We thank reviewers for the constructive feedback. Reviewers think our paper is "very-well written" (R2), "significantly new" (R3), and "of great use to practitioners" (R2, R3). Some pressing concerns are addressed below.

Contribution & separable filters (R1,R3) R1 states that our contribution over prior work, for example, FlowNet [9], is not clear. Specifically, we claim that prior deep flow networks (including [9]) re-organize the 4D cost volume into a 3D array that is processed with multi-channel 2D convolutions, but R1 states that [9] "does nothing of the sort". We humbly disagree. The following is a direct quote from [9]:

> In theory, the result produced by the correlation is four-dimensional: for every combination of two 2D positions we obtain a correlation value, i.e. the scalar product of the two vectors which contain the values of the cropped patches respectively. In practice we organize the relative displacements in channels. This means we obtain an output of size $(w \times h \times D^2)$.

The re-organized costs are then convolved along the (w, h) dimensions, leaving the filtering across displacement dimensions to be fully-connected (FC). Instead, we leave the $(w \times h \times D \times D)$ cost volume as-is and directly apply 4D convolutions. R1 asks how separable 4D filters differ from 2D multi-channel FC filters. Separable 4D filters have far fewer parameters (8.2M→1.78M weights, Tab. 5), increasing efficiency and generalizability. Specifically, because (separable) 4D filters are fully-convolutional, they can be trained models with one displacement and applied at differing ones (e.g., repurposing stereo networks for flow and vice-versa). This is difficult for FC layers.

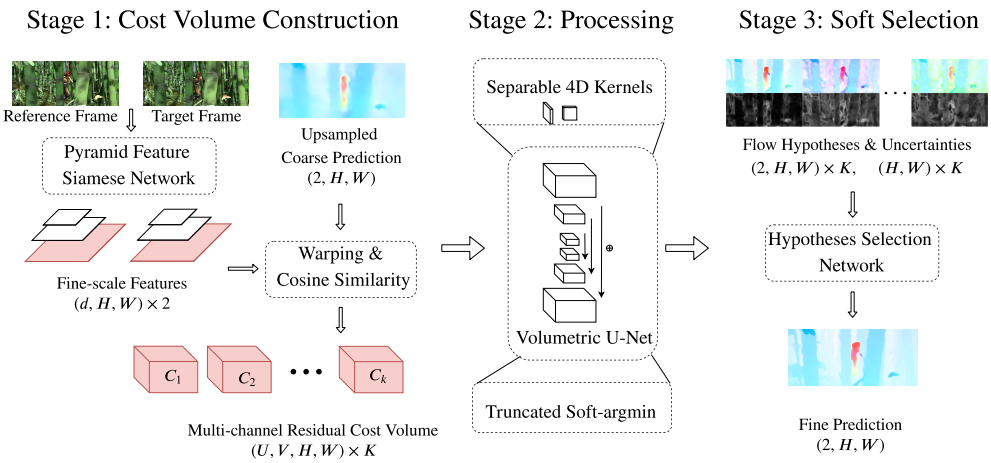

Figure 1: Illustration of volumetric processing at one pyramid level. 1) Cost volume construction: We warp features of the target image using the upsampled coarse flow and compute a multi-channel cost volume. 2) Volume processing: The multi-channel cost volume is filtered with separable 4D convolutions, which is integrated into a volumetric U-Net architecture. We predict multiple flow hypotheses using truncated soft-argmin. 3) Soft selection: The flow hypotheses are linearly combined considering their uncertainties and the appearance feature.

Revision (R1,R2,R3) The revised architecture plot is shown in Fig. 1. We will improve the readability of the method section, fix typos/format issues, and incorporate other feedback (e.g., analysis of the weakness) in the camera-ready.

Qualitative comparisons (R1) For other baselines besides PWC-Net+, we show qualitative results of a challenging example from KITTI-15 test set in Fig. 2. We will add more qualitative results to the main paper and project website.

Figure 2: Results on KITTI-15 test image 48. Color indicates the direction and magnitude of the displacements following Middlebury color wheel, as shown in the top-left image. While prior methods predict the dark wall (circled) as moving to the right together with the front vehicle, our method correctly predicts it as moving to the left.

Variations of results (R3) We run five trials of the pre-training stage of "Ours-small" model and compute the mean/std of EPE on KITTI, which becomes 9.43 ± 0.18 (used to be 9.31 in Tab. 3). More results will be added to the camera-ready.