3D Point Cloud Registration Based on Cascaded Mutual Information Attention Network

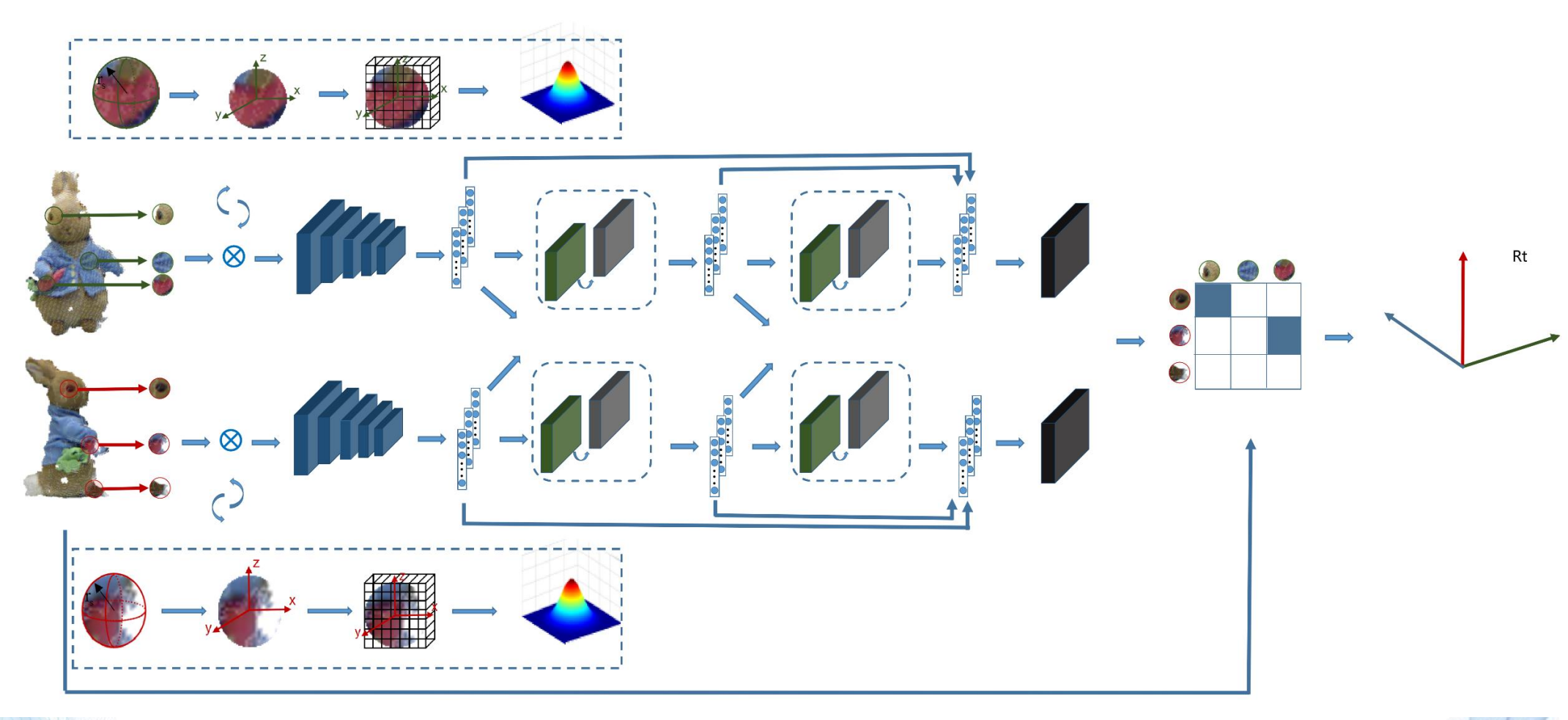
Pan, Xiang; Ji, Xiaoyi
Zhejiang University of Technology

CONTENTS

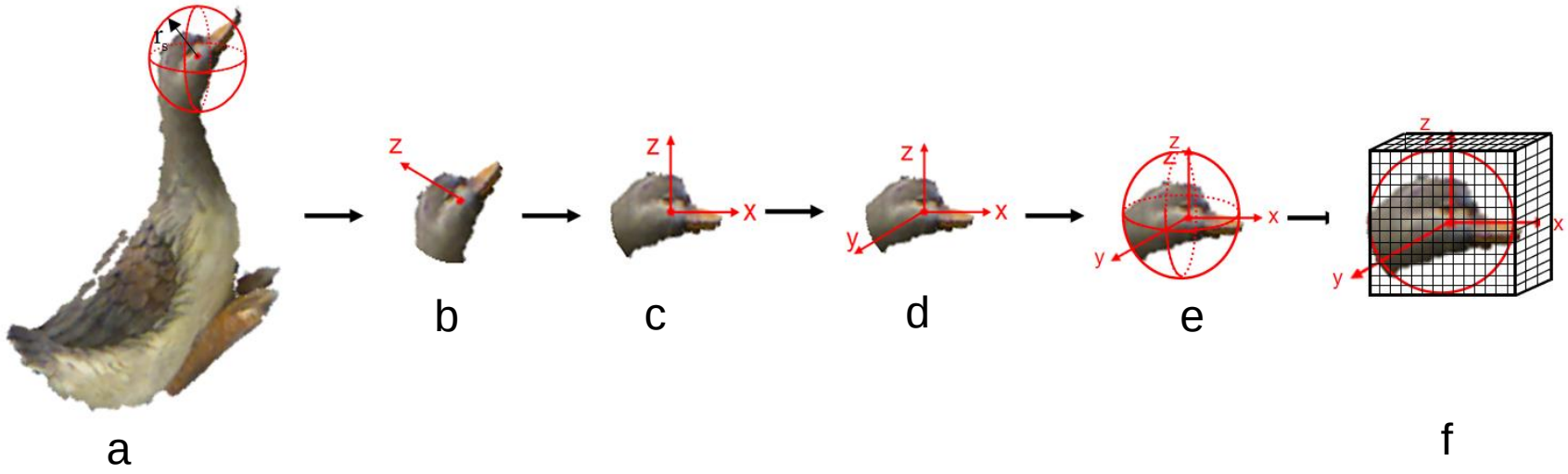
1 Contributions

- For the local point set of the 3D model, we use the spherical representation to define the local reference coordinate system, which improves the stability and reliability of the local features under the influence of rotation, transformation and noise.
- We adopts the attention response structure to improve the network depth through the cascade structure and ensure the convergence of the network.
- In order to avoid feature loss in the cascading process, our network introduces a lateral connection, and the original convolution feature is introduced as an offset to each cascade layer.

2 Network Algorithm-Overall Framework Diagram



2 Network Algorithm-Calculation of Local Coordinate System



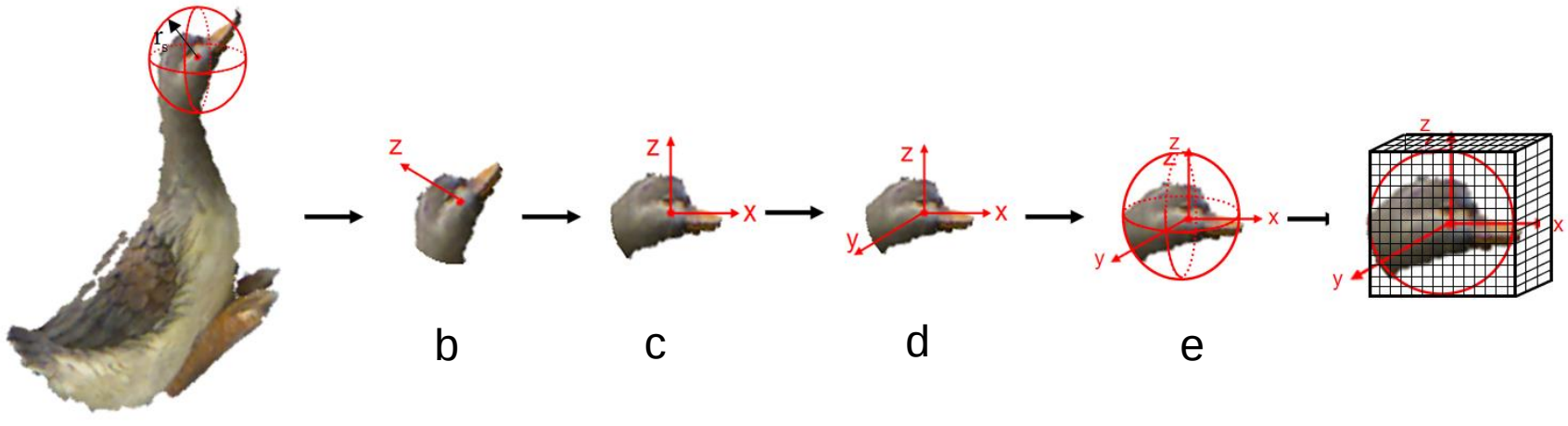
1 Randomly sample the point cloud model $P=\{p_1,...,p_i,...,p_n\}$, suppose there is a sampling point p_i , take the sampling point as the center of the sphere, and surround it with a radius r_{LRF} to obtain a sphere S_{p_i}

2 Use the points within the sphere to solve the characteristic equation to obtain the coordinates of the z-axis vector

3 Use the projection of the sampling point to the tangent plane to calculate the x-axis coordinates

4 By expressing the vector product of the z-axis and x-axis vectors, the vector representation of the y-axis can be obtained

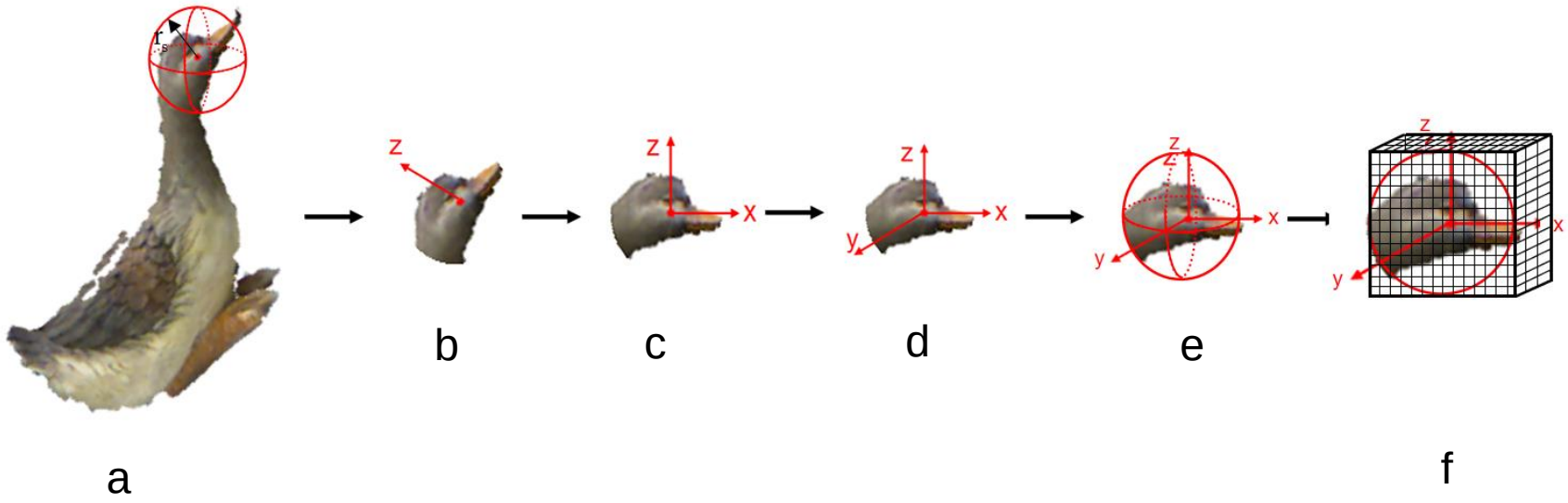
2 Network Algorithm-Calculation of Local Coordinate System



5 Construct a third-order feature description sub-matrix $M \in R^{L\times W\times H}$ for each sampling point p_i to specifically describe the spatial information and geometric characteristics near the sampling point. Each m_{lwh} is calculated by a one-dimensional Gaussian smoothing filter with a bandwidth of δ :

$$m_{lwh} = \frac{1}{n_{jkl}} \sum_{i=1}^{n_{jkl}} \frac{1}{\sqrt{2\pi}\delta} \exp \frac{-\left\| q_{lwh} - p_i' \right\|_2^2}{2\delta^2}$$

2 Network Algorithm-Calculation of Local Coordinate System



Advantage:

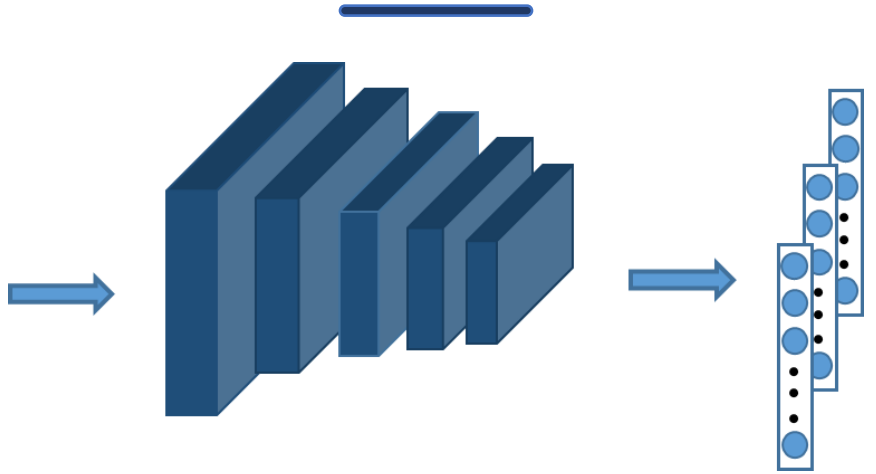
1) It can ensure that the described feature information has rotational symmetry

2) It can ensure that the feature information in any direction will not be missed in the subsequent feature calculation process

3) By using the weighted average of the neighbors to replace the feature representation of the sampling points, and the closer the neighbors have higher weights

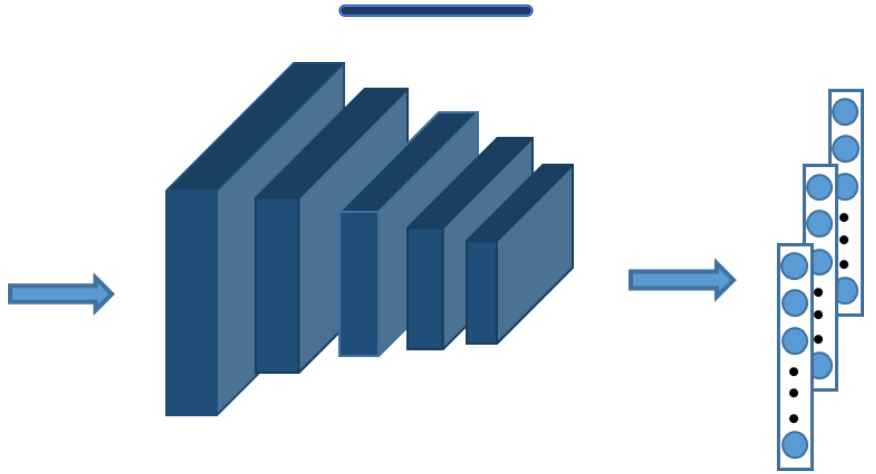
4) Partial information can produce more accurate matches in a lower query time.

2 Network Algorithm-Compute Feature Descriptor



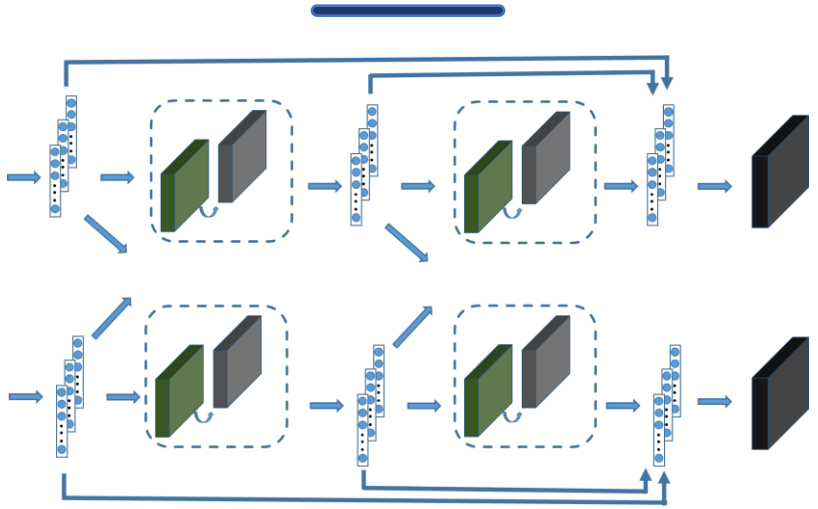
Composed of N = 7 identical layers, each layer adopts down-sampling through a full convolutional network structure with a step size of 2, followed by the batch normalization layer (BN) after the convolutional layer, and uses ReLU activation function. During the experiment, the affine parameters of the batch normalization layer are fixed to 1 and 0, and they are not trained in the training of the network.

2 Network Algorithm-Compute Feature Descriptor



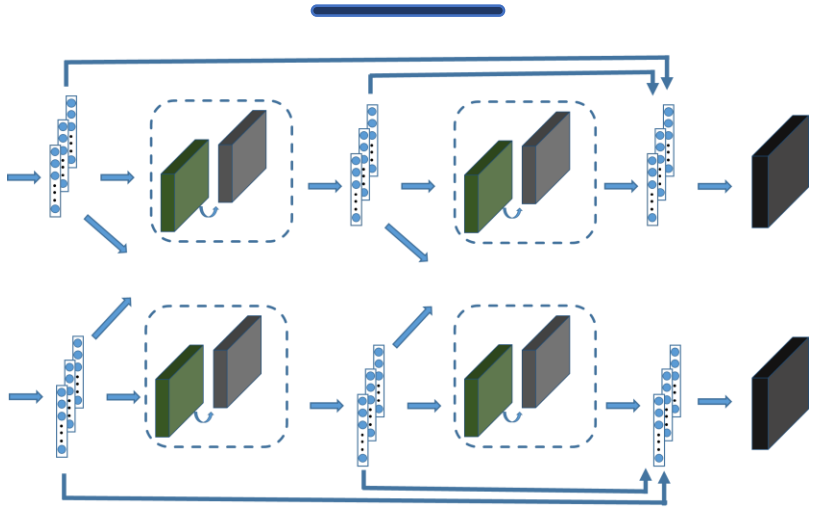
At the same time, in order to avoid the phenomenon of network overfitting. In order to avoid network overfitting, the convolutional layer with N=7 is slightly different from the product layer between them. We add a discarding rule with a discarding rate of 0.3 before the last convolutional layer. Finally, l2 normalization is used to generate local feature descriptors per unit length.

2 Network Algorithm-Cascaded Mutual Information Attention Structure



The network improves the accuracy of point cloud registration by stacking residual structures and adopting lateral connections. The attention response structure is adopted to increase the depth of the network through the cascade structure and to ensure the convergence of the network. At the same time, in order to avoid feature loss in the cascading process, the network introduces a lateral connection and introduces the original convolution feature as an offset to each cascade layer.

2 Network Algorithm-Cascaded Mutual Information Attention Structure



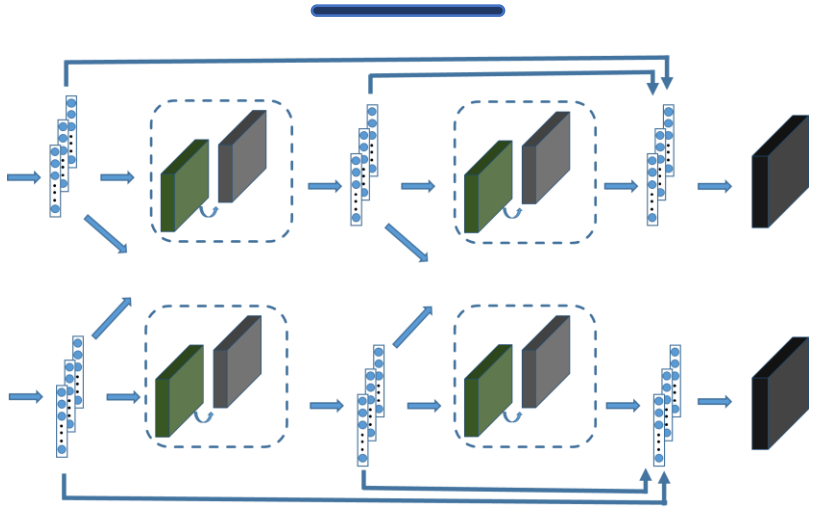
Then use F(x) and F(y) as the input of the cascaded mutual information attention structure, and implement error approximation. For the cascaded mutual information network structure, the following excitation functions are used: local features F(x) and F(y) Close to each other.

$$\Phi^1(x) = F(x) + attention(F(x), F(y)).$$

$$\Phi^1(y) = F(y) + attention(F(y), F(x)).$$

where attention() represents the dynamic characteristics obtained through the attention structure.

2 Network Algorithm-Cascaded Mutual Information Attention Structure



In order to further capture the correlation between local features and avoid problems such as the disappearance of the gradient of the convolutional network, the paper uses progressive excitation in the network structure, and uses the feature pooling layer between each layer to achieve feature fusion. It is defined as follows:

$$\Phi^2(x) = \Phi^1(x) + attention(\Phi^1(x), \Phi^1(y)).$$

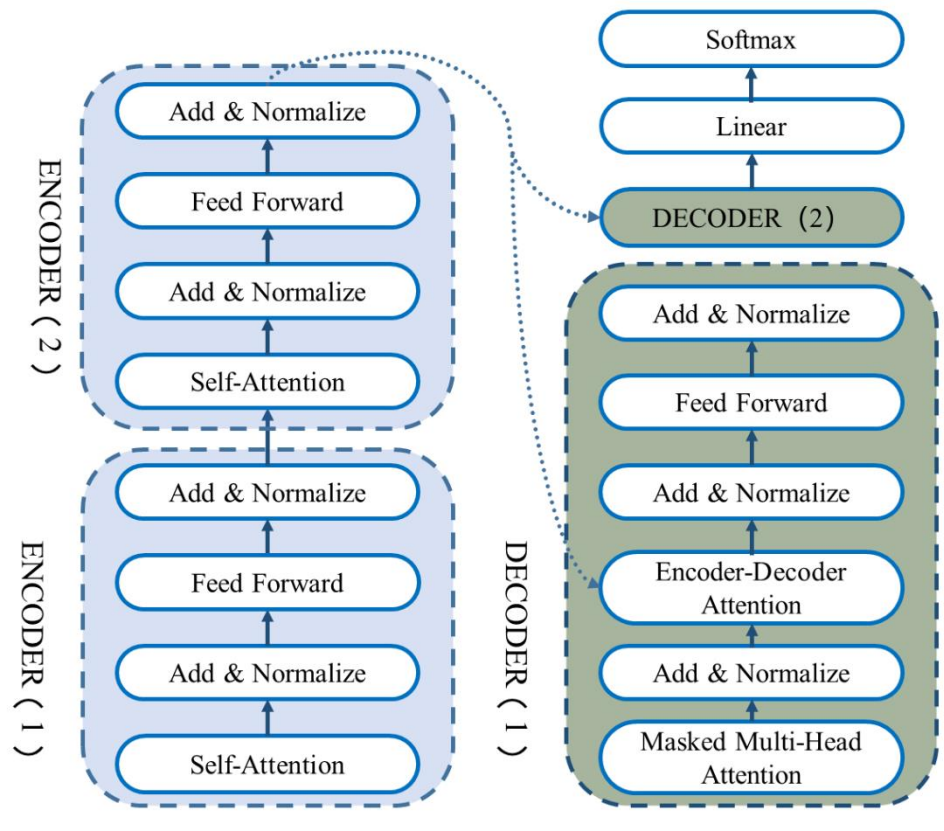
$$\Phi^2(y) = \Phi^1(y) + attention(\Phi^1(y), \Phi^1(x)).$$

$$\phi(x) = F(x) + \Phi^1(x) + \Phi^2(x).$$

$$\phi(y) = F(y) + \Phi^1(y) + \Phi^2(y).$$

where, $\phi()$ represents the final output feature in the cascaded mutual information attention structure.

2 Network Algorithm- Attention Structure



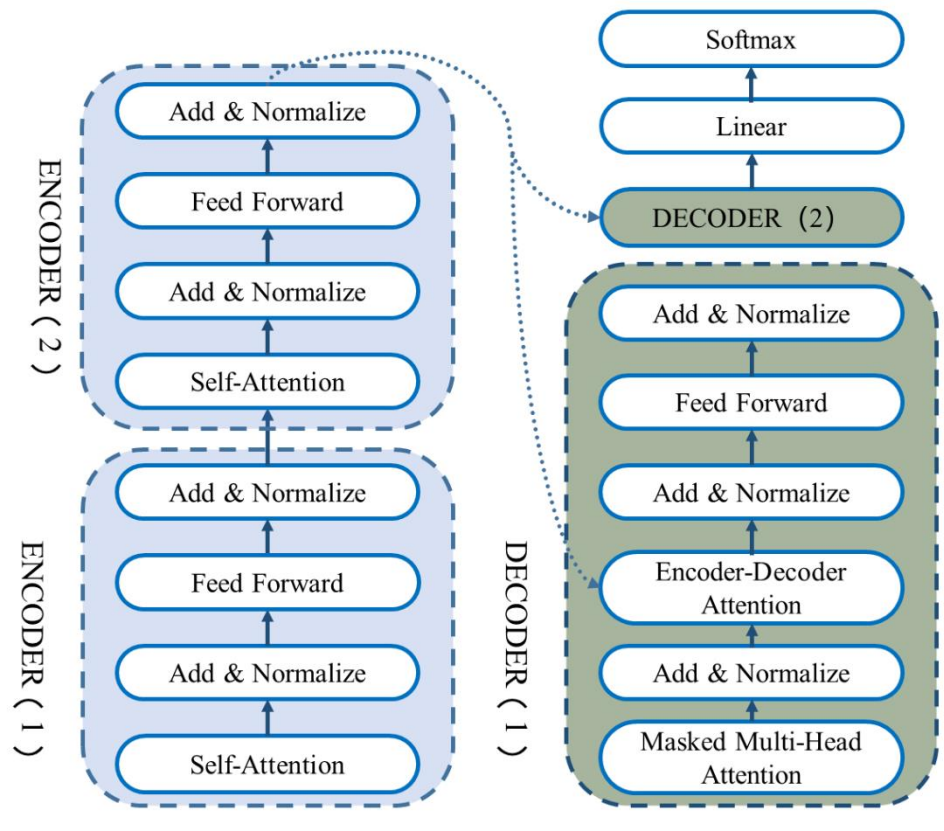
Attention Structure:

In order to better obtain the matching results, the feature information of the other party is incorporated between the respective features. Our attention model will learn a mapping function $\phi : R^{N\times P} \times R^{N\times P} \to R^{N\times P}$ to complete the fusion of feature information. After the fusion feature information is obtained, the result of this part is added as a residual item to the original feature to obtain new feature information. Among the features obtained in this way, the original information is retained, and the constraints of the matching relationship are additionally added to make the features more descriptive. The function of this mapping function is realized by a converter.

$$\sigma_x = F_x + \phi(F_x, F_y)$$

$$\sigma_y = F_y + \phi(F_y, F_x)$$

2 Network Algorithm- Attention Structure



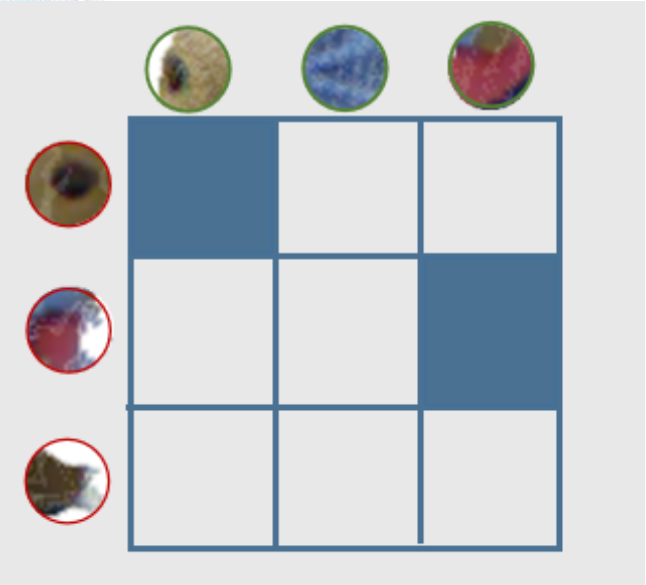
Encoder:

Encode information by using self-attention and shared multilayer perceptrons.

Decoder:

one part accepts the input from the decoder, the other part accepts the input from the encoder after encoding, uses a common attention mechanism to operate on the two, and finally outputs features with matching relationships.

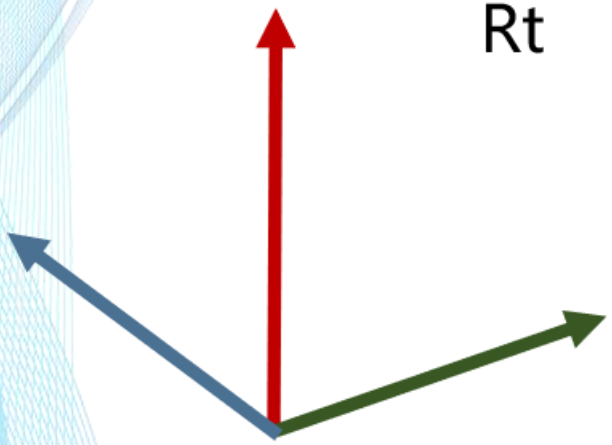
2 Network Algorithm- Attention Structure



When the gap between the local optima is relatively large, the rigid transformation estimated by the ICP method deviates significantly from the real transformation, leading to false local optima. In order to solve this problem, we proposed the use of transformer converter to learn the feature that combines the feature information of the two to solve this kind of problem. At the same time, in order to avoid a single match, we use the probability method to calculate the final matching correspondence matrix

$$m(x_i, y) = soft\max(F_y F_{x_i}^T).$$

2 Network Algorithm-Calculate Rotation and Translation



On the basis of the mapping matrix, we can calculate the rigidity change matrix.

$$\hat{y}_i = Y^T m(x_i, y) \in R^3$$

By using singular value decomposition, we can get:

$$R_{xy} = VU^T. \qquad t_{xy} = -R_{xy}\bar{x} + \bar{y}.$$

2 Network Algorithm-Loss Function

$$Loss = \frac{1}{N}\sum_{i}^{N}\left\| R_{xy}x_i + t_{xy} - y_i \right\|^2.$$

Use the mean square error to measure the effect of the rigid transformation matrix [R, t].

3 Results Visualization-data Results (1)

The Model	MSE(R)	RMSE (R)	MAE (R)	MSE (t)	RMSE (t)	MAE (t)
ICP	892.601135	29.876431	23.62611	0.086005	0.293266	0.251916
GO-ICP	192.258636	13.865736	2.914169	0.000491	0.022154	0.006219
FGR	97.002747	9.848997	1.44546	0.000182	0.013503	0.002231
PointNetLK	306.323975	17.502113	5.280545	0.000784	0.028007	0.007203
3dsmoothnet	45.897267	13.414393	5.567198	1.976948	0.000137	0.003568
DCP-v2	9.923701	3.150191	2.00721	0.000025	0.005039	0.003703
Our	**5.469285**	**2.33865**	**1.571396**	**0.000113**	**0.005024**	**0.001287**

Table 1 evaluates the experimental results of different algorithms of ModelNet40-A. It can be found that compared with traditional registration algorithms such as classic ICP and GO-ICP, whether it is DCP or paper algorithm, due to the use of convolutional network for learning, the registration accuracy is significantly improved. But compared with the DCP algorithm, the paper algorithm also has a significant improvement.

3 Results Visualization-data Results (1)

The Model	MSE(R)	RMSE (R)	MAE (R)	MSE (t)	RMSE (t)	MAE (t)
ICP	894.897339	29.914835	23.544817	0.084643	0.290935	0.248755
GO-ICP	140.477325	11.852313	2.588463	0.000659	0.025665	0.007092
FGR	87.661491	9.362772	1.99929	0.000194	0.013939	0.002839
PointNetLK	227.870331	15.095374	4.225304	0.000487	0.022065	0.005404
3dsmoothnet	5.684466	3.251960	1.415611	0.000111	0.004392	0.002219
DCP-v2	1.307329	1.143385	0.770573	0.000003	0.001786	0.001195
Our	**0.096655**	**0.816501**	**0.629797**	**0.000094**	**0.000733**	**0.000539**

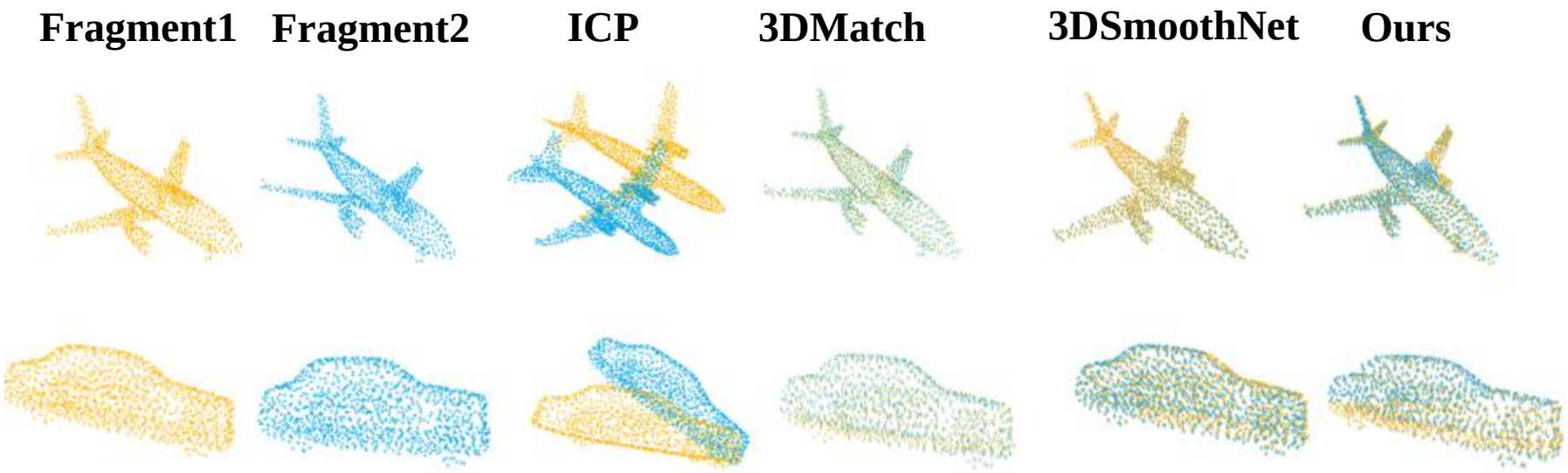
Compared with the ModelNet-A data set, the ModelNet-B data set emphasizes the generalization ability of convolutional networks. Because only some categories of 3D point cloud data participate in training. It can be found that both the DCP and the paper algorithm have a significant increase in error. But compared with the DCP algorithm, the increased error of the paper algorithm is far lower than the DCP algorithm. Therefore, it can be seen from Table 2 that the paper algorithm has better generalization than the DCP algorithm.

3 Results Visualization-data Results (2)

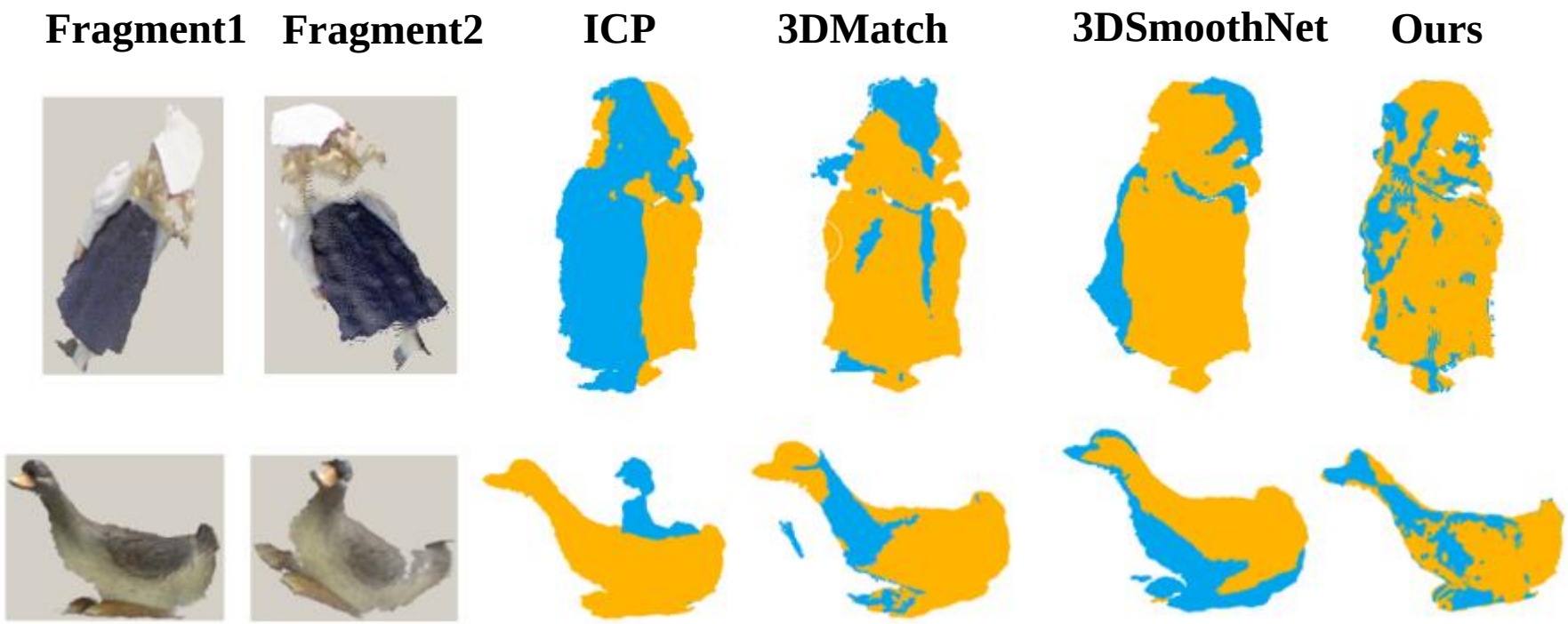
The Model	Doll	Duck	Frog	Mario	PeterRabbit	Squiller
3Dsmoothnet	3.70439	7.31273	3.91157	2.37581	5.83632	3.86516
DCP	0.226357	0.273002	0.313963	0.164783	0.206134	0.260809
Our	**0.02237**	**0.03481**	**0.03376**	**0.03233**	**0.016584**	**0.017461**

Compared with the above two data sets, the multi-view data set is more challenging. The main reason is that these data are all scanned data obtained from the real environment, with obvious noise. Therefore, this data set puts forward higher requirements on the separability of local features. Table 3 shows the experimental results of different algorithms. It can be found that the error of the paper algorithm is significantly lower than that of the DCP network. Compared with the DCP network, the registration accuracy of the paper algorithm has increased by 88.7%

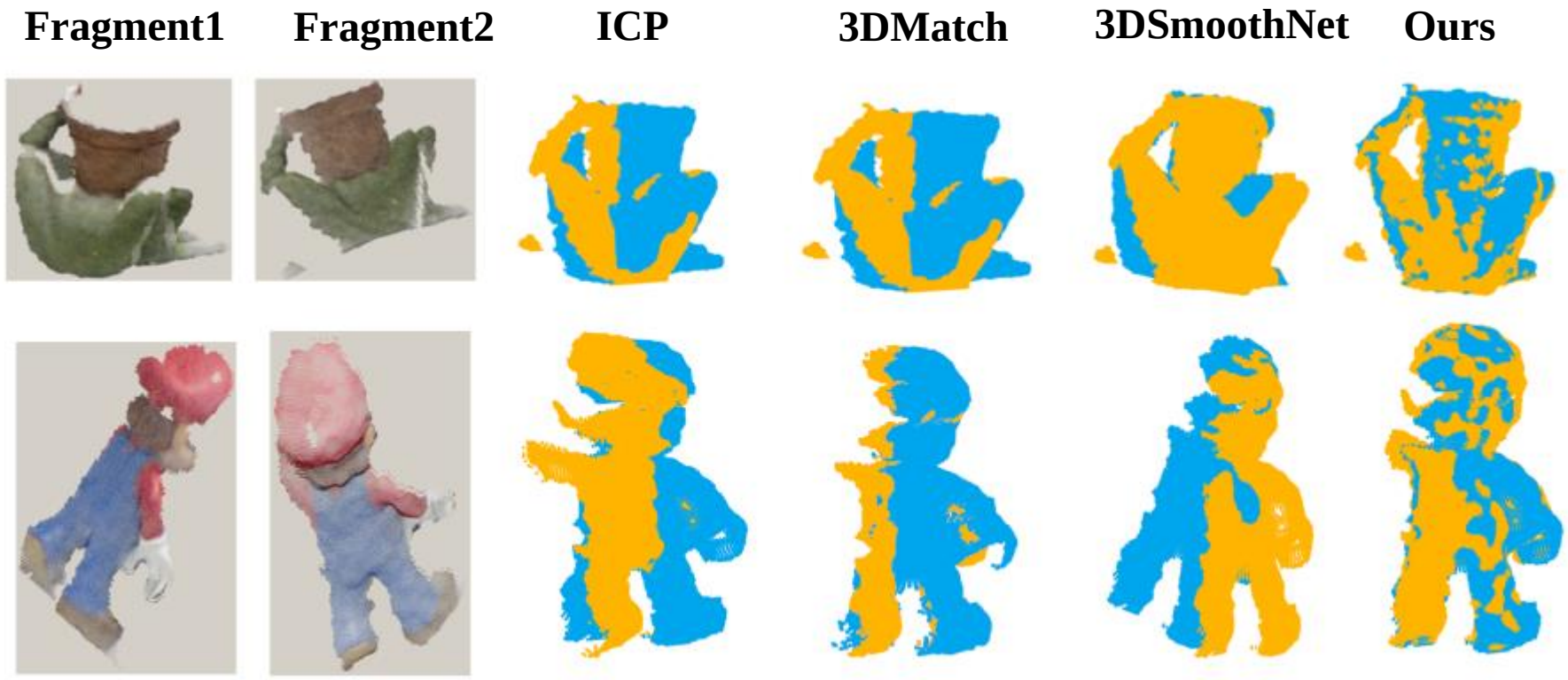
3 Results Visualization-model Display



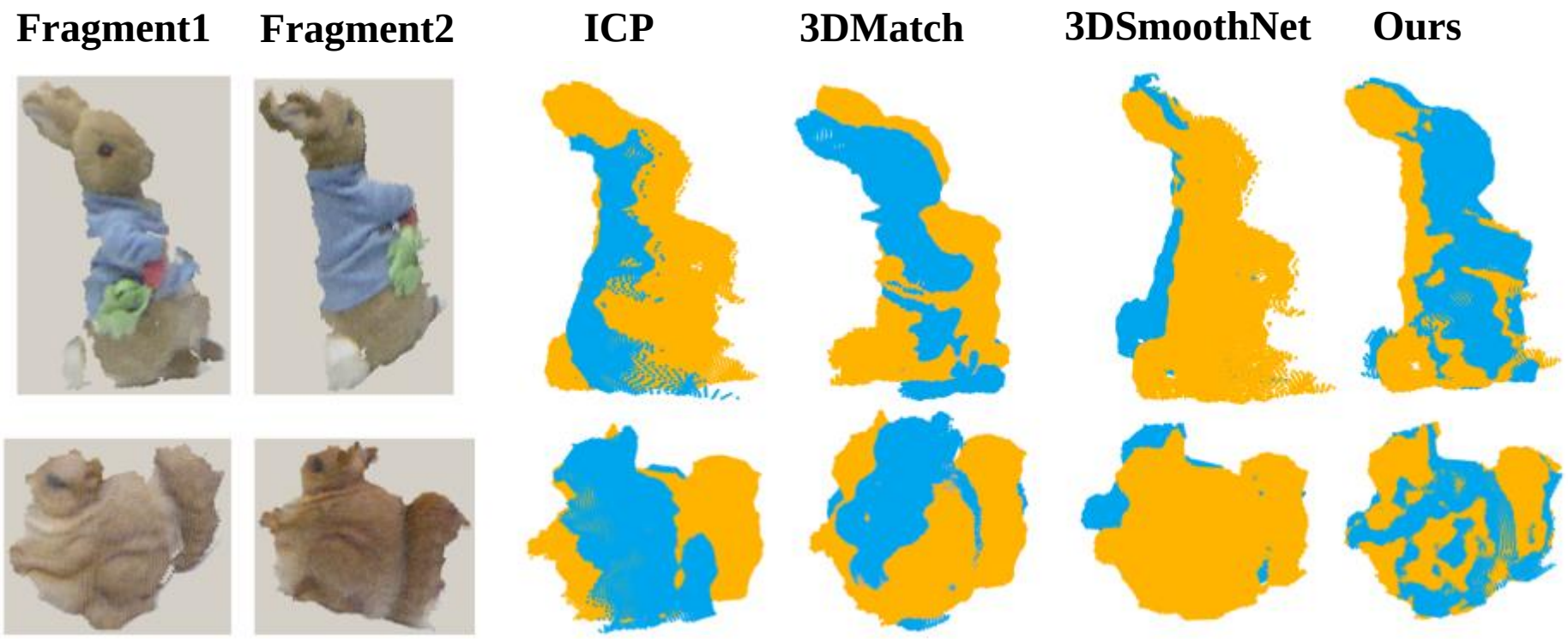
3 Results Visualization-model Display



3 Results Visualization-model Display



2 Results Visualization-model Display



THANKS!