**CMG-Net: Robust Normal Estimation for Point Clouds via Chamfer Normal Distance and Multi-Scale Geometry**

Yingrui Wu\textsuperscript{1,2}\textsuperscript{*}, Mingyang Zhao\textsuperscript{3}\textsuperscript{*}, Keqiang Li\textsuperscript{4}, Weize Quan\textsuperscript{1,2}

Tianqi Yu\textsuperscript{5}, Jianfeng Yang\textsuperscript{5}, Xiaohong Jia\textsuperscript{6,2}, Dong-Ming Yan\textsuperscript{1,2}\textsuperscript{†}

\textsuperscript{1}MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China

\textsuperscript{2}University of Chinese Academy of Sciences, Beijing, China

\textsuperscript{3}Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China

\textsuperscript{4}SenseTime Research, Shanghai, China

\textsuperscript{5}School of Electronic and Information Engineering, Soochow University, Suzhou, China

\textsuperscript{6}AMSS, Chinese Academy of Sciences, Beijing, China

\texttt{wuyingrui2023@ia.ac.cn}, \{migyangz, likeq98, qweizework, yandongming\}@gmail.com

\{tqyu, jfyang\}@suda.edu.cn, xhjia@amss.ac.cn

**Abstract**

This work presents an accurate and robust method for estimating normals from point clouds. In contrast to predecessor approaches that minimize the deviations between the annotated and the predicted normals directly, leading to direction inconsistency, we first propose a new metric termed Chamfer Normal Distance to address this issue. This not only mitigates the challenge but also facilitates network training and substantially enhances the network robustness against noise. Subsequently, we devise an innovative architecture that encompasses Multi-scale Local Feature Aggregation and Hierarchical Geometric Information Fusion. This design empowers the network to capture intricate geometric details more effectively and alleviate the ambiguity in scale selection. Extensive experiments demonstrate that our method achieves the state-of-the-art performance on both synthetic and real-world datasets, particularly in scenarios contaminated by noise. Our implementation is available at \url{https://github.com/YingruiWoo/CMG-Net}. Pytorch.

**Introduction**

Normal estimation is a fundamentally important task in the field of point cloud analysis, which enjoys a wide variety of applications in 3D vision and robotics, such as surface reconstruction (Fleishman, Cohen-Or, and Silva 2005; Kazhdan, Bolitho, and Hoppe 2006), denoising (Lu et al. 2020b) and semantic segmentation (Grilli, Menna, and Remondino 2017; Che and Olsen 2018). In recent years, many powerful methods have been developed to enhance the performance of normal estimation. However, these approaches involving both traditional and learning-based ones often suffer from heavy noise and struggle to attain high-quality results for point clouds with complex geometries.

Traditional methods (Hoppe et al. 1992; Levin 1998; Cazals and Pouget 2005) typically encompass fitting local planes or polynomial surfaces and inferring normal vectors from the fitted outcomes. Although straightforward, these approaches are vulnerable to noise and encounter challenges when attempting to generalize to complex shapes. Furthermore, their performance hinges significantly on the meticulous tuning of parameters.

In comparison with traditional approaches, learning-based proposals (Guerrero et al. 2018; Ben-Shabat et al. 2019; Hashimoto and Saito 2019; Zhou et al. 2020; Wang and Prisacariu 2020; Lenssen, Osendorfer, and Masci 2020; Ben-Shabat et al. 2020; Cao et al. 2021; Zhu et al. 2021; Zhou et al. 2022b; Zhang et al. 2022; Li et al. 2022a,b; Du et al. 2023; Li et al. 2023a) have better generalization and less dependency on parameter tuning. There are two types of learning-based normal estimators comprising deep...
We propose a new method that integrates the CND metric for robust normal estimation, which solves the direction inconsistency problem effectively and significantly boosts network training and inference.

• We design a novel network that incorporates multi-scale feature extraction along with hierarchical inference combined with intricate geometry information fusion, which is capable of capturing intricate geometric details and addressing the challenge of scale selection ambiguity.

• We perform comprehensive experiments to demonstrate the enhancements brought by our proposed method, thereby pushing the boundaries of SOTA performance, especially on noisy normal estimation scenarios.

Related Work

Traditional Methods
Principal Component Analysis (PCA) (Hoppe et al. 1992) stands as the most widely adopted point cloud normal estimation method, which fits a plane to the input surface patch. Subsequent variants involving Moving Least Squares (MLS) (Levin 1998), truncated Taylor expansion fitting (n-jet) (Caiafa, Pouget 2005), local spherical surface fitting (Guennebaud and Gross 2007) and multi-scale kernel (Aoudj et al. 2017) are proposed to reduce the noisy influence through selecting larger patches and employing more intricate energy functions. Nevertheless, these approaches typically tend to oversmooth sharp features and geometric details. To circumvent these issues, Voronoi diagram (Amenta and Bern 1998; Alliez et al. 2007; Mérigot, Ovsjanikov, and Guibas 2010), Hough transform (Boulch and Marlet 2012), and plane voting (Zhang et al. 2018) are deployed in normal estimation. However, these techniques depend on manual parameter tuning heavily, which hinders their practical applications.

Learning-based Methods
With the powerful development of neural network, learning-based normal estimation achieves better performance and less dependence of parameter tuning than traditional approaches. They can be generally divided into two categories: deep surface fitting and regression-based approaches.

Deep surface fitting methods. These methods typically employ a deep neural network to predict point-wise weights and then fit a polynomial surface to input patches using WLS such as IterNet (Lensen, Osendorfer, and Masci 2020) and DeepFit (Ben-Shabat et al. 2020). Analogously, Zhang et al. (2022) adopted the predicted weights as the guiding geometric information. AdaFit (Zhu et al. 2021) proposed a novel layer to aggregate features from multiple global scales and then predicted point-wise offset to improve the normal estimation accuracy. To learn richer geometric features, GraphFit (Li et al. 2022a) combined graph convolutional layers with adaptive modules, while Du et al. (2023) analyzed the approximation error of these methods and suggested two fundamental design principles to further improve the estimation accuracy. However, due to the constant order of the objective polynomial functions, deep surface fitting methods typically suffer from overfitting and underfitting.
Regression-based methods. This type casts the normal estimation problem as a regression process and predicts the point cloud normals via the network straightforward. For instance, HoughCNN (Boulch and Marlet 2016) transformed point clouds into a Hough space and then utilized Convolutional Neural Networks (CNN) to directly infer normal vectors, whereas Lu et al. (2020a) projected point clouds into a height map by computing distances between scatter points and the fitted plane. However, these approaches sacrifice the 3D geometry unavoidably when executing in 2D spaces. PCPNet (Guerrero et al. 2018) directly adopted the unstructured point clouds as input and then used the PointNet (Qi et al. 2017a) to capture multi-scale features instead. Hashimoto et al. (2019) combined PointNet with 3D-CNN to extract local and spatial features, and NestiNet (Ben-Shabat et al. 2019) employed mixture-of-experts framework to determine the optimal normal estimation scale. To provide more information of the input patch, Refine-Net (Zhou et al. 2022a) additionally calculated the initial normals and the height map. Recent work involve HSurf-Net (Li et al. 2022b) and SHS-Net (Li et al. 2023a) first transformed point clouds into a hyper space through local and global feature extractions and then performed plane fitting in the constructed space. NeAF (Li et al. 2023b) inferred an angle field around the ground truth normal to make it learn more information of the input patch. Benefiting from the strong feature extraction abilities of the network architectures, recent regression-induced approaches demonstrate promising results on clean point clouds. However, they have yet made significant progress in normal estimation on noisy point clouds, which are often emerged in practical scenarios.

Aiming at improving the robustness to noise, we identify a crucial inconsistency between the annotated normal and the neighborhood geometry of the noisy point and introduce CND to address this problem. Besides, compared with the recent regression methods, we propose a network that combines various geometric information extraction with a hierarchical architecture to make the complex information capture more effectively.

Rethinking Noisy Normal Estimation

Direction Inconsistency

Previous learning-based approaches directly minimize the deviations between the predicted normals and the annotated ones for training and evaluation. This is reasonable for noise-free scenarios, however, for the noisy point clouds, due to the noise-caused relative coordinate changes, the annotated normals indeed are inconsistent with the neighborhood geometry of the query points. As presented in Fig. 2(a), given a set of noisy points cloud $P$, suppose the ground truth position locating on the surface of the noisy point $p_i$ is $\hat{p}_i$. The annotated normal of $p_i$ is $n_{p_i} \in \mathbb{R}^3$, which is the same as the one of the point before adding noise, and the normal of $\hat{p}_i$ is $n_{\hat{p}_i} \in \mathbb{R}^3$. If we optimize the typically defined normal estimation loss $\|n_{p_i} - \hat{n}_{p_i}\|_2^2$ as predecessors, where $\hat{n}_{p_i}$ is the predicted normal, this will unavoidably lead to inconsistency between the annotated normal $n_{p_i}$ and the input patch $P_i$. What’s worse, this inconsistency greatly decreases the quality of the training data and thus lowers down the estimation ability of the network on noisy point clouds.

Moreover, this inconsistency also degrades downstream tasks such as denoising and 3D reconstruction. For instance, Fig. 2(c) shows the denosing principle for point clouds. If we utilize the predicted normal vector $\hat{n}_{p_i}$, which closely resembles the annotated normal vector $n_{p_i}$ (indicating a highly accurate estimation), then the introduced offset $d_{p_i}$ will not align or bring $p_i$ closer to the noise-free underlying surface. Anonymously, in the context of reconstruction tasks, as shown in Fig. 2(d), the regenerated mesh face $\hat{F}_i$ in relation to the normal vector $\hat{n}_{p_i}$ significantly deviates from the authentic mesh fact $F_i$.

Scale Ambiguity

Another challenge in current normal estimation approaches is the ambiguity regarding the optimal scale in both local and global feature extraction. Concerning local structures, using large scales typically improves robustness against noise but can lead to oversmoothing of shape details and sharp features. Conversely, small scales can preserve geometric details but are relatively sensitive to noise. When it comes to global features, large scales include more structure information from the underlying surface but may also incorporate irrelevant points, thus degrading the geometry information of the input patch. On the other hand, small scales reduce irrelevant points but are less robust to noise. Previous works have struggled to effectively extract and combine multi-scale local and global features, making them highly dependent on scale selection and resulting in unsatisfactory performance on both noisy point clouds and complex shape details.

Proposed Method

To solve the aforementioned issues, we propose a novel normal estimation approach that is robust against noise and less sensitive to scale selection. Concrete technical contributions are presented in the following.
Chamfer Normal Distance
To bridge the direction inconsistency between the annotated normal and the predicted one of the input patch, instead of using the conventional metric $\| \hat{n}_{p_i} - n_{p_i} \|^2_2$, inspired from the Chamfer Distance (CD)

$$C(\mathcal{P}, \hat{\mathcal{P}}) = \frac{1}{N_1} \sum_{p_i \in \mathcal{P}} \min_{\hat{p}_i \in \hat{\mathcal{P}}} (\| p_i - \hat{p}_i \|^2_2) + \frac{1}{N_2} \sum_{\hat{p}_j \in \hat{\mathcal{P}}} \min_{p_i \in \mathcal{P}} (\| p_i - \hat{p}_j \|^2_2),$$

(1)

where $N_1$ and $N_2$ represent the cardinalities of the point cloud $\mathcal{P}$ and $\hat{\mathcal{P}}$, we formulate the Chamfer Normal Distance (CND) as

$$CND(\mathcal{P}, \hat{\mathcal{P}}) = \frac{1}{N} \sum_{i=1}^{N} \arccos \langle n_{\hat{p}_i}, \hat{n}_{p_i} \rangle,$$

(2)

where $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors and $\hat{p}_i$ is the closest point of $p_i$ in the noise-free point cloud $\hat{\mathcal{P}}$. In contrast to previous approaches that relied on annotated normal correspondence, our proposed CND manner assures consistency with the underlying geometric structure of the input patch (Fig. 2(b)). The CND metric not only faithfully captures the prediction errors in noisy point clouds, but also eliminates the direction inconsistency during network training, thus substantially improving the network robustness and facilitating the subsequent assignments.

CMG-Net
To capture more fruitful multi-scale structure information and solve the scale ambiguity issue simultaneously, we develop a network combining various geometric information extraction with a hierarchical architecture termed CMG-Net. Given a patch $\mathcal{P} = \{p_i \in \mathbb{R}^3\}_{i=1}^{N}$ centralized at a query point $p$, as shown in Fig. 3(a), CMG-Net first normalizes the input points and rotates $\mathcal{P}$ by PCA and QSTN (Qi et al. 2017a; Du et al. 2023) to initialize the normal vectors. Then, we group the local features by $k$-nearest neighbors ($k$-NN) with different scales and aggregate them together. Besides, we design a hierarchical structure with intricate geometry information fusion, followed by the decoding of the embedding features. Our loss function modified by CND enables the network jumping out of the annotation inconsistency.

Multi-scale Local Feature Aggregation. Previous methods group the local features by $k$-NN and capture the geometric information by MLP and maxpooling (Li et al. 2022b). However, this manner often suffers from scale ambiguity and results in unsatisfactory robustness against noise. To solve this issue, as presented in Fig. 3(b), we construct graphs by $k$-NN with small and large scales and employ the skip-connection and maxpooling to capture the local structures. The Local Feature Extraction (LFE) can be formulated as

$$f^{n+1}_{i,k} = \text{MaxPool} \left( \varphi_1 \left( f^n_{i,1}, \varphi_1 \left( f^n_{i,2} \right) \right) \right),$$

(3)

where $f^n_{i,k}$ is the neighbor feature of the feature $f^n_i$, $\varphi_1$ is the MLP layer, $\varphi_1$ is the skip-connection layer, and $s_l$ represents the scale of $k$-NN with $l = 1, 2$ in default. Moreover, we use an Attentional Feature Fusion (AFF) architecture to aggregate the features which can benefit both the small and large scales. The AFF can be formulated as

$$M \left( f^{i_1}_i, f^{i_2}_i \right) = \text{sigmoid} \left( \varphi_2 \left( \text{AvgPool} \left( f^{i_1}_i + f^{i_2}_i \right) \right) \right),$$

(4)

$$f_i = \varphi_3 \left( f^{i_1}_i \cdot M \left( f^{i_1}_i, f^{i_2}_i \right) + f^{i_2}_i \cdot (1 - M \left( f^{i_1}_i, f^{i_2}_i \right)) \right),$$

(5)

where $f^{i_1}_i$ and $f^{i_2}_i$ are the local structures with different scales of feature $f_i$, $\varphi_2$ and $\varphi_3$ are the MLP layers, $N$ represents the cardinality of the input point cloud patch.

Hierarchical Geometric Information Fusion. Recent approaches have proven the effectiveness of multi-scale global feature extraction (Qi et al. 2017b; Li et al. 2022b; Qin et al. 2022), however, large scale global information and local structures may be lost after point cloud downsampling. To alleviate this problem, as shown in Fig. 3(c), we propose a hierarchical architecture that combines the multi-scale global features with the local structures. During the
Hierarchical Geometric Information Fusion, the global feature $G_{N_h}$ of current scale $N_h$ can be formulated as

$$G_{N_h} = \varphi_5 \left( \text{MaxPool} \left\{ \varphi_4 \left( f_{i,N_h}^h \right) \right\}_{i=1}^{N_h} \right),$$

where $\varphi_4$ and $\varphi_5$ are MLP layers. Meanwhile, the local structures $g_{i,N_{h+1}}$ are captured by

$$g_{i,N_{h+1}} = \text{MaxPool} \left\{ \varphi_6 \left( s_{i,j,s}^N \right) \right\}_{j=1}^s + g_{i,N_{h}}, i = 1, \ldots, N_{h+1},$$

where $g_{i,N_{h}}$ is the neighborhood feature of point $p_i$ in the scope of the scale $N_{h+1}$, $s$ is the scale of the neighborhood features, and $\varphi_6$ represents the MLP layer. Then, we downsample the patch by decreasing the patch size. Moreover, we integrate the global features of the current scale and the last scale with the local structures by

$$f_{i,N_{h+1}} = \varphi_7 \left( G_{N_h}, G_{N_h-1}, g_{i,N_{h+1}} \right) + f_{i,N_{h}}, i = 1, \ldots, N_{h+1},$$

where $\varphi_7$ is the MLP layer, and $N_{h+1} \leq N_h \leq N_{h-1}$.

Decoder. Note that the point coordinates are important basic attributes for point cloud processing and the spatial relationship between them such as distance can guide the inference process of the network (Zhao et al. 2021; Zhang et al. 2022). To explore this idea, we introduce two modules including Position Feature Fusion (PFF) and Weighted Normal Prediction (WNP) into the decoder part. As shown in Fig. 3(d), during the PFF, we embed the neighborhood coordinates of each point and fuse them with the extracted feature by skip-connections, which can be formulated as

$$F_i = \text{MaxPool} \left\{ \varphi_8 \left( f_i - p_{i,j} - p_i, s \left( p_{i,j} - p_i \right) \right) \right\}_{j=1}^s,$$

where $p_{i,j}$ is the neighbor coordinate of the point $p_i$, $f_i$ is the extracted feature of $p_i$, $s$ represents the neighborhood scale, $\varphi_8$ is the MLP layer and $\varphi_8$ is the skip-connection. As shown in Fig. 3(e), we predict weights based on the geometry information of each point and use the weighted features to predict the normal vector of the query point:

$$n = \varphi_{11} \left( \text{MaxPool} \left\{ \varphi_{10} \left( F_i \cdot \text{softmax}_{M} \left( \varphi_9 \left( F_i \right) \right) \right) \right\}_{i=1}^M \right),$$

where $\varphi_9$, $\varphi_{10}$ and $\varphi_{11}$ are the MLP layers, and the normalized $n$ is the finally predicted unit normal vector.

**Table 1:** Quantitative comparisons in terms of RMSE and CND on the PCPNet dataset. Bold values indicate the best estimator.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>CND</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noise (σ)</td>
<td>Density</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>0.12%</td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n-jet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCPNet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nesti-Net</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepFit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaFit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GraphFit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSurf-Net</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Du et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHS-Net</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Loss function.** To bridge the gap between the annotated normal and the noise-caused neighborhood geometry variation of the query point, we reformulate the sine loss by CND, namely, taking the normal $n_p$ of the nearest neighbor point $\tilde{p}$ in the corresponding noise-free point cloud $\tilde{P}$ as the ground truth

$$L_1 = \| n_{\tilde{p}} \times \hat{n}_{\tilde{p}} \|.$$

Meanwhile, we use the transformation regularization loss and the z-direction transformation loss to constrain the output rotation matrix $R \in \mathbb{R}^{3 \times 3}$ of the QSTN (Du et al. 2023)

$$L_2 = \| I - RR^T \|_2,$$

$$L_3 = \| n_{\tilde{p}} R \times z \|,$$

where $I \in \mathbb{R}^{3 \times 3}$ represents the identity matrix, $z = (0, 0, 1)$. Additionally, to make full use of the spatial relationships between data points, we adopt the weight loss similar to Zhang et al. (2022)

$$L_4 = \frac{1}{M} \sum_{i=1}^{M} (\hat{w}_i - \hat{w}_i)^2,$$

where $\hat{w}_i$ are the predicted weights for each data point, $M$ represents the cardinality of the downsampled patch, $w_i = \exp(-p_i \cdot n_{\tilde{p}}^2 / \delta^2)$ and $\delta = \max \left( 0.05^2, 0.3 \sum_{i=1}^{M} (p_i \cdot n_{\tilde{p}}^2) / M \right)$, where $p_i$ is the point in the downsampled patch. Therefore, our final loss function is defined as

$$L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4,$$

where $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.5$, and $\lambda_4 = 1$ are weighting factors.

**Experimental Results**

**Datasets.** As predecessor approaches, we first adopt the synthetic dataset PCPNet (Guerrero et al. 2018) for comparison, in which we follow the same experimental setups including train-test split, adding noise, and changing distribution density on test data. To test the generalization capability of our method, we then evaluate and compare the models trained on the PCPNet on the real-world indoor SceneNN dataset (Hua, Tran, and Yeung 2018).
Table 2: Quantitative comparisons of CND on the PCPNet dataset with gradually increased noise.

<table>
<thead>
<tr>
<th>Method</th>
<th>Noise (σ)</th>
<th>0.125%</th>
<th>0.25%</th>
<th>0.5%</th>
<th>0.75%</th>
<th>1.25%</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td></td>
<td>14.46</td>
<td>15.09</td>
<td>17.75</td>
<td>21.40</td>
<td>31.81</td>
<td>20.10</td>
</tr>
<tr>
<td>n-jet</td>
<td></td>
<td>14.40</td>
<td>14.98</td>
<td>17.67</td>
<td>21.50</td>
<td>32.08</td>
<td>20.12</td>
</tr>
<tr>
<td>PCPNet</td>
<td></td>
<td>13.42</td>
<td>15.52</td>
<td>18.12</td>
<td>20.02</td>
<td>23.92</td>
<td>18.20</td>
</tr>
<tr>
<td>AdaFit</td>
<td></td>
<td>11.70</td>
<td>12.83</td>
<td>15.62</td>
<td>17.64</td>
<td>24.01</td>
<td>16.36</td>
</tr>
<tr>
<td>DeepFit</td>
<td></td>
<td>11.42</td>
<td>12.95</td>
<td>15.52</td>
<td>17.02</td>
<td>21.74</td>
<td>15.73</td>
</tr>
<tr>
<td>GraphFit</td>
<td></td>
<td>11.01</td>
<td>12.40</td>
<td>15.05</td>
<td>16.66</td>
<td>20.56</td>
<td>15.14</td>
</tr>
<tr>
<td>HSurf-Net</td>
<td></td>
<td>11.04</td>
<td>12.67</td>
<td>15.33</td>
<td>16.79</td>
<td>20.67</td>
<td>15.30</td>
</tr>
<tr>
<td>Du et al.</td>
<td></td>
<td>10.97</td>
<td>12.48</td>
<td>15.17</td>
<td>16.77</td>
<td>21.04</td>
<td>15.29</td>
</tr>
<tr>
<td>SHS-Net</td>
<td></td>
<td>10.90</td>
<td>12.66</td>
<td>15.18</td>
<td>16.59</td>
<td>20.89</td>
<td>15.24</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>10.60</td>
<td>12.56</td>
<td>14.89</td>
<td>16.31</td>
<td>19.62</td>
<td>14.79</td>
</tr>
</tbody>
</table>

Figure 4: Qualitative comparisons on the PCPNet datasets. We use the heat map to visualize the CND error.

Implementation details. We set the input patch size \( N = 700 \) and the downsampling factors \( \rho = \{2/3, 2/3, 2/3, 1\} \). The scales of \( k\)-NN in the LFE are equivalent to 16 and 32, and \( s = \{32, 32, 16, 16\} \) in the Hierarchical Geometric Information Fusion. The number of the neighbor points during the PPF is 16. We adopt the AdamW (Loshchilov and Hutter 2017) optimizer with initial learning rate \( 5 \times 10^{-4} \) for training. The learning rate is decayed by a cosine function. Our model is trained with a 64 batch size on an NVIDIA A100 GPU in 900 epochs. More implementation details are reported in Supplementary Materials (SM).

Evaluation. The commonly used traditional methods, i.e., PCA (Hoppe et al. 1992) and n-jet (Cazals and Pouget 2005) and the latest learning-based methods, i.e., PCPNet (Guerrero et al. 2018), Nesti-Net (Ben-Shabat et al. 2019), DeepFit (Ben-Shabat et al. 2020), AdaFit (Zhu et al. 2021), GraphFit (Li et al. 2022a), HSurf-Net (Li et al. 2022b), Du et al. (2023) and SHS-Net (Li et al. 2023a) are taken as baselines. To make thorough comparison, we adopt the proposed CND metric to assess the normal estimation results and compare it with the RMSE. Moreover, the error distribution analysis can be found in SM.

Results on Synthetic Data

PCPNet. Table 1 reports the statistical results of all compared approaches on the PCPNet dataset, measured in terms of both RMSE and CND metrics. As observed, our method achieves the overall highest normal estimation accuracy across different scenarios, particularly in scenarios with noise. In comparison to RMSE, the CND metric allows for more accurate and faithful prediction evaluations while mitigating the annotation inconsistency. Qualitative comparison results are presented in Fig. 4. Notably, our method exhibits the smallest errors in regions characterized by noise and intricate geometry.

Robustness to noise. Subsequently, we specifically employ five representative models from the PCPNet dataset to assess the robustness against noise. We introduce varying levels of noise to these data which encompass one CAD model and four scanned point clouds. The quantitative outcomes displayed in Table 2 indicate that our method exhibits superior performance compared to competitors, particularly in scenarios contaminated by high levels of noise.

Generalization to Real-world Data

Next, we investigate the generalization capability using the real-world indoor SceneNN dataset. Results in Table 3 suggest that our method has the highest normal estimation accuracy in an average sense. The qualitative results presented in Fig. 5 exhibit our superiority. It is noticeable that our method successfully preserves more geometric details, such as the handle of the refrigerators. Additionally, more results on different real-word datasets can be found in SM.

Ablation Study

Network architecture. CMG-Net comprises three key components: Multi-scale Local Feature Aggregation, Hierarchical Geometric Information Fusion, and Decoder. We delve into the functions of them on the PCPNet dataset. (1). In the Multi-scale Local Feature Aggregation, we capture the local structure using two scales and integrate them by AFF. Table 4(a) reports the results of 1) without LFE; 2) with single-scale LFE, and 3) integrating multi-scale local features directly by MLP instead of AFF. As observed, compared with Ours, the multi-scale local features with AFF can effectively improve the network performance.

Table 3: Statistical CND results on the SceneNN dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Noise (σ)</th>
<th>0.125%</th>
<th>0.25%</th>
<th>0.5%</th>
<th>0.75%</th>
<th>1.25%</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td></td>
<td>14.46</td>
<td>15.09</td>
<td>17.75</td>
<td>21.40</td>
<td>31.81</td>
<td>20.10</td>
</tr>
<tr>
<td>n-jet</td>
<td></td>
<td>14.40</td>
<td>14.98</td>
<td>17.67</td>
<td>21.50</td>
<td>32.08</td>
<td>20.12</td>
</tr>
<tr>
<td>PCPNet</td>
<td></td>
<td>13.42</td>
<td>15.52</td>
<td>18.12</td>
<td>20.02</td>
<td>23.92</td>
<td>18.20</td>
</tr>
<tr>
<td>AdaFit</td>
<td></td>
<td>11.70</td>
<td>12.83</td>
<td>15.62</td>
<td>17.64</td>
<td>24.01</td>
<td>16.36</td>
</tr>
<tr>
<td>DeepFit</td>
<td></td>
<td>11.42</td>
<td>12.95</td>
<td>15.52</td>
<td>17.02</td>
<td>21.74</td>
<td>15.73</td>
</tr>
<tr>
<td>GraphFit</td>
<td></td>
<td>11.01</td>
<td>12.40</td>
<td>15.05</td>
<td>16.66</td>
<td>20.56</td>
<td>15.14</td>
</tr>
<tr>
<td>HSurf-Net</td>
<td></td>
<td>11.04</td>
<td>12.67</td>
<td>15.33</td>
<td>16.79</td>
<td>20.67</td>
<td>15.30</td>
</tr>
<tr>
<td>Du et al.</td>
<td></td>
<td>10.97</td>
<td>12.48</td>
<td>15.17</td>
<td>16.77</td>
<td>21.04</td>
<td>15.29</td>
</tr>
<tr>
<td>SHS-Net</td>
<td></td>
<td>10.90</td>
<td>12.66</td>
<td>15.18</td>
<td>16.59</td>
<td>20.89</td>
<td>15.24</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>10.60</td>
<td>12.56</td>
<td>14.89</td>
<td>16.31</td>
<td>19.62</td>
<td>14.79</td>
</tr>
</tbody>
</table>

Figure 5: Qualitative comparisons on the SceneNN datasets.
(2) To validate the effectiveness of the Hierarchical Geometric Information Fusion, we carry out experiments using the model with a fixed global scale that is equivalent to the output scale of CMG-Net. Additionally, we compare the results of the models without the global feature of the last scale or the local feature in the hierarchical architecture. Results shown in Table 4(b) demonstrate that the Hierarchical Geometric Information Fusion operation can also boost the normal estimation performance.

(3) Table 4(c) shows the ablation studies of the Decoder part, suggesting the effectiveness of PFF and WNP. Besides, more ablation results on QSTN, the input patch sizes $N$, and the downsampling factors $\rho$ can be found in SM.

**CND-modified loss function.** To demonstrate the effectiveness and generalization of the newly introduced CND-modified loss function, we conduct experiments on the PCPNet dataset, comparing the results with and without its incorporation. We employ representative methods, including the deep surface fitting method DeepFit (Ben-Shabat et al. 2020), as well as the regression methods Hsurf-Net (Li et al. 2022b), and Ours. Table 5 highlights the impact of the CND component, demonstrating its significant enhancement in normal estimation accuracy for both deep surface fitting and regression methods.

### Application of the Proposed Method

We also demonstrate the application of our method on downstream tasks. Fig. 6 presents the Poisson surface reconstruction (Kazhdan, Bolitho, and Hoppe 2006) results using the normal vectors predicted by competing approaches. Compared with ground-truth surfaces, our method achieves the best reconstruction quality (quantified by the Symmetric Mean Hausdorff Distance (SMD) $(10^{-4})$), especially in shape details of noisy regions, underscoring the higher accuracy of our normal estimation. We provide more reconstruction instances and highlight the application of our newly developed method to point cloud denoising in the SM.

### Limitations

While our method has demonstrated remarkable normal estimation accuracy across diverse 3D models, it is not yet real-time capable and still depends on annotated training data, as is the case with previous approaches. Therefore, it is highly desirable in the future to reduce the computation time and delve into unsupervised frameworks.

### Conclusions

We propose a novel method for robust normal estimation in unorganized point clouds, which shows superiority across various datasets and scenarios. We identify the issue of direction inconsistency in predecessor approaches and introduce the CND metric to address this concern. This not only boosts the network training and evaluation, but also greatly enhances the network robustness against noisy disturbance. Additionally, we design an innovative architecture that combines multi-scale local and global feature extraction with hierarchical information fusion to deal with scale selection ambiguity. Extensive experiments validate that our method outperforms competitors in terms of both accuracy and robustness for normal estimation. Moreover, we demonstrate its ability to generalize in real-world settings and downstream application tasks.
Acknowledgements
This work was partially funded by the National Natural Science Foundation of China (62172415, 12022117, 62102418), the CAS Project for Young Scientists in Basic Research (YSBR-034), and the Beijing Science and Technology Plan Project (Z231100005923033).

References
Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; and Xu, K. 2022. Geometric transformer for fast and robust point


