Supplemental Material:
Attentive and Contrastive Learning for Joint Depth and Motion Field Estimation

Seokju Lee Francois Rameau Fei Pan In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)
{seokju91,rameau.fr,feipan664}@gmail.com, iskweon77@kaist.ac.kr

In this supplementary material, we present additional details on the models, training scheme, and experiments that could not be included in the main text due to space constraints. All tables, figures, equations, and references in this supplementary file are self-contained.

Contents of the supplementary material The supplementary material is composed of the following. (1) Details of our models. (2) Additional ablation study such as networks design and hyper-parameter settings. (3) Additional experimental results.

A. Details of Architectures

The detailed architectures of DepthNet, DAM, and MotionNet are specified in Table A1, Table A2, and Table A3, respectively. Generic layers, *e.g.*, *conv*, *fc*, *pool*, *etc.*, are specified starting with a lowercase letter. Nonlinear activation functions, *e.g.*, *ReLU*, *ELU*, *etc.*, are abbreviated for visibility. The encoders of DepthNet and MotionNet are both based on ResNet18 [4]. We declare the residual block as *ResBlock* for each encoder.

DepthNet For the decoder of DepthNet, we adopt the structure of MonoDepth2 [2]. DepthNet can predict depth maps with five scales, but we empirically found that single-scale training produces better performance than multi-scale training. Thus, we predict the depth map from the last layer of the decoder, which is activated with a *sigmoid* function.

MotionNet with DAM For DAM, we design the weighted context and transforming operation for each ego-motion and residual motion feature, respectively. We attach DAM after the 2^{nd}, 3^{rd}, and 4^{th} residual layer in the ResNet encoder of MotionNet. In the case of the ResNet18 encoder, each residual layer is composed of two residual blocks. GCNet [1] attached the attention module to each residual block, however, we empirically found that deploying DAM only after the residual layer leads to a better trade off between performance and computational time. The details of this ablation is specified in Sec. B and examples of the spatial attention map are visualized in Sec. C. The decoder of MotionNet has multiple refining steps, which has been proposed by Gordon *et al*. [3]. Every refinement step aggregates the features from the encoding layer and predictions from the lower layer. Similar to DepthNet, we predict the residual motion field with a single-scale, which is refined from the last layer of the decoder.

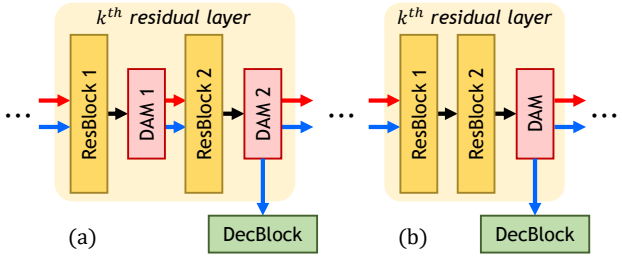

Figure A1. **Different positional design options for DAM in a residual layer.** We validate two design choices for DAM: (a) after every residual block (GCNet [1] style), and (b) after the last residual block in each residual layer.

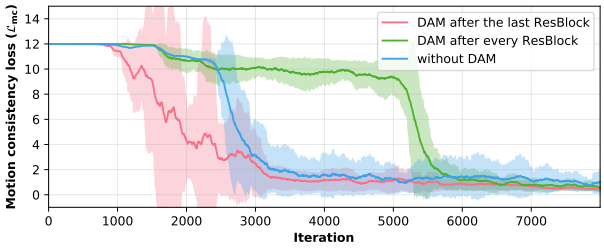

Figure A2. **Trends of motion field consistency loss in *phase-2* for different DAM configurations.** We train MotionNet on the KITTI dataset from scratch, and plot its training loss.

B. Additional Ablation Study

In this section, we discuss additional ablation studies of different design choices and hyper-parameter settings. We follow the same configurations of experiments, *e.g.*, dataset and training scheme, which are described in Sec. 4.2. of our main paper.

Different positional configurations of DAM As discussed

Type	Name	Kernel	Stride	Channel I/O	In. resol.	Out. resol.	Input
Encoder	conv_0	7×7	2	3/64	256×832	128×416	image
	maxpool_1	3×3	2	64/64	128×416	64×208	conv_0
	ResBlock_1_1	–	1	64/64	64×208	64×208	maxpool_1
	ResBlock_1_2	–	1	64/64	64×208	64×208	ResBlock_1_1
	ResBlock_2_1	–	2	64/128	64×208	32×104	ResBlock_1_2
	ResBlock_2_2	–	1	128/128	32×104	32×104	ResBlock_2_1
	ResBlock_3_1	–	2	128/256	32×104	16×52	ResBlock_2_2
	ResBlock_3_2	–	1	256/256	16×52	16×52	ResBlock_3_1
	ResBlock_4_1	–	2	256/512	16×52	8×26	ResBlock_3_2
	ResBlock_4_2	–	1	512/512	8×26	8×26	ResBlock_4_1
Decoder	conv_1_1	3×3	1	512/256	8×26	8×26	ResBlock_4_2
	upsample_1	–	–	256/256	8×26	16×52	conv_1_1
	concat_1	–	–	(256,256)/512	16×52	16×52	upsample_1, ResBlock_3_2
	conv_1_2	3×3	1	512/256	16×52	16×52	concat_1
	conv_2_1	3×3	1	256/128	16×52	16×52	conv_1_2
	upsample_2	–	–	128/128	16×52	32×104	conv_2_1
	concat_2	–	–	(128,128)/256	32×104	32×104	upsample_2, ResBlock_2_2
	conv_2_2	3×3	1	256/128	32×104	32×104	concat_2
	conv_3_1	3×3	1	128/64	32×104	32×104	conv_2_2
	upsample_3	–	–	64/64	32×104	64×208	conv_3_1
	concat_3	–	–	(64,64)/128	64×208	64×208	upsample_3, ResBlock_1_2
	conv_3_2	3×3	1	128/64	64×208	64×208	concat_3
	conv_4_1	3×3	1	64/32	64×208	64×208	conv_3_2
	upsample_4	–	–	32/32	64×208	128×416	conv_4_1
	concat_4	–	–	(32,64)/96	128×416	128×416	upsample_4, conv_0
	conv_4_2	3×3	1	96/32	128×416	128×416	concat_4
	conv_5_1	3×3	1	32/16	128×416	128×416	conv_4_2
	upsample_5	–	–	16/16	128×416	256×832	conv_5_1
	conv_5_2	3×3	1	16/16	256×832	256×832	upsample_5
	dispconv	3×3	1	16/1	256×832	256×832	conv_5_2
	sigmoid	–	–	1/1	256×832	256×832	dispconv

Table A1. Details of DepthNet.

Type	Name	Kernel	Stride	Channel I/O	In. resol.	Out. resol.	Input
Spatial attention	conv_1_1	1×1	1	c/$\frac{c}{2}$	$h\times w$	$h\times w$	feature
	conv_1_2	1×1	1	$\frac{c}{2}$/1	$h\times w$	$h\times w$	conv_1_1
	softmax	–	–	1/1	$h\times w$	$h\times w$	conv_1_2
Transform	matmul	–	–	c/c	$h\times w$	1×1	feature, softmax
	conv_2_1	1×1	1	c/$\frac{c}{4}$	1×1	1×1	matmul
	conv_2_2	1×1	1	$\frac{c}{4}$/c	1×1	1×1	conv_2_1
	sum	–	–	c/c	$h\times w$	$h\times w$	feature, conv_2_2

Table A2. Details of DAM (single attention module).

in the previous section, there are two configuration options where to deploy DAM in the encoder of MotionNet: DAM after every residual block, or residual layer. Fig. A1 illustrates each configuration. In Table A4, we compare them with the performance of monocular depth estimation after *phase-1* and *phase-3*. The results after *phase-1* show that the configuration of DAM after the last ResBlock produces the lowest error. However, after *phase-3*, both positional configurations of DAM show similar performance. Additionally, in Fig. A2, we provide the training trends of motion field consistency loss ($\mathcal{L}_{mc}$) for each DAM configuration: without DAM, DAM after every ResBlock, and DAM after the last ResBlock. We follow the training scheme of *phase-2*, however, to see distinctive results, we train MotionNet from random initialization with DepthNet pretrained from *phase-1*. The results show that the configuration with DAM after the

Type	Name	Kernel	Stride	Channel I/O	In. resol.	Out. resol.	Input
Encoder	conv_0	7×7	2	(3,3,1,1)/64	256×832	128×416	images, depth maps
	maxpool_1	3×3	2	64/64	128×416	64×208	conv_0
	ResBlock_1_1	–	1	64/64	64×208	64×208	maxpool_1
	ResBlock_1_2	–	1	64/64	64×208	64×208	ResBlock_1_1
	ResBlock_2_1	–	2	64/128	64×208	32×104	ResBlock_1_2
	ResBlock_2_2	–	1	128/128	32×104	32×104	ResBlock_2_1
	DAM_2	–	1	128/128	32×104	32×104	ResBlock_2_2
	ResBlock_3_1	–	2	128/256	32×104	16×52	DAM_2
	ResBlock_3_2	–	1	256/256	16×52	16×52	ResBlock_3_1
	DAM_3	–	1	256/256	16×52	16×52	ResBlock_3_2
	ResBlock_4_1	–	2	256/512	16×52	8×26	DAM_3
	ResBlock_4_2	–	1	512/512	8×26	8×26	ResBlock_4_1
	DAM_4	–	1	512/512	8×26	8×26	ResBlock_4_2
Ego-motion decoder	conv_1	1×1	1	512/256	8×26	8×26	DAM_4
	conv_2	3×3	1	256/256	8×26	8×26	conv_1
	conv_3	3×3	1	256/256	8×26	8×26	conv_2
	conv_4	1×1	1	256/6	8×26	8×26	conv_3
	avgpool	–	–	6/6	8×26	1×1	conv_4
Res-motion decoder	mofconv_0	1×1	1	512/3	8×26	8×26	DAM_4
	concat_1_1	–	–	(3,512)/515	8×26	8×26	mofconv_0, DAM_4
	conv_1_a	3×3	1	515/512	8×26	8×26	concat_1_1
	conv_1_b	3×3	1	515/512	8×26	8×26	concat_1_1
	concat_1_2	–	–	(512,512)/1024	8×26	8×26	conv_1_a, conv_1_b
	mofconv_1	1×1	1	1024/3	8×26	8×26	concat_1_2
	sum_1	–	–	(3,3)/3	8×26	8×26	mofconv_0, mofconv_1
	upsample_2	–	–	3/3	8×26	16×52	sum_1
	concat_2_1	–	–	(3,256)/259	16×52	16×52	upsample_2, DAM_3
	conv_2_a	3×3	1	259/256	16×52	16×52	concat_2_1
	conv_2_b	3×3	1	259/256	16×52	16×52	concat_2_1
	concat_2_2	–	–	(256,256)/512	16×52	16×52	conv_2_a, conv_2_b
	mofconv_2	1×1	1	512/3	16×52	16×52	concat_2_2
	sum_2	–	–	(3,3)/3	16×52	16×52	mofconv_2, upsample_2
	upsample_3	–	–	3/3	16×52	32×104	sum_2
	concat_3_1	–	–	(3,128)/131	32×104	32×104	upsample_3, DAM_2
	conv_3_a	3×3	1	131/128	32×104	32×104	concat_3_1
	conv_3_b	3×3	1	131/128	32×104	32×104	concat_3_1
	concat_3_2	–	–	(128,128)/256	32×104	32×104	conv_3_a, conv_3_b
	mofconv_3	1×1	1	256/3	32×104	32×104	concat_3_2
	sum_3	–	–	(3,3)/3	32×104	32×104	mofconv_3, upsample_3
	upsample_4	–	–	3/3	32×104	64×208	sum_3
	concat_4_1	–	–	(3,64)/67	64×208	64×208	upsample_4, ResBlock_1_2
	conv_4_a	3×3	1	67/64	64×208	64×208	concat_4_1
	conv_4_b	3×3	1	67/64	64×208	64×208	concat_4_1
	concat_4_2	–	–	(64,64)/128	64×208	64×208	conv_4_a, conv_4_b
	mofconv_4	1×1	1	128/3	64×208	64×208	concat_4_2
	sum_4	–	–	(3,3)/3	64×208	64×208	mofconv_4, upsample_4
	upsample_5	–	–	3/3	64×208	128×416	sum_4
	concat_5_1	–	–	3/67	128×416	128×416	upsample_5, conv_0
	conv_5_a	3×3	1	67/64	128×416	128×416	concat_5_1
	conv_5_b	3×3	1	67/64	128×416	128×416	concat_5_1
	concat_5_2	–	–	64/128	128×416	128×416	conv_5_a, conv_5_b
	mofconv_5	1×1	1	128/3	128×416	128×416	concat_5_2
	sum_5	–	–	(3,3)/3	128×416	128×416	mofconv_5, upsample_5
	upsample_6	–	–	3/3	128×416	256×832	sum_4
	concat_6_1	–	–	(3,3,3,1,1)/11	256×832	256×832	upsample_6, images, depth maps
	conv_6_a	3×3	1	11/8	256×832	256×832	concat_6_1
	conv_6_b	3×3	1	11/8	256×832	256×832	concat_6_1
	concat_6_2	–	–	(8,8)/16	256×832	256×832	conv_6_a, conv_6_b
	mofconv_6	1×1	1	16/3	256×832	256×832	concat_6_2
	sum_6	–	–	(3,3)/3	256×832	256×832	mofconv_6, upsample_6

Table A3. Details of MotionNet.

Models	*phase-1*		*phase-3*	
	all	*obj*	*all*	*obj*
without DAM	0.126	0.202	0.113	0.190
DAM after every ResBlock	0.122	0.199	0.111	**0.181**
DAM after the last ResBlock	**0.121**	**0.196**	**0.109**	0.182

Table A4. **Ablation study on different positional configurations of DAM.** We follow the same ablation scheme of Sec. 4.2. in our main paper, and measure the AbsRel errors after *phase-1* and *phase-3* on both *all* and *obj* areas. In *phase-3*, we regularize the residual motion field with CSAC.

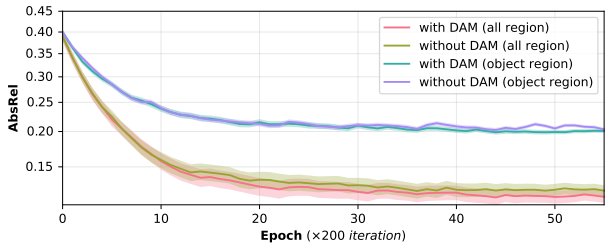

Figure A3. **Validation error trends in *phase-1* depending on the presence or absence of DAM in the motion encoder.** Models are trained and tested on the KITTI dataset, and we measure the AbsRel errors on both all and object areas.

$\{\alpha, \beta\}$	D1 (*fg*)	D2 (*fg*)
$\{50, 0.1\}$	34.1	41.3
$\{30, 0.2\}$	**32.5**	**35.7**
$\{10, 0.5\}$	33.8	37.2

Table A5. **Ablation study of different α and β on KITTI Scene Flow 2015 training set.** We set $\alpha = 30$ and $\beta = 0.2$ as our final model of CSAC.

last ResBlock converges fastest. From these ablation studies, we conjecture that the positional configuration of DAM has a relation with model complexity. Since the shared motion encoder performs two tasks, there could be a confusion from different self-supervisory signals. DAM helps to switch the extraction of motion feature in the encoder, however, the convergence will become slow if this switching mechanism is too complicated. Therefore, considering the training efficiency, we have chosen the configuration of DAM after the last ResBlock. Finally, we analyse the convergence trends of the networks in *phase-1* with and without DAM in Fig. A3 – applied on the objects only and on the entire image. We can notice a marginal improvement of the depth quality when DAM is employed. It underlines that during this training phase the motion encoder design only has little influence on the depth estimation. However, as underlined by other tests, DAM provides significant improvements on the overall system after the entire training process.

Different α and β for inlier score In Fig. A4, we visualize different mapping of the soft inlier score function $\mathcal{F}_{inlier}$ according to α and β to analyze their impacts. The graph shows that a high α value makes the curve sharp, and this leads to increase the discretization of the inlier scores. In Table A5, we conduct an ablation study to find the appropriate values of α and β. From the results, we conclude that both high discretization and tolerant mapping have adverse effects on residual motion learning. As a result of this ablation study, we selected $\alpha = 30$ and $\beta = 0.2$ for CSAC.

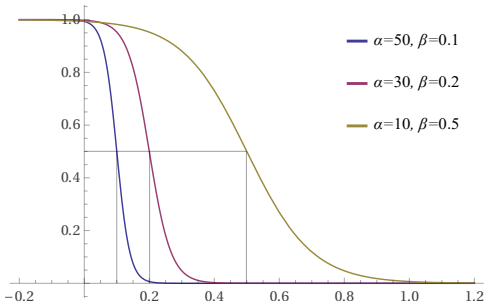

Figure A4. **Different inlier score mapping with α and β.** The steeper the curve, the more discretized, so we conduct ablations to find appropriate values.

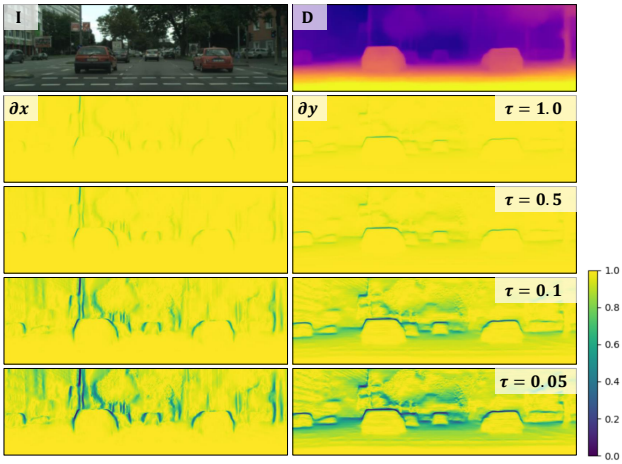

Figure A5. **Reparameterized gradient maps $e^{-\nabla D/\tau}$ depending on different τ.** We visualize image, its depth map, and gradients over the x and y axis. Smaller value of τ reduces gradients on the edges.

	$\tau = 1.0$	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.05$
Cityscapes-VIS	0.683	0.689	**0.712**	0.704

Table A6. **Motion segmentation results (mean IoU) with different τ.** This table is related (dataset and training scheme) to Table 5 of our main paper.

Reparameterized edge-aware motion smoothness We visualize the reparameterized gradients maps according to different τ in Fig. A5. The vanilla edge-aware smoothness term ($\tau = 1.0$) is not distinctive enough to leverage the boundary prior of objects. By reparameterizing with a small value of τ, we conjecture that the motion segmentation near the boundary of objects would be improved. We show motion segmentation results on different τ in Table A6.

Configuration of loss weights We summarize the learning

Phase	λ_p	λ_g	λ_s	λ_h	λ_{mr}	λ_{ms}	λ_{mp}	λ_{mc}
phase-1	1.0	1.0	0.1	0.2	–	–	–	–
phase-2	1.0	1.0	0.1	0.2	–	1.0	0.5	0.001
phase-3	1.0	1.0	0.1	0.2	0.2	1.0	1.0	0.001

Table A7. **Summarization of loss weights for each phase.**

λ_{mr}	0.1			0.2			0.3		
λ_{ms}	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0
all	0.123	0.119	0.116	0.116	**0.113**	**0.113**	0.119	0.123	0.128
obj	0.212	0.205	0.194	0.196	0.190	**0.188**	0.206	0.217	0.224

Table A8. **Ablation study of different loss weights, λ_{mr} and λ_{ms}, on the KITTI dataset.** We measure the AbsRel errors on both *all* and *obj* areas.

parameters for each dataset in Table A7. The photometric loss, $\mathcal{L}_p$, is defined as follows:

$$\mathcal{L}_p = \sum \left(\gamma_1 \left| \mathbf{I} - \hat{\mathbf{I}} \right|_1 + \gamma_2 (1 - SSIM(\mathbf{I}, \hat{\mathbf{I}})) \right), \qquad \text{(A1)}$$

where *SSIM* is the structural similarity loss [5], and $\{\gamma_1, \gamma_2\}$ is set to $\{0.3, 1.5\}$ based on cross-validation. Other parameters of our loss functions are the same as previous works as described in Sec. 3.4. of our main paper.

Different loss weights of λ_{mr} and λ_{ms} To justify the parametrization of the hyper-parameters λ_{mr} and λ_{ms}, we propose another ablation study conducted on the KITTI dataset. For this experiment, we follow the training scheme and dataset described in Sec. 4.2. of our main paper. As provided in Table A8, our motion regularization with CSAC shows stable training with $\lambda_{mr} = 0.2$ and $\lambda_{ms} = 1.0$.

C. Additional Experimental Results

In this section, we present additional experimental results as an extension of Sec. 4. of our main paper.

Full table of monocular depth estimation We provide the full results of our models for the task of monocular depth estimation in Table A9. The results consistently show that our proposed modules, DAM and CSAC, favorably work on three different dataset: KITTI, Cityscapes, and Waymo Open Dataset.

Qualitative results In addition to Fig. 7 in our main paper, we visualize the depth map and residual motion field in Fig. A6 and Fig. A7. Fig. A8 shows two representative effects of the regularization through CSAC: sharpen boundaries of object's motion, and motion hole filling in the homogeneous areas. Our module consistently preserves the shape of the moving objects, while the baseline model distorts the appearance of objects.

References

[1] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *ICCV Workshop*, 2019. 1

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 1, 7

[3] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *ICCV*, 2019. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[5] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 5

Method	Phase	Training	Test	Error metric ↓				Accuracy metric ↑		
				AbsRel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline	*phase-1*	K	K	0.126	0.975	5.244	0.211	0.849	0.946	0.979
Ours (DAM)	*phase-1*	K	K	0.123	0.920	5.212	0.205	0.854	0.945	0.978
Ours (DAM)	*phase-3*	K	K	0.120	0.895	4.972	0.196	0.859	0.950	0.980
Ours (DAM+CSAC)	*phase-3*	K	K	**0.114**	**0.876**	**4.715**	**0.191**	**0.872**	**0.955**	**0.981**
Baseline	*phase-1*	C+K	K	0.122	0.907	4.985	0.207	0.862	0.953	0.980
Ours (DAM)	*phase-1*	C+K	K	0.119	0.883	5.021	0.206	0.861	0.954	0.979
Ours (DAM)	*phase-3*	C+K	K	0.116	0.845	4.790	0.194	0.868	0.957	0.979
Ours (DAM+CSAC)	*phase-3*	C+K	K	**0.111**	**0.805**	**4.708**	**0.187**	**0.875**	**0.962**	**0.981**
Baseline	*phase-1*	C	C	0.128	1.322	6.942	0.198	0.833	0.949	0.978
Ours (DAM)	*phase-1*	C	C	0.127	1.330	6.903	0.196	0.838	0.950	0.979
Ours (DAM)	*phase-3*	C	C	0.124	1.281	6.818	0.189	0.849	**0.951**	0.981
Ours (DAM+CSAC)	*phase-3*	C	C	**0.116**	**1.213**	**6.695**	**0.186**	**0.852**	**0.951**	**0.982**
Baseline	*phase-1*	W	W	0.161	1.724	7.825	0.217	–	–	–
Ours (DAM)	*phase-1*	W	W	0.159	1.707	7.816	0.215	–	–	–
Ours (DAM)	*phase-3*	W	W	0.155	1.692	7.448	0.212	–	–	–
Ours (DAM+CSAC)	*phase-3*	W	W	**0.148**	**1.686**	**7.420**	**0.210**	–	–	–

Table A9. **Full results of our models for monocular depth estimation on KITTI (K) Eigen test set, Cityscapes (C) test set, and Waymo Open Dataset (W).** The models pretrained on Cityscapes and fine-tuned on KITTI are denoted by 'C+K'. For each partition, **Bold**: Best.

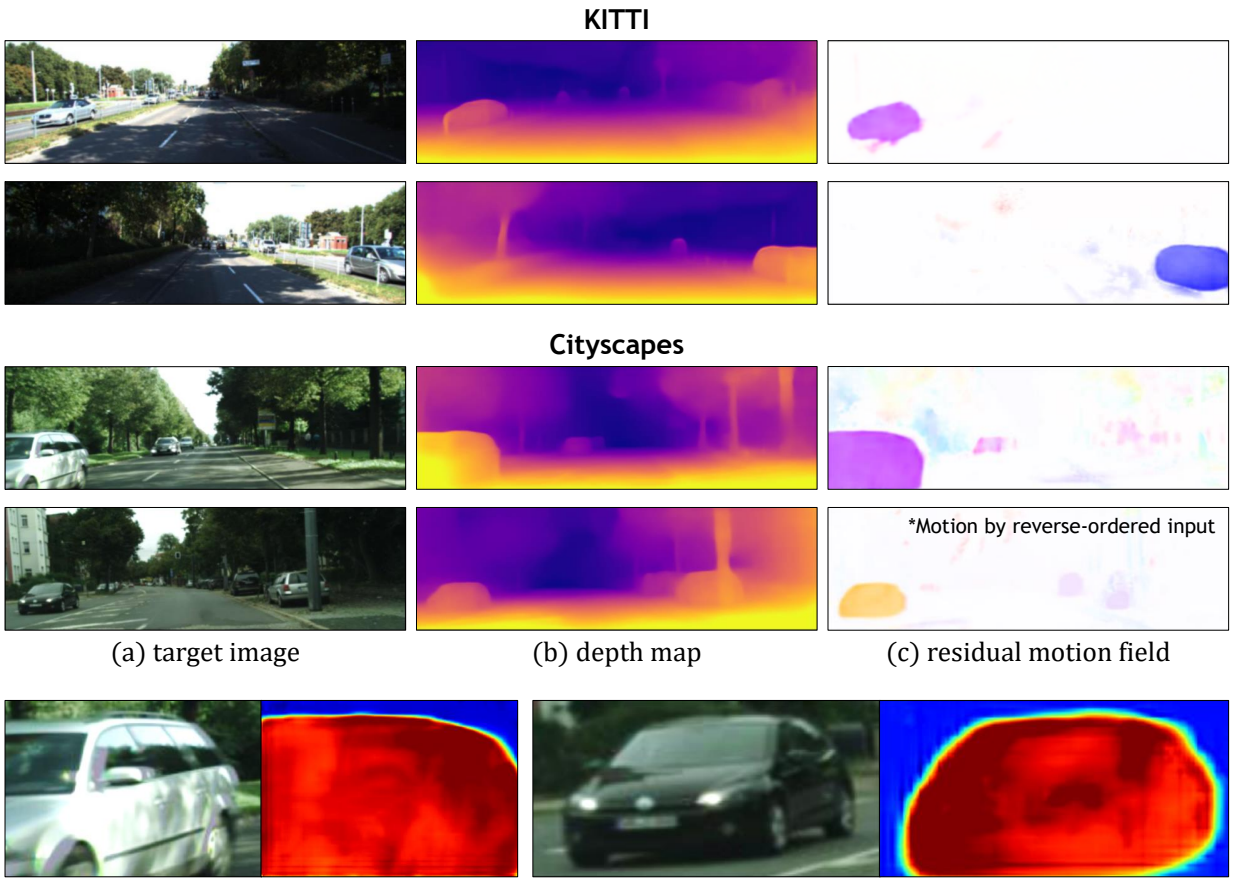

Figure A6. **Qualitative results of depth and residual motion estimation in KITTI and Cityscapes.** The residual motion field is mapped into the HSV color space.

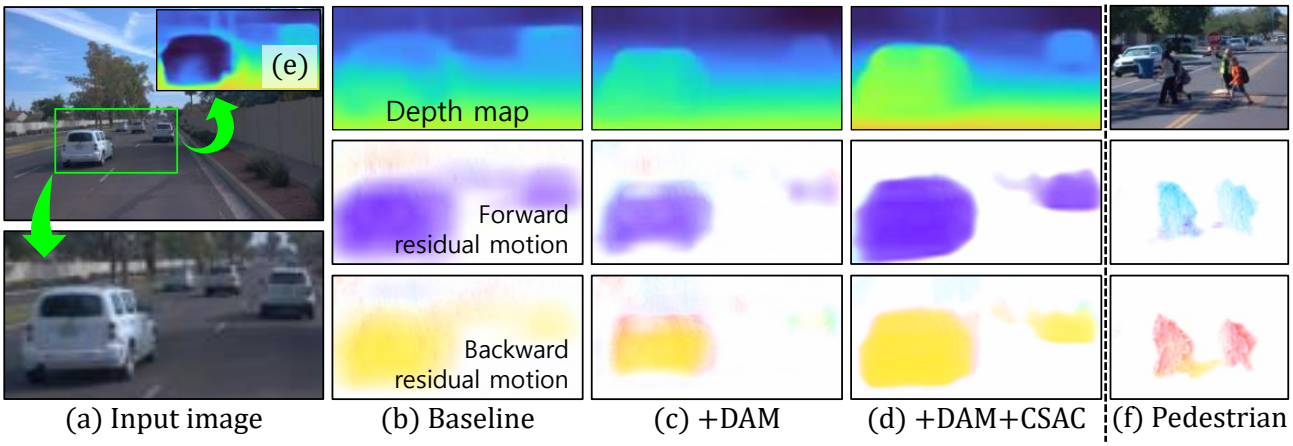

Figure A7. **Qualitative results of depth map and bidirectional residual motion field in Waymo Open Dataset.** AbsRel errors on *all* / *obj*: (b) 0.154 / 0.328, (c) 0.149 / 0.250 (improved distinction between *obj* and background), (d) 0.135 / 0.164 (sharpen object boundary). (e) Result of a diverged depth map, if auto-masking proposed by MonoDepth2 [2] fails. (f) Results of residual motion field for pedestrians.

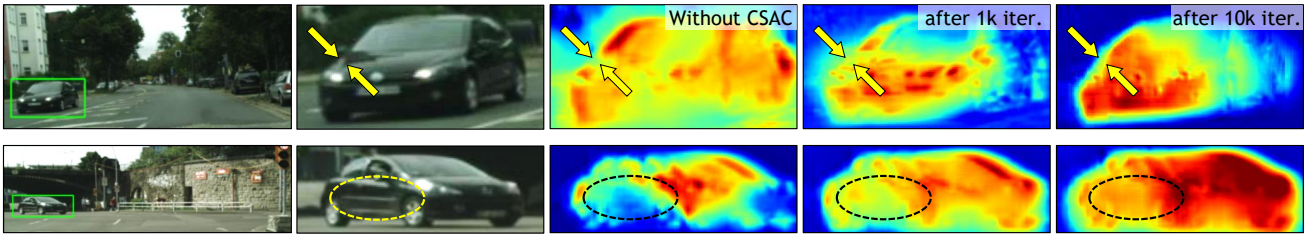

Figure A8. **Qualitative results of motion inliers in Cityscapes.** CSAC makes the motion boundaries more clear and sharper as shown in the first row, and the motion holes in homogeneous regions more consistent on the rigid objects as demonstrated in the second row.