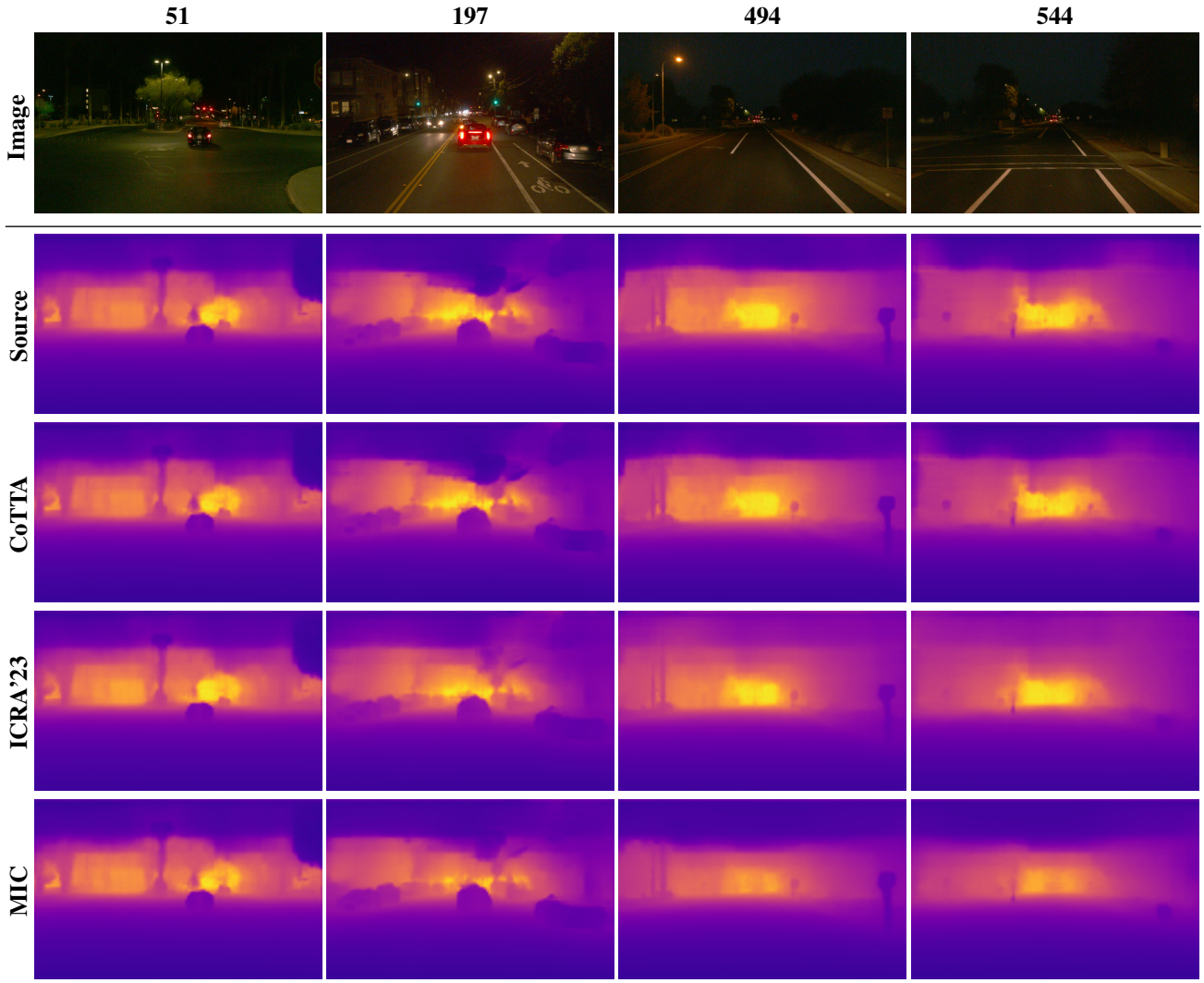


Figure A.6: Qualitative results on *Sunny-Night-5* benchmark. Column labels indicate the frame number in the benchmark.