

Article

Adaptive Curve-Guided Convolution for Robust 3D Hand Pose Estimation from Corrupted Point Clouds

Lihuang She, Haonan Sun * , Hui Zou, Hanze Liang, Xiangli Guo and Yehan Chen

School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China; shelihuang@cse.neu.edu.cn (L.S.); 2201753@stu.neu.edu.cn (H.Z.); 2401777@stu.neu.edu.cn (H.L.); 2371949@stu.neu.edu.cn (X.G.); 20237134@stu.neu.edu.cn (Y.C.)

* Correspondence: sunhaonan@stumail.neu.edu.cn

Abstract

3D hand pose estimation has achieved remarkable progress in human computer interaction and computer vision; however, real-world hand point clouds often suffer from structural distortions such as partial occlusions, sensor noise, and environmental interference, which significantly degrade the performance of conventional point cloud-based methods. To address these challenges, this study proposes a curve fitting-based framework for robust 3D hand pose estimation from corrupted point clouds, integrating an Adaptive Sampling (AS) module and a Hand-Curve Guide Convolution (HCGC) module. The AS module dynamically selects structurally informative key points according to local density and anatomical importance, mitigating sampling bias in distorted regions, while the HCGC module generates guided curves along fingers and employs dynamic momentum encoding and cross-suppression strategies to preserve anatomical continuity and capture fine-grained geometric features. Extensive experiments on the MSRA, ICVL, and NYU datasets demonstrate that our method consistently outperforms state-of-the-art approaches under local point removal across fixed missing-point ratios ranging from 30% to 50% and noise interference, achieving an average Robustness Curve Area (RCA) of 30.8, outperforming advanced methods such as TriHorn-Net. Notably, although optimized for corrupted point clouds, the framework also achieves competitive accuracy on intact datasets, demonstrating that enhanced robustness does not compromise general performance. These results validate that adaptive curve guided local structure modeling provides a reliable and generalizable solution for realistic 3D hand pose estimation and emphasize its potential for deployment in practical applications where point cloud quality cannot be guaranteed.



Academic Editor: Chiman Kwan

Received: 23 September 2025

Revised: 17 October 2025

Accepted: 19 October 2025

Published: 22 October 2025

Citation: She, L.; Sun, H.; Zou, H.; Liang, H.; Guo, X.; Chen, Y. Adaptive Curve-Guided Convolution for Robust 3D Hand Pose Estimation from Corrupted Point Clouds. *Electronics* **2025**, *14*, 4133. <https://doi.org/10.3390/electronics14214133>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: 3D hand pose estimation; corrupted point clouds; curve-guided convolution; adaptive sampling; robust geometric modeling

1. Introduction

3D hand pose estimation, a core technology in human–computer interaction and computer vision, is valuable for virtual reality, augmented reality, and robotic manipulation. Compared to traditional 2D image recognition, 3D point cloud data which has inherent disorder and invariance to rigid transformations delivers consistent results regardless of rotation, translation or scaling. This enables direct capture of hand geometric structures and spatial topological information, thus significantly improving the accuracy and robustness of pose estimation. However, in real-world applications, hand point cloud data often suffers structural distortions such as partial occlusions, noise disturbances and non-uniform

distribution, caused by sensor noise, occlusion and environmental interference. These distortions severely weaken the generalization ability of existing models. Specifically, partial occlusions like the random removal of finger region point clouds can break the topological connections of the hand, while noise disturbances such as Gaussian noise introduced by sensor jitter can destroy local geometric consistency, causing significant deviations in joint localization and skeletal reconstruction in existing models. Achieving high-precision 3D hand pose estimation from corrupted point clouds has become a pressing technical challenge in this field.

In recent years, deep-learning-based point cloud methods have advanced 3D structure modeling. Global frameworks such as PointNet [1] capture overall shape via max pooling but have limited ability to represent local details, whereas hierarchical aggregation (PointNet++ [2]), attention mechanisms (PCT [3]), and graph convolutions improve local structure representation. Recent methods including CausalPC [4] for adversarial robustness, FGBD [5] for fast graph-based denoising, and SVDFormer [6] using multi-view fusion still lack curve modeling, limiting local geometric repair in noisy hand point clouds. Most approaches, such as HandPointNet [7], are developed for idealized point clouds and degrade under real-world distortions. Moreover, standardized evaluation metrics for corrupted point clouds remain scarce, which hinders quantitative comparison and technological development.

To address these challenges, this study proposes a 3D hand pose estimation framework based on curve fitting for corrupted point clouds, as illustrated in Figure 1. Unlike general-purpose 3D point cloud processing methods, this framework is tailored explicitly for hand pose estimation under distortion and incorporates a more sophisticated mechanism to capture local structural features. The bending shape of fingers and the curvature variation of joints contain rich pose semantic information, which can be effectively captured through curve modeling. Drawing inspiration from curve-based point cloud shape analysis [8], this study innovatively integrates an Adaptive Sampling (AS) module and a Hand-Curve Guide Convolution (HCGC) module, which is distinct from the universal curve learning paradigm in existing work, instead of over-relying on global feature aggregation or static structural modeling. The AS module dynamically optimizes key point selection based on density awareness, alleviating sampling bias in distorted regions. The HCGC module generates hypothetical curves describing finger geometries via a guided walking algorithm and enhances the stability of local structure modeling by combining dynamic momentum encoding and cross-suppression strategies; both of which are enhancements tailored for hand-specific topological continuity (e.g., finger joint curvature, inter-finger occlusion) not addressed in general curve-based frameworks. These curves walk directionally on the isomorphic graph of the point cloud, capturing geometric shapes and structural information to enable robust hand pose estimation under noise and local missing conditions.

Furthermore, to quantify the model's adaptability to real-world distortions, this study employs a lightweight evaluation protocol: a portion of the samples from existing datasets are used, and the finger region is partially corrupted across fixed missing-point ratios ranging from 30% to 50% to simulate sensor occlusion or environmental interference. Based on this setup, two evaluation metrics are adopted: Robustness Curve Area (RCA) and Robustness Gain (RG), which respectively measure the overall performance degradation under varying distortion levels and the improvement over baseline models. Experimental results demonstrate that the proposed method consistently achieves stronger robustness than existing state-of-the-art approaches under a spectrum of corruption severities, while also maintaining competitive accuracy on intact datasets.

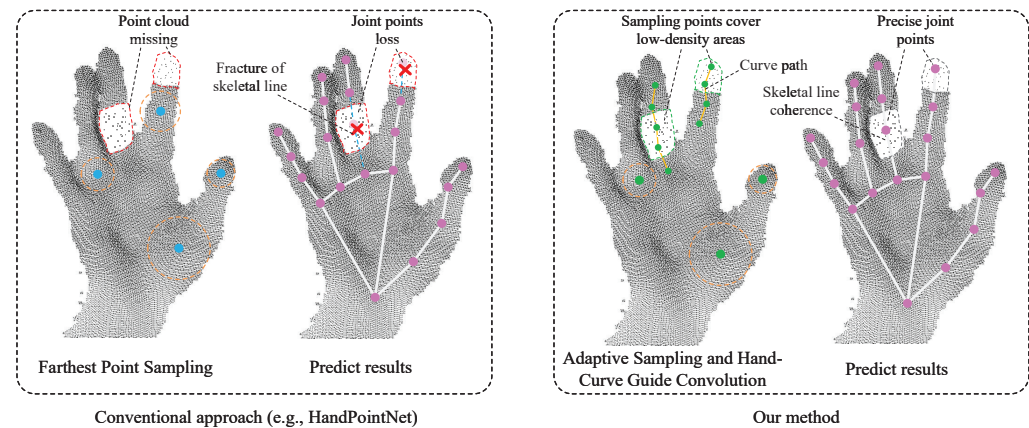


Figure 1. This figure compares the pose estimation performance of the conventional approach (HandPointNet, **left**) and the proposed method (**right**) under local point cloud missing conditions. For the conventional approach: Static Farthest Point Sampling (FPS) leads to sparse distribution of sampling points, avoiding missing regions, resulting in lost joint predictions and broken skeletal lines. Proposed method: Adaptive Sampling (AS) dynamically optimizes the distribution of sampling points, covering low-density regions (e.g., fingertips and joints), combined with curve fitting strategies to restore continuous finger geometric morphology, accurately predicting joint positions and generating smooth skeletal lines. The experimental results demonstrate that the curve fitting method effectively compensates for the impact of point cloud missing, significantly improving the robustness of hand pose estimation.

Our contributions can be summarized as follows:

- We introduce adaptive sampling into 3D hand pose estimation, enabling the network to dynamically select informative points from corrupted point clouds. This strategy highlights structural key regions while suppressing redundancy, which significantly improves robustness to missing data and sensor corruption in real-world hand scans.
- We propose a convolution module that leverages curve guidance to model the geometric continuity of hand structures and integrates it with graph-based context aggregation. The curve-guided representation preserves fine-grained finger trajectories, while the graph convolution captures long-range dependencies across joints and surfaces. This synergy effectively compensates for structural distortions and enhances feature completion in corrupted hand point clouds.
- We propose an evaluation protocol that simulates real-world distortions by removing a fixed percentage of points from existing datasets, with missing-point ratios of 30% to 50%. Two metrics, RCA and RG, are employed to provide a comprehensive quantitative assessment of model robustness under varying distortion levels, while ensuring consistent performance on intact point clouds.

2. Related Work

2.1. 3D Hand Pose Estimation from Point Clouds

3D hand pose estimation has undergone rapid development in recent years, driven by advances in deep learning and the availability of RGB-D sensors (e.g., Intel RealSense D455, Intel Corporation, Santa Clara, CA, USA) and 3D point cloud data. Early methods relied on handcrafted features, model fitting, and classical machine-learning strategies [9–14], which provided initial solutions but were sensitive to occlusion, background clutter, and illumination variations. Following these pioneering works, more advanced architectures were introduced, including 3D CNNs for depth-based hand pose estimation [15], voxel-to-voxel prediction networks (V2V-PoseNet) [16], and dense 3D regression frameworks [17]. Recent improvements further incorporate biomechanical constraints in weakly supervised

learning [18], generative losses to reduce annotation noise [19], and efficient virtual view selection for higher accuracy [20]. While these methods have achieved remarkable progress, they remain limited by sensor noise, depth ambiguities, and missing regions.

In parallel, point cloud-based methods have attracted increasing attention, as they directly exploit the 3D geometry of the hand. Ge et al. [7] proposed HandPointNet, extending PointNet to regress 3D hand joints from raw point clouds, followed by Point-to-Point Regression PointNet [21] for more accurate joint localization. Tang et al. [22] introduced hierarchical sampling optimization to improve robustness against distortions, while Chen et al. [23] designed SHPR-Net to incorporate semantic features. Beyond PointNet, convolutional architectures such as PointCNN [24] further advanced local feature learning, and Zhao et al. [25] enhanced neighborhood interactions through PointWeb. Sequence-based models like Point2Sequence [26] leveraged attention mechanisms to capture contextual relationships across points. Malik et al. [27] leveraged synthetic data for training, and Rezaei et al. [28] proposed TriHorn-Net to improve depth-based estimation. More advanced designs include Li et al. [29] with permutation-equivariant layers for point-to-pose voting, Wu et al. [30] with Capsule-HandNet to encode structural relationships, Du et al. [31] with multi-task information sharing, Zhao et al. [32] with Point Transformer for long-range feature modeling.

Very recent work has pushed these boundaries further. Wu et al. [33] proposed Point Transformer V3, which enhances backbone efficiency and representation capability for a wide range of point cloud learning tasks. Cheng et al. [34] introduced HandDiff, a diffusion-based framework that jointly refines 3D hand poses from both image and point cloud modalities, achieving superior resilience to severe noise and missing regions. Wang et al. [35] presented UniHOPE, a unified model capable of handling both hand-only and hand-object pose estimation, demonstrating strong generalization to interactive scenarios and maintaining stable performance under occlusion. Additionally, PCDM [36] employs conditional diffusion modeling to complete and refine corrupted point clouds, providing a powerful generative prior for geometric restoration. Collectively, these approaches represent the current frontier of robust 3D hand pose estimation, addressing challenges of incompleteness, noise, and occlusion, and establishing a solid foundation for future advancements in this field.

Despite these advances, current point cloud-based methods still face significant challenges under severely corrupted inputs, where missing regions and noise hinder accurate joint regression. This motivates the integration of structural priors such as curve fitting, which can interpolate across gaps, regularize noisy observations, and preserve anatomical consistency. Therefore, curve-fitting-based strategies are particularly well suited for robust 3D hand pose estimation from corrupted point clouds, complementing existing deep-learning frameworks.

2.2. Adaptive Sampling for Point Cloud Learning

Sampling is a fundamental operation in point-cloud pipelines because the chosen subset of points directly determines which local geometric cues the network can observe and exploit. Classical schemes such as Farthest Point Sampling (FPS) provide relatively uniform spatial coverage and have been widely adopted in hierarchical point networks (e.g., PointNet++ [2]), but they are agnostic to semantics and sensitive to outliers and locally missing data. To mitigate these limitations, recent work has explored task-aware and adaptive sampling strategies that steer sampling toward informative regions. Dynamic neighborhood constructions and feature-driven aggregation (for example, DGCNN [37]) improve local representation by adapting the receptive field, while methods designed for large or noisy point sets (e.g., RandLA-Net [38]) combine efficient random sampling

with attentive local aggregation to retain salient geometry. Building directly on the idea of feature-guided sampling, PointASNL [39] proposed an explicit AS module that refines an initial sample set by reweighting and relocating sample positions according to local density and learned feature cues, demonstrating improved robustness to noise and partial observations on general shape benchmarks.

In the context of corrupted hand gesture point clouds, where anatomical contours such as finger joints or fingertip arcs may be partially or wholly absent, we adopt the practical AS formulation of PointASNL as the sampling backbone. Its operation is tailored to hand gestures by biasing sampling toward anatomically critical loci and preserving contour continuity, ensuring sufficient point density in low-density but semantically important regions. Within our framework, this adaptive sampling complements the subsequent curve-guided feature modeling, enhancing robustness to structural degradation in real-world gesture data.

2.3. Curve Fitting, Structured Traversals, and Geometry-Aware Aggregation

Capturing geometric continuity in point clouds goes beyond local neighborhood aggregation: the idea of structuring points along paths or curves has gained increasing attention. A prominent example is CurveNet [8], which introduced a Curve Instance Convolution (CIC) module that groups and aggregates along curves constructed on a point graph. Such curve-based grouping captures elongated boundary or articulation structures more faithfully than purely radial neighborhoods. Alongside, geometry-aware graph methods such as GeoHi-GNN [40] learn hierarchical graph representations that are sensitive to local geometry, enabling better fitting of feature weights for complex surfaces. Edge and curve extraction methods such as NerVE [41] directly estimate parametric edges or curves from point clouds, allowing for structured geometry prediction rather than point-wise predictions.

Nevertheless, applying curve fitting or structured traversal methods in hand gesture estimation under severe degradation is challenging. Many of those works assume fairly complete point clouds or at least moderate noise and do not account for missing anatomical continuity. In particular, curve construction steps often do not encode priorities for fingers' joints or critical bends, and aggregation mechanisms usually collapse over missing data rather than compensating for it. To address these issues, our HCGC explicitly integrates joint-sensitive scoring, momentum-based curve descriptors, and curve aggregation mechanisms that respect anatomical continuity. By doing so, the proposed model can better recover discriminative geometric signals from broken hand point clouds, enhancing pose estimation accuracy in real-world settings.

3. Methodology

3.1. Network Architecture

The aim of this study is to design a robust network that can effectively process corrupted hand gesture point clouds. In practical applications, the acquired point clouds often contain missing regions and structural distortions. Since fingers exhibit complex curvatures and joint variations, it is crucial to extract features that follow these curve-like anatomical structures. A curve-guided representation allows the network to align feature extraction along finger contours, thereby capturing bending and stretching patterns that are essential for distinguishing different gestures.

The overall framework of the proposed network is illustrated in Figure 2. The model receives a raw depth image as input, which is converted into a three-dimensional point cloud through a gesture segmentation process. To standardize the input, the point cloud is normalized into an Oriented Bounding Box coordinate system. This step enhances the robustness of the model to variations in gesture orientation. The normalized point cloud is then downsampled with Farthest Point Sampling, where 1024 points are preserved to maintain a consistent input

size. During training, additional data augmentation is applied in the form of random rotations around the z-axis within the range of $[-37.5^\circ, 37.5^\circ]$, scaling factors between 0.9 and 1.1, and translations within $[-10 \text{ mm}, 10 \text{ mm}]$. These augmentations improve the generalization ability of the model and strengthen its robustness to noise and distortions.

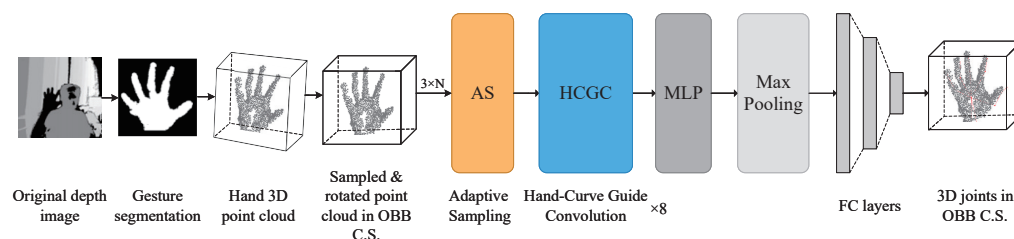


Figure 2. This figure illustrates the network architecture of the curve fitting-based 3D hand pose estimation. The input depth image undergoes hand segmentation and point cloud preprocessing, followed by the AS module and 8 Hand-Curve Guide Convolution (HCGC) layers, with the number of layers empirically set to 8. The AS module employs a density-aware strategy to dynamically select key points and combines K-Nearest Neighbors (KNN) and attention mechanisms to extract robust local features. The HCGC module generates curves describing the geometric morphology of fingers through a curve-grouping algorithm and optimizes the curve paths using dynamic momentum encoding and cross-suppression strategies. Finally, multi-layer features are aggregated through max pooling, and the joint 3D coordinates are regressed by a fully connected layer.

The network is organized into three main stages. First, an AS module is employed to select informative points and generate reliable local features, which helps the model handle noisy and incomplete point clouds. Second, the core of the architecture is built upon HCGC modules, which learn curve-based representations that align with finger anatomy and capture both local and global geometric features. Finally, the aggregated features are processed through a residual learning framework with stacked modules, followed by global pooling and a regression head that outputs the three-dimensional coordinates of finger joints in the Oriented Bounding Box coordinate system.

In summary, the proposed network integrates adaptive sampling, curve-guided feature learning, and residual aggregation to achieve reliable hand pose estimation under challenging conditions with corrupted input data.

3.2. Adaptive Sampling Module

FPS is widely used in point cloud processing, yet its sensitivity to outliers often leads to suboptimal results when handling real-world hand point clouds with missing regions or distortions. To overcome this, the proposed AS module adopts a dynamic sampling strategy inspired by PointASNL [39], adaptively selecting informative points according to local density variations. By integrating normalization and attention mechanisms, it extracts reliable geometric features from incomplete or noisy inputs, providing a stable basis for subsequent curve-based representation.

As shown in Figure 3, an initial set of uniformly distributed points $P_s \in \mathbb{R}^{N_s \times 3}$ is first sampled by FPS. For each point x_i with feature f_i , its neighborhood $N(x_i)$ is identified using a K-Nearest Neighbors (KNN) query. A self-attention mechanism updates features as

$$f_{i,k} = A(R(x_{i,k}, x_{i,j})\gamma(x_{i,j}), \forall x_{i,j} \in N(x_i)), \quad (1)$$

where R denotes the relation function, γ is a linear transformation, and A aggregates neighborhood information. The relation function is defined by dot-product similarity with Softmax normalization:

$$R(x_{i,k}, x_{i,j}) = \text{Softmax} \left(\frac{\varphi(f_{i,k})^T \cdot \theta(f_{i,j})}{\sqrt{D'}} \right), \quad (2)$$

where φ and θ are independent linear projections.

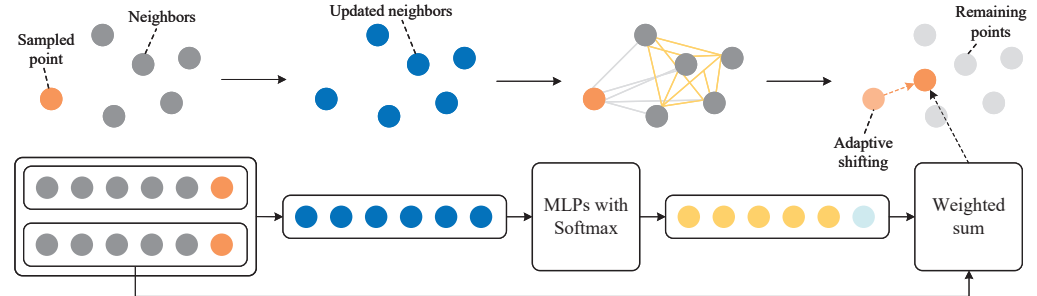


Figure 3. Architecture of the AS module. The module adaptively selects key points in the point cloud by combining KNN and attention mechanisms. Positions and features of sampled points are refined through dynamic weighting, enhancing local geometric feature extraction under noise or missing data.

To further refine sampling, point-wise MLPs compute adaptive weights for coordinates and features. The updated positions and features are obtained via weighted summation over their neighborhoods:

$$x_i^* = W_p^T X, \quad f_i^* = W_f^T F, \quad (3)$$

where W_p and W_f are Softmax-normalized weights, and X, F denote neighbor coordinates and features.

A key innovation of the AS module is its ability to emphasize regions of anatomical importance. In hand point clouds, fingers and joints exhibit high curvature and are prone to missing data. Points in low-density regions or with high curvature receive higher sampling weights. Curvature κ_i is approximated from the eigenvalues of the neighborhood covariance matrix:

$$\kappa_i = \frac{\lambda_{\max}}{\sum_{j=1}^3 \lambda_j}, \quad (4)$$

where λ_j are eigenvalues of the covariance matrix of $N(x_i)$.

Finally, adaptive normalization of relative positional information mitigates the effects of hand size, orientation, and pose variations. Combined with attention-driven feature extraction, this design allows the AS module to capture fine-grained geometric details such as finger bending and extension while maintaining robustness to noise and distortion.

3.3. Hand-Curve Guide Convolution

The HCGC module forms the core computational unit of the proposed network, specifically designed to handle corrupted hand point clouds, as illustrated in Figure 4. While the HCGC design draws inspiration from the CIC module in CurveNet [8], several adaptations have been introduced to better suit hand pose estimation. In particular, the curve construction process is modified to explicitly maintain anatomical continuity along finger joints, ensuring that each curve effectively follows the natural bending and stretching of fingers. Moreover, the aggregation stage is adapted to mitigate the influence of missing or corrupted points, which are prevalent in real-world hand scans.

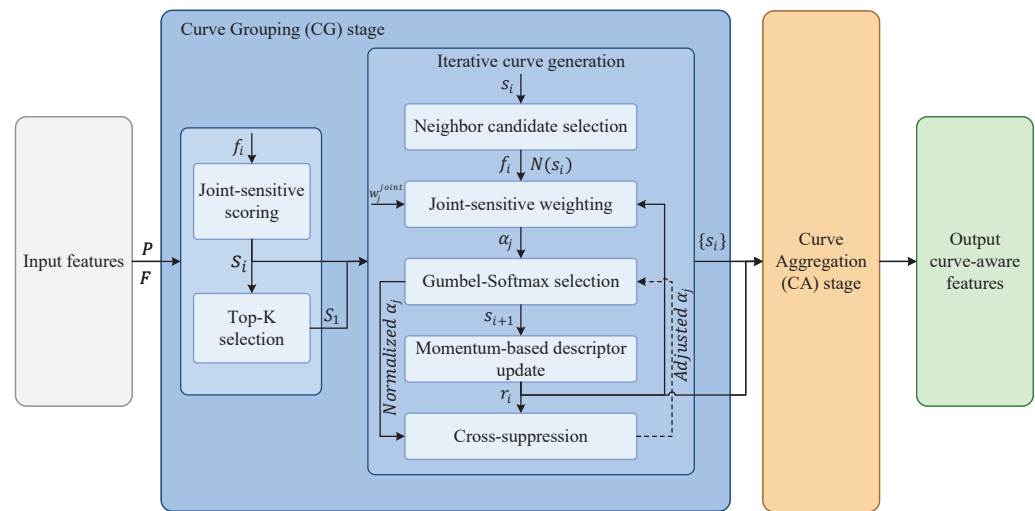


Figure 4. Detailed architecture of the HCGC module. The module receives input point features, performs local graph construction, and generates curves using the Curve Grouping (CG) stage. The CG stage explicitly incorporates joint-sensitive scoring, Top-K selection, and an iterative curve generation process with momentum-based descriptors to ensure anatomical continuity and robustness to missing points. The Curve Aggregation (CA) stage integrates features along the generated curves to produce the final curve-aware representation.

The network stacks multiple HCGC modules within a residual framework [42], capturing hierarchical features from local geometry to global hand pose. Formally, for input point cloud $P = \{p_i\}_{i=1}^N$ and features $F = \{f_i\}_{i=1}^N$

$$F^{(l+1)} = F^{(l)} + HCGC(F^{(l)}, P), \quad (5)$$

where $HCGC(\cdot)$ represents a single module's processing.

Each HCGC module integrates two core components: Curve Grouping (CG) and Curve Aggregation (CA). CG constructs curves over the isomorphic point cloud graph using a learnable walking strategy that dynamically selects neighbors based on both geometric proximity and joint-level continuity, a key improvement over CurveNet, where curves lack explicit anatomical constraints. CA aggregates features along these curves through attention mechanisms, emphasizing discriminative finger representations while suppressing irrelevant regions from missing or noisy points. To further enhance robustness, joint-sensitive weighting prioritizes points at joints and fingertips during curve construction, and continuity-preserving attention ensures aggregation respects the anatomical sequence. Stacking multiple HCGC modules within a residual framework enables stable gradient flow and hierarchical feature learning, capturing fine-grained local interactions and global hand configurations, thereby recovering semantically meaningful curves even from severely corrupted point clouds and providing a robust foundation for 3D joint regression.

3.3.1. Curve Grouping Stage

The CG stage constructs curves over the hand point cloud to preserve anatomical continuity and local geometric structures (Figures 4 and 5). Unlike CurveNet's CIC module, which relies primarily on spatial proximity, our approach integrates joint-level significance and robustness to missing points.

Initially, a Top-K selection mechanism identifies high-scoring points as starting locations for curve generation. The Top-K value determines the number of starting points chosen based on joint-sensitive scores. This parameter balances coverage of key anatomical regions with computational efficiency: a larger K increases the likelihood of capturing all

relevant joints but adds computational cost, while a smaller K focuses on the most critical points. For each point feature $f_i \in F$, a joint-sensitive score is computed as

$$s_i = \text{MLP}(f_i) \cdot w_i^{\text{joint}}, \quad (6)$$

where w_i^{joint} is a learnable weight that prioritizes anatomical keypoints, such as finger joints and fingertips. The set of starting points S_1 is obtained as

$$S_1 = \text{TopK}(\{s_i\}_{i=1}^N), \quad (7)$$

ensuring that curve initiation focuses on structurally important regions of the hand.

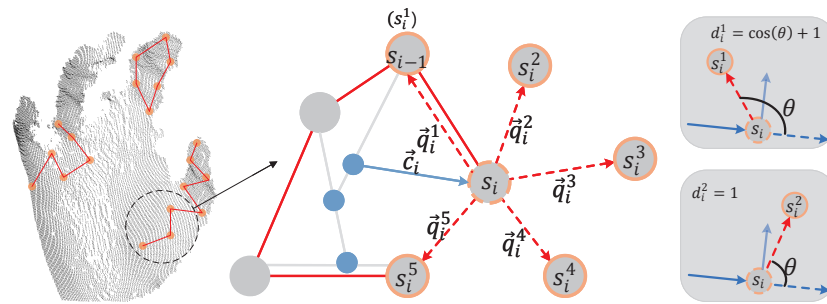


Figure 5. CG process in HCGC. Curves are generated iteratively from selected starting points using a learnable walking strategy. This preserves anatomical continuity and captures local geometric features.

From each starting point $s_i \in S_1$, curves are generated iteratively using a learnable walking policy $\pi(\cdot)$. This policy selects the next point s_{i+1} among neighboring candidates $s_j \in N(s_i)$ based on both feature similarity and anatomical importance:

$$s_{i+1} = \pi(s_i) = \sum_{j=1}^k \text{GumbelSoftmax}(\alpha_j) \cdot s_j, \quad (8)$$

$$\alpha_j = \text{MLP}([f_j, r_i]) \cdot w_j^{\text{joint}}, \quad (9)$$

where r_i is the current curve descriptor encoding the accumulated curve information. The Gumbel–Softmax operator enables differentiable selection of neighbor points, while the joint-sensitive weighting ensures that points corresponding to key anatomical locations are more likely to be included in the curve. This design improves the robustness of curve generation in the presence of missing or noisy points.

To maintain smoothness and anatomical continuity, the curve descriptor is updated dynamically through momentum-based encoding:

$$\beta = \text{Softmax}(\text{MLP}([r_{i-1}, s_i])), \quad (10)$$

$$r_i = \beta r_{i-1} + (1 - \beta) s_i, \quad (11)$$

where β is a learnable coefficient, r_{i-1} is the previous curve descriptor, and s_i is the feature of the current point. This strategy enables each curve to follow natural finger articulations while accumulating historical information to guide subsequent steps.

To prevent curves from overlapping or deviating from expected anatomical paths, a cross-suppression mechanism is applied. For each curve, a support vector $\vec{c}_i$ representing

its overall direction is computed, and candidate neighbor vectors $\vec{q}_j$ are evaluated. The suppression factor is calculated as

$$d_j = \text{clip}(\cos \theta + 1), \quad (12)$$

$$\alpha_j = d_j \cdot \alpha_j, \quad (13)$$

where θ is the angle between $\vec{c}_i$ and $\vec{q}_j$, and $\text{clip}(\cdot)$ restricts values to $[0,1]$. This ensures curves follow natural finger directions and reduces intersections.

Overall, the CG stage incorporates joint-sensitive scoring, momentum-based descriptors, and curve generation with anatomical awareness, producing curves that capture local geometry, preserve anatomical structure, and remain robust to corrupted or incomplete point clouds. This ensures that the subsequent CA stage receives geometrically and anatomically meaningful curves for effective feature aggregation.

3.3.2. Curve Aggregation Stage

Once curves are generated in the CG stage, the CA module integrates features along these curves to produce enhanced local and global representations, as illustrated in Figure 6. Each curve C_i contains a sequence of point features along the walking path, and the full set of curves is denoted by $C = \{C_1, \dots, C_n\}$. To capture both intra-curve structural details and inter-curve relationships, the module first applies attention pooling along the curve set:

$$f_{\text{inter}}, f_{\text{intra}} = AP(C), \quad (14)$$

where AP denotes attention pooling applied along multiple directions, producing an inter-curve feature vector f_{inter} and an intra-curve feature vector f_{intra} . These vectors represent global context between curves and local structural information within each curve, respectively.

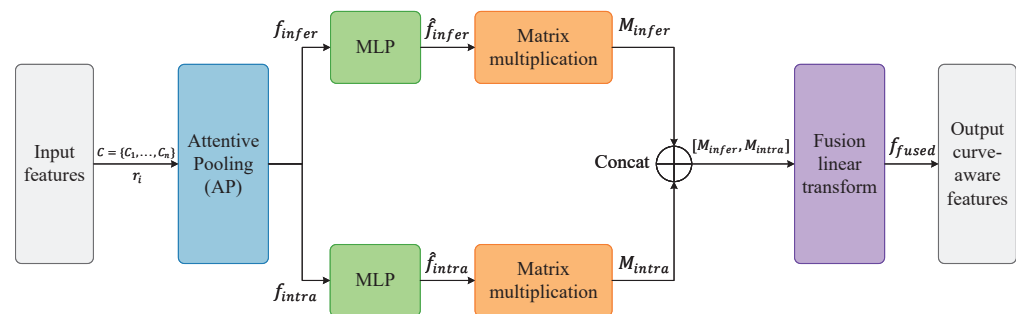


Figure 6. Detailed process of the CA stage in HCGC. Input curve features are first processed by Attentive Pooling (AP) to extract inter-curve and intra-curve information, followed by dimensionality reduction via parallel MLPs. The features are then mapped through learned linear transformations and fused using a linear fusion layer to generate the final curve-aware representation. The design emphasizes joint-sensitive weighting and anatomical continuity, ensuring robust feature extraction from corrupted point clouds.

To reduce computational complexity and facilitate subsequent fusion, f_{inter} and f_{intra} are passed through independent MLPs for dimensionality reduction:

$$\hat{f}_{\text{inter}} = \text{MLP}(f_{\text{inter}}), \quad \hat{f}_{\text{intra}} = \text{MLP}(f_{\text{intra}}). \quad (15)$$

Feature mappings are then computed via learned linear transformations

$$M_{\text{inter}} = \hat{f}_{\text{inter}} \cdot W_{\text{inter}} + b_{\text{inter}}, \quad M_{\text{intra}} = \hat{f}_{\text{intra}} \cdot W_{\text{intra}} + b_{\text{intra}}, \quad (16)$$

followed by softmax normalization to produce weights for curve fusion. Unlike the original CurveNet aggregation, our approach emphasizes joint-sensitive weighting and dynamically considers the anatomical significance of each curve, ensuring that curves representing critical regions, such as fingertips or joints, contribute more to the final feature.

Finally, the inter- and intra-curve features are fused to produce a comprehensive curve-aware representation:

$$f_{\text{fused}} = W_{\text{fuse}}[f_{\text{inter}}, f_{\text{intra}}] + b_{\text{fuse}}, \quad (17)$$

where $[\cdot, \cdot]$ denotes feature concatenation. This linear fusion layer learns to adaptively balance the contributions of inter-curve and intra-curve features through the trainable weight matrix W_{fuse} and bias b_{fuse} , resulting in a unified representation that simultaneously encodes fine-grained local geometry and global contextual dependencies across curves.

To further enhance robustness to missing or distorted points, the fused features are optionally refined using the AS module. This refinement dynamically adjusts the positions and features of key points, particularly in low-density or structurally complex regions, improving the reliability of the features for subsequent global aggregation.

Through this design, the CA module extends the original CurveNet aggregation by explicitly incorporating anatomical continuity, joint-sensitive weighting, and AS-based refinement. This ensures that the network produces a robust and discriminative feature representation that is tailored for hand pose estimation, effectively capturing both fine-grained local details and global hand structure even in the presence of incomplete or corrupted point clouds.

3.4. Evaluation Metrics

Real-world hand point clouds exhibit diverse distortions, including occlusions and background remnants, which are difficult to fully simulate. Therefore, this study focuses on the most common type of distortion: local point cloud missing.

We evaluate models on widely used 3D hand pose datasets, including MSRA [43], ICVL [44], and NYU [12], which provide diverse gestures and multiple viewpoints. Prediction accuracy is measured by Mean Per Joint Position Error (MPJPE):

$$MPJPE = \frac{1}{N} \sum_{i=1}^N \frac{1}{J} \sum_{j=1}^J \|P_j^{(i)} - \hat{P}_j^{(i)}\|_2, \quad (18)$$

where N is the number of test samples, J the number of joints, and $P_j^{(i)}, \hat{P}_j^{(i)}$ are ground truth and predicted joint coordinates.

To evaluate robustness under varying levels of corruption, we construct an error–distortion curve by plotting MPJPE as a function of the distortion severity. The overall robustness is quantified by the area under this curve, normalized by the clean performance, defined as:

$$RCA = \frac{1}{R} \sum_{r=1}^R \frac{MPJPE_{\text{corrupted}}(r) - MPJPE_{\text{clean}}}{MPJPE_{\text{clean}}}, \quad (19)$$

where R is the number of distortion levels and $MPJPE_{\text{corrupted}}(r)$ denotes the error under distortion level r . For clarity, all RCA values reported in this paper are scaled by 100, and they remain dimensionless. Thus, a smaller RCA indicates less performance degradation and stronger robustness.

The use of RCA provides a holistic measure of model robustness across varying degrees of point cloud corruption, rather than relying on single-point evaluations. By summarizing performance trends over multiple missing-point ratios, RCA captures both accuracy and

stability under progressively challenging conditions. This makes it particularly suitable for real-world hand pose estimation, where the extent and distribution of missing points can vary dynamically, offering a more faithful reflection of practical robustness than traditional per-level metrics.

Finally, to provide a comparative measure relative to a baseline, we define the RG as:

$$RG = RCA_{\text{baseline}} - RCA_{\text{proposed}}, \quad (20)$$

where RCA_{baseline} and RCA_{proposed} are the robustness curve areas of the baseline and proposed models, respectively. A positive RG indicates that the proposed model achieves better robustness improvements under varying distortion severities.

RCA and RG together provide a robust evaluation of model performance under varying levels of point cloud corruption, capturing both relative degradation and improvements over a baseline. This ensures that the practical robustness of the algorithm is thoroughly assessed in realistic scenarios.

4. Experimentation

4.1. Datasets, Setup, and Metrics

We evaluate the proposed 3D hand pose estimation framework on three benchmark datasets: MSRA, ICVL, and NYU. MSRA contains over 90,000 frames with 21-joint annotations capturing complex gestures. ICVL provides 330,000 training frames with 16-joint annotations across multiple users. NYU comprises 72,757 training frames and 8252 test frames with 14-joint annotations, recorded with a structured-light depth sensor. Figure 7 illustrates the spatial distribution of annotated hand joints. These datasets enable comprehensive evaluation of model accuracy, generalization, and robustness across diverse hand poses and viewpoints.

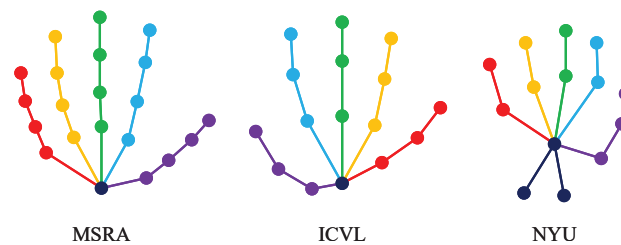


Figure 7. Spatial distribution of annotated hand joints in the MSRA (left), ICVL (middle), and NYU (right) datasets.

The network is trained using the Adam optimizer [45] with an initial learning rate of 0.001, a batch size of 32, and a weight decay of 0.0005. Input point clouds are uniformly sampled to 1024 points per frame, and in the hand geometric affine module, K is set to 64 for nearest neighbor computation to capture local geometric relationships. In addition, the Top- K selection mechanism, used in the curve-guided convolution module to identify structurally informative key points, is configured separately for each dataset: $K = 20$ for MSRA, $K = 25$ for ICVL, and $K = 30$ for NYU. The Top- K value is chosen based on the average point cloud density and gesture complexity to ensure sufficient coverage of high-quality starting points without introducing excessive computational overhead. It is important to note that this Top- K parameter is distinct from the K used in KNN for local feature aggregation.

To enhance generalization and mitigate overfitting, online data augmentation is applied, including random rotations around the z -axis within $\pm 37.5^\circ$, isotropic 3D scaling in the range $[0.9, 1.1]$, and 3D translation along each axis within $[-10, 10]$ mm. Training is

conducted for 120 epochs on MSRA, 500 epochs on ICVL, and 250 epochs on NYU, ensuring sufficient convergence across datasets of varying size and complexity.

To provide insight into computational efficiency, we also measured the model complexity and inference speed. The HCGCNet requires approximately 25 GFLOPs per forward pass for a single hand point cloud with 1024 points. On an NVIDIA RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA), the average inference time per sample is about 12 ms. These results demonstrate that the proposed framework achieves a favorable trade-off between robustness and efficiency, making it suitable for real-time 3D hand pose estimation. All FLOPs calculations consider only forward operations, and all measurements are based on full-resolution input point clouds.

4.2. Pose Estimation Visualization

To qualitatively evaluate the proposed method on intact hand point clouds, Figure 8 presents the pose estimation results on the MSRA, ICVL, and NYU datasets. Each dataset is displayed in three rows: the original point clouds, the ground truth annotations, and the predicted joint positions by our network.

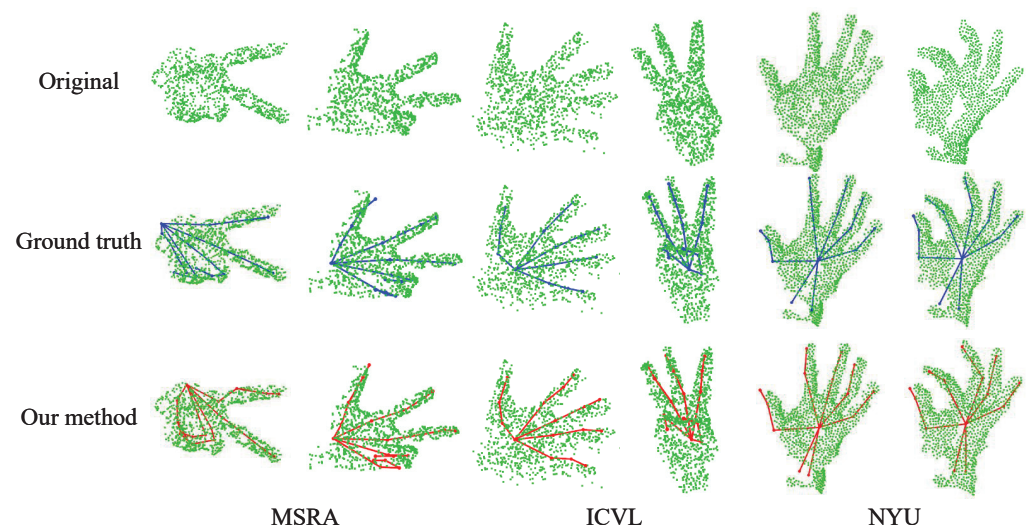


Figure 8. Visualization of pose estimation on intact hand point clouds. The first row shows the original hand point clouds, the second row presents the ground truth annotations, and the third row illustrates the predicted 3D joint positions by the proposed network for the MSRA (**left**), ICVL (**middle**), and NYU (**right**) datasets.

Visual inspection shows that the predicted 3D joint positions closely match the ground truth, accurately reconstructing hand skeletal structures across a wide variety of gestures, including open palms, clenched fists, and complex finger interactions. The network effectively captures detailed finger articulations and joint relationships, preserving anatomical continuity and maintaining consistent bone lengths. Notably, the method performs well even in challenging regions such as fingertips and thumb bases, which are prone to prediction errors.

The network demonstrates robustness to variations in viewpoint and hand shape across the three datasets. Furthermore, the predicted joints exhibit smooth and coherent trajectories, avoiding abrupt deviations and demonstrating superior local geometric fidelity. These observations confirm that, in the absence of noise or missing points, the proposed framework reliably estimates hand poses with high accuracy and preserves fine-grained anatomical details, providing a solid foundation for evaluating robustness under corrupted point cloud scenarios in the following sections.

4.3. Robustness to Corrupted Point Clouds

To evaluate the robustness of the proposed network under realistic point cloud distortions, we analyze fixed-level missing points at 30%, 40%, and 50%. Figure 9 visualizes the results for the MSRA, ICVL, and NYU datasets, showing the original point clouds and the corrupted point clouds overlaid with predictions from HandPointNet and the proposed network. Compared to existing approaches, HandPointNet exhibits noticeable distortions under partial data loss, often producing fragmented skeletal structures or significant joint deviations. In contrast, our method effectively mitigates the impact of missing points by combining adaptive sampling with curve-guided feature modeling. The generated curves guide the network to preserve anatomical continuity, maintain coherent finger shapes, and achieve accurate joint localization even under substantial point cloud degradation. Quantitative observations further confirm that the proposed network consistently reduces errors and outperforms baselines across all levels of distortion.

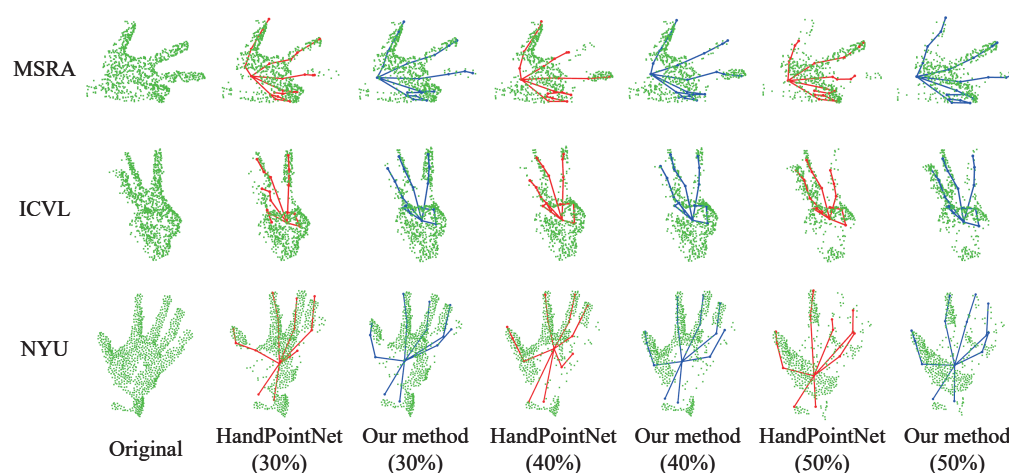


Figure 9. Visual comparison of 3D hand pose estimation under varying point cloud distortions. Each row corresponds to one dataset: MSRA (first row), ICVL (second row), and NYU (third row). The first column shows the original hand point clouds. Subsequent columns overlay the prediction results of HandPointNet (red) and our method (blue) on corrupted point clouds with 30%, 40%, and 50% missing points.

To further verify the noise resistance of the proposed framework, Gaussian perturbations with standard deviations of $\sigma = 0.02$, $\sigma = 0.03$, and $\sigma = 0.04$ are added to simulate sensor noise and spatial jitter commonly found in depth cameras. As shown in Figure 10, both methods exhibit increasing deviation as noise intensity grows, but the predictions of HandPointNet drift more significantly, showing blurred joints and unstable finger orientations. In contrast, our method maintains smoother, anatomically consistent structures and smaller positional shifts. This robustness benefits from the curve-guided representation, which promotes local geometric regularity and reduces noise propagation during feature aggregation.

These visual and quantitative analyses collectively demonstrate that the proposed framework not only reconstructs hand poses accurately on complete point clouds but also maintains high robustness under varying degrees of local distortions. The integration of dynamic key point selection and curve-guided feature completion is crucial for preserving structural integrity and enhancing generalization across diverse hand gestures and datasets.

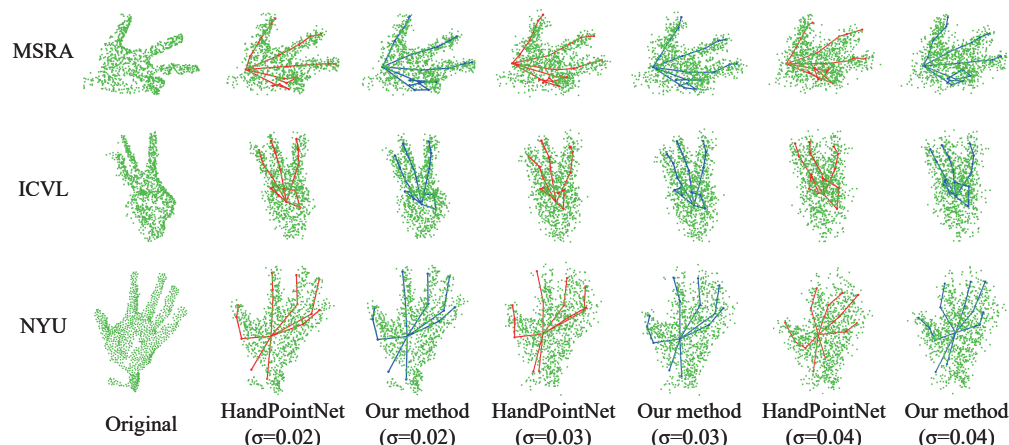


Figure 10. Visual comparison of 3D hand pose estimation under Gaussian noise with different standard deviations ($\sigma = 0.02$, $\sigma = 0.03$, and $\sigma = 0.04$). Each row corresponds to one dataset: MSRA (first row), ICVL (second row), and NYU (third row). The first column shows the original clean point clouds. Subsequent columns overlay the prediction results of HandPointNet (red) and our method (blue) under increasing noise intensity.

4.4. Comparative Analysis

We first evaluate our curve fitting-based 3D hand pose estimation method on intact point clouds. As shown in Table 1, the proposed method achieves competitive average joint errors of 7.43 mm on MSRA, 6.56 mm on ICVL, and 10.07 mm on NYU. Figures 11 and 12 further illustrate the performance. Figure 11 reports the average joint errors for each method, while Figure 12 presents the proportion of high-quality estimations within the same error threshold.

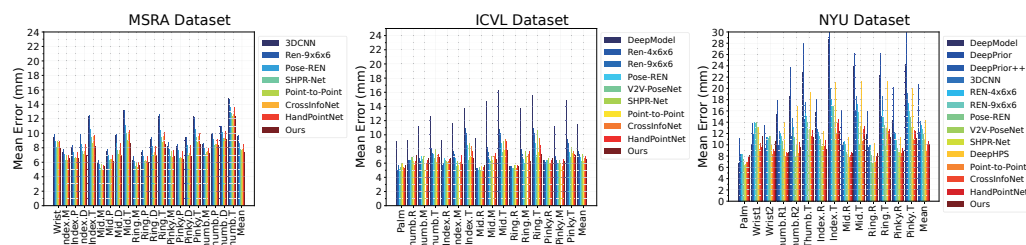


Figure 11. Comparison of average joint errors of the proposed method and state-of-the-art methods on intact point clouds across three datasets: MSRA (left), ICVL (middle), and NYU (right). The proposed method demonstrates competitive performance, achieving comparable or slightly better average joint errors while maintaining accurate 3D hand pose estimations.

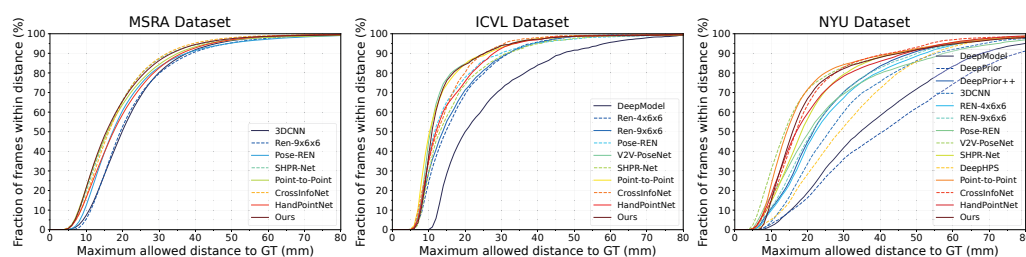


Figure 12. This figure presents a model quality analysis, evaluating the proportion of high-quality estimations. For each method, the estimation results are assessed within the same error range across the three datasets: the MSRA dataset (left), the ICVL dataset (middle), and the NYU dataset (right). The proposed method achieves a higher proportion of high-quality estimation results in all datasets, proving its robustness and reliability in 3D hand pose estimation.

Table 1. Comparison of average error distances of existing methods on the MSRA, ICVL, and NYU datasets.

Methodologies	MSRA (mm)	ICVL (mm)	NYU (mm)
DeepModel	-	11.56	17.04
DeepPrior	-	-	20.75
DeepPrior++	-	-	12.24
3DCNN	9.58	-	14.11
Ren-4x6x6	-	7.63	13.39
HandPointNet	8.51	6.94	10.54
Ren-9x6x6	9.79	7.31	12.69
Pose-REN	8.65	6.79	11.81
DenseReg	7.23	7.24	10.21
V2V-PoseNet	-	6.28	8.42
SHPR-Net	7.76	7.22	10.78
DeepHPS	-	-	14.42
Point-to-Point	7.71	6.33	9.05
CrossInfoNet	7.86	6.73	10.08
TriHorn-Net	7.13	5.74	7.68
Ours	7.43	6.56	10.07

These results confirm that our method produces accurate predictions and maintains a higher proportion of reliable estimations compared to existing approaches. Our method achieves competitive accuracy: the error is slightly higher than TriHorn-Net on MSRA (0.30 mm) and ICVL (0.82 mm), with a larger gap on NYU (2.39 mm) due to its more severe occlusion challenges. However, our primary focus is on corrupted or partially missing point clouds, where the proposed method demonstrates superior robustness. It consistently preserves accurate joint localization and coherent skeletal structures, even under severe distortions, indicating strong generalization across diverse datasets and hand gestures.

To quantitatively assess robustness, we compare the RCA and RG values of existing methods with the proposed approach for fixed missing-point ratios of 30%, 40%, and 50% (Table 2). Baseline models such as HandPointNet and TriHorn-Net exhibit higher RCA values, reflecting their greater sensitivity to missing points. In contrast, our method achieves lower RCA values across all datasets, while RG values further demonstrate its ability to maintain high-quality estimations under distortion, particularly on the MSRA dataset. These results validate that adaptive sampling combined with hand-curve guidance effectively mitigates feature deviations caused by structural distortions, while maintaining competitive performance on intact point clouds. Collectively, this analysis highlights the dual advantages of the proposed method: strong accuracy on complete point clouds and enhanced robustness under realistic corruption scenarios, confirming its practical value for real-world 3D hand pose estimation.

To further evaluate robustness under random perturbations, Table 3 summarizes the RCA values of various methods subjected to Gaussian noise with different standard deviations. Overall, all models experience gradual performance degradation as noise intensity increases, reflecting the inherent sensitivity of depth-based point clouds to spatial jitter. Our method consistently demonstrates comparatively strong performance across the three datasets, although the advantage over the best-performing baselines such as TriHorn-Net and Point-to-Point is relatively small. This moderate but consistent gain indicates that the curve-guided representation helps preserve local geometric consistency and mitigates error accumulation under noisy conditions. These results complement the findings in Table 2, confirming that our approach maintains reliable accuracy not only against missing-point corruption but also under realistic sensor noise, demonstrating its broad-spectrum robustness for 3D hand pose estimation.

Table 2. Robustness Curve Area (RCA) and Robustness Gain (RG) values of state-of-the-art methods and the proposed method on the MSRA, ICVL, and NYU datasets under fixed missing-point ratios of 30%, 40%, and 50%. Lower RCA indicates stronger robustness, and higher RG indicates greater improvement over baselines.

Method	MSRA	ICVL	NYU	Average
3DCNN	50.2	40.5	45.8	45.5
Ren-9x6x6	48.3	38.1	43.2	43.2
Pose-REN	45.6	36.4	40.1	40.7
DenseReg	40.8	35.2	38.5	38.2
SHPR-Net	42.7	34.9	37.3	38.3
Point-to-Point	38.9	33.6	36.4	36.3
CrossInfoNet	43.5	37.2	39.7	40.1
TriHorn-Net	33.2	31.5	34.3	33.0
HandPointNet	49.4	35.6	39.8	41.6
Our Method	28.7	30.7	33.1	30.8
RG	20.7	4.9	6.7	10.8

Table 3. RCA values of state-of-the-art methods and the proposed method under different levels of Gaussian noise on the MSRA, ICVL, and NYU datasets. Lower RCA indicates stronger robustness to noise-induced perturbations.

Method	MSRA	ICVL	NYU	Average
3DCNN	24.6	21.8	23.1	23.2
Ren-9x6x6	23.8	20.7	22.3	22.3
Pose-REN	22.5	19.9	21.2	21.2
DenseReg	20.7	19.1	20.0	19.9
SHPR-Net	21.1	18.6	19.5	19.7
Point-to-Point	19.8	18.3	19.0	19.0
CrossInfoNet	22.2	19.5	20.4	20.7
TriHorn-Net	18.9	17.6	18.3	18.3
HandPointNet	23.4	19.3	20.1	20.9
Our Method	18.4	18.1	17.9	18.1
RG	5.0	1.2	2.2	2.8

4.5. Key Modules Ablation Study

To systematically evaluate the contributions of the key components in our framework, we conducted a comprehensive ablation study on the MSRA, ICVL, and NYU datasets. The experiments are grouped into two main categories (Table 4): the AS module, and the HCGC module, which inherently includes the curve-guided refinement strategies such as dynamic momentum encoding and cross-suppression.

The AS module dynamically selects structurally informative key points according to local density and anatomical importance, mitigating sampling bias in distorted regions. Replacing the AS module with traditional FPS sampling results in an increase in average joint errors by 0.65 mm, 0.69 mm, and 1.55 mm on the MSRA, ICVL, and NYU datasets, respectively. This demonstrates that adaptive selection allows the network to better capture local geometric features, particularly in regions with complex finger articulations or partial occlusions, significantly enhancing robustness and accuracy under challenging point cloud conditions.

We further analyzed the effect of the HCGC module, which integrates curve grouping, curve aggregation, dynamic momentum encoding, and cross-suppression. Using traditional feature extraction without curve guidance leads to substantial performance degradation across all datasets. Introducing only curve grouping partially improves performance by capturing local structural information along the fingers. The full HCGC module, which encompasses

curve grouping, curve aggregation, and refinement strategies, achieves the lowest errors consistently. This highlights that curve-guided local feature extraction and refinement are critical for modeling continuous anatomical patterns, preserving joint relationships, and accurately estimating hand poses even under occlusion or noise. The benefits are particularly pronounced on the NYU dataset, where severe occlusions challenge pose estimation.

Table 4. Ablation study results on MSRA, ICVL, and NYU datasets. Lower values indicate better performance. Horizontal lines separate the ablation groups.

Configuration	MSRA (mm)	ICVL (mm)	NYU (mm)
AS Module Ablation			
FPS Sampling	8.08	7.25	11.62
AS Module	7.43	6.56	10.07
HCGC Module Ablation			
Traditional Feature Extraction	9.52	8.12	12.28
Curve Grouping Only	8.53	7.33	11.24
Full HCGC Module	7.43	6.56	10.07

Overall, these ablation experiments quantitatively demonstrate that each component of the proposed framework plays a crucial role: the AS module enhances sampling quality and local feature representation, while the HCGC module with integrated curve-guided refinement effectively captures fine-grained geometric variations. Their combined effect ensures superior performance on both intact and corrupted point clouds, confirming the robustness and generalizability of the proposed method across diverse datasets.

5. Conclusions

This study presents a novel curve fitting-based framework for 3D hand pose estimation specifically tailored to corrupted and distorted point clouds. Unlike conventional point cloud-based methods that often fail under sensor noise or missing data, our approach emphasizes robustness by explicitly modeling local geometric structures and leveraging curve-guided feature aggregation. The integration of curve-based reasoning allows the network to capture continuous anatomical patterns, preserving finger articulations and joint relationships even in severely corrupted point clouds.

The proposed framework incorporates two key components: the AS module and the HCGC module. The AS module dynamically selects structurally informative key points, mitigating sampling bias in distorted regions, while the HCGC module uses curve construction, momentum encoding, and cross-suppression strategies to guide feature aggregation along anatomically meaningful paths. This design ensures that local structural features are effectively captured, enabling accurate joint localization despite missing or noisy points, which is a critical challenge in real-world hand gesture capture scenarios.

Extensive experiments on the MSRA, ICVL, and NYU datasets demonstrate that our framework consistently outperforms state-of-the-art methods under various corruption scenarios, including partial point removal and sensor noise. Quantitative evaluations show that the proposed method achieves lower error rates and stronger robustness compared to baseline approaches. Specifically, on corrupted MSRA datasets with missing-point ratios ranging from 30% to 50%, our method achieves an average RCA of 30.8, surpassing advanced methods such as TriHorn-Net and maintaining stable joint localization under severe degradation. Meanwhile, on intact point clouds, it also achieves competitive average joint errors of 7.43 mm on MSRA, 6.56 mm on ICVL, and 10.07 mm on NYU, confirming that enhanced robustness does not come at the cost of accuracy. Visualizations further confirm that the AS module effectively selects structurally informative key points, while the HCGC module preserves coherent skeletal structures even under distortions. Ablation studies verify the contribution of each component

to overall performance. These results collectively confirm that the proposed curve fitting-based local structure modeling provides a reliable and generalizable solution for realistic 3D hand pose estimation under varying point cloud quality.

Despite its strong performance, the proposed method has inherent limitations. In extreme scenarios where large portions of the hand are fully occluded or when point clouds are extremely sparse, some fine-grained articulations may be less accurately recovered. Additionally, highly unusual hand poses not seen during training may still pose challenges. As future work, we plan to explore the integration of temporal information from sequential frames to further improve robustness in dynamic scenarios, investigate self-supervised or domain-adaptive learning to handle unseen poses, and extend the framework to multi-hand or hand-object interaction cases. These directions aim to enhance the generalization and practical applicability of the proposed method, while building on the strong foundational robustness demonstrated in this study.

Author Contributions: Conceptualization, L.S., H.S., H.Z., H.L., X.G. and Y.C.; methodology, L.S., H.S., H.Z., H.L., X.G. and Y.C.; software, L.S., H.S., H.Z., H.L., X.G. and Y.C. (corresponding to model implementation); writing—review and editing, L.S., H.S., H.Z., H.L., X.G. and Y.C. (all authors commented on earlier versions of the manuscript), supervision: L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Universities of the Liaoning Provincial Department of Education, China (grant number LJKZ0011), and by the Liaoning Provincial Science and Technology Innovation Project, China (grant number 2021JH1/1040011).

Data Availability Statement: The data presented in this study are openly available in [FigShare] at [10.6084/m9.figshare.30184705].

Acknowledgments: The authors would like to acknowledge the support from the Fundamental Research Funds for the Universities of the Liaoning Provincial Department of Education and the Liaoning Provincial Science and Technology Innovation Project. All experiments were conducted using publicly available datasets, and the authors sincerely appreciate the providers of these valuable resources.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The authors have no relevant financial or non-financial interests to report, and no affiliation or involvement with any organization or entity that has any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

AS	Adaptive Sampling
HCGC	Hand-Curve Guide Convolution
CIC	Curve Instance Convolution
FPS	Farthest Point Sampling
KNN	K-Nearest Neighbors
CG	Curve Grouping
CA	Curve Aggregation
AP	Attentive Pooling
MPJPE	Mean Per Joint Position Error
RCA	Robustness Curve Area
RG	Robustness Gain

References

1. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
2. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
3. Guo, M.H.; Cai, J.X.; Liu, Z.N.; Mu, T.J.; Martin, R.R.; Hu, S.M. Pct: Point cloud transformer. *Comput. Vis. Media* **2021**, *7*, 187–199. [\[CrossRef\]](#)
4. Huang, Y.; Zhang, M.; Ding, D.; Jiang, E.; Wang, Z.; Yang, M. Causalpc: Improving the robustness of point cloud classification by causal effect identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 19779–19789.
5. Watanabe, R.; Nonaka, K.; Pavez, E.; Kobayashi, T.; Ortega, A. Fast graph-based denoising for point cloud color information. In Proceedings of the ICASSP 2024—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 4025–4029.
6. Zhu, Z.; Chen, H.; He, X.; Wang, W.; Qin, J.; Wei, M. Svdformer: Complementing point cloud via self-view augmentation and self-structure dual-generator. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 14508–14518.
7. Ge, L.; Cai, Y.; Weng, J.; Yuan, J. Hand pointnet: 3d hand pose estimation using point sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8417–8426.
8. Xiang, T.; Zhang, C.; Song, Y.; Yu, J.; Cai, W. Walk in the cloud: Learning curves for point clouds shape analysis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 915–924.
9. Erol, A.; Bebis, G.; Nicolescu, M.; Boyle, R.D.; Twombly, X. Vision-based hand pose estimation: A review. *Comput. Vis. Image Underst.* **2007**, *108*, 52–73. [\[CrossRef\]](#)
10. Oikonomidis, I.; Kyriazis, N.; Argyros, A. Efficient model-based 3d tracking of hand articulations using kinect. In Proceedings of the British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011.
11. Keskin, C.; Kırac, F.; Kara, Y.E.; Akarun, L. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In Proceedings of the Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part VI 12; Springer: Berlin/Heidelberg, Germany, 2012; pp. 852–863.
12. Tompson, J.; Stein, M.; Lecun, Y.; Perlin, K. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. Graph. ToG* **2014**, *33*, 1–10. [\[CrossRef\]](#)
13. Oberweger, M.; Wohlhart, P.; Lepetit, V. Hands deep in deep learning for hand pose estimation. *arXiv* **2015**, arXiv:1502.06807.
14. Oberweger, M.; Lepetit, V. Deepprior++: Improving fast and accurate 3d hand pose estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 585–594.
15. Ge, L.; Liang, H.; Yuan, J.; Thalmann, D. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1991–2000.
16. Moon, G.; Chang, J.Y.; Lee, K.M. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5079–5088.
17. Wan, C.; Probst, T.; Van Gool, L.; Yao, A. Dense 3d regression for hand pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5147–5156.
18. Spurr, A.; Iqbal, U.; Molchanov, P.; Hilliges, O.; Kautz, J. Weakly supervised 3d hand pose estimation via biomechanical constraints. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 211–228.
19. Wang, J.; Mueller, F.; Bernard, F.; Theobalt, C. Generative Model-Based Loss to the Rescue: A Method to Overcome Annotation Errors for Depth-Based Hand Pose Estimation. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG), Buenos Aires, Argentina, 16–20 November 2020; IEEE: New York, NY, USA, 2020; pp. 101–108. [\[CrossRef\]](#)
20. Cheng, J.; Wan, Y.; Zuo, D.; Ma, C.; Gu, J.; Tan, P.; Wang, H.; Deng, X.; Zhang, Y. Efficient virtual view selection for 3d hand pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2022; Volume 36, pp. 419–426.
21. Ge, L.; Ren, Z.; Yuan, J. Point-to-point regression pointnet for 3d hand pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 475–491.

22. Tang, D.; Ye, Q.; Yuan, S.; Taylor, J.; Kohli, P.; Keskin, C.; Kim, T.K.; Shotton, J. Opening the black box: Hierarchical sampling optimization for hand pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2161–2175. [[CrossRef](#)] [[PubMed](#)]
23. Chen, X.; Wang, G.; Zhang, C.; Kim, T.K.; Ji, X. Shpr-net: Deep semantic hand pose regression from point clouds. *IEEE Access* **2018**, *6*, 43425–43439. [[CrossRef](#)]
24. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on x-transformed points. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018.
25. Zhao, H.; Jiang, L.; Fu, C.W.; Jia, J. Pointweb: Enhancing local neighborhood features for point cloud processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5565–5573.
26. Liu, X.; Han, Z.; Liu, Y.S.; Zwicker, M. Point2sequence: Learning the shape representation of 3d point clouds with an attention-based sequence to sequence network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8778–8785.
27. Malik, J.; Elhayek, A.; Nunnari, F.; Varanasi, K.; Tamaddon, K.; Heloir, A.; Stricker, D. DeepHPS: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; IEEE: New York, NY, USA, 2018; pp. 110–119.
28. Rezaei, M.; Rastgoo, R.; Athitsos, V. TriHorn-net: A model for accurate depth-based 3D hand pose estimation. *Expert Syst. Appl.* **2023**, *223*, 119922. [[CrossRef](#)]
29. Li, S.; Lee, D. Point-to-Pose Voting Based Hand Pose Estimation Using Residual Permutation Equivariant Layer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11927–11936. [[CrossRef](#)]
30. Wu, Y.; Ma, S.; Zhang, D.; Sun, J. 3D Capsule Hand Pose Estimation Network Based on Structural Relationship Information. *Symmetry* **2020**, *12*, 1636. [[CrossRef](#)]
31. Du, K.; Lin, X.; Sun, Y.; Ma, X. Crossinfonet: Multi-task information sharing based hand pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9896–9905.
32. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 16259–16268.
33. Wu, X.; Jiang, L.; Wang, P.S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; Zhao, H. Point transformer v3: Simpler faster stronger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 4840–4851.
34. Cheng, W.; Tang, H.; Gool, L.V.; Jeon, J.H.; Ko, J.H. HandDiff: 3D Hand Pose Estimation with Diffusion on Image-Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 2274–2284. [[CrossRef](#)]
35. Wang, Y.; Xu, H.; Heng, P.A.; Fu, C.W. Unihope: A unified approach for hand-only and hand-object pose estimation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Denver, CO, USA, 3–7 June 2025; pp. 12231–12241.
36. Zhang, C.; Qi, Z.; Yuan, W.; Qi, W.; Yang, Z.; Su, Z. PCDM: Point Cloud Completion by Conditional Diffusion Model. *C. Neural Process. Lett.* **2025**, *57*, 51. [[CrossRef](#)]
37. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph. TOG* **2019**, *38*, 1–12. [[CrossRef](#)]
38. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
39. Yan, X.; Zheng, C.; Li, Z.; Wang, S.; Cui, S. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5589–5598.
40. Li, N.; Li, X.; Zhou, J.; Jiang, D.; Liu, J.; Qin, H. GeoHi-GNN: Geometry-aware hierarchical graph representation learning for normal estimation. *Comput. Aided Geom. Des.* **2024**, *114*, 102390. [[CrossRef](#)]
41. Zhu, X.; Du, D.; Chen, W.; Zhao, Z.; Nie, Y.; Han, X. Nerve: Neural volumetric edges for parametric curve extraction from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13601–13610.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Sun, X.; Wei, Y.; Liang, S.; Tang, X.; Sun, J. Cascaded hand pose regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 824–832.

44. Tang, D.; Jin Chang, H.; Tejani, A.; Kim, T.K. Latent regression forest: Structured estimation of 3d articulated hand posture. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3786–3793.
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.