Supplementary for MAMo: Leveraging Memory and Attention for Monocular Video Depth Estimation

Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli
Qualcomm AI Research*
{ryasarla, hongcai, jisojeon, yunxshi, rgarrepa, fporikli}@qti.qualcomm.com

Contents

1. Architecture Details **1**
1.1. NeWCRFs + MAMo 1
1.2. PixelFormer + MAMo 1
1.3. ResNet-DPT + MAMo 2

2. Training Details **2**
2.1. Temporal consistency 2

3. Additional Results **3**
3.1. Additional Comparison on KITTI and DDAD 3
3.2. Additional Ablation Studies 3
3.2.1 Token Channels 3
3.2.2 Augmentation of Frame Subsampling 3
3.3. Qualitative Results 3

4. Optical Flow Estimation Models **4**

*Qualcomm AI Research, an initiative of Qualcomm Technologies, Inc.

1. Architecture Details

In this section we explain in more detail how we apply MAMo to the latest SOTA monocular depth estimation methods to perform video depth estimation, including PixelFormer [2], NeWCRFs [27], and a strong convolutional baseline which is a variant of DPT [17] with a ResNet encoder (referred to as ResNet-DPT).

1.1. NeWCRFs + MAMo

We apply our proposed MAMo approach to NeWCRFs [27], and refer to it as NeWCRFs + MAMo. We use follow same encoder and decoder architectures in [27]. For the encoder, Swin transformer [13] is employed to extract the features. Pyramid Pooling Module [16] is used to extract global information. Pairwise potential module (PPM) head aggregates the global and local information. For the decoder, Neural Window FC-CRFs modules are employed to compute depth D_t.[1]. Since we concatenate optical flow O_t, the previous frame's decoder features F_{t-1}, and the current frame's encoder features E_t as input to the decoder, we adjust the input channels of each Neural FC-CRF module of the decoder accordingly. Fig. 1 shows a more detailed architectural view of NeWCRFs + MAMo.

Fig. 2 provides an illustration of the Memory Attention part in MAMo. For self-attention and cross-attention layers in NeWCRFs + MAMo, we use Neural Window FC-CRFs.

1.2. PixelFormer + MAMo

We apply MAMo to PixelFormer [2] and refer to it as PixelFormer + MAMo. We use the same architectures from [2] for the encoder and decoder of PixelFormer + MAMo. For the encoder, Swin transformer [13] is employed to extract the features. Pixel Query Initialise (PQI) is used to extract global information using pyramid spatial pooling [6], and compute the initial pixel queries Q_t. For the decoder, Skip Attention Modules (SAM) are employed to compute depth D_t.[2] The input channels of SAM modules are adjusted according to the concatenation of E_t, F_{t-1} and

[1] See [27] for more details on Neural Window FC-CRFs
[2] See [2] for more details on SAM.

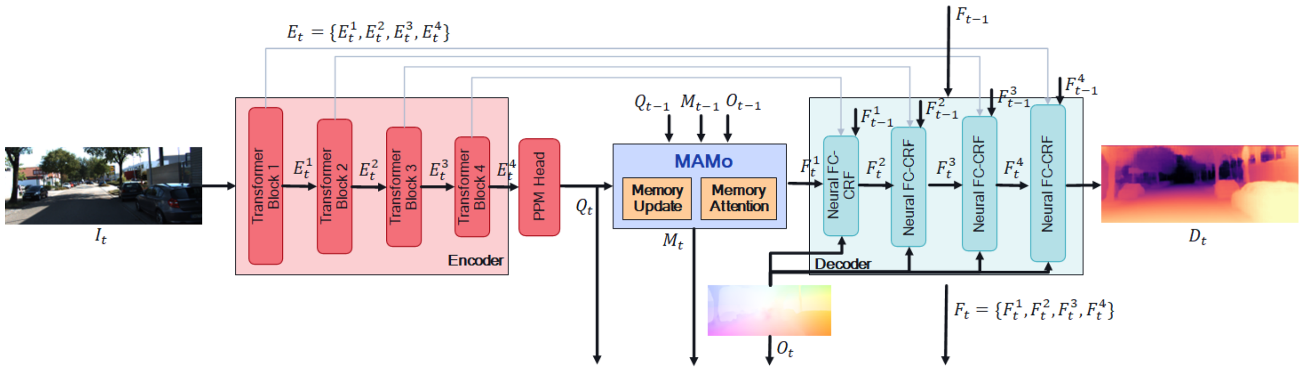

Figure 1. Detailed Architecture of NewCRFs + MAMo.

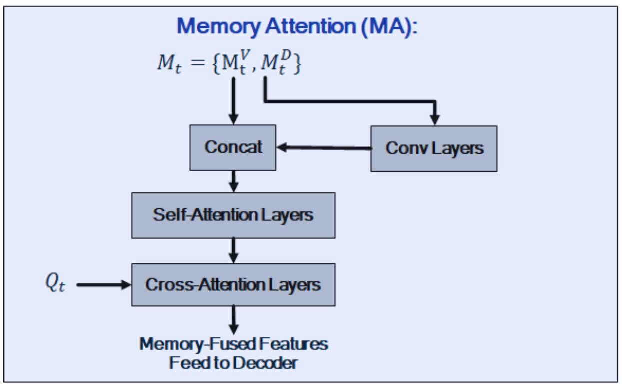

Figure 2. Overview of proposed Memory Attention in MAMo. For Self-attention and cross-attention, we use Neural FC-CRFs for NeWCRFs + MAMo, Skip Attention Module (SAM) for PixelFormer + MAMo, and LinFormer for ResNet-DPT + MAMo.

O_t. We use SAM for the self-attention and cross-attention layers in the Memory Attention of PixelFormer + MAMo.

1.3. ResNet-DPT + MAMo

We apply MAMo to ResNet-DPT [17], and refer to it as ResNet-DPT + MAMo. For the encoder, ResNet50 [7] is employed to extract the features. For the decoder, we use the fusion module from [17] to compute depth D_t. For self-attention and cross-attention layers in the Memory Attention of ResNet-DPT + MAMo, we use LinFormer attention modules [21].

2. Training Details

Detailed training steps are provided in Algorithm 1. Note, we train the networks PixelFormer, NeWCRFs, and ResNet-DPT for first 5 epochs without MAMo, and train PixelFormer+MAMo, NeWCRFs+MAMo, and ResNet-DPT+MAMo with MAMo for the rest 20 epochs.

Algorithm 1 Training MAMo video depth model

Input: Training dataset $\mathcal{D}_V$ consisting of training videos and depth ground truths. For each training video, $V = \{I_0, ..., I_T\}$ and $D^{gt} = \{D_0^{gt}, ..., D_T^{gt}\}$
Model: $h(\cdot)$ and $g(\cdot)$: encoder and full depth network
for every epoch **do**
  for $V, D^{gt} \in \mathcal{D}_V$ **do**
    Initialization
     $Q_0 \leftarrow h(I_0)$, $O_0 \leftarrow \mathbf{0}$, $F_{-1} \leftarrow \mathbf{0}$
     Update M_0 (repeat Q_0 *and* O_0 for L times)
     $D_0 \leftarrow g(I_0; M_0, O_0, F_{-1})$
    for $I_t, D_t^{gt} \in V, D^{gt}$ **do**
      Estimate O_t
     Memory Update (Sec. 3.2 in the main paper)
      $\widetilde{M}_t^V \leftarrow \{M_{t-1}^V, Q_{t-1}\}$, $\widetilde{M}_t^D \leftarrow \{M_{t-1}^D, O_{t-1}\}$
      $\widetilde{M}_t \leftarrow \{\widetilde{M}_t^V, \widetilde{M}_t^D\}$
      $I_t^w \leftarrow Warp(I_{t-1}, O_t)$
      $\widetilde{D}_t \leftarrow g(I_t; \widetilde{M}_t, O_t, F_{t-1})$
      $\widetilde{D}_t^w \leftarrow g(I_t^w; \widetilde{M}_t, O_t, F_{t-1})$
      SILogLoss $(\widetilde{D}_t, \widetilde{D}_t^w)$
      Backpropagation
      Update M_t (Eq. 2 in the main paper)
     Depth Estimation
      $D_t \leftarrow g(I_t; M_t, O_t, F_{t-1})$, $Q_t \leftarrow h(I_t)$
      Compute $\mathcal{L}_d$ *between* D_t *and* D_t^{gt}
         (Eq. 5 in the main paper)
      Update parameters of $h(\cdot)$, $g(\cdot)$
    end for
  end for
end for

2.1. Temporal consistency

We evaluate temporal consistency using the metrics from Li et al. [10],

$$aTC_t = \frac{1}{\sum(K_t == 1)} K_t \| \frac{D_t - D_t^w}{D_t} \|,$$

$$rTC_t = \frac{1}{\sum(K_t == 1)} K_t \left[\text{Max} \left(\frac{D_t}{D_t^w}, \frac{D_t^w}{K_t} \right) < \text{thr} \right],$$

where K_t is a depth validity mask, D_t is predicted depth for I_t and D_t^w is warped from D_{t-1} using optical flow; we use the latest SOTA FlowFormer [8]. Table 3 shows

Table 1. Quantitative results on KITTI (Eigen split) for distances up to 80 meters. † means methods uses multiple networks to estimate depth. ManyDepth-FS, and TC-Depth-FS means ManyDepth and TC-Depth are trained in fully-supervised fashion using ground-truths respectively. MF means multi frame methods, SF means single frame methods, and VD means extending MDE to VDE methods.↑ means higher the better, and ↓ means lower the better.

Type	Method	Encoder	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE_{log} ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
MF	NeuralRGB [12]	CNN based†	0.100	–	2.829	–	0.931	–	–
	ST-CLSTM [28]	Resnet18	0.101	–	4.137	–	0.890	0.970	0.9890
	FlowGRU [5]	CNN [5]	0.112	0.700	4.260	0.184	0.881	0.962	0.9830
	Flow2Depth [25]	CNN [14]†	0.081	0.488	3.651	0.146	0.912	0.970	0.9883
	RDE-MV [15]	ResNet18†	0.111	0.821	4.650	0.187	0.821	0.961	0.9823
	Patil *et.al.* [15]	ResNet18†+ConvLSTM	0.102	–	4.148	–	0.884	0.961	0.9824
	Cao *et.al.* [4]	–	0.099	–	3.832	–	0.886	0.968	0.9890
	STAD [9]	CNN † [12]	0.109	0.594	3.312	0.153	0.889	0.971	0.9890
	FMNet [22]	ResNeXt-101	0.099	–	3.832	0.129	0.886	0.968	0.9893
	ManyDepth-FS [23]	ResNet50	0.069	0.342	3.414	0.111	0.930	0.989	0.9970
	ManyDepth-FS [23]	Swin-large	0.060	0.248	2.747	0.099	0.955	0.993	0.9981
	TC-Depth-FS [18]	ResNet50	0.071	0.330	3.222	0.108	0.922	0.993	0.9970
SF	AdaBins [3]	EfficientNet-B5+mViT [20]	0.058	0.190	2.360	0.088	0.964	0.995	0.9991
	BinsFormer [11]	Swin-large	0.052	0.151	2.098	0.079	0.975	0.997	0.9992
	DepthFormer [1]	MiT-B4 [24]	0.058	0.187	2.285	0.087	0.967	0.996	0.9991
	SwinV2-MIM [26]	Swin-large	0.050	0.139	1.966	0.075	0.977	0.998	**0.9995**
	URCDC [19]	Swin-large	0.050	0.142	2.032	0.076	0.977	0.997	0.9994
VD	ResNet-DPT	ResNet50	0.085	0.383	3.242	0.130	0.913	0.981	0.9960
	ResNet-DPT+MAMo (ours)	ResNet50	0.071	0.301	2.984	0.121	0.926	0.990	0.9971
	NeWCRFs [27]	Swin-Base	0.054	0.157	2.140	0.081	0.973	0.997	0.9993
	NeWCRFs+MAMo (ours)	Swin-Base	0.051	0.149	2.090	0.078	0.976	0.998	0.9994
	NeWCRFs	Swin-large	0.053	0.154	2.118	0.080	0.974	0.997	0.9994
	NeWCRFs+MAMo (ours)	Swin-large	0.050	0.141	2.003	0.076	**0.977**	**0.998**	0.9994
	PixelFormer [2]	Swin-large	0.052	0.152	2.093	0.079	0.975	0.997	0.9994
	PixelFormer+MAMo (ours)	Swin-large	**0.049**	**0.130**	**1.884**	**0.072**	**0.977**	**0.998**	**0.9995**

Table 2. Quantitative results on DDAD dataset for distances up to 200 meters, and input frame resolution is 1216 × 1936.

Method	Encoder	Sq Rel↓	RMSE↓	$\delta < 1.25$ ↑
ManyDepth-FS [23]	Swin-large	4.211	13.899	0.784
SwinV2-MIM[26]	Swin-large	3.505	11.641	0.853
NeWCRFs	Swin-large	4.041	11.956	0.816
NeWCRFs+MAMo (ours)	Swin-large	**2.990**	**10.462**	0.867
PixelFormer	Swin-large	4.474	12.467	0.802
PixelFormer+MAMo (ours)	Swin-large	3.349	11.094	**0.870**

Table 3. Temporal consistency evaluation on KITTI. We use Swin-Large encoder for NeWCRFs and NeWCRFs + MAMo.

Metrics	ManyDepth	TC-Depth	NeWCRFs	NeWCRFs + MAMo		
				L=2	L=4	L=6
rTC ↑	0.920	0.901	0.914	0.952	0.963	0.966
aTC ↓	0.111	0.122	0.116	0.091	0.088	0.086

that MAMo is more temporally consistency than both the monocular baseline, as well as SOTA ManyDepth and TC-Depth.

3. Additional Results

In this section, we provide additional comparison results with latest, unpublished methods, as well as additional ablation studies.

3.1. Additional Comparison on KITTI and DDAD

In Table 1, we provide a more comprehensive comparison that includes latest unpublished methods, such as Swin-MIM [26] and and URCDC [19] on KITTI.

In Table 2, we further include Swin-MIM [26] in the comparison on DDAD, where the models are trained on KITTI and tested on DDAD.

3.2. Additional Ablation Studies

3.2.1 Token Channels

We perform an ablation study for different number of feature channels in the visual memory tokens. As shown in Table 4, when using NeWCRFs + MAMo, the model's accuracy is almost the same for token channels of 256 and 512 (we use 512 in the main paper). This allows one to improve computational efficiency as needed with slight accuracy drops.

3.2.2 Augmentation of Frame Subsampling

In the paper, we use frame subsampling as an augmentation when training the video depth model (c.f. Section 3.5 in the main paper). Table 5 provides an ablation study for not using and using frame subsampling, with drop rates r equal to 0 and 4, respectively. It can be seen that frame subsampling leads to lower depth estimation errors, since it allows the network to see more variety of motion and scene changes.

3.3. Qualitative Results

We provide additional visual results. Figures 3, 4, and 5 show that MAMo considerably improves depth estimation over baselines PixelFormer and NeWCRFs in several regions: (i) traffic sign and telephone booth in Fig. 3, (ii) person in Fig. 4, and (iii) railway tracks and car in Fig. 5.

Table 4. Ablation experiment for number of channels in visual memory token on KITTI dataset. We perform this experiment using NeWCRFs + MAMo with Swin-Large encoder.

Token Channels	Abs Rel↓	Sq Rel↓	RMSE↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$
256	0.050	0.140	2.025	0.977	0.998
512	0.050	0.141	2.003	0.977	0.998

Table 5. Ablation experiment for Frame sampling on KITTI dataset. We perform this experiment using NeWCRFs + MAMo with Swin-Large encoder.

Drop Rate	Abs Rel↓	Sq Rel↓	RMSE↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$
r = 0	0.050	0.142	2.032	0.977	0.998
r = 4	0.050	0.141	2.003	0.977	0.998

4. Optical Flow Estimation Models

We use the official codes and pre-trained checkpoints from RAFT.[3] We use Sintel-trained checkpoint for indoor scenarios like NYU-Depth V2 and KITTI-trained checkpoint for outdoor scenarios like KITTI and DDAD.

[3] https://github.com/princeton-vl/RAFT

References

[1] Ashutosh Agarwal and Chetan Arora. Depthformer: Multiscale vision transformer for monocular depth estimation with global local information fusion. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3873–3877, 2022. 3

[2] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5861–5870, January 2023. 1, 3

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, 2021. 3

[4] Yuanzhouhan Cao, Yidong Li, Haokui Zhang, Chao Ren, and Yifan Liu. Learning structure affinity for video depth estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 190–198, 2021. 3

[5] Chanho Eom, Hyunjong Park, and Bumsub Ham. Temporally consistent depth prediction with flow-guided memory units. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4626–4636, 2019. 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[8] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 668–685. Springer, 2022. 2

[9] Hyunmin Lee and Jaesik Park. Stad: Stable video depth estimation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3213–3217. IEEE, 2021. 3

[10] Siyuan Li, Yue Luo, Ye Zhu, Xun Zhao, Yu Li, and Ying Shan. Enforcing temporal consistency in video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1145–1154, 2021. 2

[11] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 3

[12] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10986–10995, 2019. 3

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In

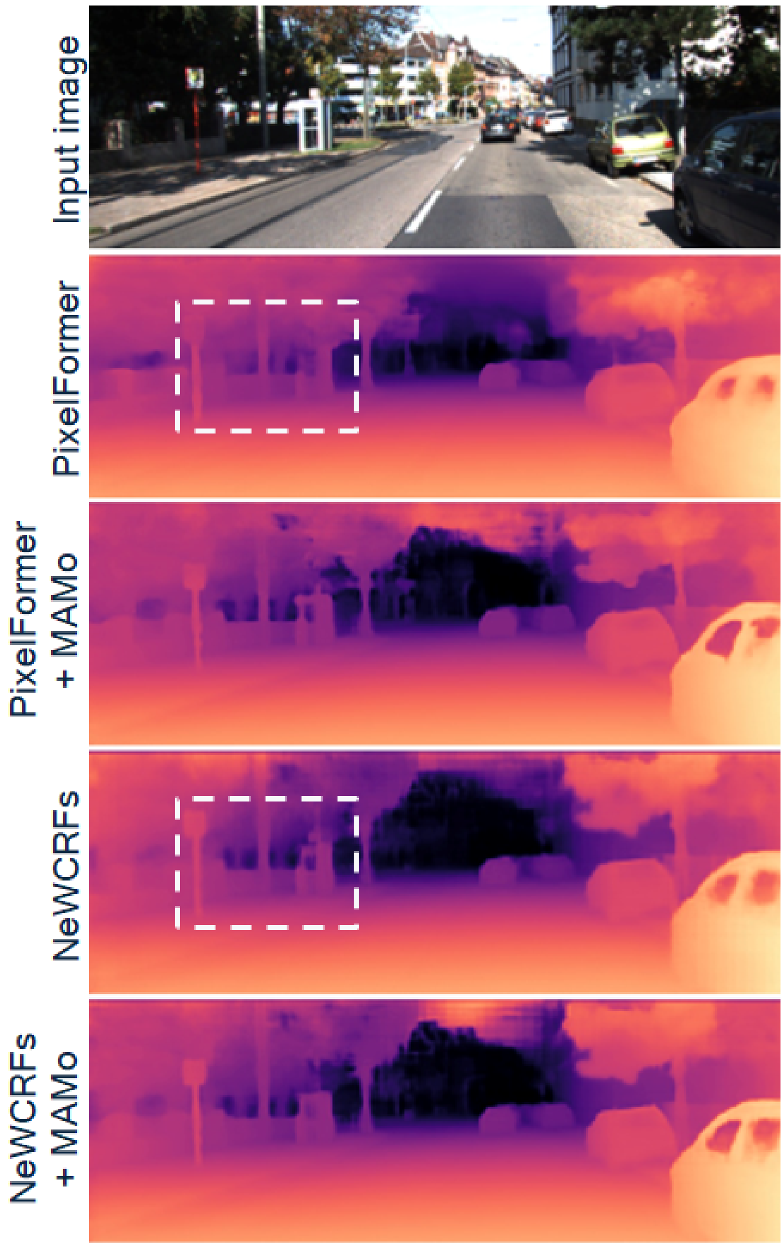

Figure 3. Qualitative results on KITTI. We highlight (white boxes) regions where MAMo significantly improves depth estimation quality.

Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021. 1

[14] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 3

[15] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *IEEE Robotics and Automation Letters*, 5(4):6813–6820, 2020. 3

[16] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11545, 2021. 1

[17] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 1, 2

[18] Patrick Ruhkamp, Daoyi Gao, Hanzhi Chen, Nassir Navab, and Beniamin Busam. Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation. In *Proceedings of*

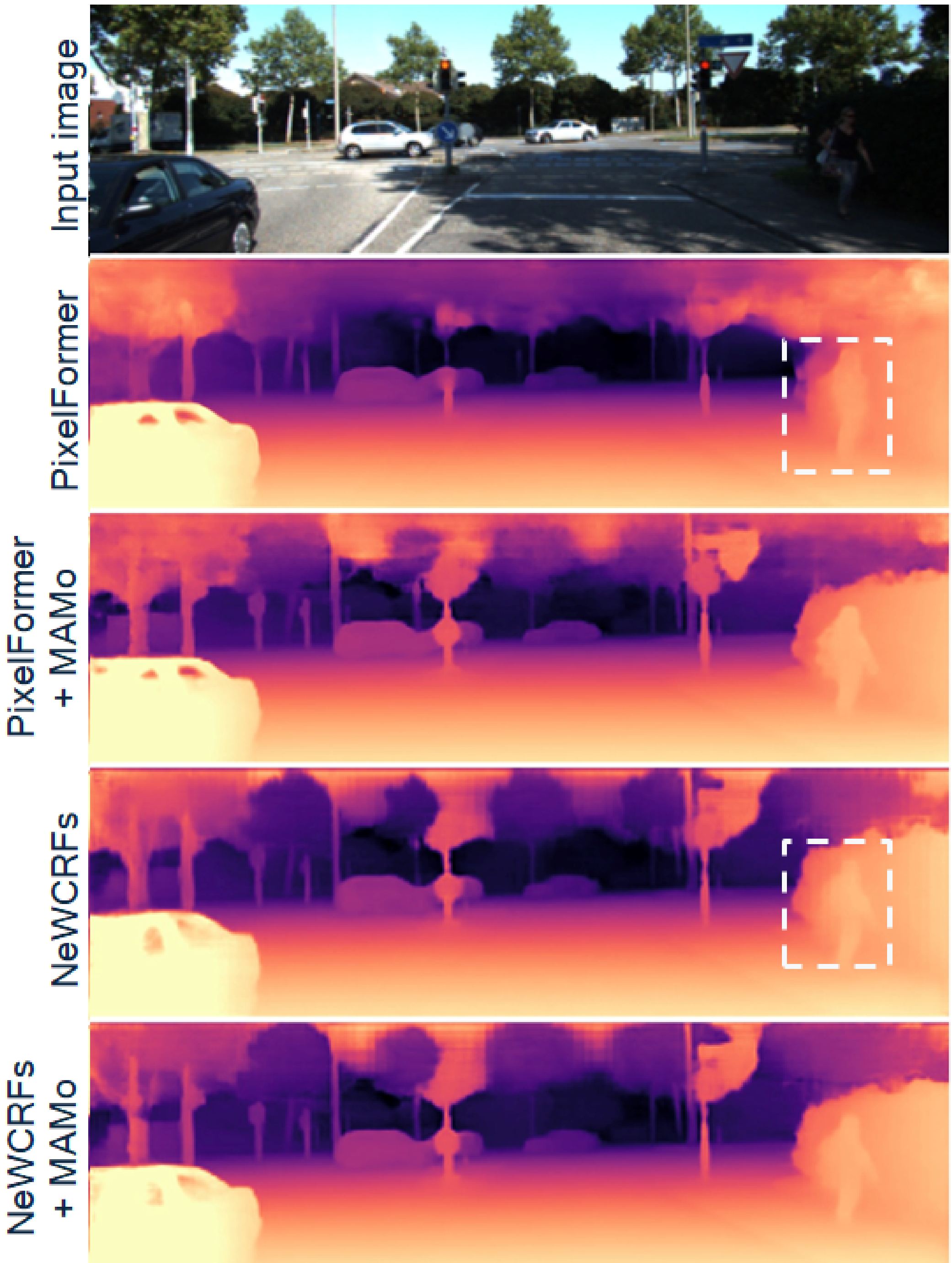

Figure 4. Qualitative results on KITTI. We highlight (white boxes) regions where MAMo significantly improves depth estimation quality.

the International Conference on 3D Vision (3DV), pages 837–847, 2021. 3

[19] Shuwei Shao, Zhongcai Pei, Weihai Chen, Ran Li, Zhong Liu, and Zhengguo Li. Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimatione. *https://arxiv.org/abs/2302.08149*, 2023. 3

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019. 3

[21] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2

[22] Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian,

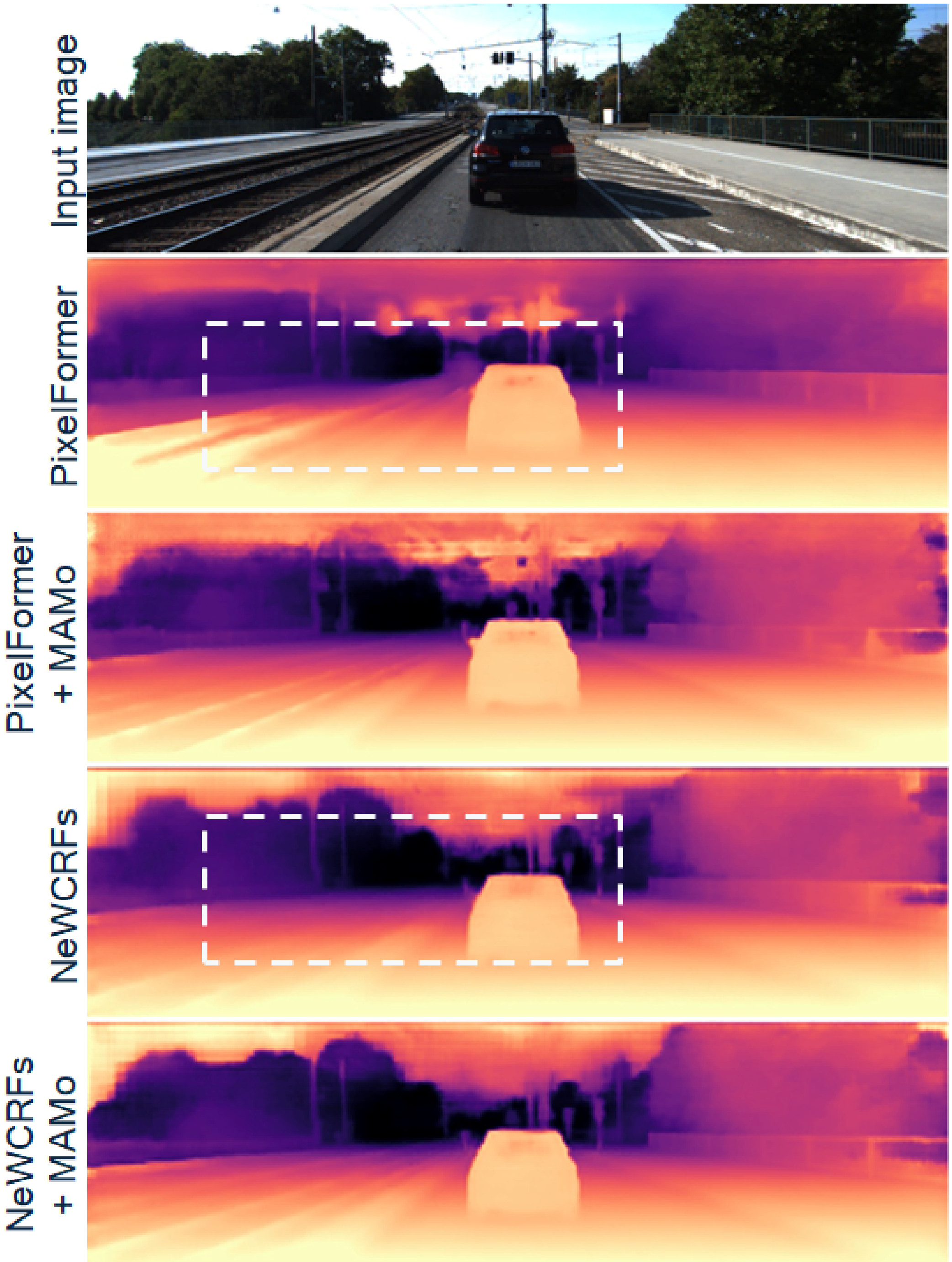

Figure 5. Qualitative results on KITTI. We highlight (white boxes) regions where MAMo significantly improves depth estimation quality.

and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6347–6358, 2022. 3

[23] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, 2021. 3

[24] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[25] Jiaxin Xie, Chenyang Lei, Zhuwen Li, Li Erran Li, and Qifeng Chen. Video depth estimation by fusing flow-to-depth proposals. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 10100–10107, 2020. 3

[26] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. pages 14475–14485, 2023. 3

[27] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Newcrfs: Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[28] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1725–1734, 2019. 3