We thank all reviewers for their comments and suggestions. The reviewers have acknowledged that the method is simple and effective; it has demonstrated superior performance on standard benchmarks and 'will have an impact for the general case of computing cost volumes'.

R1: Q1. DICL versus Reduced DICL:

To respond to R1's comment, we conducted additional experiments to compare the original DICL with the reduced DICL. Results are given in Table 1. It can be seen that the reduced DICL results in slightly improved performance than the MLP (1.72 *vs* 1.76 on the Chairs dataset), but still has a large gap with the original DICL (1.72 *vs* 1.33).

Table 1: **Ablation study on cost computation metrics.** The models for 'Chair' were trained on the Chairs dataset. The models for 'K-15' and 'S' were trained on Things dataset.

Method	Chair	K-15 train		S-train (EPE)	
	EPE	EPE	Fl-all	Clean	Final
Dot Product	1.86	10.39	31.1	2.57	4.06
Cosine Simi	1.84	10.45	30.2	2.55	4.03
3-Layer MLP	1.76	9.83	28.9	2.45	3.98
Reduced DICL	1.72	9.77	28.3	2.42	3.99
DICL	1.33	8.78	23.8	2.11	3.85

R1: Q2. Image guided MAP layer: We have also tested the image guided MAP layer in our network but only achieved minor performance improvement, *e.g.* less than 0.02 pixel in EPE on the Chairs dataset. Therefore we remove the image guidance in the MAP layer.

R1: Q3. Minor corrections: We will rephrase the words as R1 suggested, including line 156, line 139, and a different acronym for the cost re-weighting process, *e.g.* Displacement-Aware Projection (DAP) layer.

R2: Q1. Apply DICL to other existing pipelines:

We replace the non-learned metrics of two well-known pipelines *i.e.* PWCNet and VCN with our DICL module and report the results on the Chairs dataset in Table 2. With our DICL module, both PWCNet and VCN achieve a notable improvement: 8.5% for PWCNet (2.00 *vs* 1.83) and 13.7% for VCN (1.68 *vs* 1.45).

Table 2: **PWCNet and VCN with our DICL module.** Models were trained and evaluated on the Chairs dataset.

Method	PWCNet	PWCNet + DICL	VCN	VCN + DICL
Chair EPE	2.00	1.83	1.68	1.45

R2: Q2. Is 5D processing a problem? One of the largest challenge for the optical flow problem is the large search space, although it has been largely alleviated by the coarse-to-fine techniques. Unlike stereo matching that a disparity is always positive, a displacement in optical flow can be either negative or positive. Therefore, when setting the max displacement to 3 on each scale, the corresponding searching window is 7×7, which has already matched the searching range of deep stereo matching methods (48 for PSMNet on quarter resolution). Moreover, since 4D convolutions will occupy much more GPU memories than 3D convolutions used in stereo, solving optical flow with 5D feature volumes and 4D convolutions is impractical. Similar to deep stereo matching, the cost volume plays a crucial role in ensuring the network to learn matching rather than context-flow mapping. Therefore, we keep the cost volume in our network.

R2: Q3. Relevant papers: These are a few related CVPR 2020 papers that were officially published after the deadline of NeurIPS. Upon the reviewer's request, we will include those papers in a revised version for the sake of completeness.

R4: Q1. Ablation study on Sintel and KITTI: Per R4's comment, we perform an extra ablation study of our method on the Sintel and KITTI 2015 datasets. As provided in Table 1, our DICL module performs consistently better than other cost computation variants with a large margin.

R4: Q2. Memory usage and resource intensive: We agree that, in the training phase the gradients need to be stored in full $K \times H \times W \times U \times V$ grid, but our method needs not to store the full feature volume. We will clarify this part in Table 1 of the paper. Also, as R5 suggested, we will replace the theoretical memory consumption with the actual memory usage. Compared with VCN, our method requires slightly more iterations (150K *vs* 140K) to train on the Chairs dataset, but much faster in inference (0.08s *vs* 0.18s). It is worth noting that our training iterations are much fewer than PWCNet (150K *vs* 1200K).

R4: Q3. Minor corrections: We will add a further discussion of our method versus VCN, change the term 'hand-crafted' metrics to 'non-learned', and tighten up language as suggested.

R5: Q1. Real resource usages: Upon the reviewer's suggestion, we will replace the theoretical resource usage with the real one, *e.g.* training with a crop size of [256, 384] on the Chairs dataset, it requires 1.9G memory for VCN and 1.1G (58% of the former) for ours to process a pair of images.

R5: Q2. Revisit title: Thanks for the suggestion, we may change the title to 'Displacement-invariant matching cost learning for accurate optical flow estimation'.