Learning Hand Articulations by Hallucinating Heat Distribution

Chiho Choi Sangpil Kim Karthik Ramani
Purdue University
West Lafayette, IN 47907, USA
{chihochoi, kim2030, ramani}@purdue.edu

Abstract

We propose a robust hand pose estimation method by learning hand articulations from depth features and auxiliary modality features. As an additional modality to depth data, we present a function of geometric properties on the surface of the hand described by heat diffusion. The proposed heat distribution descriptor is robust to identify the keypoints on the surface as it incorporates both the local geometry of the hand and global structural representation at multiple time scales. Along this line, we train our heat distribution network to learn the geometrically descriptive representations from the proposed descriptors with the fingertip position labels. Then the hallucination network is guided to mimic the intermediate responses of the heat distribution modality from a paired depth image. We use the resulting geometrically informed responses together with the discriminative depth features estimated from the depth network to regularize the angle parameters in the refinement network. To this end, we conduct extensive evaluations to validate that the proposed framework is powerful as it achieves state-of-the-art performance.

1. Introduction

Rapid advances in human-computer interaction interfaces have been promising a realistic environment for gaming and entertainment in the last few years. However, a comprehensive hand tracking technology that would enhance virtual and augmented reality experiences still does not exist. Extensive and lengthy researches [20, 17, 35, 23, 28, 18, 4, 19, 26, 7, 36] have been directed toward identifying (i) the articulation complexity of the hand, (ii) self-similarity and self-occlusion of the fingers, and (iii) data acquisition artifacts such as depth noise. Although these research efforts have provided a coarse interpretation of hand movements, the current hand pose estimation approaches do not include: (i) an understanding of the geometric consistency of complex kinematic poses of the articulated hand and (ii) an additional input modality (besides a single depth image) to produce a better estimation model. In this paper, we demonstrate that better hand pose estimation can be attainable when these gaps are addressed.

We propose a promising method for 3D hand pose estimation that achieves performance higher than or comparable to the state-of-the-arts. Specifically, we exploit a convolutional neural network (ConvNet) model which can extract the property of heat distribution over a 3D hand mesh model from a single depth image. The proposed method incorporates a heat distribution network to learn a geometrically informative representation of hand articulations as an additional modality. At training time, our modality hallucination network takes as input a depth image and is trained to capture the corresponding heat distribution modality. Thus, our method produces both the depth and heat distribution features from a single depth image at test time.

With the boom of interest in deep learning, 3D hand pose estimation is increasingly becoming a part of the learning and development processes of mid-level features learned from a large dataset. In view of this, previous approaches have been proposed for estimating hand poses by incorporating a prior model [18], regressing heatmap features from a single view [35] and multiple views [7], synthesizing a hand pose in a closed loop [19], and training cascaded networks following the structural hierarchy of the hand [26] using a convolutional neural network architecture. Different from these approaches which operate in a single depth modality input, we follow the success of a multi-modal learning framework [9, 37] which uses complementary features of the different modalities. To address it, we employ a function of geometric properties on the surface of the hand described by heat distribution as an additional modality to depth data.

The behavior of heat diffusion on the surface of a shape has generally been considered to be geometric features by analyzing a shape operator computed from the heat kernel matrix. The operator investigates the local geometry of the shape at small time scales and captures the global structure at large scales to be insensitive to non-rigid deformation, topological changes, and noise present in 3D models. Con-

sequently, the shape signatures/descriptors built on such descriptive representations have been extensively studied in the geometry community [30, 22, 2] for shape matching and retrieval. The robustness for identifying the points on the mesh surface naturally motivates us to pursue 3D keypoint retrieval (*i.e.,* hand joint positions) in the hand pose estimation problem. In this work, we build a heat distribution descriptor that incorporates the deformation invariant properties of heat diffusion over an articulated hand at multiple scales. Therefore, our method is robust to the changes of the topology of the hand and noise present in input data. The proposed ConvNet architecture is trained to hallucinate the multi-scale heat distribution descriptors using paired depth and heat distribution data.

The concept of modality hallucination has been previously presented in [5, 27] to produce a more informed model on visual recognition tasks. Our work shares analogies with [8] which transfers mid-level depth features extracted from an RGB image across domains. The potential for modality hallucination motivates us to consider learning an additional representation which is informed by analysis of the multi-scale heat distribution property, in the form of the articulated hand. Our main insight is that a geometrically consistent representation of the heat distribution modality can be learned from a single depth image, in addition to mid-level depth features. We use the resulting geometric responses together with depth features to further enhance the regression accuracy of the system. In practice, we found this step implicitly penalizes the initial estimates to be more effective and robust than the depth-alone framework. Our main contributions are summarized as follows:

1. Pixel-wise segmentation of an articulated hand using a ConvNet architecture which is robust to the cluttered background and efficient to compute in real-time.
2. Multi-scale geometric representations of the hand as a heat distribution descriptor which compactly encodes the information of hand articulations.
3. Modality hallucination using a single depth image, which transfers additional feature representations to produce a more informed estimation model.
4. The penalization of the initially predicted joint angle parameters with the guidance of the end-effectors (*i.e.*, the coordinates of the fingertips) in a feature space.

The rest of the paper is structured as follows. In the next section, we discuss the related work. In section 3, we briefly describe our synthetic 3D hand model. A multi-scale heat distribution descriptor is discussed in section 4. Section 5 presents a detailed explanation of the proposed pose estimation framework. In section 6, we present experimental results that illustrate the robustness of our approach. Finally, section 7 concludes the paper with a brief discussion and future directions of work.

2. Related Work

We overview the most relevant works on depth sensor based 3D hand pose estimation in the literature. Two different approaches are prevalent for categorizing hand pose estimation: discriminative methods and generative methods.

Discriminative methods A system for 3D hand pose estimation has been developed through the use of a large database. This group of approaches provides a trained classifier or regressor [11, 34] to find a mapping between image features and corresponding hand configurations. However, these methods can be susceptible to self-occlusion and ambiguous for low-resolution input data. In [31, 33], local pose regression methods are presented, demonstrating the efficacy of their approach against occlusions. While successful in many cases, they may experience jitters between frames when image features are insufficient to discriminate different poses. Also in [4], a collaborative filtering model is presented to regress the unknown pose parameters using similar poses. Recently, a convolutional neural network framework has been employed to improve the robustness to occlusions and jitters replacing hand-crafted features. Hand poses are estimated by incorporating a prior model [18], regressing the heatmaps from a single view [35] and multiple views [7], synthesizing a hand pose in a closed loop [19], and training networks following the structural hierarchy of the hand [26]. To our knowledge, we present the first work for 3D hand pose estimation using complementary geometric features as an additional modality to depth data.

Generative methods The optimization of an objective function has been a mainstream approach to recover the hand configurations using a deformable 3D hand model. Initially, particle swarm optimization (PSO) was successfully applied in [20, 21] to find a best fit model from a population of candidate solutions. In addition, gradient-based optimization was considered in [17, 32] to achieve faster convergence. While straightforward to implement, they iteratively update the initial pose parameters toward the local best solution. Hence, these methods may fail to track the hand when a prior estimate is inaccurate or to provide real-time performance. More recently, hybrid approaches [40, 23, 25, 12] have been introduced to recover loss of tracking using a per-frame reinitializer. Although these methods avoid model drift, the system achieves low frame rates [40], requires clear fingertip detection [23], or is heavily dependent on random forest [25] which shows relatively lower performance and higher memory requirements [26].

3. Preliminaries

3D hand model A kinematic hand model with 21 degrees of freedom (DOF) was well studied in [4, 26] for 3D hand pose estimation. This model imposes functional constraints

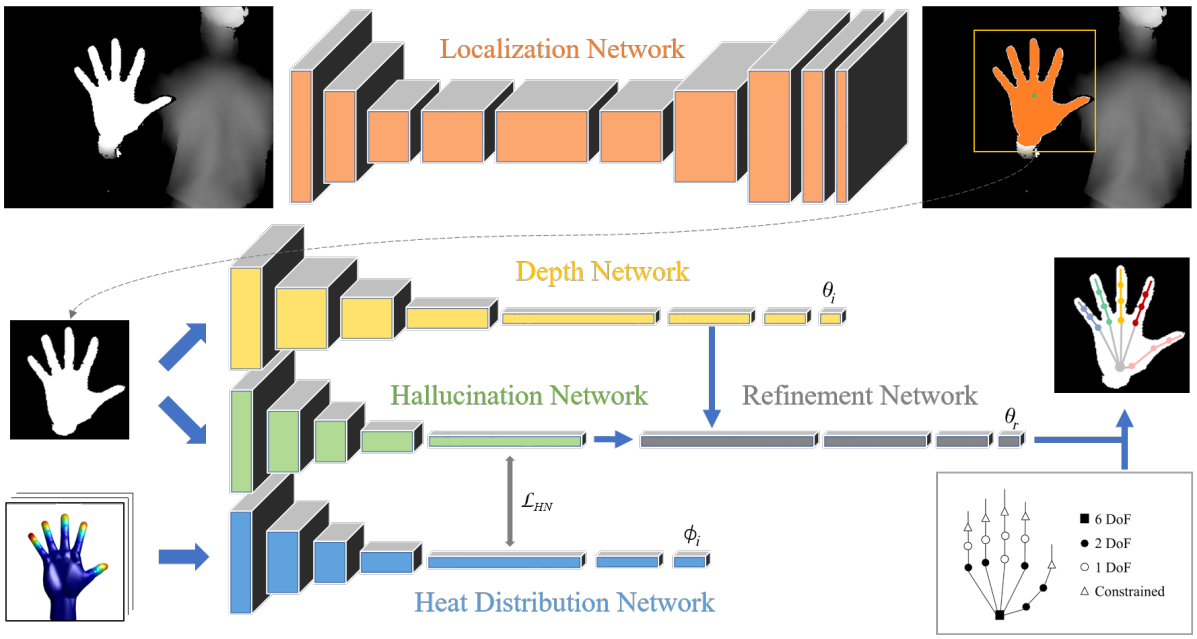


Figure 1: The pipeline overview. At training time, the hallucination network is trained to mimic heat distribution features using depth data. At testing time, the localization network takes as input a depth image to localize the hand. The identified hand is used to extract complementary features from the depth and hallucination network. The refinement network regularizes an initial pose estimate using the given feature representations.

[13, 16] for joint configurations and finger movement to reduce the complexity of hand articulations, and therefore simulates more realistic hand poses. We additionally construct a lower arm segment to help to identify the global hand orientation, regularizing jitters of the estimated pose [25]. Note that this portion is individually rotated with 2 DOFs along its longitudinal and lateral axes.

Dataset creation In order to train our ConvNet framework, we generate a synthetic dataset accurately annotated with ground truth labels $\mathcal{Y}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{D})$, where $\boldsymbol{\theta}$ denotes a set of 18 joint angle parameters; $\boldsymbol{\phi}$ is a set of 21 joint position triplets $\{x, y, z\}$; and $\mathbf{D}$ denotes multi-scale heat distribution descriptors. We render 300K hand poses by uniformly sampling each of the 18 joint angle parameters, covering a full range of hand articulations. Furthermore, the imposed hand motion constraints enable the effective simulation of realistic poses in the restricted configuration space.

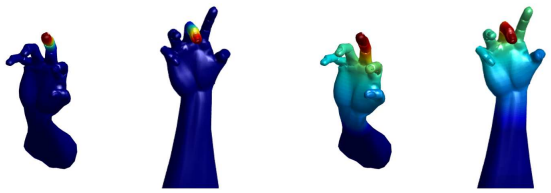
Approach overview An overview of our approach is depicted in Figure 1. We first preprocess depth data obtained from both 3D sensors and our pose simulator. A depth map of size 320×240 is center cropped to be the size 240×240. Then we generate a depth image (range [-1, 1]) with depth normalization and mean subtraction. This image is fed into our localization network to output a centroid of the hand in the uv-coordinates and a pixel-wise segmentation of the image. We draw a depth-dependent bounding box around the centroid of the hand using its corresponding triplet $\phi_c = \{x_c, y_c, z_c\}$. Finally, we resize the bounding box to obtain a 64×64 depth image that only contains the hand segment and remove depth noise using a median filter. For the NYU dataset, we first rescale the depth map to be the size 320×240 before cropping it. At training time the 64×64 depth image is used to train the depth network with the joint angle labels, whereas the heat distribution network learns the fingertip position triplets from the paired heat distribution descriptors $\mathbf{d}$ (detailed in the subsequent section). To learn heat distribution features through the hallucination network, we use a hallucination loss $\mathcal{L}_{HN}$ between two networks. Hence, the intermediate responses of the heat distribution descriptors can be extracted from the corresponding depth image. Our refinement network uses the resulting heat distribution features to regularize the estimated joint angle parameters $\boldsymbol{\theta}_i$. The hand skeleton is used to illustrate the quality of the estimated parameters $\boldsymbol{\theta}_r$ in Figure 1.

4. Heat Distribution

We briefly discover a heat operator derived in [30] and introduce the heat distribution descriptor to be used to train our hallucination network.

4.1. Heat Flow on the Hand Surface

Our hand model $\mathcal{M}$ is a compact Riemannian manifold without boundaries, which consists of 3,869 mesh vertices



(a) (b)

Figure 2: Visualization of the heat distribution descriptor on different hand poses over time. (a) The point heat source (red-colored) is placed at the tip of the middle finger at time $t = 0$. (b) For a large value $t = 40$, the behavior of heat distribution is geometrically consistent on both poses.

and 7,734 triangular faces. Thus, we can write the heat diffusion equation on the surface of the hand:

$$\left(\Delta + \frac{\partial}{\partial t}\right) u(i,t) = 0, \tag{1}$$

where Δ is the Laplace-Beltrami operator and $u(i,t)$ is heat distribution at vertex i at time t. In addition, let H_t be the heat operator which satisfies $\mathrm{H}_t = e^{-t\Delta}$. Then the solution to Eqn. 1 is $u(i,t) = \mathrm{H}_t(f)$, where $f : \mathcal{M} \rightarrow \mathbb{R}$ denotes the amount of heat available at $t = 0$. Therefore, the heat flowing through the mesh surface from source vertex j to i at a given diffusion time t for all $i, j \in \mathcal{M}$ can be denoted by the heat kernel $\mathrm{H}_t(i,j)$:

$$\mathrm{H}_t(i,j) = \sum_k e^{-\lambda_k t} \mathrm{v}_{ki} \mathrm{v}_{kj}, \tag{2}$$

where λ_k and v_k is the k-th eigenvalue and the k-th eigenfunction of the Laplace-Beltrami operator Δ, respectively.

4.2. Heat Distribution Descriptor

Figure 2 illustrates heat distribution on the hand surfaces over time. A unit heat source is given at the tip of the middle finger (marked in red) at time $t = 0$, and the amount of diffused heat to the rest of the surface is visualized in Figure 2b. At small time scales, the local geometry of the hand can be investigated, while the global structure can be encoded at large scales. Note that the analogy of heat distribution on different hand poses validates the geometrically consistent property of the diffusion process. This property motivates us to design a heat distribution descriptor which is invariant to shape deformation and topological changes.

We employ the characterization of heat distribution at each point $i \in \mathcal{M}$ heat transferred from a set of key sources $J = \{j_1, ..., j_5\}$ in $T = \{t_1, t_2, t_3\}$ time steps. Let P be the number of vertices of $\mathcal{M}$, then the proposed heat distribution descriptor $\mathbf{d}_t \in \mathbb{R}^{P \times 1}$ is as follows:

$$\begin{aligned} \mathbf{d}_t^j &= [\mathrm{H}_t(i_1, j), ..., \mathrm{H}_t(i_p, j), ..., \mathrm{H}_t(i_P, j)]^{\mathrm{T}} \\ &\qquad \forall j \in J \ \text{ and } \ \forall t \in T, \\ \mathbf{d}_t &= \sum_j \mathbf{d}_t^j \quad \forall t \in T. \end{aligned} \tag{3}$$

Here each entry of the P-dimensional vector $\mathbf{d}_t$ corresponds to the cumulative amount of heat available at each vertex i at time $t \in T$ diffused from source vertices $j \in J$. Consequently, we compute the heat distribution matrix $\mathbf{D} = [\mathbf{d}_{t_1}, \mathbf{d}_{t_2}, \mathbf{d}_{t_3}] \in \mathbb{R}^{P \times T}$, where $\{t_1, t_2, t_3\} = \{10, 30, 50\}$ in practice[1]. We further process the descriptor matrix $\mathbf{D}$ by rendering each column vector as an image format. A hidden point removal [10] strategy determines the visible vertices from the viewpoint of a camera. Our pose simulator investigates the visibility of the vertices and linearly interpolates the amount of heat distribution between neighboring vertices using the Phong interpolation method. As a result, we generate T-channel descriptors to feed into our hallucination network described in the following section.

Note that we use a heuristic to determine the heat sources J. We uniformly sample each of the source points from P vertices. These indices are fixed while generating our training dataset so that every hand poses share consistent geometric representations. Also, note that we do not find a significant difference in regression accuracy when we choose another set of J. At test time, the input point cloud is not indexed for the heat sources, and this is the main reason we hallucinate heat distribution features from a depth image.

5. Learning Hand Articulations

Our system follows the approach of [4, 33, 26] that estimates the *joint angle parameters* on a per-frame basis. Unlike the other pose estimation methods, this approach directly employs the motion constraints guided by the physical anatomy of the hand. In this setting, all estimated poses are kinematically valid and follow a natural sequence, and this is why we choose the angle parameters over the joint positions. Now we discuss how the proposed method learns hand articulations from depth and auxiliary modality features, in the form of the joint angles.

5.1. Localization Network

Hand localization has been heuristically solved in the literature [34, 23, 31, 29, 4, 26] by assuming (i) the hand appears largest in front of the sensor or (ii) the wristband can be identified by color segmentation. However, the underlying assumptions would be further from real scenarios, such as those at far-range or with a cluttered background. To achieve robust performance for localization, we divide the

[1] We observe that the behavior of heat diffusion is local (finger-level) at $t = 10$ and becomes global (hand-level) at $t = 50$.

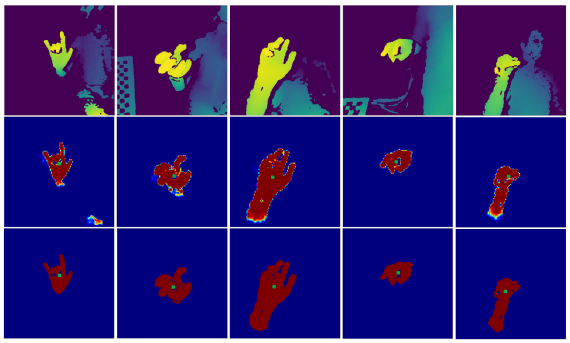


Figure 3: Visual analysis of hand localization. First row: input 240×240 depth images cherry-picked from the HandNet [38]. Second row: estimated hand probability map and centroid (green square). Third row: ground truth labels.

problem into two sub-tasks: hand segmentation and hand center regression.

We present a ConvNet architecture specifically designed to solve these tasks at one go. The graph of the network architecture is visualized in the supplementary material. Our main insight is that the deep neural network effectively identifies pixel-wise class labels through the convolution process [1]. To achieve this from our hand segmentation problem, the first three convolutional layers with a following max pooling layer down-sample the input 240×240 image to be the size 30×30. The next four convolutional layers capture the low-level image features in depth to distinguish the hand and background. Then we perform two unpooling operations in between convolutions to up-sample the given depth features (to be the size 120×120). The unpooling uses the original activations stored from the previous max pooling layers, which is critical for our system for the following reasons: (i) the unpooling process consistently increases the spatial size of the feature map to reconstruct the detailed hand segment, and (ii) it balances computational time and segmentation accuracy by generating sparse representations. Note that the deconvolution method [24] was also considered, which showed similar accuracy but required higher processing time because of its convolution operation. In addition, we employ intermediate convolutional features to regress the hand center. This branch is comprised of four additional convolutions and one inner product, estimating the centroid of the hand $\{u_c, v_c\}$[2]. It is further converted into the triplet ϕ_c to draw the bounding box around the hand. Figure 3 visualizes our hand localization.

5.2. Multi-modal Learning

Our system learns complementary features about hand articulations from different modalities. We train the depth network with the joint angle labels, taking into account the process of knowledge transfer across fingers. Moreover, multi-scale convolutional features are encoded through the heat distribution network to identify the fingertip positions.

[2]In practice, we achieved the mean distance error of 14.56 *pixels* in an image of size 320×240 on the HandNet dataset [38].

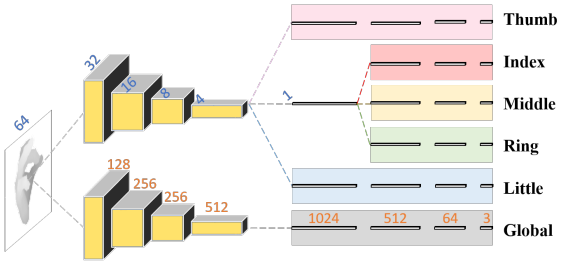


Figure 4: The proposed depth network consists of two streams: the top stream for the five fingers and the bottom stream for the global orientation parameters. Numbers in blue indicate the width & height of the feature map, and those in orange represent the number of kernels.

Depth network (DN) The success of multi-task learning in [3] has caused immense effects on the deep learning models (*e.g.* natural language processing in [6, 39], face detection in [41, 42], and human pose estimation in [14, 15]). These works all aim to achieve improved performance and prevent overfitting by transferring shared knowledge. Aligned with these works, we estimate the joint angle parameters of five fingers θ_i from a single network. The architecture of our multi-task depth network is shown in Figure 4. The first four convolutional layers share knowledge of the hand. This is crucial for learning a perceptual set of attributes, such as self-occlusions or self-similarities of the fingers across domains and hence leads to further improvements in the regression performance (see Section 6.3). For our specific operation, we group the fingers according to the anatomical position (*i.e.*, three groups: thumb, index-middle-ring, little) before passing the fifth convolutional layer. This insight allows us to achieve higher regression accuracy by learning structural representations from a correlation of adjacent fingers. In addition, we explore the global orientation of the hand from a separate network initiated in parallel using the same network configuration. For the proposed depth network (DN), we introduce the loss weights α, β, and γ to properly scale the loss function:

$$\mathcal{L}_{DN} = \alpha\mathcal{L}_T + \beta(\mathcal{L}_I + \mathcal{L}_M + \mathcal{L}_R + \mathcal{L}_L) + \gamma\mathcal{L}_G, \quad (4)$$

where the subscript denotes each finger (T: thumb, I: index, M: middle, R: ring, L: little, G: global). In practice, we observe that the thumb finger contributes less to the total loss $\mathcal{L}_{DN}$. Thus, we set the loss weights $\alpha = 3$, $\beta = 1$, and $\gamma = 1$ which balance the optimization process.

Heat distribution network (HDN) Our heat distribution network is trained with the fingertip position labels using T-channel descriptors that represent the multi-scale heat distribution property of the hand. The bottom of Figure 5 il-

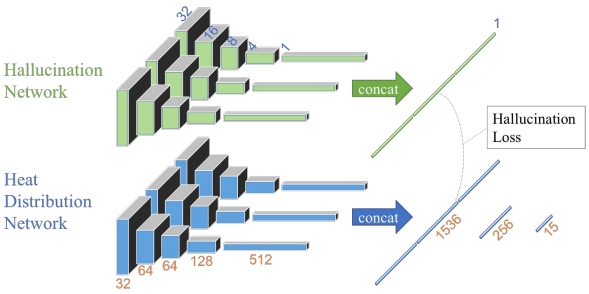


Figure 5: The architecture of the hallucination network (top) and heat distribution network (bottom). The concat layer concatenates multiple features to one blob. Numbers in blue indicate the width & height of the feature map, and those in orange represent the number of kernels.

lustrates the overall architecture. Each of the parallelized networks independently learns hand articulations from local to global geometric features through the first five convolutional layers. Then we utilize a feature concatenation step to aggregate three convolutional features into a single composition along the depth dimension. This step allows us to encode both the finger-level local geometry and global hand structure into more informative representations generated across the time scales. As a result, our network is capable of learning a better mapping function between input hand poses and the corresponding fingertip positions ϕ_i.

5.3. Modality Hallucination and Refinement

Hallucination network (HN) The parameter values (i.e., weights and bias) of the hallucination network (HN) are initialized using the network parameters of the pre-trained heat distribution network (HDN). We then fine-tune these values with a Euclidean loss $\mathcal{L}_{HN}$ between the intermediate feature vectors, similarly to [8]. However, we do not use the whole structure of the HDN from our HN. Instead, our HN has only the first five convolutional layers as illustrated at the top of Figure 5. Note that the choice of the number of layers is empirically determined in the next part. As a result, our hallucination network outputs the geometrically descriptive responses learned from the heat distribution descriptors using a corresponding depth image.

Refinement network (RN) Conceptually, the resulting triplets $\boldsymbol{\phi}_i$ together with the joint angles $\boldsymbol{\theta}_i$ estimated from the DN are used to regularize the angle parameters in the refinement network (RN). In practice, however, the direct use of $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_i$ does not achieve performance improvements. Alternatively, our RN takes as input a feature vector that is well-informed to predict the joint angle parameters $\boldsymbol{\theta}_i$ and fingertip positions $\boldsymbol{\phi}_i$. Hence, we generate a concatenated vector of depth feature F^l_{DN} and mimicked hallucination feature F^l_{HN}. The input feature maps for concatenation can be extracted from any layer l in the network, so we empirically determine where to extract these features with respect to regression accuracy. Figure 6 compares the performance of various combinations of feature concatenation. Note that we conduct these experiments using the depth activations and heat distribution activations F^l_{HDN} to eliminate the effect of hallucination error. It shows that the concatenation of depth features extracted after the sixth convolutional layer and heat distribution features extracted after the fifth convolutional layer achieves the highest performance (red bar). The RN consists of the four inner product layers with a following non-linear (ReLU) layer. We progressively reduce the dimension of the vector as a factor of 4, that is 2048-512-128-n (where $n = 18$ is the number of angles).

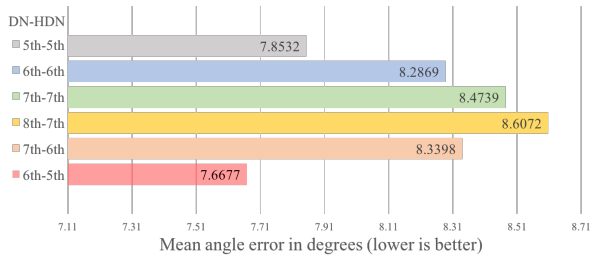


Figure 6: The mean angle error is used to evaluate different combinations of feature concatenation on our synthetic dataset. A number associated with each colorbar denotes layers of the DN (left) and the HDN (right), respectively. The best accuracy is achieved when we concatenate depth features extracted after 6th conv layer and heat distribution features extracted after 5th conv layer (red bar).

Network optimization Finally, we have three sets of network parameters independently learned from the depth network (DN), hallucination network (HN), and refinement network (RN). We further fine-tune the given networks using depth data and the corresponding angle labels θ. Then the total loss can be drawn as follows:

$$\mathcal{L}_{Optimize} = \zeta\mathcal{L}_{DN} + \eta\mathcal{L}_{RN}. \quad (5)$$

We set the loss weights $\zeta = 1$ and $\eta = 5$ so that the depth network and refinement network to be properly optimized with input depth data without updating the heat distribution network. Note that the same loss weights (α, β, γ) are used for the depth network as discussed previously.

6. Experiments

We conduct evaluations using a synthetic and public dataset to validate the efficacy of the proposed approach.

6.1. Datasets

We first introduce a self-generated synthetic dataset. This dataset is mainly used to evaluate our design choices. As discussed in Section 3, we use a hand model with

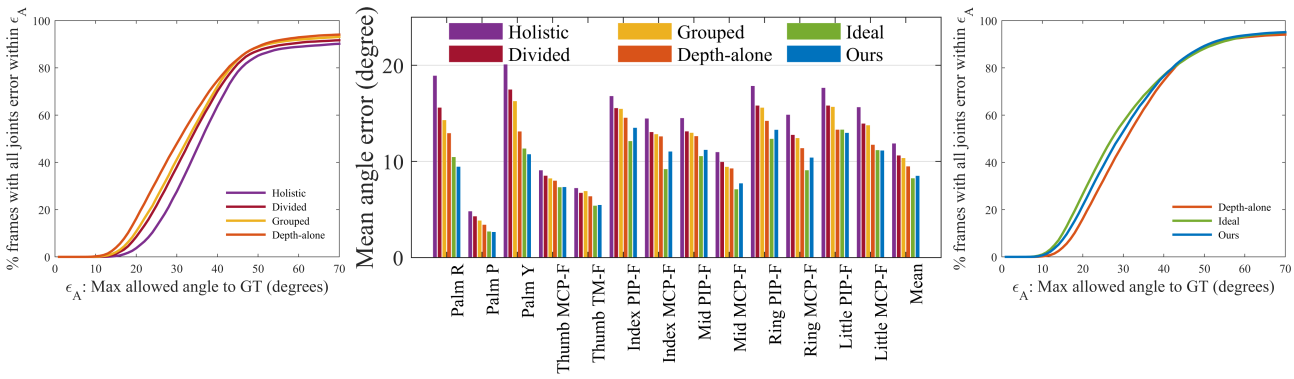


Figure 7: Quantitative evaluation of our method with repect to the self-generated baselines.

21 DOFs to render realistic hand poses with motion constraints. In the same manner, we collect the other 30K depth images with ground truth labels $\mathcal{Y}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{D})$.

Additionally, we use public datasets (NYU [35] and MSRA14 [23]) to compare the performance of our approach to the state-of-the-art methods. Two datasets are collected from different camera types across contexts. Specifically, the NYU dataset involves a continuous sequence of hand movements acquired at far-range, whereas the MSRA14 dataset contains various gesture types of 6 individuals captured from close viewpoints. Note that our system requires the joint angle parameters for training and testing. Thus, we compute the ground truth angles of these datasets using the inverse kinematics as proposed in [33].

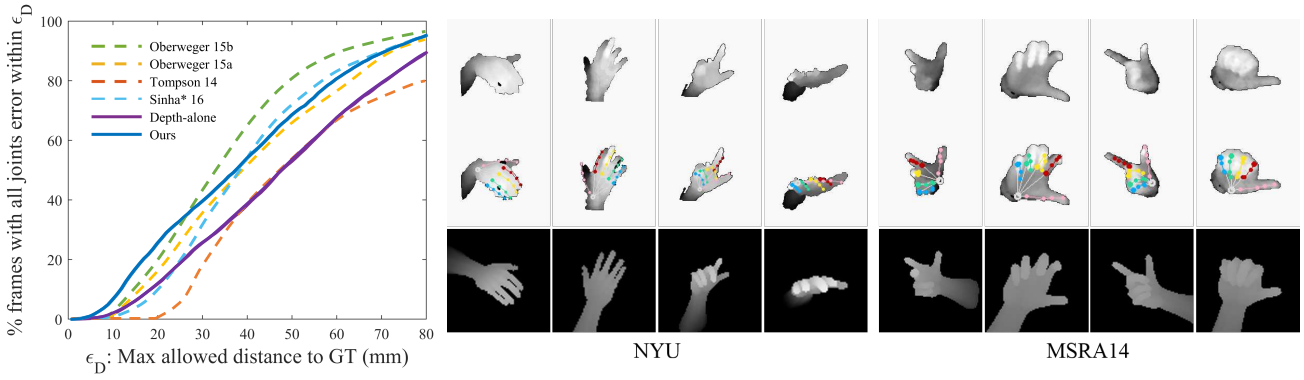
6.2. Comparison to Baselines

Why multi-task learning? We demonstrate the rationale for using the multi-task approach to estimate the joint angle parameters from the depth network. We first define four baselines: (i) *Holistic*, which estimates all joint parameters using a single network; (ii) *Divided*, which divides the *Holistic* baseline into six sub-tasks (Thumb-Index-Middle-Ring-Little-Global) after the fifth convolutional layer; (iii) *Grouped*, which groups the fingers (T-IMR-L) according to their anatomical position and also separates the global network from the finger network; and (iv) *Depth-alone*, where we set the loss weights of the *Grouped* baseline as discussed. Figure 7 (left) quantitatively compares these baselines on a synthetic dataset, where we measure the robustness of each baseline. The performance of the *Holistic* baseline is dramatically improved by simply adopting the multi-task learning approach, and it is further enhanced by grouping fingers together, as the *Grouped* achieves higher regression accuracy than does the *Divided* baseline. This indicates that the network model learns a structural correlation across fingers to anatomically constrain hand configurations. While training the network, the loss of the thumb finger fluctuated more and converged faster than did the other losses. We thus scale the loss function of the thumb by setting $\alpha = 3$ so that the contribution of the thumb will be 3 times larger than that from the *Depth-alone*. In this way, we achieve even better performance, as also demonstrated from the individual mean angle error in Figure 7 (middle). This comparison validates the rationale of our use of the multi-tasking approach as opposed to other choices.

Why regularize the initial estimation? Furthermore, we explore the efficacy of the proposed refinement process. Figure 7 (right) shows the quantitative evaluation of our method with respect to the following baselines: (i) *Depth-alone* estimates of the joint angles with the aforementioned settings and (ii) *Ideal* estimates where geometric features are directly extracted from the HDN to eliminate the effect of hallucination error. The fact that our method performs better than the *Depth-alone* baseline validates the efficacy of the RN. The individual mean angle error (see Figure 7 [middle]) shows consistent results. These results also indicate that the fingertip positions guide the initially estimated angle parameters to be more accurate. Overall, our approach resulted in comparable performance to *Ideal* or even higher accuracy for the global orientations and the little finger in terms of the mean joint angle error, demonstrating that our hallucination network is well-trained to mimic heat distribution features in detail.

6.3. Comparison with the State of the Arts

Quantitative evaluation Figure 8a quantitatively compares the performance of our approach with the state-of-the-art methods using a publicly available NYU dataset [35]. The maximum allowed joint distance error is examined in terms of the distance threshold ϵ_D. Here we observe that the overall performance of the *Depth-alone* baseline (purple line) is greatly improved in *Ours* (blue line) by hallucinating geometric features and penalizing the initial predictions. Moreover, our approach achieves performance higher than that of the state-of-the-art methods [35, 18, 26] over all the ranges. It further demonstrates that the better estimation model can



(a) (b)

Figure 8: Performance evaluations on the overall robustness. (a) Quantitative evaluation is conducted using the NYU dataset [35]. (b) Qualitative evaluations using the NYU and MSRA14 [23] dataset. The first row shows the input depth image, and the estimated poses are visualized in the second row. The third row shows pose reconstruction based on our estimates.

Sub.	1	2	3	4	5	6	Avg.
[20]	35.4	19.8	27.3	26.3	16.6	46.2	28.6
[23]	8.6	7.4	9.8	10.4	7.8	11.7	9.2
[7]	30.1	19.7	24.3	19.9	21.8	20.7	22.7
Ours	17.6	15.2	26.4	16.9	26.6	17.5	20.0

Table 1: Quantitative comparison (in mm) of our approach with the state-of-the-arts (generative methods [20, 23] and discriminative method [7]) on the MSRA14 dataset [23].

be built by learning complementary information from a different input modality. Our method also shows comparable performance to the generative approach [19] with a higher fraction of frames that have Euclidean error less than 27 mm. It indicates that our approach performs better with a smaller error tolerance.

We additionally show the comparison of our approach to the generative methods [20, 23] and discriminative [7] method using the MSRA14 dataset [23]. For this, we follow the cross-dataset experiment proposed in [7]. We fine-tune our network models using the MSRA15 dataset [31] to measure the averaged distance error (in mm) of the palm and five fingertips from the MSRA14 dataset. In Table 1, we observe that the discriminative methods (ours and [7]) show lower accuracy than that of the generative method [23]. For this, we share similar insights with [7] as follows: (i) the discriminative methods neither incorporate temporal information between frames nor use a manual initialization in the first frame and (ii) the hand is not calibrated or scaled for each subject, which is crucial to reduce errors. However, the proposed method mostly outperforms [20] and [7] as it achieves a lower error rate. Thus, we conclude that the use of geometric representations as an additional modality results in more robust hand pose estimation.

Qualitative evaluation We conduct qualitative evaluations of our method using the NYU and MSRA14 dataset. The second row in Figure 8b illustrates hand poses estimated from the depth images in the first row. In addition, we provide the corresponding hand reconstructions in the third row, demonstrating that our approach enforces kinematically valid hand configurations. Although the fourth column of the NYU input image has missing pixels (see the fingertips), our method robustly predicts the hand pose without using temporal information.

7. Conclusion

We address two important elements that have been missing in the current hand pose estimation approaches: (i) the understanding of geometric properties of the articulated hand and (ii) the use of an additional input modality to produce more informative representations. To incorporate these factors into the pose estimation system, we present a multi-scale heat distribution descriptor specifically designed to encode the local geometry as well as the global structural features of the hand. This descriptor is used to learn the convolutional responses, and our system hallucinates them using a corresponding depth image. Consequently, we use the geometrically informed features together with the discriminative depth representations extracted from the depth network to accurately estimate hand articulations. The extensive evaluations conducted using both the synthetic and real dataset validate the robustness of the proposed approach as we achieve performance higher than or comparable to the state-of-the-art methods.

Acknowledgements This work was partially supported by the NSF Award No.1235232 from CMMI and 1329979 from CPS, as well as the Donald W. Feddersen Chaired Professorship from Purdue School of Mechanical Engineering. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 5

[2] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikov. Shape google: Geometric words and expressions for invariant shape retrieval. *ACM Transactions on Graphics (TOG)*, 30(1):1, 2011. 2

[3] R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998. 5

[4] C. Choi, A. Sinha, J. H. Choi, S. Jang, and K. Ramani. A collaborative filtering approach to real-time hand pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2336–2344, 2015. 1, 2, 4

[5] C. M. Christoudias, R. Urtasun, M. Salzmann, and T. Darrell. Learning to recognize objects from unseen modalities. In *European Conference on Computer Vision*, pages 677–691. Springer, 2010. 2

[6] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. 5

[7] L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3D hand pose estimation in single depth images: from single-view CNN to multi-view CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3601, 2016. 1, 2, 8

[8] J. Hoffman, S. Gupta, and T. Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016. 2, 6

[9] H. Izadinia, I. Saleemi, and M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2013. 1

[10] S. Katz, A. Tal, and R. Basri. Direct visibility of point sets. In *ACM Transactions on Graphics (TOG)*, volume 26, page 24. ACM, 2007. 4

[11] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Real time hand pose estimation using depth sensors. In *Consumer Depth Cameras for Computer Vision*, pages 119–137. Springer, 2013. 2

[12] P. Krejov, A. Gilbert, and R. Bowden. Guided optimisation through classification and regression for hand pose estimation. *Computer Vision and Image Understanding*, 155:124–138, 2017. 2

[13] J. Lee and T. L. Kunii. Model-based analysis of hand posture. *IEEE Computer Graphics and applications*, 15(5):77–86, 1995. 2

[14] S. Li and A. B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014. 5

[15] S. Li, Z.-Q. Liu, and A. B. Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 482–489, 2014. 5

[16] J. Lin, Y. Wu, and T. S. Huang. Modeling the constraints of human hand motion. In *Human Motion, 2000. Proceedings. Workshop on*, pages 121–126. IEEE, 2000. 2

[17] S. Melax, L. Keselman, and S. Orsten. Dynamics based 3D skeletal hand tracking. In *Proceedings of Graphics Interface 2013*, pages 63–70. Canadian Information Processing Society, 2013. 1, 2

[18] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands Deep in Deep Learning for Hand Pose Estimation. In *Proc. Computer Vision Winter Workshop (CVWW)*, 2015. 1, 2, 7

[19] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a Feedback Loop for Hand Pose Estimation. In *Proceedings of the International Conference on Computer Vision*, 2015. 1, 2, 8

[20] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3D tracking of hand articulations using kinect. In *BMVC*, volume 1, page 3, 2011. 1, 2, 8

[21] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Tracking the articulated motion of two strongly interacting hands. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1862–1869. IEEE, 2012. 2

[22] M. Ovsjanikov, Q. Mérigot, F. Mémoli, and L. Guibas. One point isometric matching with the heat kernel. In *Computer Graphics Forum*, volume 29, pages 1555–1564. Wiley Online Library, 2010. 2

[23] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1106–1113, 2014. 1, 2, 4, 7, 8

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 5

[25] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642. ACM, 2015. 2, 3

[26] A. Sinha, C. Choi, and K. Ramani. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4150–4158, 2016. 1, 2, 4, 7

[27] L. Spinello and K. O. Arras. Leveraging RGB-D data: Adaptive fusion and domain adaptation for object detection. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4469–4474. IEEE, 2012. 2

[28] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[29] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2015. 4

[30] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009. 2, 3

[31] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 824–832, 2015. 2, 4, 8

[32] A. Tagliasacchi, M. Schröder, A. Tkach, S. Bouaziz, M. Botsch, and M. Pauly. Robust articulated-ICP for real-time hand tracking. In *Computer Graphics Forum*, volume 34, pages 101–114. Wiley Online Library, 2015. 2

[33] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3325–3333, 2015. 2, 4, 7

[34] D. Tang, T.-H. Yu, and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3224–3231. IEEE, 2013. 2, 4

[35] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169, 2014. 1, 2, 7, 8

[36] C. Wan, A. Yao, and L. Van Gool. Direction matters: hand pose estimation from local surface normals. In *European Conference on Computer Vision*. Springer, 2016. 1

[37] A. Wang, J. Lu, J. Cai, T.-J. Cham, and G. Wang. Large-margin multi-modal deep learning for RGB-D object recognition. *IEEE Transactions on Multimedia*, 17(11):1887–1898, 2015. 1

[38] A. Wetzler, R. Slossberg, and R. Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. *arXiv preprint arXiv:1507.05726*, 2015. 5

[39] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4460–4464. IEEE, 2015. 5

[40] C. Xu and L. Cheng. Efficient hand pose estimation from a single depth image. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3456–3462. IEEE, 2013. 2

[41] C. Zhang and Z. Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1036–1041. IEEE, 2014. 5

[42] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014. 5