Iterative Residual Refinement for Joint Optical Flow and Occlusion Estimation
– Supplementary Material –

Junhwa Hur Stefan Roth
Department of Computer Science, TU Darmstadt

Here, we provide additional details on IRR-PWC, our occlusion upsampling layer, more qualitative examples on the ablation study, as well as a qualitative comparison with the state of the art.

A. IRR-PWC

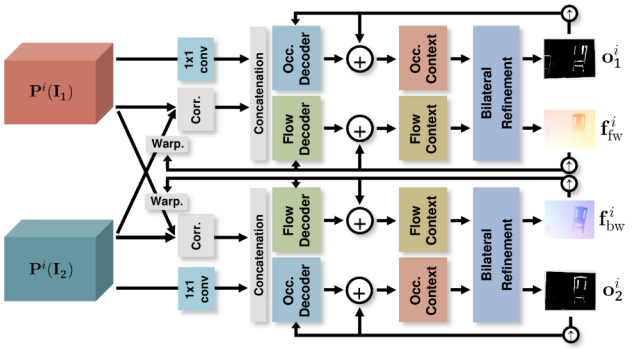
Fig. 9 shows our IRR-PWC model that jointly estimates optical flow and occlusion using bi-directional estimation, bilateral refinement, and the occlusion upsampling layer. Given a 7-level feature pyramid as in the original PWC-Net [52], our IRR-PWC first iteratively and residually estimates optical flow and occlusion up to a quarter resolution of the input image, as shown in Fig. 9a. Then, given the estimates at the 5$^{\text{th}}$ level, we show how we use our occlusion upsampling layer in Fig. 9b to scale the estimates up to the original resolution. The upsampling layer upscales the resolution by $2\times$ at once, and applying the upsampling layer at the 6$^{\text{th}}$ and 7$^{\text{th}}$ level scales the quarter resolution estimate back to the original resolution.

Fig. 7 in the main paper shows the detailed structure of the upsampling layer. In the following, we describe the details on the residual blocks in the upsampling layer.

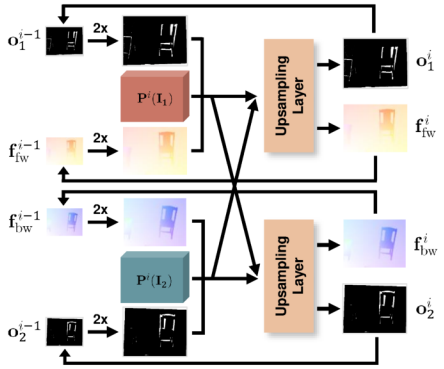
B. Details on the Occlusion Upsampling Layer

In the occlusion upsampling layer shown in Fig. 7 in the main paper, the *residual blocks* [35] are fed a set of feature maps as input and output residual occlusion estimates to refine the upscaled occlusion map from the previous level. Fig. 10 shows the details of the *residual blocks*. As shown in Fig. 10a, the subnetwork consists of 3 residual blocks (*i.e.* 3 *ResBlocks*) with 3 convolution layers. One *ResBlock* consists of *Conv+ReLu+Conv+Mult* operations as shown in Fig. 10b, *cf*. [35]. This sequence of 3 *ResBlocks* with one convolution layer afterwards estimates the residuals over one convolution output of the input feature maps, and the final convolution layer of the *residual blocks* outputs the residual occlusion. The number of channels for all convolution layers here is 32, except for the final convolution layer, which has only 1 channel for the occlusion output.

We use weight sharing also on the upsampling layers between bi-directional estimations and between pyramid levels or iteration steps. Furthermore, the *ResBlocks* in Fig. 10a also share their weights, which is different from [35], where they are not shared. With this efficient weight-sharing scheme, the occlusion upsampling layer im-

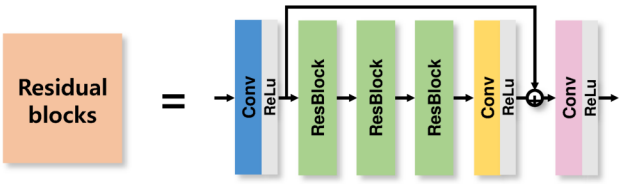


(a) Jointly estimating optical flow and occlusion up to a quarter resolution of the original input, *i.e.* pyramid levels $1 \leq i \leq 5$.

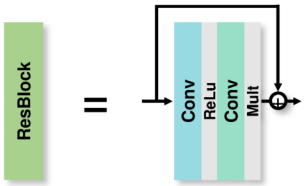


(b) Upsampling optical flow and occlusion using the upsampling layer, *i.e.* pyramid levels $6 \leq i \leq 7$.

Figure 9. **IRR-PWC**: Our PWC-Net variant with joint optical flow and occlusion estimation based on bi-directional estimation, bilateral refinement, and the occlusion upsampling layer. *(a)* Our IRR-PWC model jointly estimates optical flow and occlusion up to a quarter resolution of the input image (*i.e.* up to the 5$^{\text{th}}$ level), the same as the original PWC-Net. *(b)* Then, we use our occlusion upsampling layer to upscale the outputs back to the original resolution while improving accuracy.



(a) Residual blocks subnetwork.



(b) One residual block.

Figure 10. **Residual blocks in the upsampling layer**: *(a)* The *residual blocks* consist of 3 weight-shared *ResBlocks* with 3 convolution layers. *(b)* One *ResBlock* consists of *Conv+ReLu+Conv+Mult* operations [35].

proves the occlusion accuracy by 2.99% on the training domain (*i.e.* the FlyingChairsOcc dataset) and 4.08% across datasets (*i.e.* Sintel) with only adding 0.031 M parameters.

C. Additional Qualitative Examples

Occlusion upsampling layer. Fig. 11 provides qualitative examples of occlusion estimation and demonstrates the advantage of using the occlusion upsampling layer. The models used here are trained on the FlyingChairsOcc dataset only (no fine-tuning on the FlyingThings3D-subset dataset or Sintel) and tested on Sintel Train Clean. The occlusion upsampling layer enhances the occlusion estimates to be much sharper along motion boundaries and refines coarse estimates. Also, the upsampling layer further detects thinly-shaped occlusions that were not detected at the quarter resolution. Unlike optical flow, where a quarter resolution estimate is largely sufficient, we can see from these qualitative examples that estimating occlusions up to the original resolution is very critical for yielding high accuracy.

Ablation study on PWC-Net. In addition to Fig. 8 in the main paper, we here give more qualitative examples for the ablation study. In Fig. 12, all models are also trained on the FlyingChairsOcc dataset and tested on Sintel Train Clean. Our proposed schemes significantly improve the accuracy over the baseline model (*i.e.* PWC-Net [52]), yielding better generalization across datasets.

D. Qualitative Comparison

D.1. Occlusion estimation

Figure 13 demonstrates a qualitative comparison with the state of the art on occlusion estimation. Qualitatively, MirrorFlow [25] misses many occlusions in general, and FlowNet-CSSR-ft-sd [27] is able to detect fine details of occlusion. In contrast, our method tries not to miss occlusions, which results in a better recall rate but somewhat lower precision than those of FlowNet-CSSR-ft-sd [27]. Overall, our method demonstrates better F1-score than FlowNet-CSSR-ft-sd [27], achieving state-of-the-art results on the evaluation dataset (*i.e.* Sintel Train Clean and Final). Note that FlowNet-CSSR-ft-sd [27] is additionally trained on the ChairsSDHom dataset [26] for handling small-displacement motion, which is related to thinly-shaped occlusions. Our approach is not trained further.

D.2. Bi-directional flows and occlusion maps

MirrorFlow [25] is one of the most recent related works that estimates bi-directional flow and occlusion maps. Fig. 14 provides a qualitative comparison with MirrorFlow [25] on the Sintel and KITTI 2015 datasets. In this comparison, we use our model fine-tuned on the training set of each dataset and display qualitative examples from the validation split. Comparing to MirrorFlow [25], our model demonstrates far fewer artifacts and fewer missing details for both flow and occlusion estimation. Although there is no ground truth for backward flow nor an occlusion map for the second image available for supervision, our bi-directional model is able to estimate the backward flow and the second occlusion map well while only using the ground truth of forward flow and the occlusion map for the first image (latter is only available on Sintel) during fine-tuning.

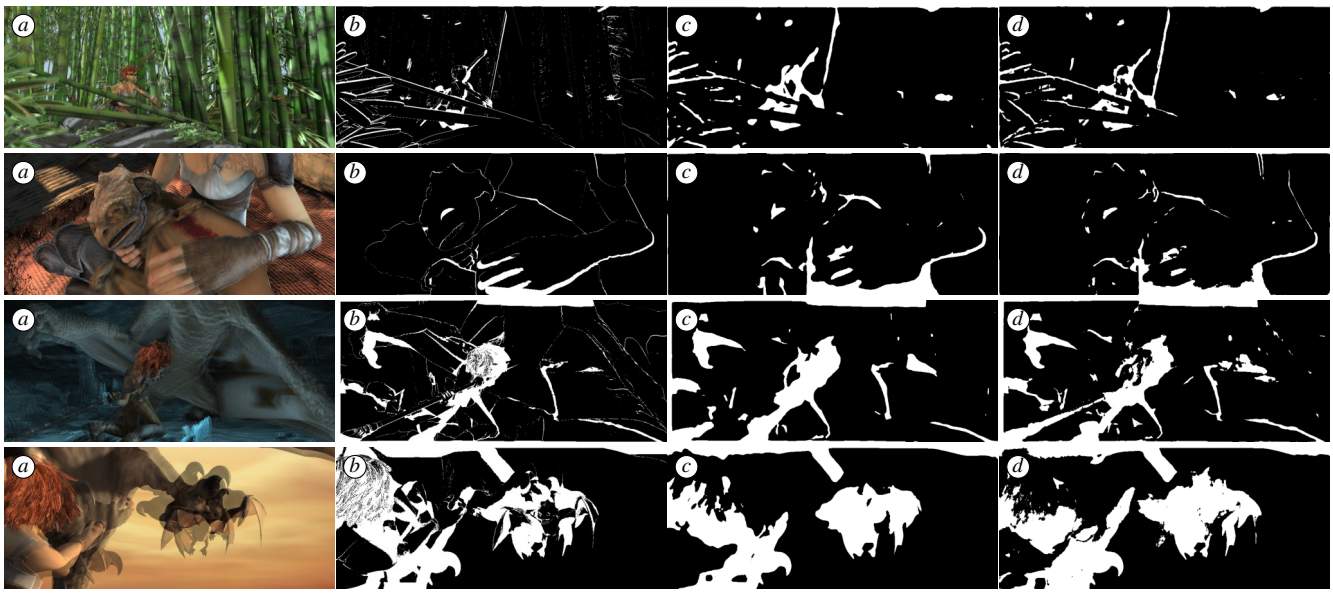


Figure 11. **Qualitative examples of using the occlusion upsampling layer**: *(a)* overlapped input images, *(b)* occlusion ground truth, *(c)* without using the occlusion upsampling layer, and *(d)* with using the occlusion upsampling layer. The occlusion upsampling layer makes occlusion estimates much sharper along motion boundaries and detects additional thinly-shaped occlusions.

Figure 12. **More qualitative examples from the ablation study on PWC-Net**: *(a)* overlapped input images, *(b)* the original PWC-Net [52], *(c)* PWC-Net with Bi, *(d)* PWC-Net with Occ, *(e)* PWC-Net with Bi-Occ, *(f)* optical flow ground truth, *(g)* PWC-Net with IRR, *(h)* PWC-Net with Occ-IRR, *(i)* PWC-Net with Bi-Occ-IRR, and *(j)* our full model (i.e. IRR-PWC). Our full model significantly improves flow estimation over the original PWC-Net with fewer missing details and clearer motion boundaries. Note that there are gradual improvements when combining several of the proposed components.

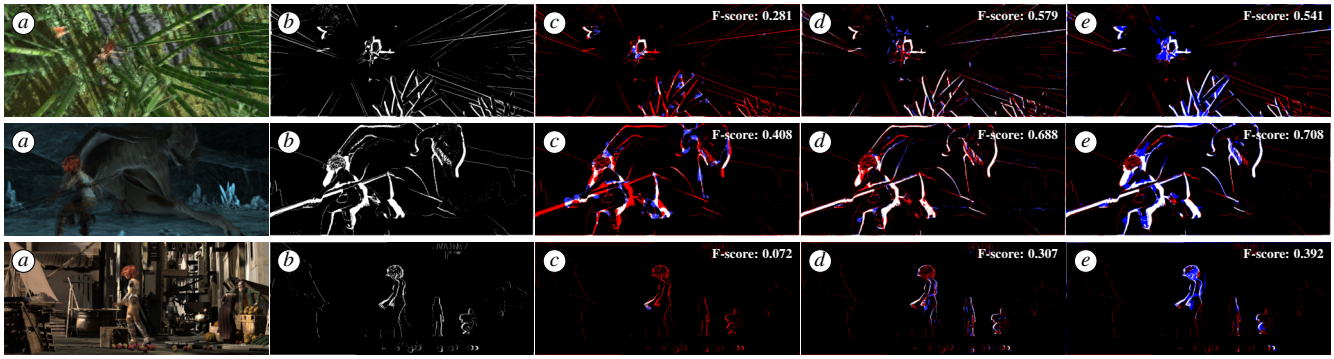


Figure 13. **Qualitative comparison of occlusion estimation with the state of the art**: *(a)* overlapped input images, *(b)* occlusion ground truth, *(c)* MirrorFlow [25], *(d)* FlowNet-CSSR-ft-sd [27], and *(e)* ours. In the result image of each method, blue pixels denote **false positives**, red pixels denote **false negatives**, and white ones denote true positives (*i.e.* correctly estimated occlusion). We include the F-score of each method in the top-right corner. Our model yields a better F-score on the second and the third scene than FlowNet-CSSR-ft-sd [27].

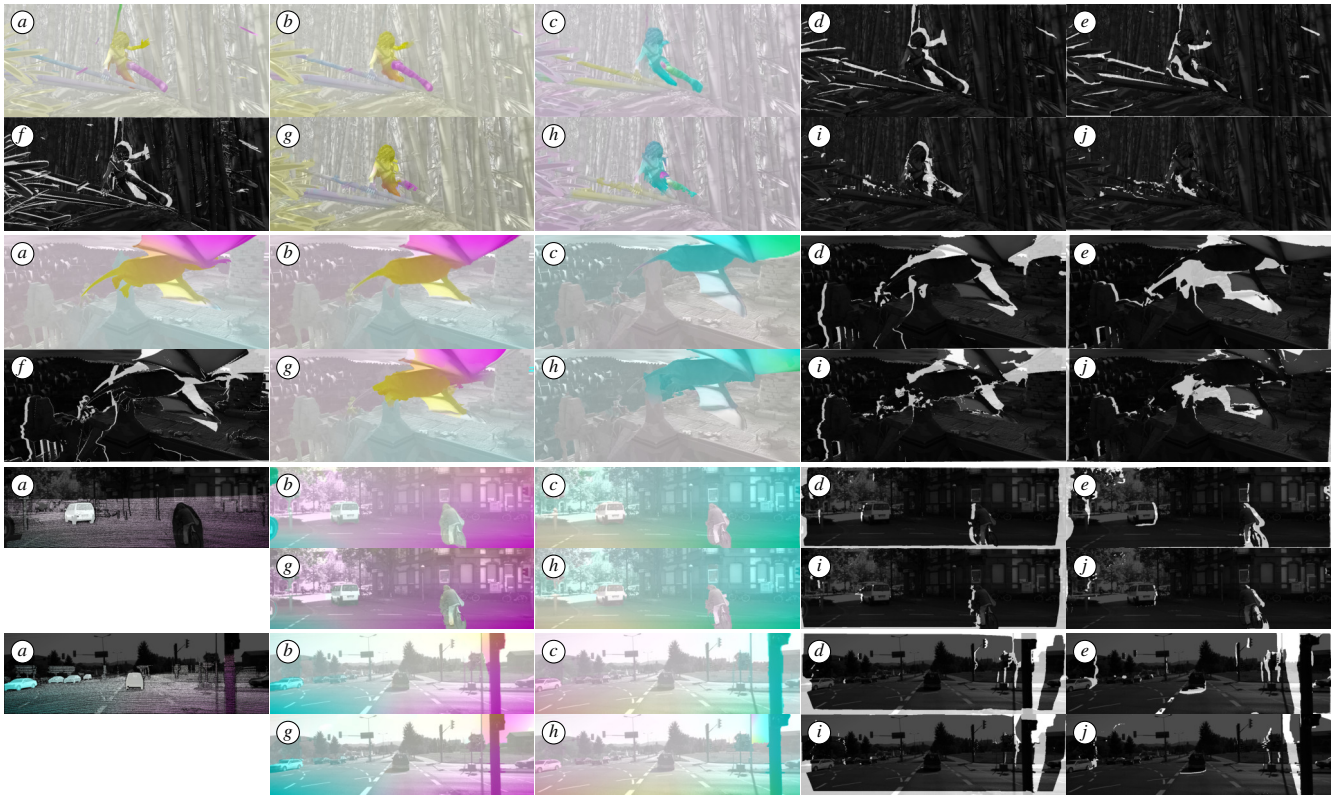


Figure 14. **Qualitative comparison of the bi-drectional optical flows and occlusion maps in both views with MirrorFlow [25]**: All results are overlayed on the corresponding image, either the first frame or the second frame. *(a)* Ground truth optical flow, *(b)* our forward flow, *(c)* our backward flow, *(d)* our occlusion map for the first frame, *(e)* our occlusion map for the second frame, *(f)* ground truth occlusion map, *(g)* forward flow of MirrorFlow, *(h)* backward flow of MirrorFlow, *(i)* occlusion map of MirrorFlow for the first frame, *(j)* occlusion map of MirrorFlow for the second frame. Note that KITTI has only sparse ground truth for optical flow and does not provide ground truth for occlusion.