XPose: Towards Extreme Low Light Hand Pose Estimation

Green Rosh Meghana Shankar Prateek Kukreja Anmol Namdev
B H Pawan Prasad
Samsung R&D Institute India, Bangalore
greenrosh.ks@samsung.com

Abstract

Recent advances in deep learning have enabled considerable strides in hand pose estimation in well-lit conditions. However, to the best of our knowledge, there is no existing method for hand pose estimation from RGB images captured in low-light conditions. This task is highly challenging due to the overwhelming amount of noise which plague image capture in low-light conditions (<1 lux). In this paper, we propose XPose, the first method for extreme low light hand pose estimation from RGB images. We also introduce the first dataset for low light hand pose estimation consisting of ∼120k images along with accurate hand pose labels. Our dataset consists of images captured in low light and well-lit conditions from multiple viewpoints. We propose an innovative deep learning based methodology for monocular low-light hand pose estimation using guidance from well-lit and multi-view images available in our dataset, during training time. We show that our method, using the proposed LLPose dataset, significantly outperforms existing methods for hand pose estimation both qualitatively and quantitatively in low light conditions.

1. Introduction

Hand pose estimation aims at accurately localizing the 3D joints of a hand from input images. There have been tremendous advances in hand pose estimation primarily fueled by the success of deep neural networks [1–3, 6–8, 10, 11, 16, 18, 20, 24, 47, 55, 56]. With the availability of several large-scale datasets [35, 44, 60], the latest deep learning methods can estimate hand keypoints very accurately in well-lit conditions. However, hand pose estimation in low light conditions is still under-explored. Existing methods for low light hand pose estimation requires additional sensors such as Infrared (IR) [40] and Time-of-Flight (ToF) [52]. To the best of our knowledge, there is currently no method that can estimate hand poses in extremely low light conditions using a standard digital camera such as smartphones and DSLRs.

This problem is extremely challenging due to two major reasons. Firstly, images captured in extremely low light conditions get corrupted by excessive amount noise. Hence it is challenging to accurately localize hand joints from these images even for a human. Secondly, deep learning based hand pose estimation requires a large amount of labeled data. However, to the best of our knowledge, there exists no dataset consisting of low-light images from a digital camera along with associated hand pose labels.

In this paper, we propose the first method for extreme low light hand pose estimation, named XPose. We propose to estimate hand pose from images captured in low light conditions in linear domain, instead of the traditional non-linear sRGB domain. The efficacy of linear image processing has been explored in fields such as image denoising [5] and HDR [23]. However, to the best of our knowledge, none of the existing methods have explored linear images for hand pose estimation. We also introduce a large-scale dataset for hand pose estimation in extremely low light conditions, named LLHands. Our dataset consists of low light images in both linear demosaiced and non-linear sRGB domains, along with accurate ground-truth labels. We also captured paired multi-view and well-lit images for every image in our dataset. The proposed LLHands dataset consists of 120K images in linear and non-linear domains along with hand pose labels. To the best of our knowledge, this is the first dataset for hand pose estimation in extremely low light conditions using images captured from a digital camera.

Finally, we propose a novel algorithm to guide the training of the proposed monocular low-light hand pose network using additional knowledge from well-lit multi-view images available in our dataset. In order to leverage the paired well-lit images, we propose a knowledge distillation approach to transfer rich hand pose representation learned from well-lit images to low-light training. Further, we utilize the multi-view data by constraining our network to predict hand poses that are consistent across all the views and lighting conditions (Fig. 1). We show that the proposed methodology utilizing multi-view well-lit guidance enables our network

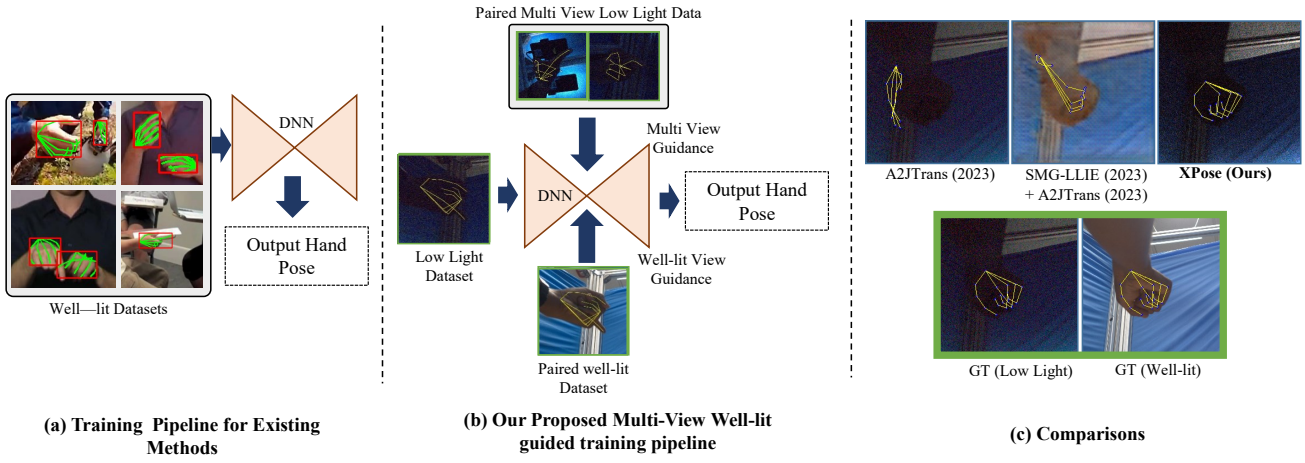


Figure 1. Extreme Low Light Hand Pose Estimation. **(a)** Existing methods for hand pose estimation use well-lit images for training, using guidance only from the ground-truth label. **(b)** We propose a novel method wherein additional guidance is provided using multi-view, well-lit images for low light hand pose estimation. **(c)** Our method significantly outperforms the state of the art method A2JTrans [28], even using a low light enhancement (SMG-LLIE) based pre-processing. Low light and paired well-lit ground-truths shown for reference

to predict high quality hand poses from a single linear image captured in extremely low light conditions. A sample result of the our proposed method is shown in Fig. 1 (c). In this figure, we provide an example of an extremely low light capture, where even human vision is limited. Existing hand pose methods completely fail to estimate hand pose in such scenarios even using a low light image enhancement based pre-processing (SMG-LLIE [48]). However, our method is able is able to faithfully estimate the hand pose in such extreme low light conditions.

To summarize, our major contributions are:

- We propose XPose, the first method to the best of our knowledge, for monocular hand pose estimation in extremely low light conditions using images captured from a digital camera.
- We propose an innovative methodology for this challenging problem by leveraging information from multi-view well-lit images during training.
- We introduce a large scale dataset, named LLHands, consisting of both linear demosaiced and non-linear sRGB images captured in extremely low-light conditions, along with accurate 3D hand pose labels. To the best of our knowledge, this is the first hand pose dataset captured in extremely low light conditions.

2. Related Works

2.1. Hand Pose Estimation

Hand pose estimation methods can be broadly classified into skeleton-based [25, 36, 39, 45, 46, 50, 51, 59] and mesh-based approaches [1–3, 6–8, 10, 16, 18, 20, 24, 47, 55, 56]. Skeleton-based approaches aims to estimate the locations of hand joints in 2d or 3d. Some of the earlier works propose to learn 2.5D [25] or 3D [59] representations using labelled datasets. Mesh-based approaches aim to estimate the complete dense hand pose from one or more images. Most of these methods [1–3, 24, 56] assume a parametric model for the hand using the popular MANO representation [41]. A recent method [10] propose to eliminate the MANO representation by directly regressing the hand mesh using graph CNNs [10]. There have also been several recent methods aimed at estimating hand poses in two-hand interacting scenarios [13, 28–30, 33, 35] and hand-object interaction scenarios [2, 12, 24, 27, 29, 34, 49]. There have also been methods to estimate 3D hand pose using multiple cameras [21, 22, 57]. However, all these methods assume well-lit conditions for accurate pose estimation. Methods specifically designed for low light is under-explored in computer vision. Recently, [31] introduced the first method and dataset for human pose estimation from RGB images captured in extreme low light. However, to the best of our knowledge, there is no existing method for hand pose estimation from images captured in low-light conditions using digital cameras.

2.2. Hand Pose Datasets

There have been several datasets proposed for hand pose estimation. Methods such as [32, 37, 59] introduce synthetic datasets for hand pose estimation. One of the earliest real-image dataset for hand pose estimation was proposed by [38]. However, they use manual annotations to label the hand poses, which is extremely inconvenient and not scalable. To alleviate this challenge, methods such as

[17, 22, 53] use markers attached to the hand to track the location of the joints. However, this results in biased data consisting of visible markers. To address these challenges, a fully automated hand pose annotation method was proposed by [44]. Inspired by this method, most of the later datasets propose semi-automated method to label the hand pose [21, 39, 60]. Recent methods have also introduced datasets for interacting hands [35] and hand-object interactions [4, 14, 19, 27, 29, 43, 49]. However, all these datasets are captured in well-lit conditions in sRGB domain. To the best of our knowledge, there is no existing dataset for hand pose estimation in low-light conditions. In this paper, we introduce the first hand-pose dataset consisting of images captured in low-light conditions. Our dataset is also the first hand pose dataset comprising of images in linear demosaiced and non-linear sRGB domains for both low and well-lit conditions.

3. LLHands Dataset

We propose LLHands, the first dataset for hand pose estimation in extremely low light capture conditions. The proposed LLHands dataset consists of low light images in both linear RGB and non-linear sRGB formats, along with accurate 2D and 3D hand pose labels. We also provide aligned well-lit images for each of the low-light images. We capture a large scale dataset using a multi-camera capture rig consisting of 60 smartphones. We also propose a fully-automated approach for obtaining ground-truth hand pose labels for low-light images captured using the multi-camera capture rig. The proposed LLHands dataset consists of 120K images from 10 subjects along with associated hand pose labels. We believe that the proposed dataset will pave the way for new research in both single and multi-view hand pose estimation in extremely low light conditions. We provide details on the data capture methodology and automated data labeling in the following subsections.

3.1. Low Light Data Capture

We construct a custom multi-camera rig consisting of 60 smartphones to capture the proposed LLHands dataset. Our multi-camera rig is depicted in Fig. 2 (a). The smartphones are mounted along a multi-level hexagonal rig, directed towards its center. Our capture rig also consists of a highly responsive wirelessly controlled uniform lighting system which can alter the lighting from 0.01 lux to 1000 lux. We also developed an associated software which allows us for synchronized capture using all the 60 smartphones. During capture, the subjects are asked to place their hands in a static manner at approximately the center of the rig. We then perform two captures in quick succession: a low light capture and an associated well-lit capture. We capture the low-light images at 0.5, 1 and 2 lux respectively, and the well-lit capture at 1000 lux. The subjects were requested to perform a variety of common hand poses such as pinch, pointing, open palm and fist using both left and right hands. We use COLMAP [42] to obtain camera calibration parameters from each captured set of images. COLMAP is fast and can automatically calibrate cameras using a single set of multi view images. This makes calibration process easier compared to other methods which uses cumbersome methods requiring 3d calibration targets and a large number of captures [35].

3.2. Automated Hand Pose Labelling

We label 21 hand keypoints similar to the format followed by [44]. Manually labelling all the keypoints for the entire dataset is extremely challenging due to complex hand articulation and the extreme low light capture conditions. Methods such as [44] and [35] proposed RANSAC based methods for automated hand pose labeling. However, these methods assume highly accurate camera calibration. Since, we use COLMAP for automated calibration, certain cameras may not be well-calibrated. Hence naively using these methods would impact the process of RANSAC based inlier detection. Hence we propose a novel methodology for automated hand labelling which accounts for camera views with poor calibration parameters.

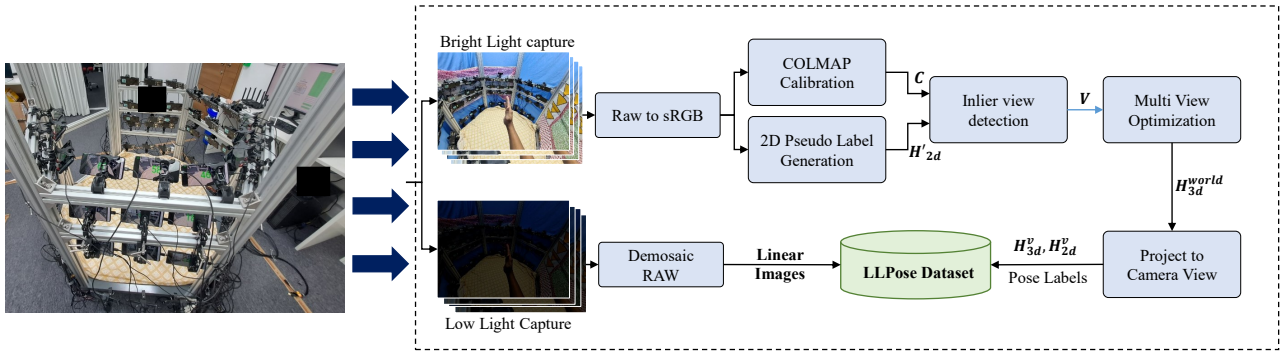
For each low-light capture, we compute the ground-truth labels using the paired bright-light data, as shown in Fig. 2. Let $\mathbf{I_n^l}$ and $\mathbf{I_n^b}$ denote the set of n-view images captures in low light and bright light respectively. We first pass $\mathbf{I_n^b}$ through a pre-trained 2D hand pose estimator [54] to obtain a set of 2D pseudo labels, $\mathbf{H'_{2d}}$. We propose an iterative approach to determine the views with accurate camera calibration parameters as well as pseudo hand pose labels. Let $\mathbf{C_i}$ denote the set of views being used during iteration i. For the first iteration, we set $\mathbf{C_i}$ to the total number of cameras in our multi-capture rig. During each iteration (i), we use RANSAC [15] to determine the inlier views for every keypoint. We choose the set of views with reprojection error less than 10 pixels as the inlier set for iteration i. Let $\mathbf{C_i^{inlier,k}}$ denote the set of inlier views for keypoint k after iteration i. Next we compute the 3D location of the keypoint as follows:

$$H_{3d}^{k,i} = \arg\min_{\mathbf{x}} \Sigma_{v \in \mathbf{C_i^{inlier,k}}} \|\mathcal{P}_v(\mathbf{x}) - H_{2d}^{\prime(k,v)}\|_2 \quad (1)$$

where $H_{3d}^{k,i}$ is the required 3d location of the k^{th} keypoints after the i^{th} iteration, $H_{2d}^{\prime(k,v)}$ denote the pseudo label of the k^{th} keypoint for view v and $\mathcal{P}_v$ denote a projection function to view v. We use BFGS algorithm for this optimization.

Next we compute the average error of all the keypoints for all the views as follows:

$$\mathcal{E}_{(i,v)} = \Sigma_{k=1}^{21} \|\mathcal{P}_v(H_{3d}^{k,i}) - H_{2d}^{\prime(k,v)}\|_2 \quad (2)$$



(a) Multi View Data Capture Rig **(b) Automated Hand Pose Labelling**

Figure 2. Data Capture and Labeling Pipeline. We develop a fully-automated hand pose labeling method using RANSAC based multi-view triangulation from well-lit images.

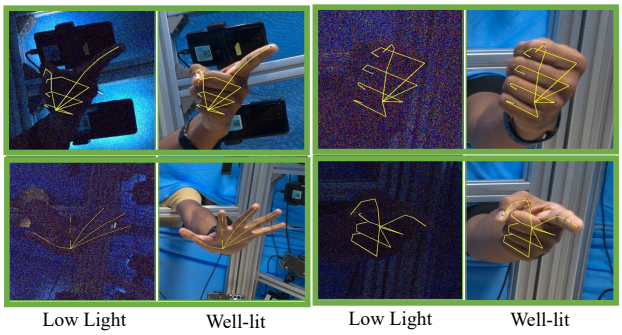


Figure 3. Samples from proposed dataset. Our dataset consists of images in low light and well-lit conditions. Note that our automated labelling pipeline yields accurate hand pose labels

We denote the views with $\mathcal{E}_{(i,v)}$ greater than 30 pixels as the views with erroneous camera calibration parameters and remove them from $\mathbf{C_i}$ for the next iteration of the algorithm. We observe that two iterations of our algorithm provides a good estimate of the 3D and 2D labels. We finally perform a manual verification to further reject images with poor labels. Some examples from our dataset is provided in Fig. 3. It can be seen that our dataset consists of images captured in extremely low-light conditions and corrupted by excessive noise. We provide accurate ground-truth hand pose labels as shown in Fig. 3.

4. XPose: Extreme Low Light Hand Pose Estimation

Hand pose estimation from a single low light image is very challenging due to the presence of high amount of noise. This noise gets further aggravated due to the multitude of non-linear operations performed by the camera Image Signal Processor (ISP). In order to alleviate the challenges associated with extreme low-light image capture, we propose to perform hand pose estimation using linear images. As observed by Low Light Image Enhancement methods such as [5], image processing is easier in the linear domain as opposed to the non-linear sRGB domain. Hence, we process the RAW images from the camera sensors using only linear operations such as black level correction and demosaicing and forego all the non-linear operations.

Even though the linearity provided by raw image helps in low light image processing, estimating hand pose from a single noisy image is still ill-posed. In very low light conditions, the signal-to-noise ratio reduces significantly even for images in linear domain. This makes it challenging for a neural network to learn low light hand pose estimation using supervision from only the ground-truth hand pose labels. Hence we propose to provide additional guidance during training time using the multi-view well-lit images in the proposed LLPose dataset.

An overview of the proposed method is given in Fig. 4. During every iteration of the training, we randomly sample n low-light views ($\mathbf{I^l}$) and n well-lit views ($\mathbf{I^w}$) of the same hand pose. Each of the images in $\mathbf{I^l}$ and $\mathbf{I^w}$ are passed through low-light hand pose estimators ($\mathcal{N}^l$) and well-lit hand pose estimators ($\mathcal{N}^w$) respectively. The network weights are shared across multiple views, but not between $\mathcal{N}^w$ and $\mathcal{N}^l$. The outputs from all the views are jointly optimized to provide additional multi-view guidance for training $\mathcal{N}^l$. During inference $\mathcal{N}^l$ is used as single image low light hand pose estimator. We use CPN [9] as the backbone architecture for both $\mathcal{N}^l$ and $\mathcal{N}^w$, as shown in Fig. 4 (a). Our network consists of a resnet50 based feature extractor followed by GlobalNet and RefineNet modules [9] to estimate the hand keypoints. To enable our network to predict 3D locations of the hand keypoints, our RefineNet consists of two prediction heads. One of the prediction heads estimates the 3D locations of every hand keypoint relative to the wrist, while the other head estimates the

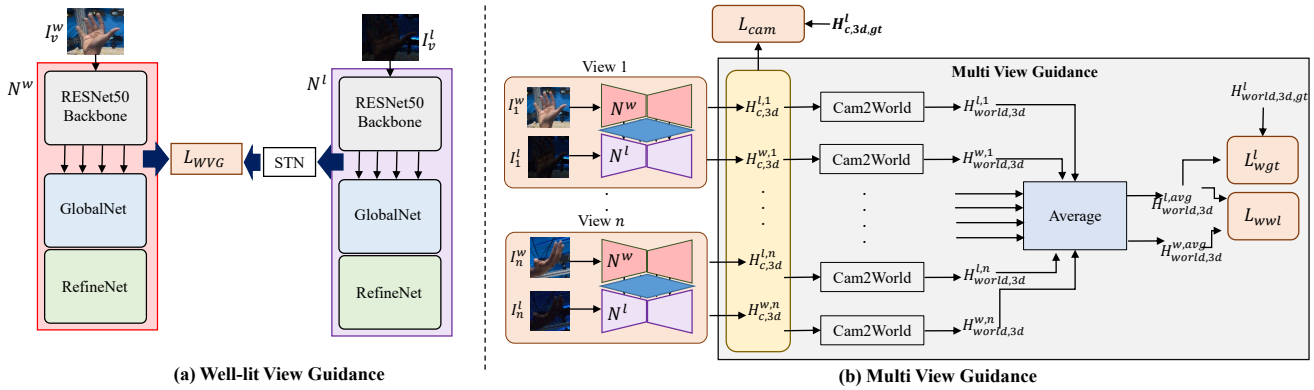


Figure 4. Proposed Method. (a) The proposed well-lit view guidance to utilize the well-lit images in our dataset. (b) The proposed multi-view guidance to generate consistent results across all the views

absolute depth of the wrist keypoint. The final 3D location is then obtained by combining these outputs.

4.1. Well-lit view Guidance

An overview of the proposed well-lit view guidance training is provided in Fig. 4 (a). For a given view v, we first linearly boost the intensity of the low-light image (I^l_v) to match the intensity of that of the well-lit image (I^w_v). During training, the boosted I^l_v and I^w_v are provided as inputs to $\mathcal{N}^l$ and $\mathcal{N}^w$ respectively. The goal of $\mathcal{N}^w$ is to provide guidance to $\mathcal{N}^l$ using the rich hand pose feature representation learned from well-lit images. The objective of $\mathcal{N}^l$ is to learn to predict 3d hand pose from the noisy and boosted low-light image.

We first pre-train $\mathcal{N}^w$ using linear well-lit images and their corresponding hand pose labels. The intermediate feature maps of the trained $\mathcal{N}^w$ contains a rich representation of semantic information required to estimate hand pose. Next, we distill this rich hand-pose representation to guide the training of $\mathcal{N}^l$. During each iteration of training, we provide I^w_v to the pre-trained $\mathcal{N}^w$ and boosted I^l_v to $\mathcal{N}^l$. The weights of $\mathcal{N}^w$ are frozen during this stage of this training. We extract m intermediate feature maps $\mathcal{F}^l_{1..m}$ and $\mathcal{F}^w_{1..m}$ from the low-light and bright-light images respectively (denoted in Fig. 4 (b)). We aim to minimize the distance between $\mathcal{F}^l_{1..m}$ and $\mathcal{F}^w_{1..m}$ so that the rich hand-pose representation contained in $\mathcal{F}^w_{1..m}$ can be distilled into $\mathcal{F}^l_{1..m}$. However, trivially minimizing the error between $\mathcal{F}^l_{1..m}$ and $\mathcal{F}^w_{1..m}$ would not provide effective distillation. This is due to minor misalignments between I^w_k and I^l_k incurred during the sequential capture process detailed in section 3.1. To alleviate this issue, we propose to learn a feature level alignment between $\mathcal{F}^l_{1..m}$ and $\mathcal{F}^w_{1..m}$. To this end, we pass $\mathcal{F}^l_{1..m}$ through a Spatial Transformer Network (STN) [26] to learn an affine transformation between the features extracted from the low-light images and the well-lit images. STN consists of differentiable layers allowing back propagation and consequently distillation of information from $\mathcal{N}^w$ to $\mathcal{N}^l$. We minimize the error between $\mathcal{F}^w_{1..m}$ and the affine transformed $\mathcal{F}^l_{1..m}$ using a well-lit view guidance (wvg) loss as follows:

$$L_{wvg} = \Sigma^n_{j=1}\Sigma^m_{i=1}\|\mathcal{F}^w_i - STN(\mathcal{F}^l_i)\|_2 \tag{3}$$

where n denotes the number of views and $\|.\|_2$ denotes the L_2 distance.

We also use the ground-truth 3D hand pose labels in camera coordinate system to guide the output of $\mathcal{N}^l$ as follows:

$$L_{cam} = \Sigma^n_{v=1}\|H^{l,v}_{3d} - H^{gt,v}_{3d}\|_2 \tag{4}$$

where $H^{l,v}_{3d}$ and $H^{gt,v}_{3d}$ denotes the 3d hand pose obtained from $\mathcal{N}^l$ and the ground-truth hand pose labels respectively, in the camera coordinate system of view v.

The final loss function for well-lit view guidance is as follows:

$$L_{well} = \alpha.L_{wvg} + \beta.L_{cam} \tag{5}$$

where α and β are empirically chosen weighing parameters.

4.2. Multi View Guidance

As detailed in section 3.1, our dataset contains multiple views for every hand pose. We further leverage this information to provide multi-view guidance during the training of $\mathcal{N}^l$. An overview of the proposed multi-view guidance is provided in Fig. 4 (b). We design a set of loss functions to constrain the training of $\mathcal{N}^l$ such that it generates consistent results across multi-view data. Our proposed methodology enforces collaboration between all the views using various consistency loss functions. This enables our network to leverage multi-view information during back-propagation of loss. Each of these loss functions are detailed below:

4.2.1 World Consistency Loss

We introduce World Consistency Loss to enforce multi-view consistency in the world coordinate system, representing the collective coordinate system for the all the cameras. Let $H^{l,v}_{camera}$ and $H^{w,v}_{camera}$ denote the 3d hand pose estimations by $\mathcal{N}^l$ and $\mathcal{N}^w$ respectively in the camera coordinate system of view v. We first transform these 3d keypoints into world coordinate system using the extrinsic parameters obtained using COLMAP based camera calibration. Let $H^{l,v}_{world}$ and $H^{w,v}_{world}$ denote the corresponding 3D hand pose in world coordinate system. For consistent low-light predictions, $H^{l,v}_{world}$ should be the same for all values of v. To enforce this constraint, we minimize the distance between the average of the low-light predictions and the ground-truth hand pose label in world coordinate system as follows:

$$L_{wgt} = \| \frac{\Sigma_{v=1}^{n} H^{l,v}_{world}}{n} - H^{gt}_{world} \|_2 \tag{6}$$

where n denotes the number of views and H^{gt}_{world} denotes the ground-truth hand pose in world coordinate system.

We also enforce consistency with hand pose estimations obtained from well-lit images in world coordinates. Since I^w contains much less noise compared to I^l, predictions from $\mathcal{N}^w$ is more reliable than those from $\mathcal{N}^w$. Hence adding consistency loss with H^{w}_{world} forces $\mathcal{N}^l$ to predict hand poses similar to that of $\mathcal{N}^w$. This loss is defined as follows:

$$L_{wwl} = \| H^{l,avg}_{world} - H^{w,avg}_{world} \|_2 \tag{7}$$

we use the notations $H^{l,avg}_{world}$ and $H^{w,avg}_{world}$to denote hand pose predictions respectively from the low light and well-lit images in world coordinates, averaged over all the views

4.2.2 Pixel Reprojection Loss

We also use a pixel level reprojection loss to ensure the predictions from all views are consistent in the 2D pixel space. As observed by [57], pixel level 2D predictions are generally more accurate than depth predictions for single image hand pose estimation. Hence reprojecting the estimated 3D hand pose to ensure consistency across multi-view 2D hand poses enables the network to predict more accurate 3D hand pose using multi-view collaboration. To enforce this constraint, we first project $H^{l,v}_{camera}$ to all the n views. We then minimize the error between the predicted 2D hand poses and the projected hand poses as follows:

$$L_{pr} = \Sigma_{v=i}^{n} \Sigma_{i=1}^{n} \| \mathcal{P}_i(H^{l,v}_{camera}, i) - H^{l,i}_{2d} \|_2 \tag{8}$$

where n is the number of views, $\mathcal{P}_i(p^v_{camera}, i)$ denotes the 2D projection of a point p in the camera coordination system of view v to view i, and $H^{l,i}_{2d}$ denotes the 2D prediction for view i by $\mathcal{N}_l$.

4.3. Full Loss Function

The final objective of the network is to minimize the error between 3D predictions and ground-truth labels, while ensuring consistency across all the views and well-lit predictions as follows:

$$L = \gamma.L_{well} + \delta.L_{wgt} + \eta.L_{wwl} + \sigma.L_{pr} \tag{9}$$

where γ, δ, η and σ are empirically chosen weighing parameters.

5. Experiments and Results

5.1. Comparisons against state-of-the-art

We use the testing dataset of the proposed LLPose dataset consisting of 1200 low light images to evaluate our method. We use MPJPE (Mean Per Joint Position Error) as the objective metric for our evaluations. We compare the proposed method against the state-of-the-art methods in 3D hand pose estimation, Zhou *et.al* [58], Keypoint Transformer [20] and A2J Transformer [28], under different light conditions, as summarized in Table 1. First, we evaluate using the original checkpoints provided by these methods. Since these methods are originally designed for well-lit images, they perform significantly worse on low-light images compared to our method. From the table, it can be seen that our method is $\sim$12$\times$ better in terms of MPJPE compared to the original checkpoint of the next best method (KptTrans - Row 3).

We also perform another experiment wherein the low light images are first enhanced using a state-of-the-art low light image enhancer (SMG-LLIE [48]). This method enhances the low-light input images to make them appear as if they were captured in bright-light conditions. These resulting images are then provided to the state-of-the-art methods for hand pose estimation. From Table 1, (rows 4-6), it can be seen that this approach outperforms the original checkpoints by a significant margin. However, the results obtained by this approach is still significantly inferior compared to our method. Our method outperforms the MPJPE scores obtained using this approach by $\sim$9$\times$. This result suggests that hand-pose aware training, using paired labels is required for effective low-light hand pose estimation, instead of using image enhancement based pre-processing steps.

We also finetune [58] and [20] using our proposed LLPose dataset. As shown in Table 1 (rows 7-9), our method outperforms the state-of-the-art methods even after finetuning using low light images by $\sim$1.9$\times$. This result suggests that our method, which leverages multi-view well-lit images during training, is able to perform hand pose estima-

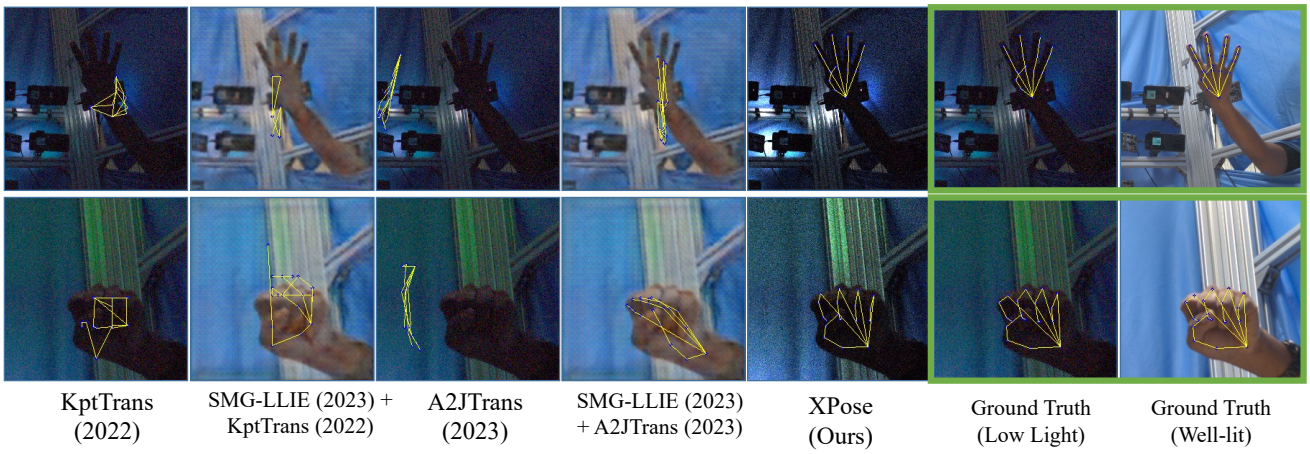


Figure 5. Comparisons against state of the art methods for extreme low light hand pose estimation (0.5 lux). KptTrans [20] and A2JTrans [28] completely fails in low light conditions. Pre-processing using SMG-LLIE [48] improves the output, but it is significantly degraded compared to our method. We provide ground-truth labels on low light and well-lit images for reference

Method	0.5 lux	1 lux	2 lux	Average	MACs/pixel (millions)
Zhou *et.al* '20 [58]	153.22	151.79	150.12	151.71	0.21
A2JTrans '23 [28]	95.45	94.82	94.72	95.00	0.44
KptTrans '22 [20]	76.38	76.17	76.09	76.21	0.19
SMG-LLIE '23 [48] + Zhou *et.al* '20 [58]	61.94	60.42	60.14	60.83	0.57
SMG-LLIE '23 [48] + A2JTrans '23 [28]	58.43	56.65	56.59	57.22	0.80
SMG-LLIE '23 [48]+ KptTrans '22 [20]	55.58	53.52	53.01	54.03	0.55
Zhou *et.al* '20 (finetuned) [58]	28.83	28.55	27.03	28.14	0.21
KptTrans '22(finetuned) [20]	11.98	11.16	10.80	11.32	0.19
XPose (Ours)	**6.39**	**6.29**	**5.65**	**6.12**	**0.15**

Table 1. Quantitative comparisons (MPJPE) against state-of-the-art methods on LLHands testing set with various lighting conditions. Lower value is better. Our method outperforms all the other methods on all lighting conditions. Our method has less computational complexity compared to other methods.

tion in challenging low-light scenarios better than the other methods, which are dependent only on single images during training. Further, it can be seen that finetuned methods outperforms the original checkpoints by a significant margin. This shows the necessity of low-light data for hand pose estimation.

We also provide qualitative comparison results of the our method in Fig. 5. We provide examples of images captured in extremely low light conditions (0.5 lux) in this comparison. We have chosen samples where even humans cannot easily identify the hand keypoints. It can be seen that the existing methods completely fail to predict hand poses in extreme low light conditions. While SMG-LLIE based preprocessing improves the results slightly, the results are still heavily degraded. On the other hand, our method is able to predict accurate hand poses in challenging low-light conditions. Our combination of low light dataset and guidance from multi-view well-lit images enables our network to effectively learn hand pose estimation in extreme low light conditions.

Training Data	MPJPE
Non Linear Well-lit	34.74
Non Linear Low Light	9.67
Linear Low Light	**6.79**

Table 2. Impact of the proposed low light linear dataset. We evaluate baseline architecture using low light test data. Using low light images in linear domain yields the best results

We also analyze the computational complexity of our method using Multiply-Accumulate (MACs/pixels). From Table 1, it can be seen that our method is faster than all the other state-of-the-art methods.

5.2. Impact of Linear Low Light Dataset

In this section, we analyze the impact of the proposed LLHands dataset on low light hand pose estimation (Table 2). Specifically, we investigate the necessity of linear low light images. We use only the CPN based single-view baseline architecture for all experiments in this section. For this analysis, we train the network using three different datasets: a) non-linear (sRGB) well-lit images; b) non-linear (sRGB) low light images; and c) linear low light images. The training set consists of exactly the same hand poses, with the only difference being the lighting conditions and the linearity. We evaluate all of these models using our low-light testing set. From Table 2, it can be seen that the model trained using non-linear low light images outperforms the model trained using non-linear well-lit images by $\sim$3.5$\times$. This result shows the necessity of a domain specific low-light hand pose dataset. Further, it can be seen that the model trained using linear low light images outperforms the MPJPE value of the model trained using non-linear low light images by $\sim$1.5$\times$. This result suggests that performing hand pose estimation in linear domain alleviates the challenges associated with low light image capture, resulting in considerable improvement in accuracy.

5.3. Ablation Studies

In this section, we analyze the impact of the various novel components proposed in this paper. All the experiments in this section are evaluated using linear low-light data. We analyze the impact of Spatial Transformer Networks (STN), well-lit view guidance (WV) and multi view guidance (MV) in Table. 3. For this, we first train a baseline model consisting of only the CPN backbone without STN, WV or MV guidances (Row 1). Next we introduce well-lit view guidance using the loss function L_{well} (Row 2). In this experiment, we do not use STN to align the intermediate feature maps between $\mathcal{N}_l$ and $\mathcal{N}_w$. Next we introduce STN (Row 3) and finally we introduce multi view guidance (Row 4). It can be seen that using all the proposed components (STN, WV, MV) improves the MPJPE score by $\sim$ 0.7 compared to the baseline architecture. Further, it can be seen that using only the WV guidance also improves upon the baseline model. It can also be seen that the use of STN helps improve the MPJPE scores due to better alignment of features.

WV Guidance	STN	MV Guidance	MPJPE
$\times$	$\times$	$\times$	6.79
$\surd$	$\times$	$\times$	6.58
$\surd$	$\surd$	$\times$	6.41
$\surd$	$\surd$	$\surd$	**6.12**

Table 3. Network Component Analysis. Our method which uses STN, Well-lit View (WV) and Multi View (MV) guidance yields the best results.

$\mathbf{L_{well}}$	$\mathbf{L_{pr}}$	$\mathbf{L_{wwl}}$	$\mathbf{L_{wgt}}$	MPJPE
$\surd$	$\times$	$\times$	$\times$	6.41
$\surd$	$\surd$	$\times$	$\times$	6.29
$\surd$	$\surd$	$\surd$	$\times$	6.27
$\surd$	$\surd$	$\surd$	$\surd$	**6.12**

Table 4. Impact of Multi-view loss functions. Our method using all the proposed loss components yields the best results.

We also investigate the impact of the various components of the loss function used in multi-view guidance. For this experiment, we first train a baseline model by removing all the components related to multi view guidance (L_{wgt}, L_{wwl}, L_{pr}). Next we sequentially introduce the pixel reprojection error (Ł$_{pr}$), and the world consistency losses w.r.t to well-lit images (L_{wwl}) and w.r.t the ground-truth (L_{wgt}). The results on this experiment is summarized in Table. 4. It can be seen that the proposed method which uses all the loss components yields the best accuracy scores.

6. Conclusion

In this paper, we propose XPose, the first method for hand pose estimation from digital camera images captured in low light conditions. We also introduce the first dataset for extreme low light hand pose estimation, consisting of 120K images in linear demosaiced and non-linear sRGB domains. Along with low light images and hand pose labels, our proposed method also consists of paired well-lit images captured from multiple views. This multi-view well-lit data consists of additional information which can be utilized during training time to improve single image low light hand pose estimation. To this end, we propose a novel multi-well-lit view guided network architecture for effective single-image low light hand pose training. We use specially designed loss functions to effectively use rich hand pose representation from well-lit images to guide low-light hand pose estimation. Further, we also constrain our network to predict hand pose consistent across multiple views. Our method, using the proposed LLPose dataset, achieves a $\sim$12$\times$ improvement in MPJPE score in low-light conditions, compared to the state-of-the-art methods and $\sim$9$\times$ improvement over these methods after a low light enhancement based pre-processing. We also conduct ablation studies to show the efficacy of our dataset and the novel components introduced in our method. We believe that our dataset and methodology will pave way for new research in low light hand pose estimation.

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 1, 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2020. 1, 2

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 1, 2

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 3

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 1, 4

[6] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 1, 2

[7] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13283, 2021. 1, 2

[8] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021. 1, 2

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 4

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 769–787. Springer, 2020. 1, 2

[11] Xiaoming Deng, Dexin Zuo, Yinda Zhang, Zhaopeng Cui, Jian Cheng, Ping Tan, Liang Chang, Marc Pollefeys, Sean Fanello, and Hongan Wang. Recurrent 3d hand pose estimation using cascaded pose-guided 3d alignments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):932–945, 2022. 1

[12] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020. 2

[13] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *2021 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2021. 2

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 3

[15] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[16] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. *arXiv preprint arXiv:2303.04991*, 2023. 1, 2

[17] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 3

[18] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 1, 2

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 3

[20] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 1, 2, 6, 7

[21] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020. 2, 3

[22] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang,

Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 3

[23] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 1

[24] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 1, 2

[25] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 5

[27] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14713–14724, 2023. 2, 3

[28] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8846–8855, 2023. 2, 6, 7

[29] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 2, 3

[30] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21169–21178, 2023. 2

[31] Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 704–714, 2023. 2

[32] Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20395–20405, 2023. 2

[33] 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. 2

[34] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12989–12998, 2023. 2

[35] Gyeongsik Moon, Shoou-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. 1, 2, 3

[36] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–59, 2018. 2

[37] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1154–1163, 2017. 2

[38] Markus Oberweger, Gernot Riegler, Paul Wohlhart, and Vincent Lepetit. Efficiently creating 3d training data for fine hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4957–4965, 2016. 2

[39] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12999–13008, 2023. 2, 3

[40] Gabyong Park, Tae-Kyun Kim, and Woontack Woo. 3d hand pose estimation with a single infrared camera via domain transfer learning. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 588–599. IEEE, 2020. 1

[41] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[42] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

[43] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 3

[44] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference*

on Computer Vision and Pattern Recognition, pages 1145–1153, 2017. 1, 3

[45] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 2

[46] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–98, 2018. 2

[47] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11698–11707, 2021. 1, 2

[48] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9893–9903, 2023. 2, 6, 7

[49] Yue Xu, Yong-Lu Li, Zhemin Huang, Michael Xu Liu, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. Egopca: A new framework for egocentric hand-object interaction understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5273–5284, 2023. 2, 3

[50] Linlin Yang, Shicheng Chen, and Angela Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11364–11373, 2021. 2

[51] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9877–9886, 2019. 2

[52] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao 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. 1

[53] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4866–4874, 2017. 3

[54] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 3

[55] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11281–11292, 2021. 1, 2

[56] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 1, 2

[57] Xiaozheng Zheng, Chao Wen, Zhou Xue, Pengfei Ren, and Jingyu Wang. Hamuco: Hand pose estimation via multi-view collaborative self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20763–20773, 2023. 2, 6

[58] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 6, 7

[59] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 2

[60] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 1, 3