
Hand Puppets: 3D Hand Pose Prediction from Shadow Puppet Images

Abhishek Tandon
atandon2

Gaini Kussainova
gainik

Rakshith Srinivasa Murthy
rakshits

Rutika Moharir
rmoharir

Abstract

Hand shadow puppetry is an ancient form of entertainment that uses complex hand interactions to create shapes and figures that resemble animals, birds and other objects. But what if you see the puppeteer create a shadow of a rabbit and you wish to learn the hand pose that brought about this shadow. To this end, we propose a learning-based approach for 3D interacting hand reconstruction from a single shadow puppet image. Prior works in hand pose estimation are centered around estimating 3D hand pose from RGB images. In this project, we make the first attempt to reconstruct 3D hands from monocular shadow images. For a given input shadow image, our method generates 3D hand meshes with precise 3D pose. To achieve this, we employ a two-stage framework. In the first stage, we use a deep network to generate coarse pose predictions. This is followed by the second stage of iterative pose optimization to refine the predicted pose such that its rendered shadow exactly matches the one in the input. We investigate the importance of the second stage of optimization and compare results in the absence of this stage. Extensive qualitative results demonstrate the effectiveness of our approach. The code, data and trained models are available at <https://github.com/Tandon-A/Hand-Puppets>.

1 Introduction

Reconstructing two interacting hands in 3D is an active research area, as it enables applications in various fields of vision and graphics, including augmented and virtual reality, robotics, and sign language translation. Prior works focus on hand pose estimation from RGB images and/or depth sensors or multi-camera setups. However, these methods cannot be easily applied to shadow images.

An intuitive solution to learn hand pose parameters from shadow images is to use a deep network to extract possible features and supervise using constraints on the pose parameters to avoid unnatural poses.

In this work, we employ a two-stage framework for interacting hand pose predictions. In the first stage, we use a CNN (Pose Prediction Network) to generate coarse predictions of the pose parameters. Specifically, the CNN takes a shadow image as input and regresses the pose parameters of the two hands. These initial predictions provide a good initialization for the second stage. Some results of the predicted mesh generated from only the Pose Prediction Network are shown in the third column in Fig 1. You can see that the resulting mesh is able to capture the coarse structure of the ground truth mesh, however there is still some misalignment in the predicted shadow of the reconstructed mesh.

Hence in the second stage, we employ conventional gradient-descent based optimization (Iterative Pose Optimization) that refines the first-stage predicted pose output in several steps. The second column in Fig 1 shows several reconstructed 3D hand meshes using our two stage framework. We










| Ground truth | Pose Prediction Network + Iterative Pose Optimization | Pose Prediction Network |
|---|---|--|
|  |  |  |
|  |  |  |
|  |  |  |

Figure 1: **Interacting Hand Pose Estimation from Shadow Images:** 1) The first column is the ground-truth shadow image and hand mesh. 2) The second column shows the hand mesh and its corresponding predicted shadow obtained from our two-stage pipeline. 3) To depict the importance of the iterative pose optimization stage, the third column shows hand mesh predicted only from the Pose Prediction Network.

can see that using the iterative pose optimization predicts hand poses whose shadow image closely matches the input shadow image.

Our contributions are as follows: 1) We propose a novel two-stage framework for 3D interacting hand pose estimation from shadow images. 2) Our pipeline does not need any additional inputs like depth maps or heat maps.

2 Prior Work

We briefly cover related prior works in 3D hand reconstruction from images.

2.1 3D Single Hand Pose Estimation

Existing literature for 3D single hand pose estimation approaches mainly use RGB images. Regression-based approaches [14, 13] take depth images as input and directly map the image to the hand joint locations. Detection-based methods [7, 13] estimate a probability density map for every joint. Recent monocular RGB image based techniques [5, 12, 3] primarily use a CNN to predict the MANO [11] model parameters.

2.2 3D Interacting Hand Pose Estimation

Reconstructing two hands is challenging due to severe self-occlusions. Some approaches address this problem using multi-view setups, depth maps or marker gloves. *Kyoung Mu Lee et. al.*[8] proposes InterNet that implements interacting and single hand pose estimation from a single RGB image. The network predicts the handedness (existence of left and right hand), 2.5D right and left hand poses and right hand-relative left hand depth. The final 3D interacting hand poses are refined using 2.5D poses and relative depth between the hands. *Yu Rong et. al.*[6] reconstruct 3D interacting hand poses from a monocular RGB image by optimizing the initial pose and penalizing collisions using penetration loss.

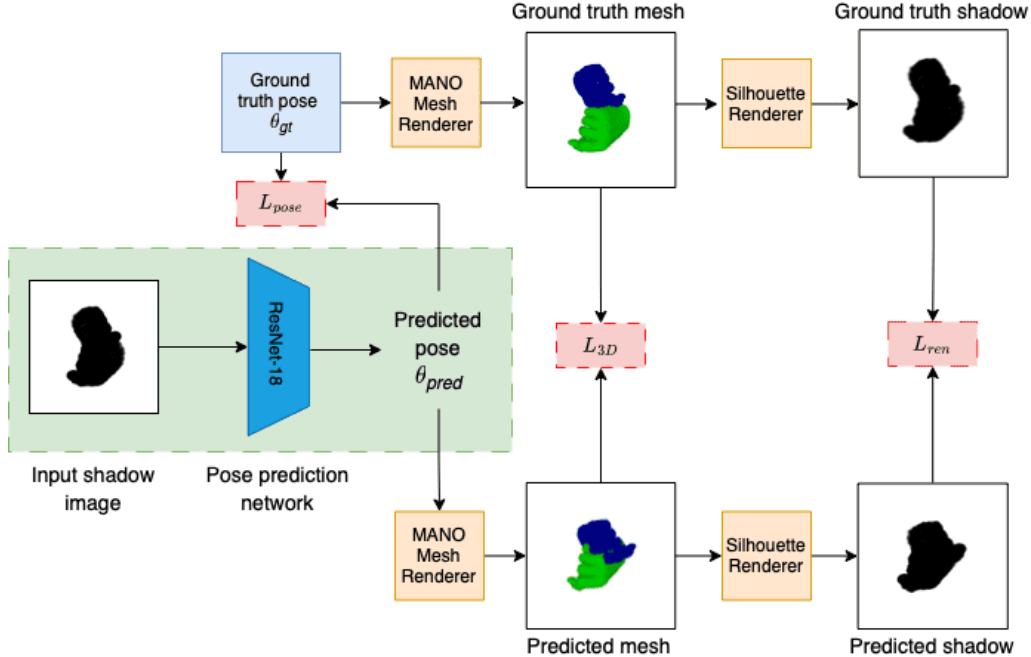


Figure 2: **Pose Prediction Network:** Overview of our first stage in the two-stage framework. Given only a shadow image as input, the Pose prediction network computes a coarse pose estimate. This stage adopts a convolutional neural network encoder (ResNet-18) to regress the hand pose parameters.

2.3 Hand Pose Estimation from Interacting Hand Shadow Images

Our problem statement does not fit into the domain of the approaches discussed above since we wish to estimate 3D interacting hand pose from just shadow images which are not as feature rich as their RGB counterparts. Thus we do not have a clear separation between the two hands making this a more difficult task.

3 Proposed Method

3.1 Hand Mesh Representation

The MANO[11] model is a low-dimensional parametric hand mesh model that captures hand shape and pose variations. The hand surface is represented by a 3D mesh with vertices V where the number of vertices is $N_v = 778$ and the vertices are mapped to joints J where $N_J = 16$ is the number of joints. The 3D position of the i^{th} vertex and the j^{th} joint is denoted by $v_i, J_j \in (\beta, \theta) \in \mathbb{R}^3$ where $\beta \in \mathbb{R}^{N_s}$ is a shape parameter vector and $\theta \in \mathbb{R}^{N_p}$ is the pose parameter vector. Here $N_s = 10$ and $N_p = 51$.

The MANO model defines a function

$$v, J : \mathbb{R}^{N_s} \times \mathbb{R}^{N_p} \longrightarrow \mathbb{R}^{N_v \times 3} \times \mathbb{R}^{N_J \times 3} \quad (1)$$

that computes the 3D position of all the N_v vertices and N_J joints.

Since we are only concerned about the hand pose and not the shape, we regress only the θ values. We use independent hand models for the left and right hand and concatenate the parameters for each of them to get $\theta = (\theta_{left}, \theta_{right})$.

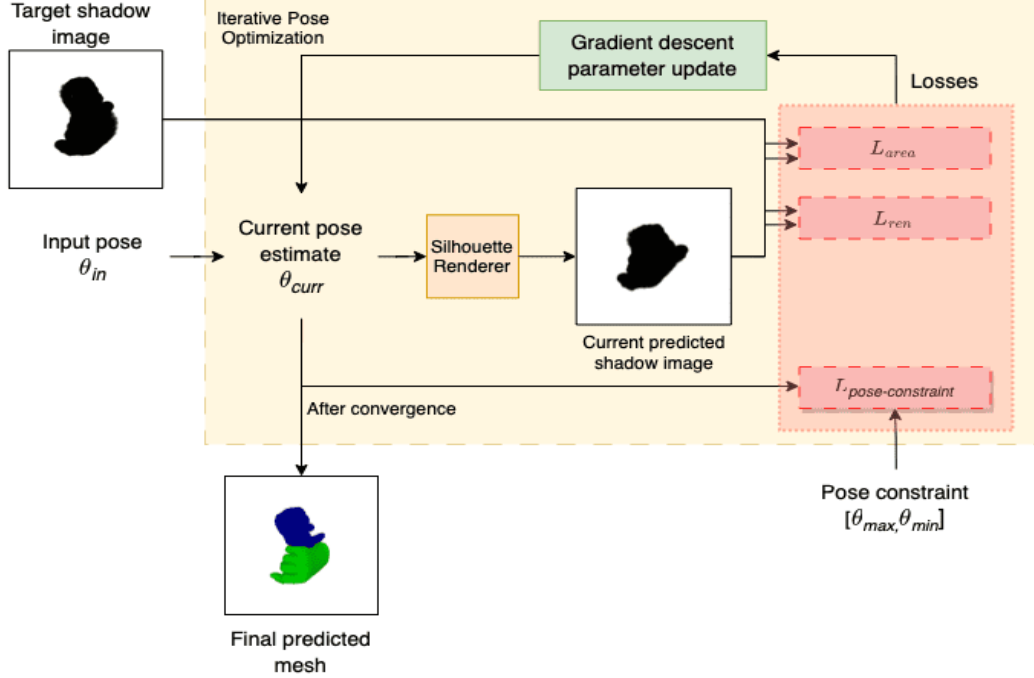


Figure 3: **Iterative Pose Optimization:** Given an initial prediction from our Pose Prediction Network, the optimization pipeline tries to modify the pose parameters in such a way that the rendered shadow of the predicted mesh is as close as possible to the input shadow image along with other regularization terms.

3.2 Pose Prediction Network

Architecture: The Pose Prediction Network takes a shadow image \mathbf{S} as input and extracts image features \mathbf{F} using ResNet-18[2] and outputs the estimated pose parameters θ . The network and training details are depicted in Fig. 2. Below, we mention the significance of each of the loss functions used in the training phase of the network.

Geometric Losses: In addition to imposing supervision on the predicted pose parameters θ , we also constrain the predicted 3D joints. Specifically, we obtain the ground-truth 3D joint locations \hat{J}_{3D} and the predicted 3D joint locations J_{3D} by using the ground-truth and predicted pose parameters $\hat{\theta}$ and θ respectively as input to the MANO Model.

$$L_{pose} = \|\theta - \hat{\theta}\|_2^2 \quad (2)$$

$$L_{J3D} = \|J_{3D} - \hat{J}_{3D}\|_2^2 \quad (3)$$

Rendering Loss We also use a pixelwise rendering loss between the predicted shadow image S and input shadow image \hat{S} .

$$L_{ren} = \|S - \hat{S}\|_2^2 \quad (4)$$

Loss Functions The entire pipeline is fully differentiable with respect to the learnable parameters, thus making the Pose Prediction Network end-to-end trainable. The overall loss function is summarized as

$$L = \lambda_{pose}L_{pose} + \lambda_{J3D}L_{J3D} + \lambda_{ren}L_{ren} \quad (5)$$

where λ_{pose} , λ_{J3D} and λ_{ren} are tunable hyper-parameters to adjust the trade-off among different types of supervision on the network.













| Ground Truth | Pose Prediction Network + Iterative Pose Optimization |
|---|---|
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Figure 4: **Results of our two-stage framework:** In each of the grid cells, the top image is the rendered shadow and bottom image depicts the hand mesh.

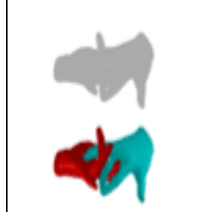



| Ground Truth | Pose Prediction Network + Iterative Pose Optimization |
|--|---|
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Figure 5: **Failure cases of our two-stage framework:** In each of the grid the top image is predicted shadow while at the bottom we have the hand mesh.

3.3 Iterative Pose Optimization

In the Pose Prediction Network, we tried to regress the mesh parameters θ_{mesh} from just an input shadow image which makes it difficult for the network to predict accurate pose parameters for challenging poses with interacting fingers.

Hence we implement an test time optimization framework wherein we refine the estimated pose in an iterative algorithm. At each refinement step we impose constraints on the pose parameters and the rendered shadow of the predicted mesh such that this shadow is as close as possible to the input shadow. Specifically, we use the rendering loss as defined in equation 4. Additionally we use an Area loss which compares the number of pixels occupied by the hands in the predicted shadow N_p and input shadow N_i . This loss enables the optimization to better deal with local minima. We also constrain each of the 51 predicted parameters θ_i to be in specific range $\theta_{i_{min}}, \theta_{i_{max}}$ where i goes from 0 to 51. These $\theta_{i_{min}}$ and $\theta_{i_{max}}$ values are obtained from the dataset.

$$L_{area} = |N_p - N_i| \quad (6)$$

$$L_{pc} = \sum_{n=1}^{51} \max(0, \theta_i - \theta_{i_{max}}) + \max(0, \theta_{i_{min}} - \theta_i) \quad (7)$$

We use gradient-descent optimization which updates the parameters by minimizing the following objective.

$$L = \lambda_{ren}L_{ren} + \lambda_{area}L_{area} + \lambda_{pc}L_{pc} \quad (8)$$

4 Experiments

To evaluate the effectiveness of our approach, we present the qualitative results on the recovered mesh from a single shadow image input. We also perform ablation studies where we compare the results of the 3 baselines - first using just the Iterative Pose Optimization pipeline (with random pose initialization), second using just the Pose Prediction Network and finally combining the Pose Prediction Network with the Iterative Pose Optimization to show the results of our two-stage framework.

4.1 Experimental Settings

4.1.1 Dataset

The *InterHand2.6M dataset*[8] contains large-scale multi-view single and interacting hand sequences under various poses. We use the MANO[11] annotations and generate shadow images corresponding to the interacting hand poses in the dataset by using the Silhouette Renderer from Pytorch3D[10] from a fixed viewing direction. We generated such shadow images for 10000 hand poses with complex interactions and used these mapping of shadow image and pose parameters for the supervised learning of our pose prediction network.

4.1.2 Training

We implement our framework with Pytorch [9]. The hyper-parameters in equation 5 are empirically set to $\lambda_{ren} = 0.01$, $\lambda_{pose} = 1$ and $\lambda_{J3D} = 1$. Adam Optimizer[4] is used to optimize the framework. We begin with a learning rate of $1e-4$ and decay the learning rate whenever validation loss plateaus for more than five epochs.

4.2 Results

Visualizations on a few samples from the Interhand2.6M dataset are shown in Fig 4.

4.3 Failure Cases

Fig 5 shows some of the failure cases of our framework. As seen from these results, it is evident that images with complex interacting hand poses are difficult to handle. Since our framework needs to

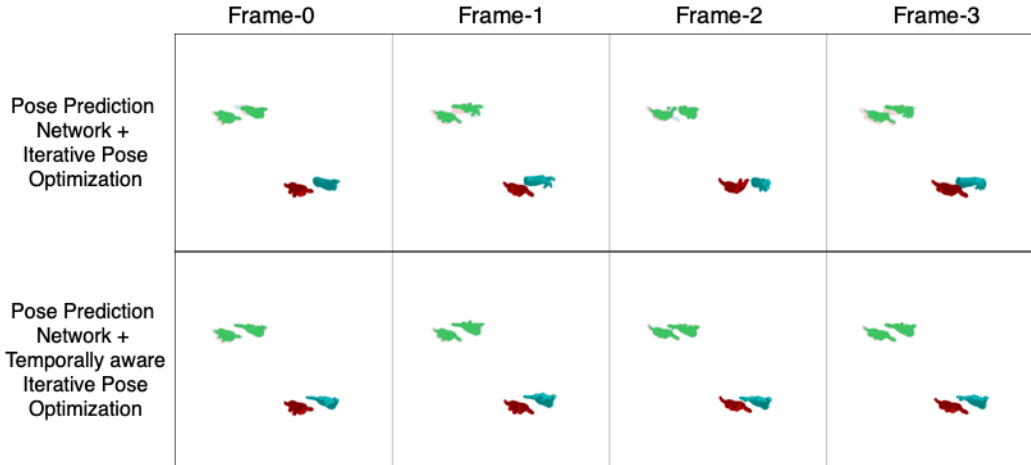


Figure 6: **Pose Prediction Network + Temporally-aware Iterative Pose Optimization:** Here for each frame of a video sequence, the bottom right is the reconstructed mesh. In the top left we show the input shadow in blue, the predicted shadow rendered from the reconstructed mesh in pink and the overlap between the two shadows in green. Consequently, higher green coverage indicates higher overlap between the predicted shadow and the input shadow.

regress pose from just interacting hand shadow image, it is not able to differentiate between the two hands in the image. This results in unnatural pose predictions wherein the fingers are intersecting with each other.

4.4 Ablation Study

Baseline 1: Iterative Pose Optimization only: To verify the need of a learning based network to generate the initial coarse pose predictions, we test an optimization only pipeline where we directly regress the hand pose through Iterative Pose Optimization starting with a random initialization. From the results we can see that this generates predicted shadows similar to the input shadow at the expense of generating unnatural hand poses.

Baseline 2: Pose Prediction Network only: Here we include visualizations of the predicted pose obtained from just the Pose Prediction Network without doing Iterative Pose Optimization. The results are shown in Fig. 7. The results show that in simple cases the network is able to correctly predict the pose. However, in complex interacting hands it generates unnatural hand poses since the network is not able to differentiate between the two hands and hence generates mesh with intersecting fingers.

Baseline 3: Pose Prediction Network + Iterative Pose Optimization: Here we combine the first and second stage of the framework and observe the results. We can see that in cases where the predictions from the Pose Prediction Network are a little misaligned, the Iterative Pose Optimization stage is able to refine the predicted mesh to get the correct predicted shadow image.

Pose Prediction Network + Temporally-aware Iterative Pose optimization: Instead of testing the performance on just images, we use a sequence of image frames from a video. In a given video sequence, for the first frame at $t = 0$ we get the initial pose estimates from the Pose Prediction Network. However, for the subsequent frames, we use the second stage of the pipeline, i.e. the final predicted pose from the optimization pipeline at $t = 0$ is used as the initial estimate of pose at $t = 1$ which is then optimized using the optimization framework. We can see that, this prior in the optimization pipeline greatly helps with refining procedure and generate smooth transition in the hand meshes across multiple frames. The results are shown in Fig 6

5 Future Work

5.1 Regularization Terms

From the failure cases in Fig 5, we see cases where the two hands in the predicted mesh are colliding. To ameliorate such scenarios we plan to add a penetration loss as used in [1].

5.2 Camera Pose Estimation

Currently the Pose Prediction Network is regressing the MANO[11] pose parameters. Besides these parameters we can also regress a set of weak-perspective camera parameters $\pi \in \mathbb{R}^3$. Given these camera parameters we can obtain the orthogonal projection of the 3D joints to get the 2D joints and supervise these. This will enable us to test the performance of our framework on real world images.

5.3 Pose Sequence Optimizer

After modelling the final predicted pose θ_{final} that forms the correct shadow image upon projection, we would like to generate the sequence of transformations leading to this final hand pose from a rest pose. Due to the unavailability of ground-truth values for the intermediate steps, this can be modelled as a optimization between the initial and the final hand pose over n steps.

6 Conclusion

We have presented a novel pipeline for reconstructing 3D hand mesh from monocular shadow images. We have employed a two-stage framework, wherein we first predict coarse predictions for the hand pose parameters which were then refined using the second stage of optimization. We show several visualizations of experiments on the *InterHand2.6M dataset*[8].

7 Acknowledgments

We acknowledge Samradh Agarwal for giving us the idea for the project. We also acknowledge Dr. Deepak Pathak for many discussions and inputs about the solution.

References

- [1] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [3] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021.
- [4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [5] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.
- [6] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2761–2770, 2022.
- [7] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018.
- [8] Gyeongsik Moon, Shou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020.

- [9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [10] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020.
- [11] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017.
- [12] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5340–5348, 2019.
- [13] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Lihao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2018.
- [14] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei. Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*, 2016.

| Ground Truth | Pose Prediction Network + Iterative Pose Optimization | Pose Prediction Network | Iterative Pose Optimization |
|---|---|--|---|
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Figure 7: **Ablation study:** We compare the three baselines to verify the importance of each of our stage in the two-stage pipeline.