A Hierarchical Synthetic Benchmark pipeline for Hand-Pose Estimation

1stPreston Mann
CSE. UTA
UTA
Arligton, United States Of America
preston.mann@mavs.uta.edu

***Abstract*—Hand-pose estimation has progressed rapidly on utilizing several different benchmarking datasets, yet we still lack a systematic understanding of which factors of variation cause current models to fail and how data demands scale with task difficulty. We address this gap by generating four synthetic RGB datasets, each introducing a single new challenge while holding all others constant: hand morphology changes, in-plane rotation, unconstrained 3-D rotation and continuous finger flexion. Three networks are trained a lightweight Simple CNN, a randomly initialized ResNet18, and a pre trained ResNet18 are trained under identical environments.Results reveal a hierarchy of difficulty: shape variation is solved by all models; in-plane rotation modestly strains shallow architectures; full 3-D rotation and finger articulation trigger sharp error spikes, where pre-training cuts error by half and limited-capacity models break down entirely. These findings supply the first controlled baseline relating accuracy to specific degrees of freedom and provide a practical guide for dataset construction and model selection in future research.**

***Index Terms*—Pose Estimation, Computer Vision, Machine Learning**

I. Introduction

In computer vision, hand pose estimation is the problem of estimating the joint angles/positions of the human hand from images or video. A wide spectrum of real-world applications can benefit from highly accurate hand pose estimation. Examples include human computer interfaces, gaming, virtual and augmented reality, sign language recognition, and medical rehabilitation. However, we are still far from a general-purpose solution that would match human accuracy for this task

The field of hand pose estimation remains challenging due to the complexity of hand positions and the differences between real-world environments and the environments where data is collected. The research aims to investigate the hierarchy of problems that arise within this domain by identifying and clustering together sub-problems that occur during the joint estimation process. I propose that within a problem space, such as hand pose estimation, there exists a hierarchy of sub-problems that models need to learn and generalize, understanding this hierarchy is critical to improving the overall pipeline performance where the pipeline is independent of any specific model architectures

A fundamental difficulty of current methods is the time, cost and effort required to obtain training data. Learning accurate deep learning models can require millions of training images. However, labeling the location of multiple keypoints (four per finger, plus a few for the palm and wrist) on each image is time-consuming, taking 10-20 seconds per image in our experience. An hour of recording hand motion at 30 frames per second gives us 108,000 images. However, labeling those images would take about 300 hours (assuming 10 seconds per image), and that creates a crucial bottleneck. That is why this paper will be using a small subset of synthetically generated images. Generally, most research targets the full, highly complex task, but this makes it hard to pinpoint why a model succeeds or fails and how much data each source of variation truly demands. To expose those dependencies, we generated four synthetic datasets with DART [2], each adding one new challenge while holding other factors constant: 1 Morphology – hands vary in width, length, and overall scale, but the pose and viewpoint remain fixed. 2 In-plane rotation – the same pose is rotated around the wrist z axis and rescaled. 3 Full 3-D rotation – the hand is freely oriented in space, yet finger articulation is still fixed. 4 Finger flexion – all five fingers interpolate from open to closed across four views (front, back, and side profiles). Every image is paired with 3-D joint coordinates, and we train a convolutional network on each dataset, using mean absolute error (L1 loss)

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

measured in millimeters. By comparing performance four datasets, we reveal exactly when the data variation, orientation, and articulation of the hands push a model beyond its limits, and we will compare that to two resnet18s [3] one where we train the weights and one with pretrained weights from ImageNet [1] This structured approach can deliver a clear benchmark for future hand-pose methods and a practical guide to how task complexity, data volume, and model capacity interact. Later work will expand the size of the dataset and models tested on those datasets as this is just the preliminary experiments.

II. Related Work

Much of the progress in 3-D hand-pose estimation has revolved around three benchmarks: NYU Hand Pose [11], ICVL Hands [10], and MSRA Hands [7]. The NYU dataset [11] pairs RGB and depth images and provides dense 3-D labels for thirty-six joints (Although not all joints are always used), creating a challenging test because severe self-occlusions occur as the single training subject performs complex poses. ICVL Hands [10] is substantially larger and includes ten subject but its records depth only and covers a much narrower and more repetitive pose range and has some flaws in the ground truths, lower-resolution sensor and smaller joint set also limit dexterity analysis. MSRA Hands [7] twenty-one joints and nine subjects performing a defined set of gesture classes, yet its depth frames still exhibit limited background variety, controlled lighting, and a fixed camera-to-hand distance that simplify the viewpoint.

While these benchmarks have been used for serval approaches and pipelines they are entangle multiple sources of difficulty shape variation, viewpoint changes, and finger articulation within the same training/test split. As a result, when a model underperforms it is hard to know whether the culprit is hand morphology, global orientation, sensor noise, or pose complexity. They also provide only partial coverage of the full dexterity space: most frames capture the hand in mid-air rather than interacting with objects, and finger flexion rarely spans the continuous spectrum of the entire hand pose problem space. Most models also make assumptions like the hand is already segmented out of the image and they will also tell the model where the hand is in the image by giving it a bounding box or center of mass for the hand.

Several models such as TriHorn [8], Virtual View Selection [5], DeepPrior [6], and AWR [4] have been proposed and benchmarked as shown on table 1 on these datasets with an increase in accuracy that is getting close to human level. However, despite the increase in model performance when trying to cross validate on different datasets or use these models in the real world they tend to fail showing that the on-paper results are still far from being a true representation of how well they would do in the real world.

This study builds on the lessons of these datasets but takes a different route. By using DART to produce four synthetic datasets that isolate morphology changes, in-plane rotation, full 3-D rotation, and finger flexion in turn while retaining perfectly accurate metric labels we can pinpoint exactly when and why a standard supervised network begins to fail. This controlled hierarchy complements NYU [11], ICVL [10], and MSRA [7] by supplying a clean diagnostic lens for understanding how task complexity, data volume, and model capacity interact.

Dataset	Model	Error (mm)	Rank
ICVL Hands	TriHorn-Net	5.73	#2
MSRA Hands	TriHorn-Net	7.13	#1
NYU Hands	TriHorn-Net	7.68	#3
ICVL Hands	DeepPrior++	8.1	#15
MSRA Hands	DeepPrior++	9.5	#10
NYU Hands	DeepPrior++	12.3	#15
ICVL Hands	AWR	5.98	#4
MSRA Hands	AWR	7.15	#2
NYU Hands	AWR	7.48	#2
ICVL Hands	Virtual View Selection	4.76	#1
NYU Hands	Virtual View Selection	6.4	#1

TABLE I
Performance of models (Average 3D Error in mm)

III. Implementation

Each of the four DART datasets was divided into a 90 percent training split and a 10 percent test split. All images were resized to 224 by 224 pixels, and normalized. Because the datasets already isolated specific factors of difficulty, augmentations were kept minimal. The Datasets created are defined as in table 2

Dataset	Samples	pose variations
1 Morphology	6800	34
2 In-plane rotation	9600	48
3 Full 3-D rotation	7500	75
4 Finger flexion	4800	48
DART full dataset(sampled)	100,000	NA

TABLE II
Datasets benchamrked

Three networks were trained on every dataset under the same hyper parameters and datalaoders. SimplePoseCNN(Table 3) is a compact model that we will be analyzing to see at what point it starts to fail on these datasets, and we will compare it to the other Architectures that serve as a baseline. with four convolution layers, followed by global average pooling and two fully connected layers that directly regress the 3D coordinates of each hand joint.

ResNet18NW re uses the standard ResNet18 backbone but starts by randomly initializing weights to train them. ResNet18WL is structurally identical, yet loads ImageNet pre-trained weights before fine-tuning. We do this to see how a complex architecture can generalize the data without being pre-trained, in that way it can serve as a baseline for our simplePoseCNN model. We also want to baseline a pre-trained model to serve as a baseline for what a good model should be able to do since we are only fine-tuning the weights it should outperform everything starting out. The training was supervised using the 3D ground truths derived from the unity coordinates of DART that were assigned to MANO [9] with post-processing scripts. We used the mean absolute error L1

Layer Type	Parameters	Output Size
Input	RGB Image	$224 \times 224 \times 3$
Conv2D + ReLU	3×32, 5×5, stride 2, padding 2	$112 \times 112 \times 32$
MaxPool2D	2×2	$56 \times 56 \times 32$
Conv2D + ReLU	32×64, 3×3, padding 1	$56 \times 56 \times 64$
MaxPool2D	2×2	$28 \times 28 \times 64$
Conv2D + ReLU	64×128, 3×3, padding 1	$28 \times 28 \times 128$
MaxPool2D	2×2	$14 \times 14 \times 128$
Conv2D + ReLU	128×256, 3×3, padding 1	$14 \times 14 \times 256$
AdaptiveAvgPool2D	1×1	$1 \times 1 \times 256$
Flatten	–	256
Linear + ReLU	$256 \rightarrow 512$	512
Linear	$512 \rightarrow (21 \times 3)$	63
Output	Reshaped to $(21, 3)$	3D keypoints

TABLE III
ARCHITECTURE FOR SIMPLEPOSECNN.

loss for all joints. The Adam optimizer was employed with an initial learning rate of 1e-4, a batch size of 32 and 20 total epochs per training run. The best performing models in the validation split were retained for final testing. Each experiment was repeated multiple times to gauge for any extreme variance.

IV. RESULTS

Dataset	ResNet18WL	ResNet18NW	SimpleCNN
Morphology	1.34	1.19	1.15
In-plane rotation	0.95	1.14	3.30
Full 3-D rotation	1.29	3.45	27.96
Finger flexion	2.06	4.63	21.15
DART full dataset	3.01	NA	NA

TABLE IV
RESULTS OF THE MODELS ON THE DATASET

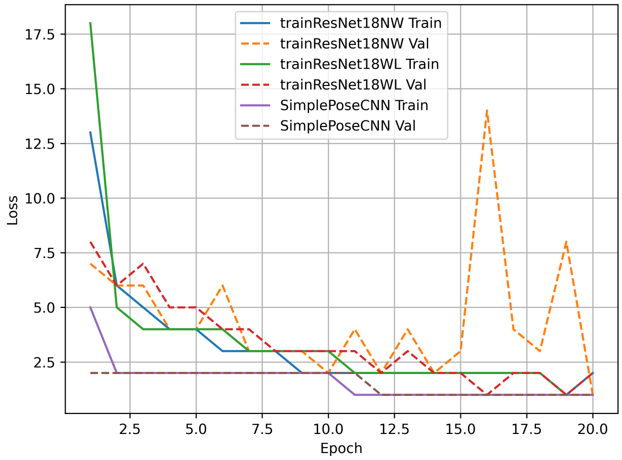

Fig. 1. 1 Morphology

We evaluated the three networks on our datasets mirror the rising difficulty of our synthetic hierarchy where our mean-absolute error is reported in millimeters. Dataset 1 morphology changes with a consistent pose. All three models solved the task with the expected precision. The pre-trained ResNet18 finished at 1.34 mm, the randomly initialized ResNet18 at 1.19 mm, and the SimplePoseCNN at 1.15 mm.

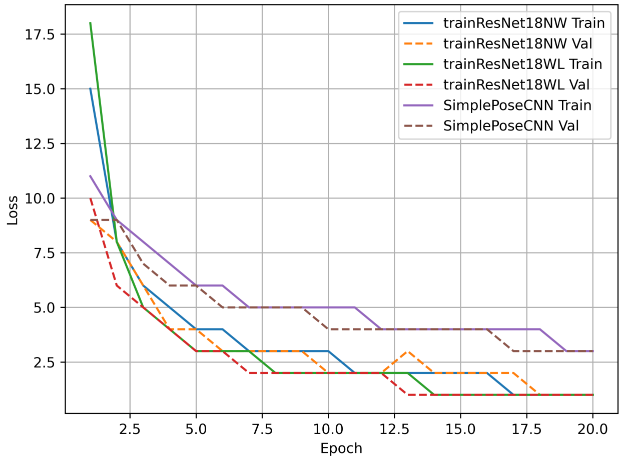

Fig. 2. 2 In-plane rotatio

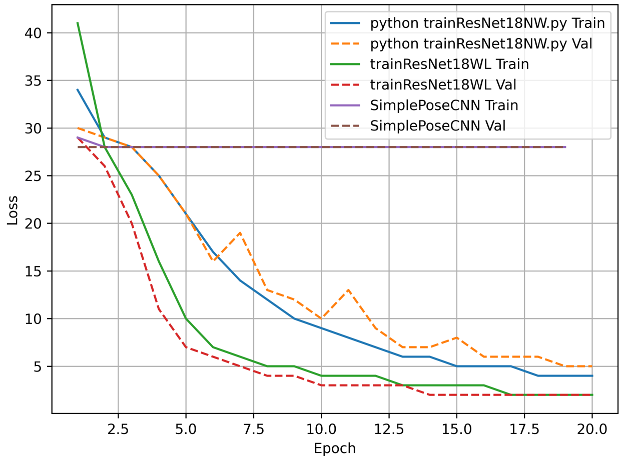

Fig. 3. 3 Full 3-D rotation

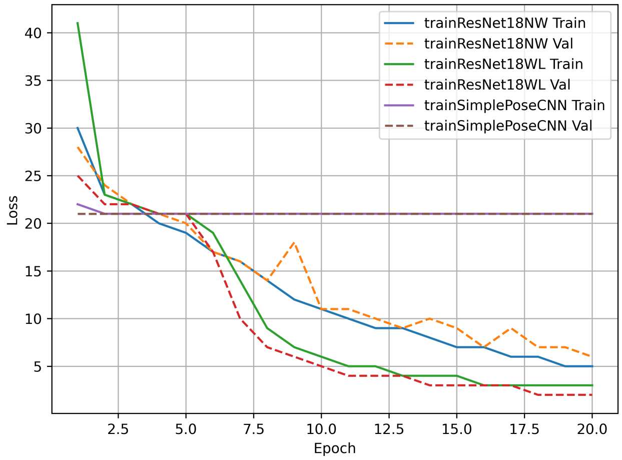

Fig. 4. 4 Finger flexion

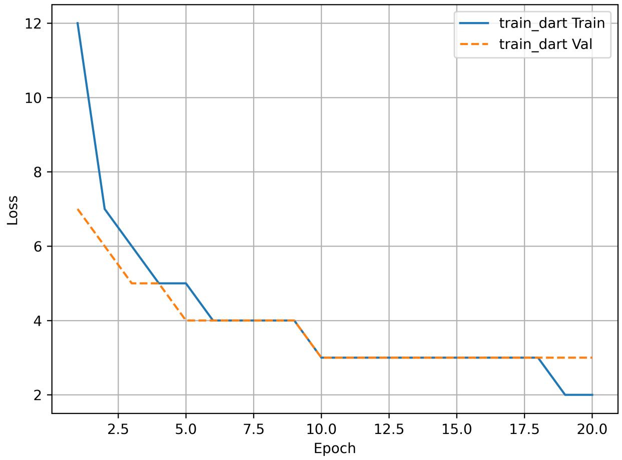

Fig. 5. full dart dataset

Tiny differences suggest that hand-shape variation alone does not require deep residual features or ImageNet pre-training. Dataset 2 in-plane rotation. Introducing rotation on the z axis of the wrists with a consistent pose: pre-trained ResNet18 dipped further to 0.95 mm, while ResNet18-rand hovered at 1.14 mm. The Simple CNN's error tripled to 3.30 mm, revealing how a simple network will start to struggle once global orientation rather than purely local shape needs to be modeled. While the other models that are more complex did had higher performance.

Dataset 3 full 3-D rotation. Allowing rotational freedom in any direction made this estimation task the dominant challenge. The pre-trained ResNet18 remained robust at 1.29 mm but the randomly initialized variant deteriorated to 3.45 mm. The SimplePoseCNN completely failed, finishing at 27.96 mm over twenty times worse than its own score on Dataset 1. Training curves show it flat lining early, indicating that its limited model size cannot resolve depth also showing that rotation is a clear challenge and will lead to a complete model failure for a simple model also indicating that more data with more rotational freedom would be helpful for constructing a better dataset.Dataset 4 finger flexing. Adding continuous articulation from open palm to close palm across four viewpoints pushed every model. The pre-trained ResNet18 rose modestly to 2.06 mm, still very good. The trained ResNet18 jumped to 4.63 mm, and SimplePoseCNN remained stuck above 21 mm indicating that it still suffers from failure at this task but it's not as hard as dealing with rotations. Across all data splits a clear hierarchy emerges. Shape variation alone is trivially learnable in-plane rotation introduces some strain; 3-D orientation is a real breaking point and continuous finger flexion is another breaking point for simple models. Pre-training on generic images buffers much of this difficulty, cutting error by roughly half whenever orientation or articulation is present. The SimplePoseCNN stops improving after just a few epochs, its validation loss mirroring the flat training curve, a sign that the network simply cannot fit to the added degrees of freedom.

V. Conclusion

This work introduces a controlled, four-level synthetic benchmark that isolates the individual challenges that might complicate hand-pose estimation in real world. By training the same three networks on each level we obtain three key insights: Shape variation alone is trivial adding in plane rotation increases error slightly arbitrary 3-D orientation and finger flexion yield the first true failure points. Pre-training on ImageNet features consistently halve the error once orientation or articulation enters the picture, confirming that generic visual priors transfer well to hand pose estimation. A shallow CNN matches large models on the easiest split and in some cases does better in the earlier stages of training but collapses when higher-dimensional kinematics must be represented, underscoring the need for deeper backbones and more data. For our future work We plan to expand the synthetic dataset to cover more hand poses and extend the dataset to larger dataset. Ideally, we will generate 100k samples for each dataset and we will also evaluate self-supervised and transformer-based architectures on the same hierarchy of datasets and explore curriculum strategies that introduce degrees of freedom progressively during training.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[2] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. DART: Articulated Hand Model with Diverse Accessories and Rich Textures. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. Awr: Adaptive weighting regression for 3d hand pose estimation, 2020.

[5] Dexin Zuo Cuixia Ma Jian Gu Ping Tan Hongan Wang Xiaoming Deng Yinda Zhang Jian Cheng, Yanguang Wan. Efficient virtual view selection for 3d hand pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[6] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation, 2017.

[7] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1106–1113, 2014.

[8] Mohammad Rezaei, Razieh Rastgoo, and Vassilis Athitsos. Trihornnet: A model for accurate depth-based 3d hand pose estimation. *Expert Systems with Applications*, page 119922, 2023.

[9] 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), November 2017.

[10] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3786–3793, 2014.

[11] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33, August 2014.