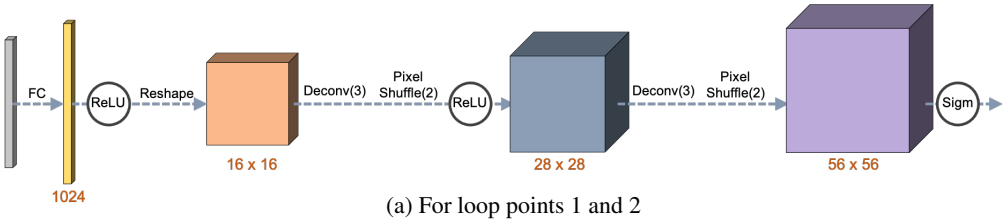
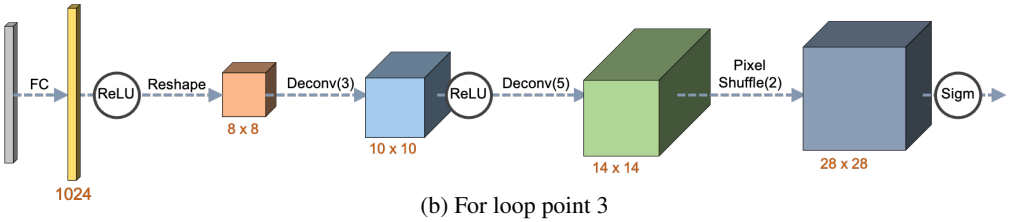
Supplementary Materials for DIR-Net

1. AMG Structure

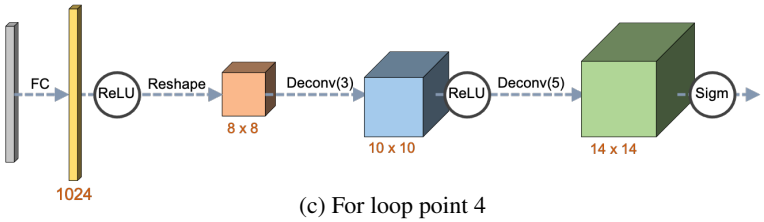
DIR-Net refines original features from FE for refined predictions by performing Hadamard product of the attention maps and the feature maps. Figure 1 depicts the structures of AMG that we utilize for various loop points of Iterative refinements. For fair comparisons, we implement two deconvolution layers in all the structures to share similar model complexity. We also perform pixel-shuffle [1] layers for up-samplings, efficiently bringing depth features to spatial ones. Figure 2 shows how we perform recursive inference without use of AMG, which is used for the experimental results of Table 4 from the paper.



(a) For loop points 1 and 2



(b) For loop point 3



(c) For loop point 4

Figure 1: Structures of AMG for various loop points

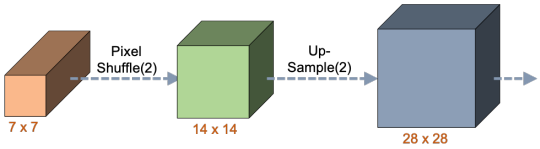


Figure 2: Reordering / up-sampling of RF's output features for recursive feeding.

2. Progressive vs. End-to-End Training Protocols

We present in Table 1 the validation accuracy of two different training protocols. For the notations, we denote AUC and error based on predictions at l-th loop of DIR-Net trained with $l_{max} = N$, respectively, as $\text{AUC}^{l}_{l_{max}=N}$ and $\text{Err}^{l}_{l_{max}=N}$. The overall accuracy of end-to-end training protocol is slightly lower than that of the progressive protocol, and its performance

of early loops (e.g. 1st and 2nd loops) is comparably lower, which implies less number of samples are likely to exit at early loops, causing larger overall computation cost. Progressive training protocol ensures to maximize the capacity of a network for each case of maximum loop allowance, performing with a higher overall validation accuracy.

Table 1: **Progressive vs. End-to-end training protocols** Progressive protocol progressively develops its prediction accuracy. Results at each loop l refer $\text{AUC}^l_{l_{max}=5}$ and $\text{Err}^l_{l_{max}=5}$. The # of paremeters of DIR-Net is **1.68M** and **460K** parameters respectively for STB and FPHA datasets.

	Progressive Training				End-to-end Training				
Loops	AUC (20-50)		Err (px/mm)		AUC (20-50)		Err (px/mm)		GFLOPs
	2D	3D	2D	3D	2D	3D	2D	3D	
0	0.722	0.987	9.08	9.48	0.672	0.981	10.34	10.91	0.46
1	0.784	0.996	6.65	7.41	0.773	0.990	6.97	9.43	0.67
2	0.796	0.997	6.23	7.09	0.781	0.995	6.69	8.29	0.88
3	0.803	0.997	6.04	6.97	0.784	0.996	6.59	8.07	1.09
4	0.805	0.998	5.94	6.89	0.785	0.997	6.56	7.89	1.30
5	0.806	0.998	5.93	6.88	0.786	0.997	6.53	7.88	1.51

(a) Results of two training protocols on STB dataset.

	Progressive Training				End-to-end Training				
Loops	AUC (0-50)		Err (px/mm)		AUC (0-50)		Err (px/mm)		GFLOPs
	2D	3D	2D	3D	2D	3D	2D	3D	
0	0.707	0.768	8.97	11.61	0.664	0.760	10.34	12.02	0.14
1	0.716	0.768	8.67	11.60	0.701	0.760	9.15	12.01	0.2
2	0.717	0.769	8.65	11.58	0.703	0.761	9.08	11.97	0.27
3	0.717	0.769	8.64	11.54	0.703	0.761	9.08	11.96	0.33
4	0.717	0.771	8.64	11.46	0.703	0.761	9.08	11.94	0.39
5	0.717	0.772	8.64	11.40	0.704	0.763	9.07	11.87	0.45

(b) Results of two training protocols on FPHA dataset.

3. Qualitative results of DIR-Net

We provide qualitative results of our method in Figures 3 and 4. The qualtitative results show that our method perform accurate pose and shape estimations while adaptively reducing the overall computational burden of hand pose estimation inference per input.

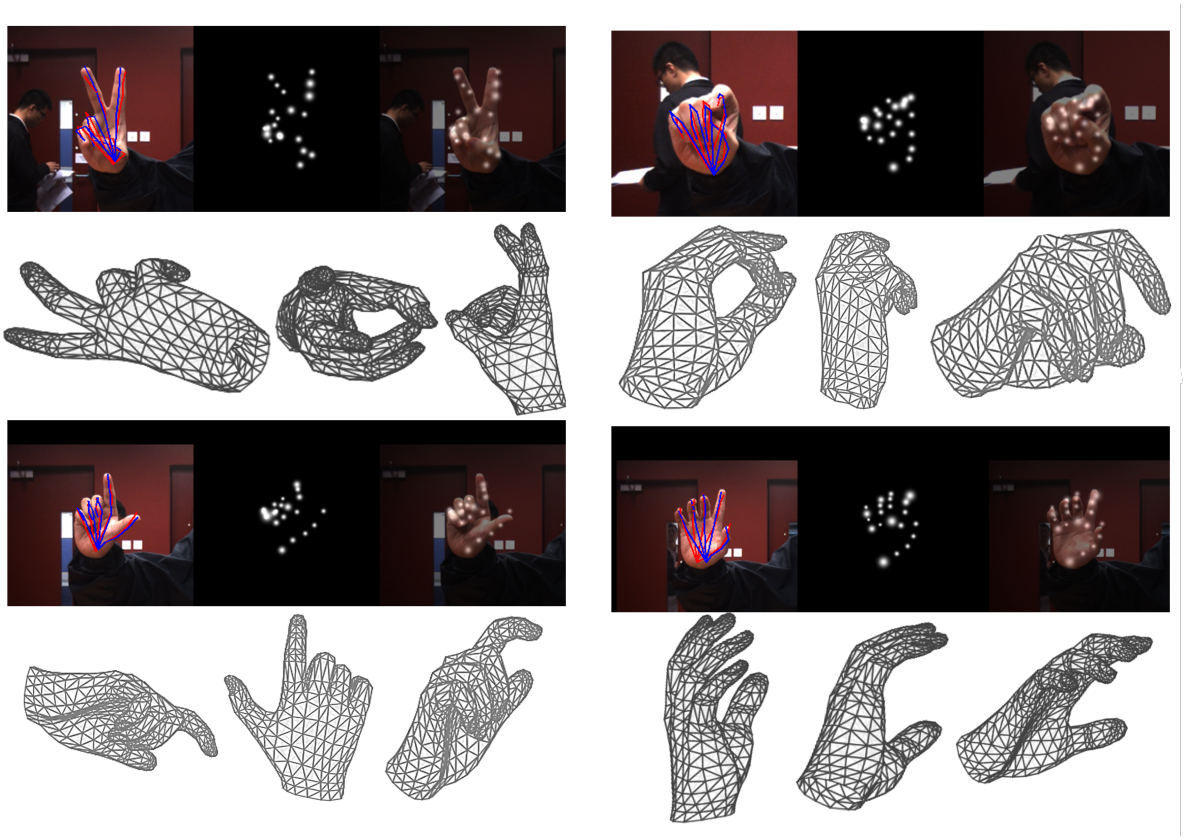


Figure 3: Qualitative results for STB dataset

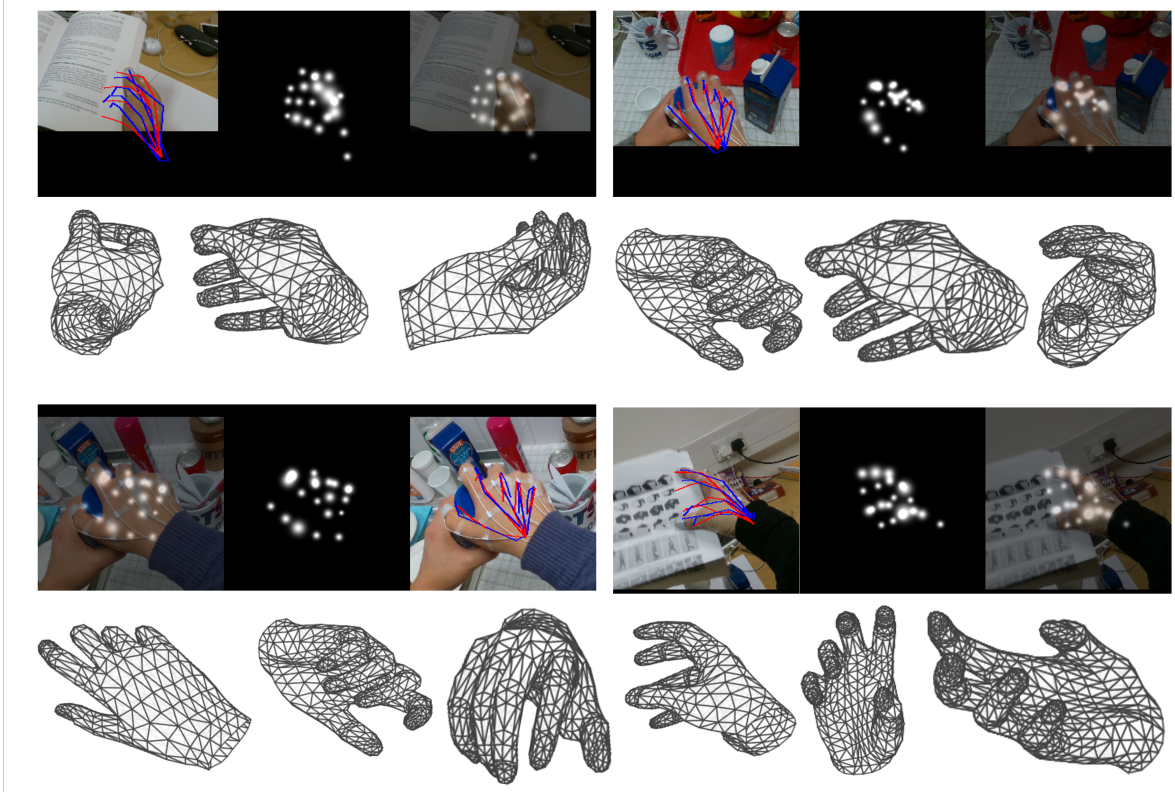


Figure 4: Qualitative results for FPHA dataset

References

[1] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.