U^{2}Flow: Uncertainty-Aware Unsupervised Optical Flow Estimation

Supplementary Material

A. Method details

A.1. Recurrent update block

Fig. 1 provides a detailed illustration of our recurrent update block, highlighting the processing flow within a single iteration k.

At each iteration, the block takes three inputs: the previous optical flow estimate $\mathbf{F}_{1\to2}^{(k-1)}$, the context feature extracted from the reference image, and the correlation feature retrieved from the 4D correlation volume using the previous flow estimate. The flow estimate and correlation feature are fused into a motion feature, which is then concatenated with the context feature. This combined tensor is fed into a Gated Recurrent Unit (GRU), which updates its hidden state to $\mathbf{h}^{(k)}$, with the initial state $\mathbf{h}^{(0)}$ initialized from the context feature.

From the updated hidden state, our prediction heads generate two outputs: a residual flow update $\Delta\mathbf{F}_{1\to2}^{(k)}$ and a corresponding per-pixel uncertainty map $\boldsymbol{\sigma}_{1\to2}^{2\,(k)}$. Finally, a learned upsampling module upsamples the refined flow and uncertainty to the full resolution of the input image.

A.2. Augmentation and uncertainty supervision

Fig. 2 provides a schematic overview of our self-supervision strategy, which generates the signals for both the augmentation loss (ℓ_{ar}) and the uncertainty supervision loss (ℓ_{unc}).

Specifically, the process begins by computing an initial flow estimate, $\mathbf{F}_{1\to2}$, for an image pair $(\mathbf{I}_1, \mathbf{I}_2)$ in a forward pass. Subsequently, we apply a set of strong appearance augmentations (e.g., color jitter, contrast adjustment, Gaussian noise, random erase) and spatial transformations (e.g., translation, rotation, rescaling) to both the images and the flow field. This produces an augmented pair $(\hat{\mathbf{I}}_1, \hat{\mathbf{I}}_2)$ and a transformed pseudo-ground-truth flow $\hat{\mathbf{F}}_{1\to2}$. The network then re-estimates the flow for the augmented pair, yielding a new prediction $\hat{\mathbf{F}}'^{(k)}_{1\to2}$ at iteration k.

The per-pixel ℓ_1 distance between these two flow fields, denoted as $\hat{D}^{(k)}(\boldsymbol{p}) = \|\hat{\mathbf{F}}_{1\to2}(\boldsymbol{p}) - \hat{\mathbf{F}}'^{(k)}_{1\to2}(\boldsymbol{p})\|_1$, serves as the self-supervised target, as it directly captures the model's predictive inconsistency under perturbation. This target inconsistency, $\hat{D}^{(k)}$, is then leveraged in two distinct ways, as detailed in Eq. 9 and Eq. 10 in the main paper:

- **For the uncertainty loss (ℓ_{unc}):** The inconsistency serves as the supervisory signal within our maximum likelihood objective to train the uncertainty head.
- **For the augmentation loss (ℓ_{ar}):** We directly minimize this inconsistency to enforce model robustness against perturbations.

In addition to these augmentations, and as described in the main paper, we also incorporate a semantic augmentation loss, ℓ_{sem} [13]. This loss follows a similar formulation to ℓ_{ar} but is computed on semantically augmented data, where object regions are randomly copied and pasted between image pairs. This process introduces more realistic occlusions and compositional variations, further enhancing model robustness.

B. Result details

B.1. Implementation details

Our model is tested on Ubuntu 18.04 with Python 3.7.16, PyTorch 1.13.1, Torchvision 0.14.1, and CUDA 11.3. We use 2 NVIDIA RTX 3090 GPUs with 24 GB of memory each. More details can be found in the code appendix.

B.2. Efficiency analysis

We evaluate the computational efficiency of our network. For each RGB sample of resolution 376×1242, our model achieves an average inference time of 0.0663 seconds ($\sim$15FPS) on an NVIDIA RTX 3090 GPU. In terms of model size, our network contains 5.22M parameters, making it marginally more compact than the original RAFT [11] architecture (5.26M). This parameter efficiency is attributed to the lightweight design of our proposed flow refinement module and uncertainty estimation head.

B.3. Ablation study on the number of recurrent iterations

To analyze the impact of the recurrent refinement process, we performed an ablation study on the number of iterations (K) using the Sintel (final pass) and KITTI-2015 training sets. The results are summarized in Table 1.

As expected, the accuracy of the optical flow estimates generally improves with an increasing number of iterations across both datasets. The performance gains exhibit diminishing returns and begin to saturate at approximately $K = 12$ iterations. This trend confirms the effectiveness of the iterative refinement process while justifying our choice of $K = 12$ for the final model to balance performance and efficiency.

Interestingly, the quality of our uncertainty estimation, measured by AUSE, remains stable across different values of K. This suggests that the uncertainty head relies on intrinsic feature properties rather than the final converged flow, demonstrating the robustness of our uncertainty learning mechanism.

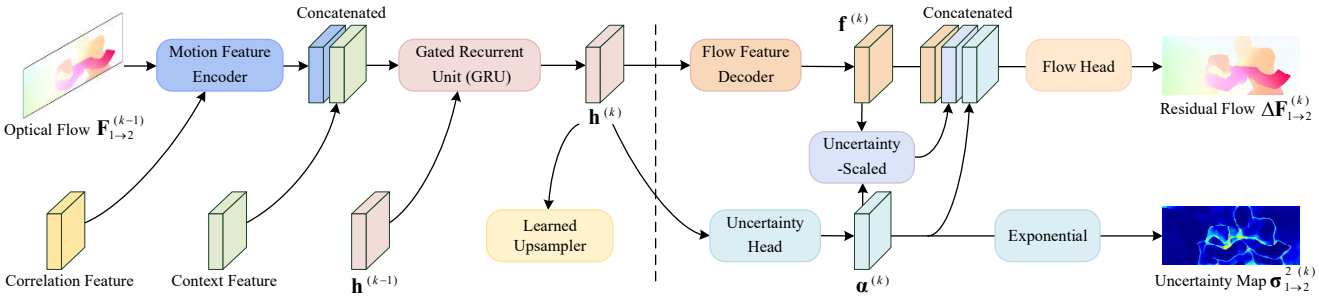

Figure 1. Detailed illustration of the recurrent update block. The diagram depicts the process flow for the k-th iteration. Components to the left of the dashed line represent the original RAFT [11] architecture, while components to the right constitute our proposed flow refinement module and uncertainty estimation head.

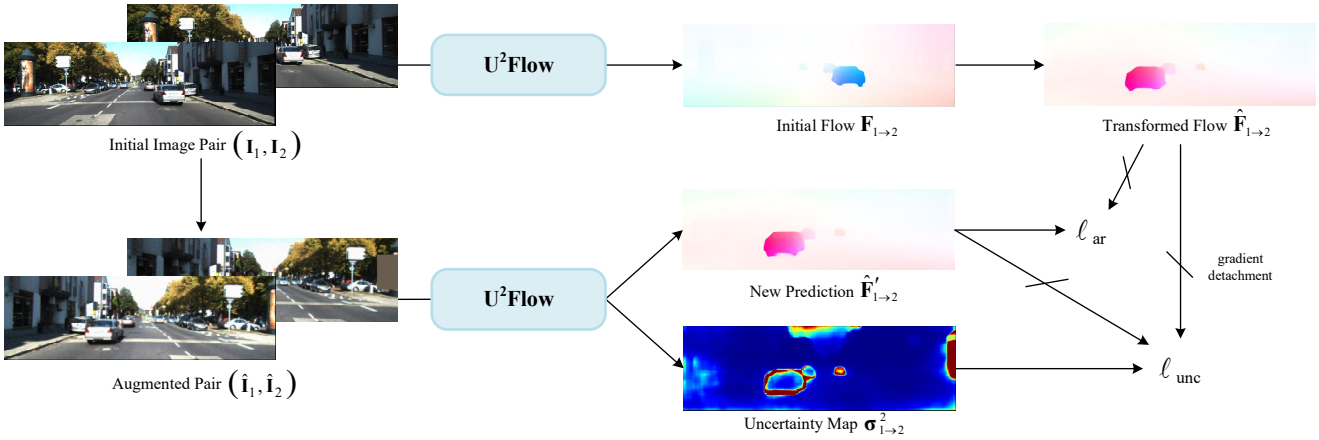

Figure 2. Schematic diagram of the self-supervision process. (1) An initial flow $\mathbf{F}_{1\to2}$ is predicted from the original image pair. (2) Strong augmentations are applied to both the images and the flow, creating an augmented image pair and a transformed pseudo-ground-truth flow $\hat{\mathbf{F}}_{1\to2}$. (3) A new flow prediction $\hat{\mathbf{F}}'_{1\to2}$ is generated from the augmented images. (4) The inconsistency $\hat{D}$ between the pseudo-ground-truth and the new prediction serves as the target signal for both the augmentation loss (ℓ_{ar}) and the uncertainty supervision loss (ℓ_{unc}).

B.4. Ablation study on augmentation sensitivity

To further analyze the impact of different data augmentation strategies on our method, we provide an ablation study on augmentation types in Tab. 2. We observe that spatial augmentation is the dominant factor for reliable uncertainty learning: using *Spatial Only* achieves performance comparable to *Full* (e.g., AUSE 0.12 vs. 0.11 on Sintel), while *Appearance Only* degrades significantly (AUSE 0.46). Appearance augmentation offers negligible gain on KITTI and a minor improvement on Sintel, likely due to its synthetic lighting.

This is consistent with optical flow being primarily a geometric correspondence task, where spatial perturbations provide stronger uncertainty supervision than photometric variations.

B.5. Ablation study on uncertainty scaling

In Eq. 4 of the main paper, we introduce an uncertainty-scaled flow feature $\tilde{\mathbf{f}}^{(k)} = \mathbf{f}^{(k)} \odot \mathbf{s}^{(k)*}$ to modulate flow representations according to the predicted reliability. This operation suppresses unreliable features before propagation, enabling the refinement module to focus on more reliable regions. To evaluate the effectiveness of this design, we compare two variants in Eq. 5:

- **w/o Uncertainty-Scaled**: directly concatenate the flow feature $\mathbf{f}^{(k)}$ with the uncertainty prediction $\boldsymbol{\alpha}^{(k)}$, leaving the network to implicitly learn how to utilize the uncertainty information.
- **With Uncertainty-Scaled (ours)**: use the uncertainty-scaled feature $\tilde{\mathbf{f}}^{(k)} = \mathbf{f}^{(k)} \odot \mathbf{s}^{(k)*}$ to explicitly modulate the flow representation before feature fusion.

As shown in Tab. 3, the proposed uncertainty scaling consistently improves performance, indicating that explicitly modulating flow features with predicted uncertainty is more effective than relying on implicit feature fusion.

B.6. Analysis of the uncertainty-enhanced homography smoothness loss

The homography smoothness loss [13] regularizes flow estimation by enforcing consistency with a planar motion model. In our framework, we employ an uncertainty-based reliability mask to select the regions where this loss is ap-

Iter. Num.	Sintel Final			KITTI 2015				
	EPE ↓	AUSE ↓	CC ↑	EPE ↓	Fl-all ↓	AUSE ↓	CC ↑	Runtime
2	3.52	0.11	0.69	3.71	12.31	0.20	0.53	18 ms
4	2.75	0.10	0.69	2.32	8.23	0.21	0.50	35 ms
8	2.41	0.11	0.68	1.90	6.75	0.22	0.48	52 ms
12	2.32	0.11	0.67	1.83	6.59	0.22	0.48	66 ms
16	2.30	0.11	0.67	1.82	6.53	0.22	0.47	97 ms

Table 1. Ablation study on the number of recurrent iterations (K). Flow accuracy improves and saturates with more iterations, while uncertainty estimation (AUSE) remains robustly stable. Runtime is measured on KITTI-2015.

		EPE all	EPE matched	EPE unmatched	d0-10	d10-60	d60-140	s0-10	s10-40	s40+
U2Flow+FF [210]	Final	4.098	1.805	22.793	3.821	1.414	1.107	0.810	2.499	24.280
U2Flow [215]	Final	4.157	1.800	23.383	3.798	1.412	1.108	0.817	2.716	24.226
U2Flow+FF [229]	Clean	2.829	0.941	18.246	2.370	0.698	0.458	0.453	1.529	17.785
U2Flow [230]	Clean	2.830	0.940	18.261	2.369	0.696	0.457	0.454	1.532	17.782

Figure 3. Detailed test results of our model on Sintel [1].

Method	Sintel		KITTI	
	AUSE ↓	CC ↑	AUSE ↓	CC ↑
Appearance Aug. Only	0.46	0.25	0.25	0.51
Spatial Aug. Only	0.12	0.65	0.12	0.64
Full (Ours)	**0.11**	**0.66**	**0.12**	**0.64**

Table 2. Ablation study of augmentation strategies for uncertainty.

Error	Fl-bg	Fl-fg	Fl-all
All / All	4.87	11.64	6.00
All / Est	4.87	11.64	6.00
Noc / All	3.53	9.00	4.52
Noc / Est	3.53	9.00	4.52

U²Flow + FF

Error	Out-Noc	Out-All	Avg-Noc	Avg-All
2 pixels	6.07 %	9.96 %	0.8 px	1.3 px
3 pixels	3.47 %	6.26 %	0.8 px	1.3 px
4 pixels	2.38 %	4.51 %	0.8 px	1.3 px
5 pixels	1.80 %	3.50 %	0.8 px	1.3 px

U²Flow + FF

Error	Fl-bg	Fl-fg	Fl-all
All / All	5.04	11.55	6.13
All / Est	5.04	11.55	6.13
Noc / All	3.58	8.94	4.56
Noc / Est	3.58	8.94	4.56

U²Flow

Error	Out-Noc	Out-All	Avg-Noc	Avg-All
2 pixels	6.08 %	10.03 %	0.8 px	1.4 px
3 pixels	3.48 %	6.37 %	0.8 px	1.4 px
4 pixels	2.40 %	4.63 %	0.8 px	1.4 px
5 pixels	1.82 %	3.64 %	0.8 px	1.4 px

U²Flow

(a) KITTI-2015 results (b) KITTI-2012 results

Figure 4. Detailed test results of our final model on KITTI [2, 8].

Model Variant	Sintel		KITTI 2015	
	Final	Clean	EPE	Fl-all
w/o Uncertainty-Scaled	2.40	1.52	2.04	7.24
w/ Uncertainty-Scaled (Ours)	**2.32**	**1.42**	**1.83**	**6.59**

Table 3. Ablation study on the uncertainty scaling mechanism. Explicitly modulating flow representations with the scaled feature $\tilde{\mathbf{f}}^{(k)}$ yields better performance than implicit direct concatenation.

plied. In this section, we further elaborate on the dataset-dependent performance of this uncertainty-enhanced homography smoothness loss (ℓ_{hg}), which, as shown in the main paper, improves results on KITTI but degrades them on Sintel.

Our investigation reveals that the core issue is not the quality of the uncertainty mask itself, but rather the nature of the high-uncertainty regions within each dataset. As illustrated in Fig. 5, the uncertainty-based reliability mask effectively identifies regions with high estimation error on Sintel. However, in the Sintel dataset, these high-uncertainty regions frequently correspond to characters and creatures undergoing complex, non-rigid, and non-planar motion (e.g., walking, running, or fighting). Applying the homography smoothness loss to these areas forces the network to regularize the flow field based on a planar motion assumption, which is geometrically invalid for such movements. This fundamental mismatch introduces erroneous constraints, ultimately degrading the overall flow estimation quality rather than refining it.

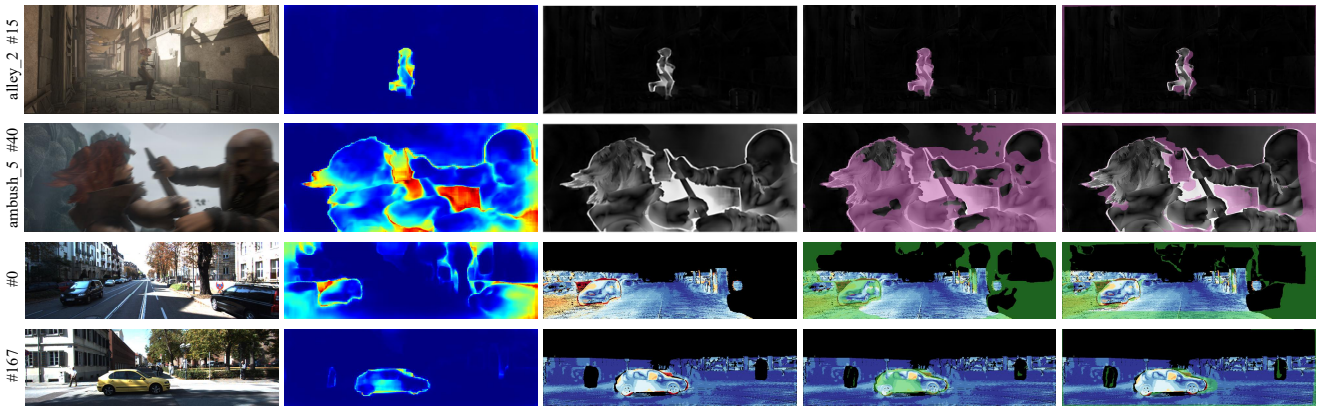

Figure 5. Visualization of selected regions for homography smoothness on Sintel and KITTI. The top two rows show samples from the Sintel dataset, while the bottom two rows are from KITTI. The uncertainty-based reliability mask successfully identifies regions with high flow error in both datasets. However, for Sintel, these high-error regions are typically non-rigid characters for which the planar homography assumption does not hold. In contrast, for KITTI, high-error regions often correspond to rigid vehicles and planar road surfaces, which are well-suited for homography regularization. This illustrates why the uncertainty-enhanced ℓ_{hg} is effective on KITTI but not on Sintel.

In contrast, the forward–backward consistency check [7], which is mainly used as an occlusion detector, tends to highlight static or slow-moving background regions that become occluded. As shown in Fig. 5, these background areas (e.g., walls, ground) are typically planar and thus suitable for homography regularization.

The situation in the KITTI dataset is markedly different. The scenes are dominated by rigid objects (vehicles) and large planar surfaces (roads, buildings). Even the primary moving objects like cars are composed of multiple planar surfaces. Consequently, the planar motion assumption holds true for many high-uncertainty regions, making the uncertainty-enhanced ℓ_{hg} a highly effective regularizer that provides a significant performance boost.

B.7. Benchmark test screenshots

The results of U^{2}Flow and U^{2}Flow (+FF) have been submitted to the KITTI and Sintel online benchmarks and can be found on the leaderboards by searching for "U^{2}Flow". Screenshots of the detailed evaluation metrics from the official websites are provided in Fig. 3 and Fig. 4.

B.8. Results on the Spring Dataset

To further demonstrate the generalization capability of our approach, we evaluate our model on the Spring [6] test set without fine-tuning. As shown in Tab. 4, our U^{2}Flow, as an unsupervised method, consistently outperforms SMURF [10] across most metrics (below the dashed line). Moreover, compared with supervised approaches (above the dashed line), which are trained in a multi-stage manner on multiple datasets, our model—trained exclusively on Sintel [1]—still demonstrates strong generalization ability and surpasses them on several key metrics.

We also provide qualitative results on the Spring dataset in Fig. 6. As illustrated, the predicted flow fields are spatially coherent and preserve sharp motion boundaries. In addition, the predicted uncertainty maps show a reasonable correlation with the estimation errors, suggesting that the model can capture the reliability of its predictions.

B.9. More qualitative examples

We present additional qualitative results from the KITTI-2015 test set (Fig. 7) and the Sintel test set (Fig. 8). We compare our method with the state-of-the-art unsupervised two-frame approach UPFlow [5] and the multi-frame trained model SMURF [10], which is also based on RAFT [11].

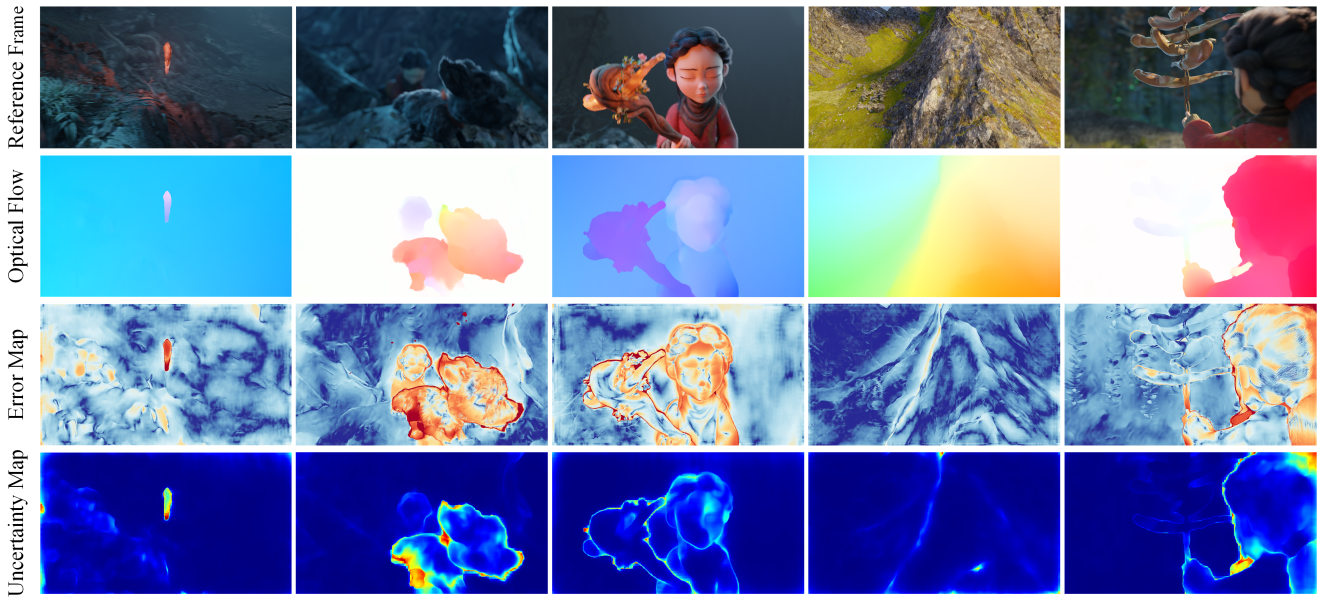

Figure 6. Qualitative results on the Spring benchmark [6] without fine-tuning.

Model	1px												EPE	Fl	WAUC
	total	low-det.	high-det.	matched	unmat.	rigid	non-rig.	not sky	sky	s0-10	s10-40	s40+			
RAFT [11]	6.79	6.43	**64.09**	6.00	39.48	4.11	**27.09**	**5.25**	30.18	**3.13**	**5.30**	41.40	1.476	3.20	90.92
GMA [4]	7.07	6.70	66.20	6.28	39.89	4.28	28.25	5.61	29.26	3.65	5.39	40.33	0.914	3.08	90.72
GMFlow [12]	10.36	9.93	76.61	9.06	63.95	6.80	37.26	8.95	31.68	5.41	9.90	52.94	0.945	2.95	82.34
FlowFormer [3]	6.51	6.14	64.22	5.77	**37.29**	**3.53**	29.08	5.50	21.86	3.38	5.53	35.34	0.723	2.38	**91.68**
SMURF [10]	6.43	6.04	66.52	5.49	45.12	3.58	27.95	5.53	20.07	3.44	6.59	30.99	0.659	2.10	90.45
U^{2}Flow (Ours)	**6.32**	**5.94**	66.02	**5.38**	45.31	3.56	27.26	5.46	**19.38**	3.41	6.54	**30.13**	**0.608**	**1.88**	89.32

Table 4. Optical flow generalization results on the Spring benchmark [6] without fine-tuning. We report the 1px outlier rate across low/high-detail, (un)matched, (non-)rigid, and (non-)sky regions, along with the EPE, Fl error, and WAUC [9] metrics. The best and second-best results are highlighted in bold and underline, respectively, while key metrics are emphasized in blue.

References

[1] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 611–625, 2012. 3, 4

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013. 3

[3] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–685, 2022. 5

[4] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9772–9781, 2021. 5

[5] Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. Upflow: Upsampling pyramid for unsupervised optical flow learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1045–1054, 2021. 4

[6] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4981–4991, 2023. 4, 5

[7] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 4

[8] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 3

[9] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE In-*

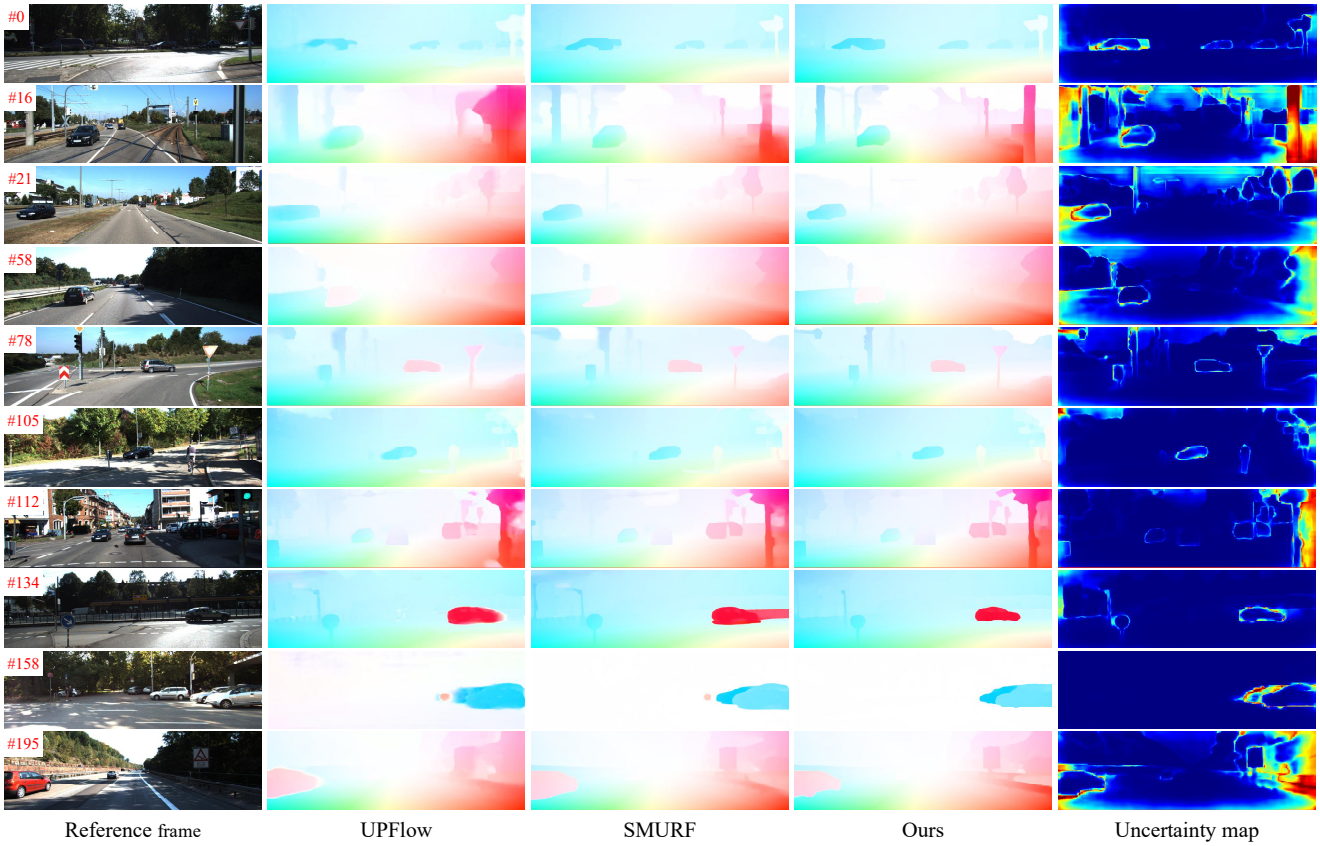

Figure 7. Additional qualitative results on the KITTI-2015 test set.

ternational Conference on Computer Vision (ICCV), pages 2213–2222, 2017. 5

[10] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3887–3896, 2021. 4, 5

[11] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 1, 2, 4, 5

[12] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8121–8130, 2022. 5

[13] Shuai Yuan, Lei Luo, Zhuo Hui, Can Pu, Xiaoyu Xiang, Rakesh Ranjan, and Denis Demandolx. Unsamflow: Unsupervised optical flow guided by segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19027–19037, 2024. 1, 2

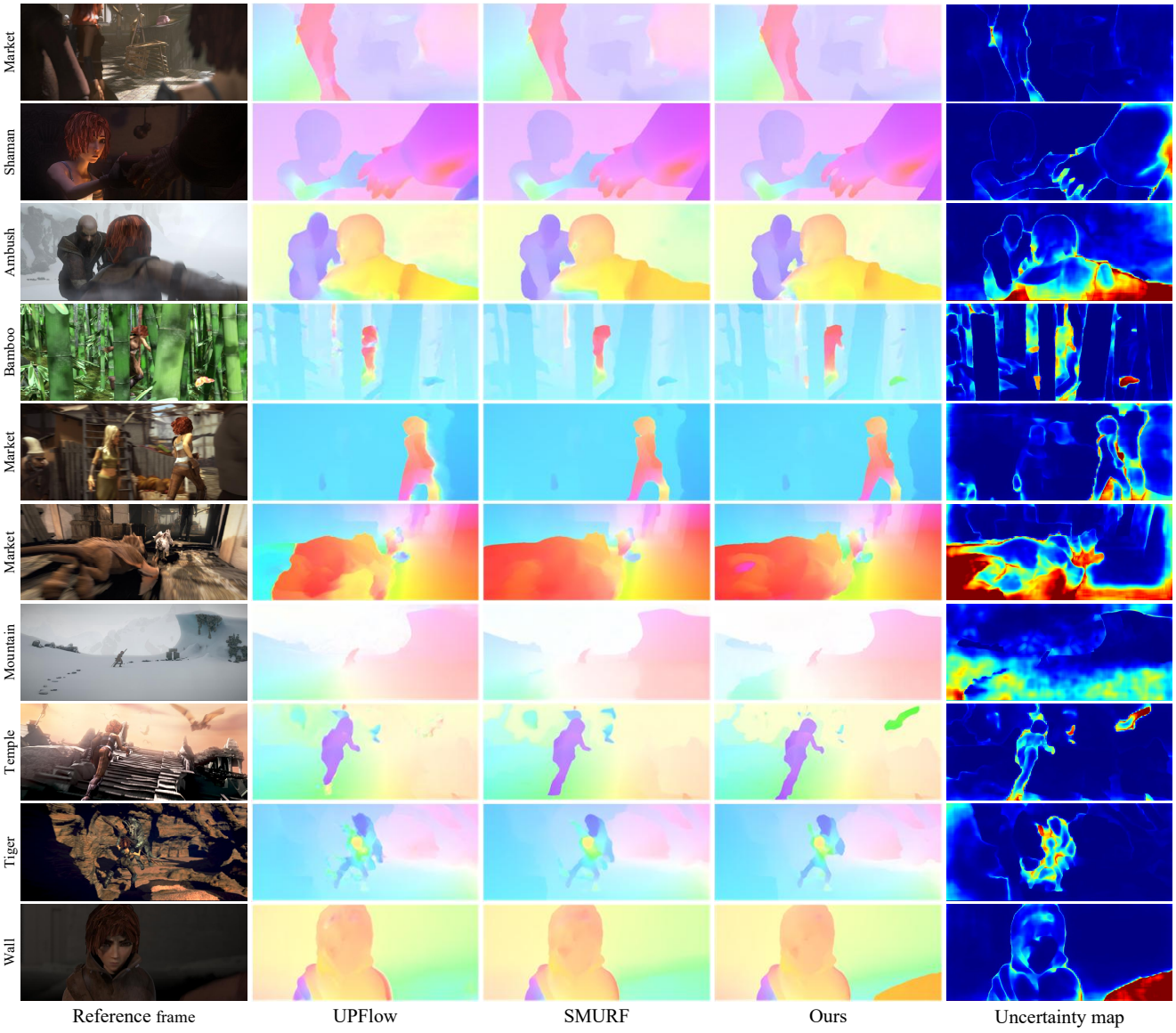

Figure 8. Additional qualitative results on the Sintel (final pass) test set.