ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447321

SELF-SUPERVISED MULTI-SCALE HIERARCHICAL REFINEMENT METHOD FOR JOINT LEARNING OF OPTICAL FLOW AND DEPTH

Rokia Abdein[⋆†], *Xuezhi Xiang*[⋆†], *Yiming Chen*[⋆†], *Mingliang Zhai*[§], *Abdulmotaleb El Saddik* [‡]

[⋆] School of Information and Communication Engineering, Harbin Engineering University,
Harbin, 150001, China.
[†] Key Laboratory of Advanced Marine Communication and Information Technology, Ministry
of Industry and Information Technology, Harbin, China.
[§] School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China.
[‡]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa,
ON K1N 6N5, Canada.

ABSTRACT

Recurrently refining the optical flow based on a single high-resolution feature demonstrates high performance. We exploit the strength of this strategy to build a novel architecture for the joint learning of optical flow and depth. Our proposed architecture is improved to work in the case of training on unlabeled data, which is extremely challenging. The loss is computed for the iterations carried out over a single high-resolution feature, where the reconstruction loss fails to optimize the accuracy particularity in occluded regions. Therefore, we propose to hierarchically refine the optical flow across multiple scales while feeding the rigid flow calculated from depth and camera pose to provide more refinement. We further propose a self-supervised patch-based similarity loss to be optimized with the reconstruction loss to improve accuracy in the occluded regions. Our proposed method demonstrates efficient performance on the KITTI 2015 dataset, with more improvement in the occluded regions.

Index Terms— Optical flow estimation, depth estimation, joint learning, self-supervised learning, occlusion handling.

1. INTRODUCTION

The joint learning of optical flow and depth is important to provide more information to reconstruct the scene flow, which is useful for different applications such as autonomous vehicles and mobile robots. However, improving them through joint learning is challenging since each task is constrained by the other, which hinders them from getting an optimal solution, as if they were estimated separately. On the other hand, resolving some issues with one task and improving its performance could also provide useful cues to enhance the other task. Various methods for training optical flow and depth methods rely on the availability of the ground truth to provide direct and accurate supervision for their models. Since providing enough annotated and real-world datasets is expensive in laboratories, training on an unannotated dataset can be considered a good alternative, but it comes at the expense of providing more constraints to bridge the gap with the supervised methods. The training on an unannotated dataset heavily depends on reducing the difference between the reference image and its reconstruction from the estimated optical flow or depth. The occluded parts are naturally excluded, as they cause a significant error.

Different network architectures that have been adopted for the supervised learning of optical flow have been adapted to work under the unsupervised situation, such as OAFlow [1] which adopts the FlowNet [2] as the baseline with occlusion handling techniques to avoid their direct influence on the estimated flow. ARFlow [3] utilizes the PWC-Net [4] with more advanced techniques to allow the network to learn the flow in occluded regions and to be more effective and reliable in challenging scenes. SMURF [5] improved the RAFT [6] architecture by proposing a full-image warping operation to reduce the occlusions that result from out-of-frame motion. Therefore, the occlusion is considered the critical part that should be carefully handled in the absence of supervision from the ground truth.

In the joint learning of optical flow and depth, this problem exacerbates since the wrongly estimated regions have a mutual negative impact on each task. Self-Mono-SF [7] uses the PWC-Net as a baseline, infers the occlusion cues via the bi-directional estimation process, and involves it in an occlusion-aware loss for scene flow and disparity estima-

This work was supported in part by National Natural Science Foundation of China under Grant 62271160, in part by the Natural Science Foundation of Heilongjiang Province of China under Grant LH2021F011, in part by the Fundamental Research Funds for the Central Universities of China under Grant 3072022TS0801, in part by the Key Laboratory Open Fund under Grant AMCIT2103-03.

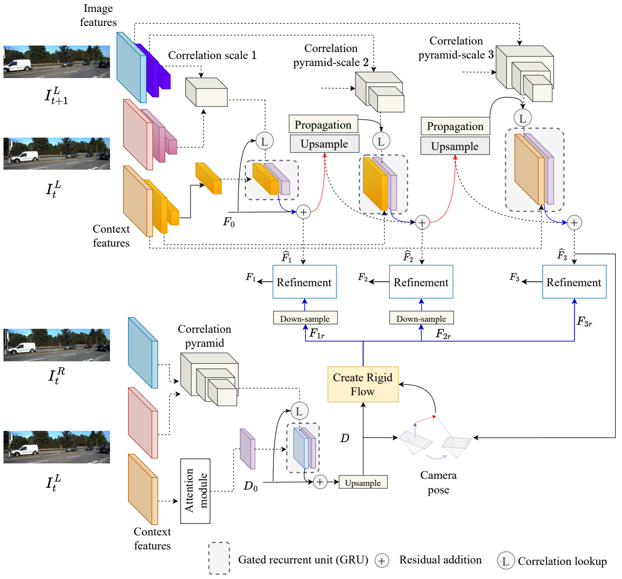

Fig. 1: Our proposed architecture for the joint learning of optical flow and depth. D is the depth, F is the optical flow, and F_r is the rigid flow.

tion. DRAFT [8] utilizes the RAFT as a baseline, initializes the triangulated depth from the optical flow, and estimates an occlusion mask to remove spurious information from the self-supervised and consistency losses. In this paper, we propose a novel architecture for the joint learning of depth and optical flow. Our architecture joins the strength of the RAFT by utilizing the global correlation, refining the optical flow using the gated recurrent unit (GRU), and the strength of the pyramid scheme of the PWC-Net by performing this refinement on different scales. The key insight is to avoid the significant error related to refining the optical flow over a single high-resolution feature that resulted from the training on unlabeled datasets using the reconstruction loss that ignores the occluded regions. In addition, our architecture adopts a single-iteration multi-scale strategy instead of a multi-iteration single-scale strategy for the training on an unlabeled dataset, where increasing the number of iterations will bring more computation overhead without significant improvement. The reason is that the error that resulted from the first iteration due to the occlusion is high, and increasing the number of iterations cannot remedy it. Instead of increasing the iterations, we use the depth to refine the optical flow, which demonstrates a remarkable enhancement in the occluded regions, as shown by our experiments. Besides this architecture, we propose a self-supervised patch-based similarity loss that is added to the reconstruction loss to reduce the error in the occluded regions. The key idea is to reduce the difference between the estimated optical flow in the occluded regions and their most similar neighbors in a patch-based scheme. We can summarize our contributions as follows:

- We propose a novel multi-scale hierarchical refinement architecture with a multi-scale single-iteration strategy for the joint learning of optical flow and depth on an unlabeled dataset that can hierarchically refine the optical flow across multiple scales, allowing the network to learn better optical flow, particularly in occluded regions.

- We propose a novel self-supervised patch-based similarity loss to ensure consistency between the matched and unmatched regions, utilizing a kernel-like searching scheme that demonstrates its efficiency in alleviating the occlusion impact on the estimated optical flow.

- Our proposed method achieves competitive results compared to the state-of-the-art methods for the joint learning of optical flow and depth tasks on the KITTI 2015 scene flow dataset.

2. METHOD

2.1. Proposed architecture

Fig. 1 shows the overall architecture of our proposed method. We jointly train two networks, the optical flow and depth networks. For the depth network, the input is a stereo image, and only single-scale single-iteration is adopted to estimate the depth map. The attention module [9] is used for flow propagation. Working on a single high-resolution scale provides a simple yet fast network. The occluded regions due to the left-right shift between the stereo pairs have no significant effect as in the occluded regions of the optical flow since the displacement is not large. In the optical flow, the occluded regions increase with increasing resolution, thereby increasing the error. Therefore, we propose to add multiple scales with only one iteration in order to allow the network to catch the large movements from the lower resolutions and use them to refine the higher resolution with the GRU. Three feature scales are extracted: $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the size of the original image. We utilize the all-to-all correlation as proposed in RAFT [6] but with different pyramid scales according to the feature size. The correlation pyramid has scales 4, 3, and 1 for the three scales, from higher to lower, respectively. Each correlation pyramid is passed along with the context features to the GRU to estimate the optical flow. The estimated flow at each scale is added to the upsampled flow of the previous scale after multiplying the upsampled flow with an estimated mask from the reference image to avoid inaccurate propagation. The estimated depth and optical flow at the higher scale are used to estimate the camera pose using the PnP method [10] which is then used along with the depth to find the rigid flow. Finally, we use the rigid flow to refine the optical flow by minimizing the difference between the optical flow and the rigid flow in the static regions where the rigid flow is more accurate, as suggested by [10].

2.2. Training losses

The reconstruction losses are utilized for training on the unlabeled dataset for both the optical flow and depth networks, as proposed by [10]. Beside these losses, we remedy the occluded parts in the optical flow by reducing the difference between the estimated flow of the occluded pixels and the estimated flow of their neighbors from the non-occluded regions that have the largest similarity with them:

$$L_{sim} = \sum_{i,j \in R} Sim(P_i^{occ}, P_j^{noc}) \ . \ \|F_i - F_j\|_2, \qquad (1)$$

where

$$Sim(P_i^{occ}, P_j^{noc}) = 1 - exp(-\frac{P_i^{occ} . \ P_j^{noc}}{\epsilon}), \qquad (2)$$

where P_i^{occ} is the occluded pixel at index i of the reference image R, P_j^{noc} is the non-occluded pixel at index j of the reference image R, and the softmax temperature ϵ is a hyperparameter and set to 0.45. F_i and F_j are the estimated optical flows for the pixels at indices i and j respectively. However, this per-pixel operation is extremely expensive and slows down the training process. Therefore, we propose patch-based loss. In this loss, we divide the image into patches, and a kernel with a size of 3×3 is moved over these patches, where each location in this kernel is a patch of pixels. Then, we check the similarity between the center patch and its neighbor patches and multiply it by the difference between the optical flow of the center patch and its neighbor patches to calculate the loss as follows:

$$L_{sim} = \sum_{j \in N} Sim(P_i, P_j) \ . \ \|F_{pi} - F_{pj}\|_2, \qquad (3)$$

where P_i is the center patch and P_j are the neighbor patches N which are 8 patches according to our 3×3 kernel. F_{pi} is the optical flow for the pixels in the center patch, and F_{pj} is the optical flow for the neighbor patches. The similarity is calculated as in the equation 2. We set the flow values of the occluded pixels to zero to avoid their influence on the neighboring pixels.

3. EXPERIMENTAL RESULTS

3.1. Settings

Our method is trained on the unlabeled KITTI raw dataset [11] that has a total of 28,968 images out of 42,382 images after excluding the images that are included in the KITTI 2015 training dataset. For optical flow and depth validation, the KITTI 2015 training dataset [12] is used with the corresponding ground truth. The input images for depth and optical flow training and testing are resized to 256×832, and we extract the features at the scales of $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$. We initially warm up the optical flow network for 215K iterations with batch size=8, and we warm up the depth network for 81K iterations with batch size=10 on the KITTI raw [11] for both tasks. Then, we start the joint training of the optical flow and depth networks for 200K iterations with batch size=10 on the KITTI 2015 scene flow dataset [12].

Table 1: Quantitative results of optical flow estimation on KITTI 2015. Bold indicates the best results.

Method	Average EPE	F1-all
GeoNet [13]	10.81	—
DF-Net [14]	8.98	26.01%
Self-Mono-SF [7]	7.51	23.49%
Bridging-Net [15]	7.13	27.13%
Wang et al. [16]	6.66	23.04%
CC [17]	6.21	26.41%
Unos [18]	5.58	—
Matteo et al. [19]	5.39	20.00%
UnRigidFlow [10]	**5.19**	14.68%
Ours	5.51	**14.60%**

3.2. Quantitative results

Table. 1 reports the results for the optical flow. For a fair comparison, we compared our method with the methods that use joint training, as they are considered more challenging than the methods that estimate the optical flow separately. All the stated methods are trained on the unlabeled KITTI datasets. Our method achieves the best F_1 and a competitive end point error compared to the previous methods. For the depth estimation, our method surpasses the previous methods, as reported in Table. 2. Fig. 2 shows a visualization comparisons on samples from KITTI 2015 scene flow dataset.

3.3. Ablation study

The proposed architecture and loss. Table. 3 reports the results compared to using the original RAFT architecture that is trained on unlabeled dataset as a baseline. Our proposed loss shows a remarkable enhancement, particularly in the unmatched (occluded) regions. Our architecture significantly improves the overall accuracy compared to the baseline by a large margin. All contributions together achieve the best accuracy, with a slight increase in the outliers.

Patch size in similarity loss. For each location in our proposed loss with the 3×3 kernel, we tried different patch sizes: 9, 15, and 35. Our ablation is carried out using the RAFT architecture without any improvements. Table. 4 reports the results. The patch size of 9 pixels improves the overall accuracy, including the accuracy in both matched (non-occluded) and unmatched (occluded) regions. The size 15 improves the accuracy more in the occluded regions, but it starts to affect the flow in non-occluded regions by bringing unreliable values from the long-range context. For the size of 35, the result shows that despite the fact that the accuracy improved over the

Table 2: Quantitative results of depth estimation conducted on KITTI 2015 training set (KITTI split). Depth errors in middle columns and prediction accuracy in right columns are used for evaluation. † indicates Eigen split. Bold indicates the best results.

Method	Stereo	Error (lower is better)				Accuracy, (higher is better)		
		AbsRel	SqRel	RMSE	RMSlog	δ<1.25	δ<1.25^2	δ< 1.25^3
CC [17]		0.140	1.070	5.326	0.217	0.826	0.941	0.975
Matteo et al. [19]	✓	0.118	0.748	4.608	0.186	0.865	0.961	0.985
Self-Mono-SF [7]		0.106	0.888	4.853	0.175	0.879	0.965	0.987
Bridging-Net [15]	✓	0.087	0.765	4.380	0.184	0.906	0.959	0.978
RAFT-MSF [20]		0.082	0.726	4.165	0.148	0.921	0.971	0.986
DRAFT [8] †		0.097	0.647	3.991	0.169	0.899	0.968	0.984
MonoDepth [21]	✓	0.068	0.835	4.392	0.146	0.942	0.978	0.989
Ours	✓	**0.062**	**0.682**	**4.169**	**0.137**	**0.947**	**0.979**	**0.990**

Table 3: Ablation on the effectiveness of each contribution on the optical flow results. Bold indicates the best results.

Baseline	Architecture	Loss	EPE-all	EPE-match	EPE-unmatch	F1-all
✓			8.12	3.91	23.61	23.56%
✓		✓	6.01	3.68	14.91	17.71%
✓	✓		5.66	3.37	15.06	**14.51**%
✓	✓	✓	**5.51**	**3.32**	**14.54**	14.60%

Table 4: Ablation on patch size for similarity loss of the optical flow task. Bold indicates the best results.

Patch_size	EPE-all	EPE-matched	EPE-unmatched	F1-all
No-loss	8.12	3.91	23.61	23.56%
9	**6.01**	**3.68**	**14.91**	**17.71%**
15	6.05	3.76	14.85	17.85%
35	6.13	3.86	14.87	18.28%

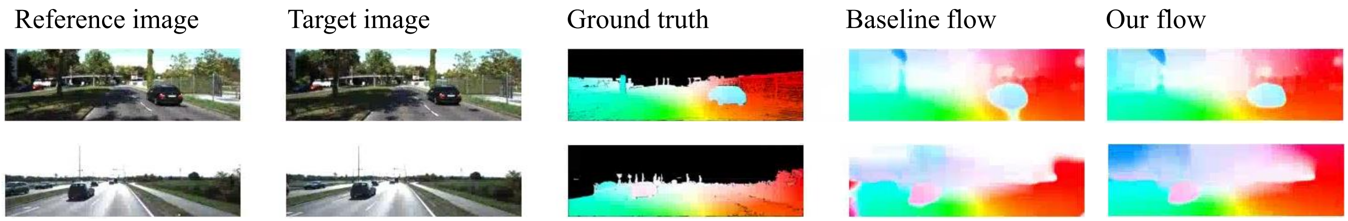

Fig. 2: Visualization comparison between the estimated flow from the baseline and from our method on samples from KITTI 2015 dataset.

baseline in the occluded regions, the error in the non-occluded regions and the outliers increased significantly. For a moderate solution, we adopted the size of 9 in our method and final results.

4. CONCLUSION

In this paper, we provide a new method for joint learning of optical flow and depth that addresses occlusion in the case of unlabeled dataset training. We suggested an innovative design that uses a single-iteration, multi-scale strategy to optimize optical flow at multiple scales. More cues learned from the estimated depth are fed into each refining layer to help improve the optical flow. Furthermore, we presented a self-supervised patch-based similarity loss that is optimized with the reconstruction loss to help the network improve its estimated optical flow, particularly in occluded regions. On the KITTI 2015 scene flow dataset, our technique considerably enhances the accuracy of the estimated depth and optical flow. More experiments and details will be provided in a future work.

5. REFERENCES

[1] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, “Occlusion aware unsupervised learning of optical flow,” InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4884-4893.

[2] A. Dosovitskiy A, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, “Flownet: Learning optical flow with convolutional networks.,” InProceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp.2758-2766.

[3] L. Liu, J. Zhang, R. He, Y. Liu, Y. Wang, Y. Tai, D. Luo, C. Wang, J. Li, and F. Huang, “Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation,” InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6489-6498.

[4] D. Sun, X. Yang, MY. Liu, J. Kautz, “Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume,” InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8934-8943.

[5] A. Stone, D. Maurer, A. Ayvaci, A. Angelova, and R. Jonschkowski, “SMURF: Self-teaching multi-frame unsupervised RAFT with full-image warping,” InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3887-3896.

[6] Z. Teed, and J. Deng, “Raft: Recurrent all-pairs field transforms for optical flow,” InEuropean Conference on Computer Vision (ICCV), 2020, pp.402-419.

[7] J. Hur, and S. Roth, “Self-supervised monocular scene flow estimation,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp.7396-7405.

[8] V. Guizilini, KH. Lee, R. Ambruş, and A. Gaidon, “Learning Optical Flow, Depth, and Scene Flow Without Real-World Labels,” IEEE Robotics and Automation Letters, 2022, pp.3491-3498.

[9] A. Luo, F. Yang, X. Li, and S. Liu, “Learning optical flow with kernel patch attention,” InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8906-8915.

[10] L. Liu, G. Zhai, W. Ye, and Y. Liu, “Unsupervised Learning of Scene Flow Estimation Fusing with Local Rigidity,” International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp.876-882.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” The International Journal of Robotics Research, 2013, p. 1231-1237.

[12] M. Menze, and A. Geiger, “Object scene flow for autonomous vehicles,” InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3061-3070.

[13] Z. Yin and J. Shi, “Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp.1983-1992.

[14] Y. Zou, Z. Luo, and J.B. Huang, “Df-net: Unsupervised joint learning of depth and flow using cross-task consistency,” in European Conference on Computer Vision (ECCV), 2018, pp.36-53.

[15] HY. Lai, YH. Tsai, and WC. Chiu, “Bridging stereo matching and optical flow via spatiotemporal correspondence,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1890-1899.

[16] G. Wang, C. Zhang, H. Wang, J. Wang, Y. Wang, and X. Wang, “Unsupervised learning of depth, optical flow and pose with occlusion from 3d geometry,” IEEE Transactions on Intelligent Transportation Systems, 2020, pp.308-320.

[17] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M.J. Black, “Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.12240-12249.

[18] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, “Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.8071-8081.

[19] F. Tosi, F. Aleotti, PZ. Ramirez, M. Poggi, S. Salti, LD. Stefano, and S. Mattoccia, “Distilled semantics for comprehensive scene understanding from videos,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4654-4665.

[20] B. Bayramli, J. Hur, H. Lu, “RAFT-MSF: Self-Supervised Monocular Scene Flow using Recurrent Optimizer,” arXiv preprint arXiv:2205.01568. 2022 May 3.

[21] C. Godard, O. Mac Aodha, and G.J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Conference on Computer Vision and Pattern Recognition (CVPR),2017, pp.270-279.