Supplementary Document:
Two-hand Global 3D Pose Estimation using Monocular RGB

1. Training Details

To obtain the experimental results for all targeting datasets, we use the same training schedule for *HandSegNet*, $PoseNet_{2D}$ and $PoseNet_{3D}$. Specifically, we use the Adam optimizer [1] with an initial learning rate of 0.001, β_1 = 0.9 and β_2 = 0.999. We use cross entropy (CE) loss for the segmentation loss of *HandSegNet* and mean squared error (MSE) loss for the training of all other parts of the networks. We set the batch size to 4 and trained *HandSegNet* for 30,000 iterations. $PoseNet_{2D}$ and $PoseNet_{3D}$ are trained for 15,000 iterations since each iteration consists of training of two separate hands. The learning rates decrease with a rate of 0.5 every 5,000 iterations.

For results obtained without segmentation and batch normalization layers using $PoseNet_{2D}$, we used standard stochastic gradient descent and an initial learning rate of 0.000001 for better convergence. We also set the weight decay to 0.0005 for $PoseNet_{3D}$.

2. Additional Qualitative Results

For evaluation on real-world data, we manually annotated a small dataset to train the networks and provide preliminary qualitative results in Fig. 1. We evaluate on simple poses due to the limited variety in our annotated data. Our preliminary results indicate that our method is capable of evaluation on real-world data when sufficient training data in the real-world domain becomes available in the future. Note that evaluation on the RGB data in the real-world domain is extremely challenging due to factors such as the vast color space, different skin color/texture, complex background noise, motion blur, lighting, shadow features, etc. As a result, evaluation on videos in the wild is beyond the scope of this paper and will be addressed in our future works. It is worth mentioning that since it is currently infeasible to quantitatively evaluate on the task of RGB-based two-hand global 3D pose estimation using real-world data due to the lack of ground truth data, Ego3DHands serves as a necessary benchmark dataset for this new task. We also provide additional qualitative results for the 4 target datasets in Fig. 2.

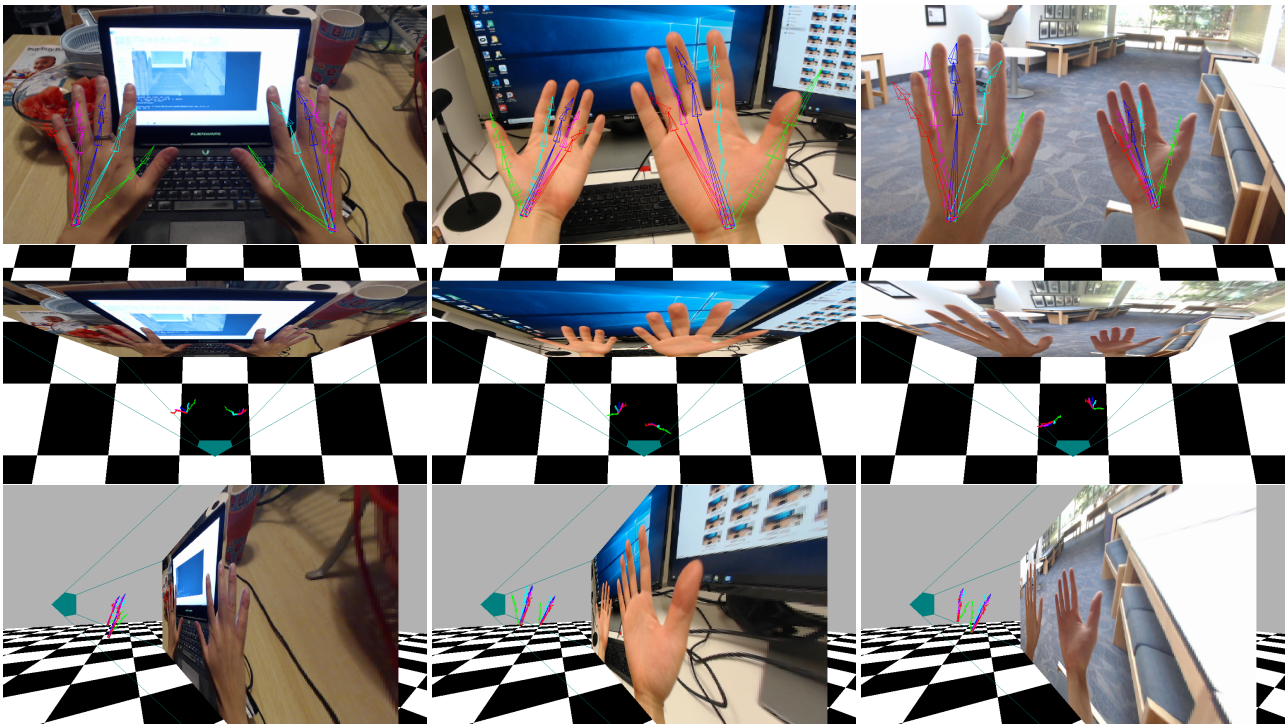
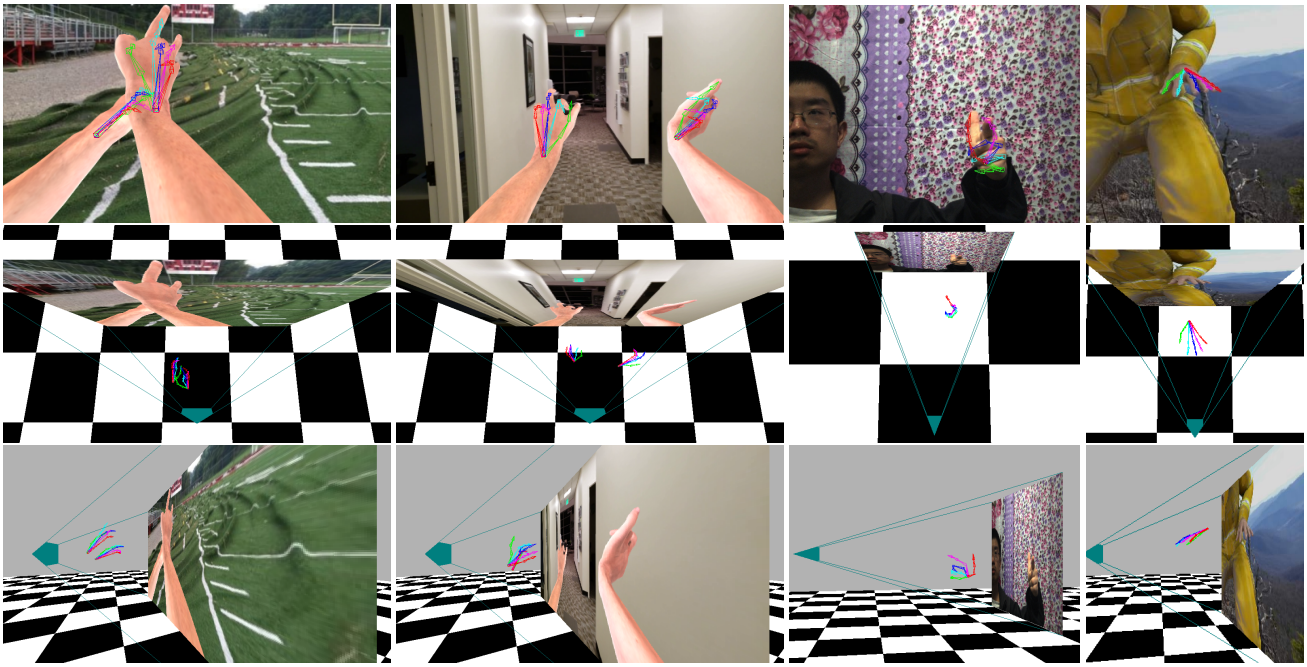
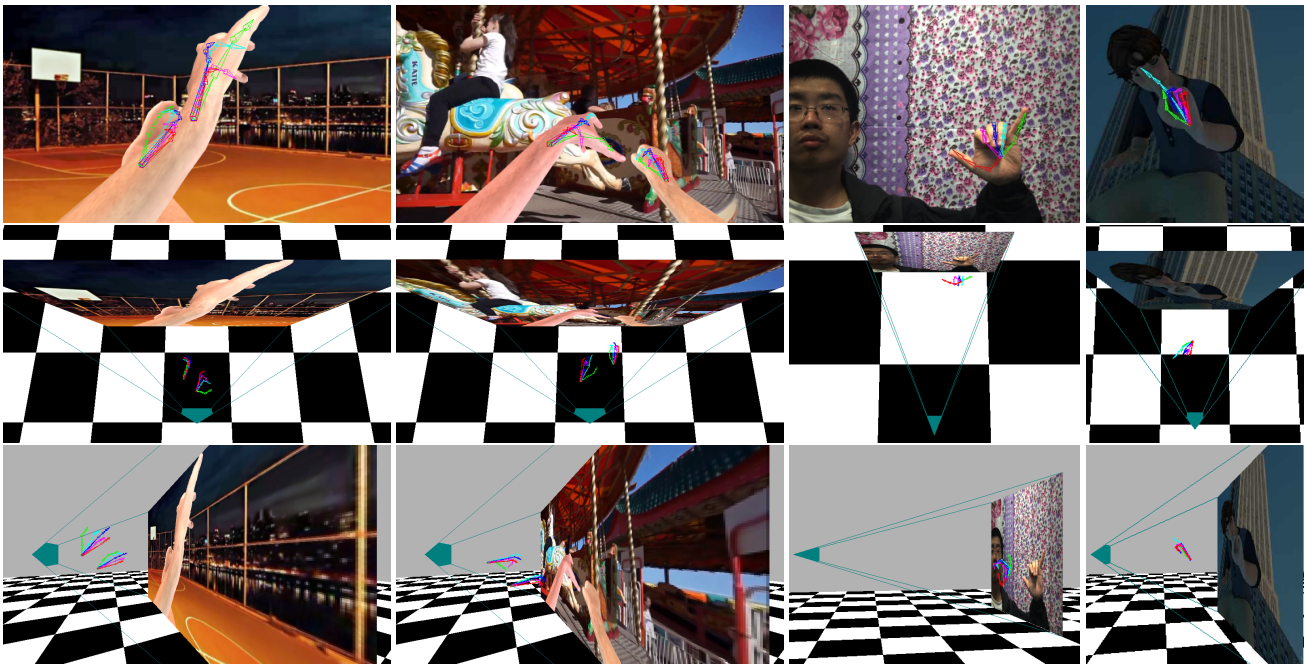


Figure 1: Preliminary qualitative results on real-world data. We collected sample test sequences using 3 different background environments. Top row visualizes the 3D global hand poses from the center camera view. Middle and bottom rows show the top and side views respectively.



(a) $Ego3D_s(img_1)$ (b) $Ego3D_d(img_1)$ (c) STB (img_1) (d) RHP (img_1)



(a) $Ego3D_s(img_2)$ (b) $Ego3D_d(img_2)$ (c) STB (img_2) (d) RHP (img_2)

Figure 2: Additional qualitative results for 3D global two-hand pose estimation on $Ego3D_s$, $Ego3D_d$, STB and RHP.

References

[1] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015. 1