DMG6D: A Depth-based Multi-Flow Global Feature Fusion Network for 6D Pose Estimation

Zihang Wang[1], Qiang Zhang[1, †], Xueying Sun[1], Jianwei Zhu[1], Hao Wei[2]

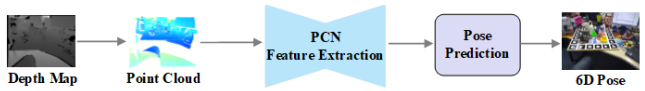
***Abstract*—The accurate estimation of the 6-degree-of-freedom pose of a target object in the environment plays a pivotal role in robot perception, providing a foundation for interaction and manipulation between robots and the surrounding environment. Nonetheless, traditional vision sensors are prone to diminished reliability in visual perception due to environmental factors such as lighting conditions and occlusions. Depth sensors, such as Time-of-Flight (TOF) and structured light sensors, offer promising opportunities for reliable target pose estimation. However, accurately determining the pose solely based on a single depth image presents significant challenges due to the limited availability of rich appearance and texture information. To comprehensively address this challenge, we investigate the mechanism of feature extraction and representation using depth images, along with the utilization of normal angle and point cloud information derived from the depth images, to achieve robust estimation of the visual target poses. By exploiting the latent information within the depth images, including normal angles and point clouds, we have developed the DMG6D robust target pose estimation framework. Within the DMG6D framework, we first employ physical methods to infer the normal angle and spatial position of each pixel in the depth image. Subsequently, we introduce a three-branch feature extraction and a global feature fusion network to enable a comprehensive depiction of the target object. Finally, a robust pose estimation for the target object is obtained utilizing the least squares method. Experimental results emphatically demonstrate that the proposed DMG6D surpasses existing algorithms in terms of its ability to estimate 6D poses using depth images, effectively underscoring the efficacy of our designed depth image feature extraction strategy. Access to the code and video is available at https://github.com/wangzihanggg/DMG6D.**

I. Introduction

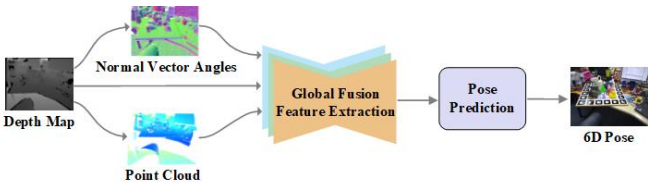
The 6-degree-of-freedom (6-DoF) pose estimation of an object involves geometrically mapping the object's coordinate system to the camera coordinate system, typically represented by a transformation matrix encompassing 3D rotation and 3D translation of the target object. In various domains such as robot interaction [1][2][3], virtual/augmented reality [4], and autonomous driving [5][6], accurate estimation of object poses is crucial. Robots require pose estimation for grasping tasks, interactions in virtual/augmented reality necessitate precise object poses for realistic interactions, and vehicles rely on pose estimation for obstacle avoidance and navigation.

Presently, most 6-DoF pose estimation algorithms are based on RGB images[7][8] or RGB-D[9][10][11] image inputs. However, these methods heavily rely on color and texture information from RGB data, often disregarding the geometric information inherent in depth images. Consequently, they may struggle in extreme scenarios such as variations in illumination or limited color textures, leading to compromised performance. Leveraging the geometric structure information of objects can offer a promising solution to these challenges. Depth sensors, characterized by their sensitivity to object geometry and insensitivity to light intensity, align well with our objectives. Moreover, with the gradual decrease in the price of depth sensors, utilizing depth sensors exclusively for 6-DoF pose estimation emerges as a viable and cost-effective option.

Figure 1. **Network Structure Comparison**



(a) **The CloudPose[12] Network.** Converting depth maps into point clouds and employing point cloud processing networks for feature extraction and pose estimation, using only point cloud modality.



(b)**The Proposed Multi-Flow Global Feature Fusion Network.** With three feature information streams—depth, angle, and spatial features—integrated throughout the entire process of the three networks, we introduce a fusion module based on global features into the full pipeline. This serves as a bridge for modality interaction, facilitating better learning representation of target objects.

To achieve stable 6D pose estimation using only depth data, a common approach involves transforming the depth image into a point cloud and utilizing a Point Cloud Processing Network (PCN) for feature extraction in pose estimation [12] (Fig. 1.a). However, conventional PCNs may not fully exploit all the information within the depth image. For the first time, method [13] introduces a feature extraction network tailored to point clouds, specifically designed for large scene segmentation tasks. In [11], there is an exploration of

*Research supported by the National Natural Science Foundation of China (grant number 61903162) and Jiangsu Province's "Double Innovation Plan":Research and Development of Flexible Cooperative Robot Technology for Intelligent Manufacturing.

†Corresponding author: qzhang@just.edu.cn (Qiang Zhang)

Zihang Wang, Qiang Zhang, Xueying Sun and Jianwei Zhu are with the College of Automation, Jiangsu University of Science and Technology, No. 666 Changhui Road, Zhenjiang 212100, China (e-mail: 202210305124@stu.just.edu.cn; qzhang@just.edu.cn; sunxueying@just.edu.cn; 221110303129@stu.just.edu.cn).

Hao Wei is with the Ocean College, Zhejiang University, Zhoushan, Zhejiang 316021, China (e-mail: isweihao@zju.edu.cn).

Manuscript 248 submitted to 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE). Received March 9, 2024.

converting the depth image into a point cloud based on the camera's depth conversion factor. The approach employs method [13] for robust feature extraction and level-by-level fusion of RGB modal features, resulting in stable 6D pose estimation. This represents a typical data-driven scheme, as point cloud data consist of discrete, point-by-point information during the feature extraction process.

Another approach [14], suggests converting depth images into surface normal vector images. This method effectively separates the object from the background, leveraging rich information about the target object's angle features. The surface normal vector angular modality captures the physical angular geometric characteristics of the target object. Superior results can be achieved by fusing this information with point cloud-based data features. As depicted in Fig. 1.b, our method integrates the three representations of depth image, normal vector angle image, and point cloud through an appropriate fusion mechanism. This integration maximizes the effective information in the depth data, leading to robust 6D pose estimation.

In this study, we introduce a novel multi-stream global feature fusion network designed to perform global fusion at each stage of the encoder-decoder, facilitating the learning of feature representations from depth data across different modalities for 6D pose estimation. The schematic overview of our proposed methodology is illustrated in Fig. 1.b. We leverage the Swin Transformer [15] to encode features from the original depth image and the normal vector angular map. Additionally, we employ RandLA-Net [15] for extracting features from the point cloud, incorporating layer-by-layer complementary feature fusion across these three modalities. Specifically, the depth image provides rich dense depth information, while the normal vector angle image furnishes angle information, both of which facilitate object-background separation. On the other hand, the point cloud captures spatial structural features of the object. By performing point-to-point fusion of dense features derived from depth, angle, and spatial structure information, we obtain a comprehensive dense feature embedding based on the depth representation. Subsequently, we adopt the 3D keypoint detection method proposed in [10], utilizing the feature embeddings acquired through the aforementioned approach for both 3D keypoint prediction of the target object and 6D pose regression.

To validate the effectiveness of our method, we conducted extensive experiments on two widely-used datasets: the LineMod dataset and the YCB-Video dataset. The experimental results demonstrate that our approach surpasses the current state-of-the-art methods relying on depth-input.

In summary, the key contributions of our work include:

- We develop the DMG6D framework, which leverages depth information along with angle and spatial structure information to enable robust feature extraction and accurate estimation of the 6 degrees of freedom (6-DoF) pose. This is achieved through the incorporation of these information sources into a data-driven neural network algorithm.
- To address the challenge of extracting robust features from multimodal information sources, we propose a novel mechanism in the DMG6D framework based on global feature fusion of multi-feature streams. This mechanism effectively integrates diverse representations of depth information to enhance the robustness and accuracy of feature extraction.
- We extensively evaluate our method on benchmark datasets, including the LineMod and YCB-Video datasets. Remarkably, our approach achieves pose estimation accuracies of 98.9% and 97.8% on these datasets, respectively. Furthermore, when compared to other methods that utilize depth information as input, our method demonstrates state-of-the-art performance.

II. Related Works

A. Depth Map Features

Mining robust information from depth images has been a prominent research focus within the realm of computer vision. For instance, [16][17] employ convolutional neural networks (CNNs) directly on Depth Maps to extract features for classification and segmentation tasks. Similarly, [14][18] compute the normal vector angle of each pixel from depth images, utilizing a form of differential histogram for 3D object recognition. Moreover, [13][19][20] elevate the depth map to point clouds using the camera's fixation matrix, facilitating spatial perspective for tasks such as semantic segmentation and target detection in large-scale scenes. Motivated by these prior works, we leverage the conversion of depth images into normal vector angle images and point clouds, respectively. By harnessing the distinctive advantages offered by each of these three modalities, we aim to achieve robust feature extraction.

B. 6D Pose Estimation from Depth Map

With advancements in large-scale instance segmentation models and point cloud representation learning, numerous 6D pose estimation algorithms leveraging depth data inputs have emerged. For instance, Liu et al.'s CATRE [21] employs Mask-RCNN [22] to isolate the object of interest from the depth map, then feeds it into a ShapeNet pre-trained point cloud network for feature extraction. This process is followed by matching it with the target object model's point cloud to determine the 6D pose. Similarly, Gao et al.'s [12] method utilizes the same segmentation network as found in PoseCNN [7], coupled with an enhanced version of PointNet [19] for point cloud feature extraction and 6D pose regression. In contrast, our proposed method thoroughly exploits the depth, angle, and spatial structure features available from depth images, and integrates them effectively to estimate the 6D pose.

C. Multi-Flow Feature Fusion for 6D Pose Estimation

In the domain of 6D attitude estimation, the synergistic fusion of multiple information streams is commonly employed, particularly in the bimodal fusion of RGB-D data. Wang et al.'s DenseFusion [9] introduces a dense fusion module for RGB-D features; however, there is a lack of information interaction during the feature extraction process. On the other hand, He et al.'s FFB6D [11] pioneers the use of an RGB-D two-module omnidirectional feature fusion during both the feature extraction coding and decoding stages, yielding commendable results. To the best of our knowledge, our method represents the first instance of applying multi-information stream fusion to 6D pose estimation based on depth images. This approach

aims to maximize the extraction of depth, angle, and spatial structure information inherent in depth images.

III. Method

Given a depth map, the objective of 6D target pose estimation involves determining the transformation matrix of the target object from its coordinate system to the camera coordinate system. This matrix comprises a rotation matrix, denoted as $R \in SO(3)$, and a translation matrix, denoted as $T \in SO(3)$. As this task solely depends on the depth image, the formulation of the pose estimation algorithm should systematically harness the geometric structural information inherent in the target object.

A. Overview

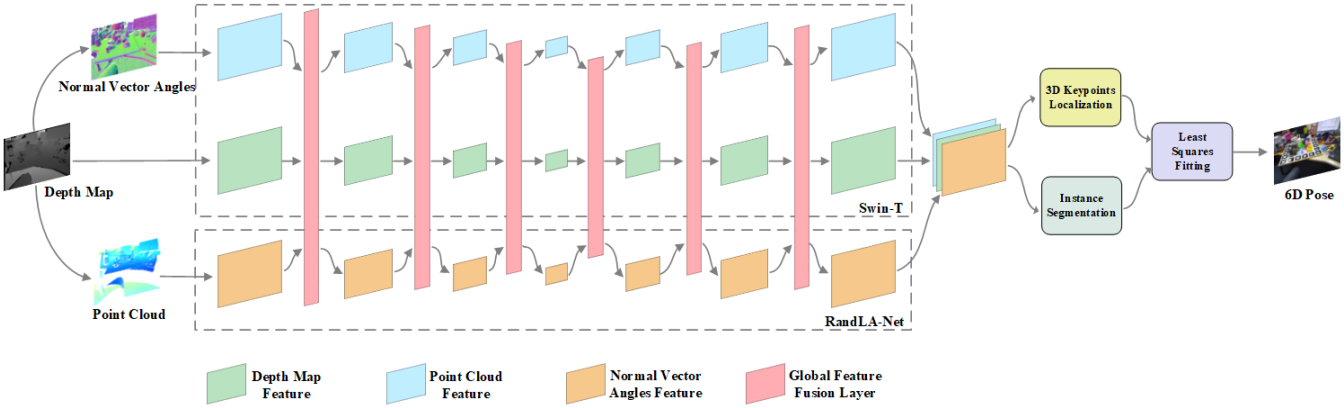
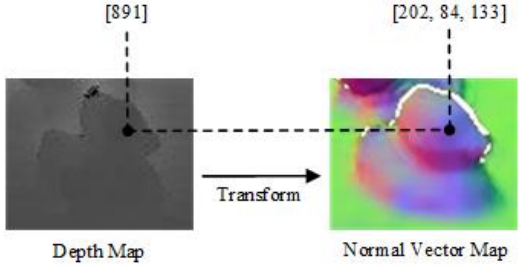


Figure 2. **The pipeline of DMG6D.** Using Swin Transformer for representation learning on depth maps and normal vector angle maps, and using PCNs for representation learning on point clouds. Within the processes of the three networks, we introduce a fusion module based on global features as a communication bridge. Subsequently, the extracted robust features are inputted into a 3D keypoint voting module to obtain 3D keypoints for each target object, followed by pose recovery through least squares fitting.

We introduce a multi-stream, hierarchical fusion network framework to address the problem of 6D attitude estimation from depth data, as illustrated in Figure 2. This framework is designed to extract features from depth data and progressively fuse them using attention networks and Point Cloud Networks (PCNs). It employs 3D keypoint localization detection heads, which leverage both spatial and channel attention mechanisms for keypoint determination, and subsequently utilizes the least squares method for pose fitting. More specifically, the Swin Transformer is utilized to extract appearance features of objects from depth images and images of normal vector angles, while the RandLA-Net is employed to extract geometric features from point clouds. Throughout the forward progression of these three feature streams, a point-to-point fusion mechanism is integrated at each level, enabling the image and point cloud modalities to independently learn each other's embedding representations. Subsequently, the learned features at each point are processed through an enhanced 3D keypoints detection module, which is based on the attention mechanism. The final step involves regressing the pose transformation matrix using the least squares method.

B. Depth image and normal vector image feature extraction

Figure 3. **Example of Normal Vector Angle Image Transformation.** Compute the angles a_x, a_y, a_z between each pixel in the depth map and the camera coordinate system, and normalize them to the range [0-255].



Previous depth-image-based 6D pose estimation algorithms tend to directly input the depth image into the feature extraction network, but this approach often neglects the orientation information of the target object implied by the depth image. In order to obtain orientation information and local geometric information from depth images, we follow the method of [14] to generate for each pixel point in each depth image its normal vector angle corresponding to the camera coordinate system N with XYZ axes, and this process can be described as:

$$\begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} = \begin{bmatrix} \arccos(N \cdot x) \\ \arccos(N \cdot y) \\ \arccos(N \cdot z) \end{bmatrix} \quad (1)$$

Subsequently, the obtained angle values (a_x, a_y, a_z) are normalized to the range of 0 to 255. These three angle values are then utilized to determine the position of each pixel point, resulting in the formation of an RGB image-like representation, as depicted in Figure 3.

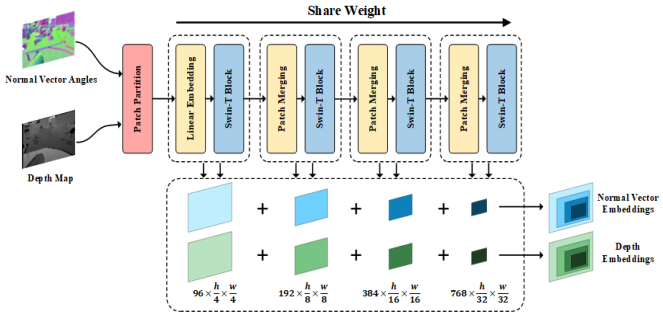
To effectively capture both the depth features from the depth image and the orientation features from the normal vector angle image, we propose a encoder-decoder structure network as the backbone for feature extraction. Recognizing the homology between the two images, we employ shared weights for feature embedding extraction. The network architecture employs Swin-T [15] (Tiny Swin Transformer) as an encoder and UperNet [23] as a decoder to learn multi-scale feature representations from both images.

The Swin-T encoder accepts both depth images and normal vector angle images as inputs, with its structure illustrated in Figure 4. It comprises four Swin Transformer modules, each consisting of an Encoder, Bottleneck, Decoder, and transition links. Taking the depth image as an example, it initially conducts Patch Partitioning into 4x4 Patches. Following a

linear embedding layer and two self-attentive Swin Transformer blocks for feature extraction, it undergoes downsampling via a Patch Merging layer, thereby completing one iteration of feature extraction. Each completed feature extraction operation interacts with the other two information streams for point-to-point feature fusion, and this process is iterated four times to complete the feature extraction for the depth image. The feature extraction process for the normal vector angle image follows a similar procedure.

Upon completion of the feature extraction process, we restore the feature map size through an upsampling operation utilizing UperNet [23]. UperNet executes upsampling based on bilinear interpolation and incorporates the Pyramid Pooling Module (PPM) from PSPNet [24] as the final layer of the feature pyramid. The upsampling operation is conducted three times, with each iteration followed by the interactive fusion of features from the three modalities (as explained in Section 3.4), along with concatenation with the encoder feature embedding. Ultimately, the last unsupervised upsampling, utilizing the function provided by MMSegmentation, is implemented to obtain the final pixel-level feature embedding of the same size as the input image.

Figure 4. **Swin-T Structure Diagram.** The encoder for depth maps and normal vector angle maps is based on Tiny Swin Transformer (Swin-T), which consists of four stages. The network processes the input images through cropping to obtain features at four different scales, which are then concatenated and fed into the decoder network, as shown in Figure 2. Throughout the entire encoding process, the depth maps and normal vector angle maps share weights.

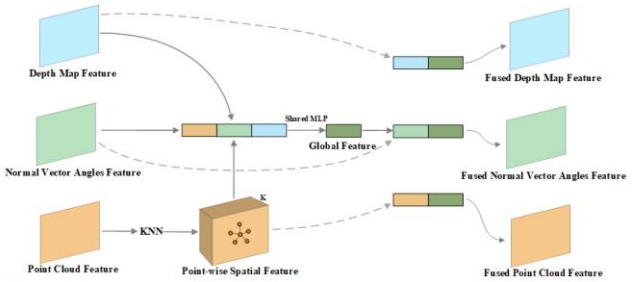


C. *Point Cloud Feature Extraction*

In the preprocessing stage of the depth image, we initially apply a multiscale filling method to address gaps in the depth image. Subsequently, we utilize the camera's internal reference matrix to transform the depth image, yielding the desired point cloud. To extract multi-scale embedded features from the point cloud, the point cloud feature extraction network comprises a segment of point cloud feature preprocessing CNN, along with four iterations of the RandLA-Net [15] feature extraction downsampling module and three iterations of the RandLA-Net feature extraction upsampling module. Throughout these seven up/down-sampling sessions, there are point-by-point feature interactions with both the depth image and the normal vector angular modality.

D. *Global Feature-Pointwise Fusion Mechanism*

Figure 5. **Global Feature Fusion Module.** For each pixel in the depth map and normal vector angle map, we locate its k-nearest neighbor pixels in the point cloud map and collect their corresponding features from the depth feature map and angle feature map. Similarly, we adopt the same approach for each point in the point cloud to obtain dense features pointwise. These features are then processed through a shared MLP to acquire global features, which are subsequently fused with the original feature maps from the three modalities, resulting in an updated fusion feature map.



In order to integrate the feature embeddings from three or more modalities, we have devised a global feature-based pointwise fusion mechanism, as illustrated in Figure 5. This study encompasses three modalities: depth map, normal vector angle image, and point cloud. Since all three modalities originate from the depth image, ensuring alignment between pixels and points, we can execute multimodal fusion on a pixel-point basis. Specifically, each pixel in the depth image and normal vector angle image corresponds to a point in the point cloud modality. We employ k-nearest neighbors (kNN) to identify the K nearest neighbors of the target point, where K is determined by the ratio of the embedding size of the depth image/normal vector angle features to the embedding size of the point cloud features. These points are then fed into a multilayer perceptron (MLP) to capture spatial features surrounding the target pixel. This process can be formulated as follows:

$$F_{p2g} = MLP(\{P_i | P_i \in \kappa\mathrm{NN}(P, \{P_i\}, k)\}) \quad (2)$$

Where F_{p2g} represents the spatial feature embedding of the target point, and P_i refers to the ith point in the nearest neighborhood of the target point P. Following feature complementation by k-nearest neighbors (kNN), the spatial feature embeddings of the point cloud modality attain the same size as the image feature embeddings of the depth image modality and the normal vector angle image modality after Swin-T feature extraction. To efficiently fuse the point-to-point features of the three modalities, we conduct global feature fusion:

$$F_{global} = MLP(F_{p2g} \oplus F_{dpt} \oplus F_{nv}) \quad (3)$$

Among them, F_{dpt} and F_{nv} represent the feature embeddings of the depth image and normal vector angle image, respectively, after each has been extracted by the Swin-T module. To mitigate overfitting, the global feature fusion section only takes the features of the three modalities as input into the multilayer perceptron after concatenation.

After acquiring the global features, they are fused with the feature embeddings of each modality. Initially, the global features are integrated with the point cloud modalities:

$$F_{fp} = MLP(Pool(F_{global}) \oplus F_{dpt}) \quad (4)$$

The spatial feature embedding of the point cloud modal data, denoted as F_{dpt}, undergoes processing by the point cloud network. As the size of F_{global} differs from that of F_{dpt}, we downsample F_{global} using the point cloud indexes utilized in obtaining F_{p2g}. Subsequently, we fuse the spatial feature

embedding of the point cloud modality with the global features through the multilayer perceptron to derive the fusion of global features denoted as F_{fp} . Following this, we proceed with the feature fusion of global features with the depth image modalities and normal vector angular modalities:

$$F_{fd} = MLP(F_{global} \oplus F_{dpt}) \quad (5)$$

$$F_{fn} = MLP(F_{global} \oplus F_{nv}) \quad (6)$$

This concludes the fusion of global features with the three modalities in a straightforward manner, providing each modality with feature embedding from depth, angle, and space, while simultaneously preventing overfitting. This global feature fusion mechanism is employed at every stage of the encoder-decoder structure to facilitate interaction among depth, angle, and spatial features.

E. 3D Key Point Pose Detection Algorithm Transformed based on CBAM Mechanism

Recently, He et al.'s PVN3D [10] introduced a 6D pose estimation algorithm based on 3D keypoints, while their FFB6D[11] work proposed SIFT-FPS, which enhances PVN3D's 3D keypoint selection algorithm by leveraging object texture and geometric information more comprehensively. In this study, we adopt the detection head algorithm from FFB6D. Specifically, we initially select the 3D keypoints of each target object using the aforementioned feature extraction module and then estimate the pose matrix through least squares fitting.

Multi-target 3D Keypoint Localization Detection Head:

Having obtained dense feature embeddings encompassing depth, angle, and spatial information from a single depth image, we proceed to fit the translation matrix, rotation matrix, and instance segmentation image. Drawing inspiration from He et al.'s FFB6D, we employ a farthest point sampling algorithm (SIFT-FPS) to sample keypoints on the target object's surface. This method ensures that keypoints are not only uniformly distributed but also possess distinctive textures for easy detection. To enhance the dense feature embedding's accuracy in instance segmentation, translation matrix keypoint regression, and rotation matrix keypoint regression, we incorporate a spatial and channel-based self-attention module into the detection head module, inspired by the CBAM [25] algorithm. This addition strengthens the dense feature embedding's capability to perform these tasks independently and enhances the stability of pose regression.

Least Squares Recovery of the Attitude Matrix:

To compute the rotation matrix R and translation matrix T of the target object relative to the camera coordinate system, we utilize the N 3D keypoints p_i^{obj} obtained from the 3D keypoint localization module, along with the corresponding N 3D keypoints p_i^{cam} in the camera coordinate system, as inputs. Subsequently, we regress R and T using the least squares method to minimize squared loss [26].

$$L_{squ} = \sum_{i=1}^{N} (p_i^{cam} - (R \cdot p_i^{obj} + T))^2 \quad (7)$$

IV. EXPERIMENTS

We assess the performance of our method using two benchmark datasets: the YCB-Video dataset[27] and the LineMod dataset[28]. Our proposed DMG6D method demonstrates significant improvements over state-of-the-art benchmark algorithms on both datasets. Additionally, we conduct comprehensive ablation experiments to illustrate the effectiveness of our proposed algorithm.

A. Benchmark Datasets

The YCB-Video dataset comprises 92 RGB-D videos, encompassing 21 target objects, with a total of 133,827 frames of 640x480 images. Following the approach outlined in [29], we partition the dataset into training and testing sets, and, as [30], we apply hole filling to the depth images.

The LineMod dataset includes 13 RGB-D videos featuring non-textured objects in cluttered scenes. We follow the methodology from [31] to divide the dataset into training and testing sets. Additionally, we generate 20,000 synthetic images for each object to augment the training process.

B. Evaluation Metrics

We employ two commonly used 6-DoF pose estimation evaluation metrics to measure the performance of the algorithm: Average Distance Distance (ADD) and Average Distance Distance-S (ADD-S) [7]. For asymmetric objects, the ADD algorithm computes the average distance between the predicted pose by the framework and the object's ground truth vertices:

$$ADD = \frac{1}{m} \sum_{p \in M} \| (R_p + T) - (R^* p + T^*) \| \quad (8)$$

Here, p represents the vertices of the target object, M denotes the set of target vertices, m is the number of vertices, R and T represent the rotation matrix and translation matrix predicted by the model, and R^*, T^* represent the ground truth rotation matrix and translation matrix. For symmetric objects, the algorithm for calculating the Average Distance Distance-S (ADD-S) is:

$$ADD - S = \frac{1}{m} \sum_{p_1 \in M} \left(\min_{p_2 \in M} \| (R p_1 + T) - (R^* p_2 + T^*) \| \right) \quad (9)$$

In the experiments conducted on the YCB-Video dataset, we follow the approach outlined in prior work [7], utilizing the Area Under the Curve (AUC) of the precision-recall curve based on the ADD(S) metric as the evaluation metric. For the LineMod experiments, we adhere to the methodology established in previous research [32], employing the Average Distance Distance metric where the distance is less than 10% of the object's diameter (ADD-0.1d) as the evaluation criterion.

C. Implemention Details

DMG6D uses two encoder-decoder structures to extract robust features from a single depth map. For image modality, we use Swin-T as the encoder and UperNet as the decoder. For point cloud modal feature extraction, we follow the PVN3D method to randomly sample 19,200 points of the point cloud and use RandLA-Net for feature representation learning. At each codec layer of the three network streams, omnidirectional feature interaction fusion is constructed using a global feature

fusion module based on shared MLP. After processing throughout the omnidirectional fusion network, 19,200 3D keypoints are generated, each with a 128-dimensional feature F_p, and these dense feature embeddings are then fed into the rotation matrix estimation module, translation matrix estimation module, and instance segmentation module. The rotation and translation matrix estimation modules use L1 Loss[33] and the instance segmentation uses Focal Loss[34], and we follow [10] for a multi-task loss scheme to jointly optimise the three tasks. The training process was done on a single NVIDIA RTX3090Ti GPU with Batch Size set to 8. Each object of LineMod was trained for 30-40 epochs, and the YCB-Video dataset was trained for 50 epochs.

D. Evaluation on Two Benchmart Datasets

We assessed the proposed model using both the LineMod dataset and the YCB-Video dataset.

Evaluation on the LineMod dataset: Table 1 presents the quantitative evaluation results of our proposed method on the LineMod dataset. These results remain unrefined by subsequent processing and are compared against other state-of-the-art methods. For asymmetric objects, we compute the Average Distance Distance (ADD) metric when the distance is less than 10% of the object's diameter (ADD<0.1d), while for symmetric objects, we compute ADD-S<0.1d. Additionally, we classify our results into three groups based on the input modality: RGB, RGB-D, and Depth Only. The findings reveal that our proposed method surpasses the current state-of-the-art methodologies by 1.4% across all methods utilizing depth data as input. Furthermore, our approach outperforms certain methods relying on RGB and RGB-D data inputs within specific categories, underscoring the efficacy of the DMG6D framework in feature extraction from depth data. Visualizations in Fig. 6 depict the reprojected effects of the proposed algorithm on select LineMod objects, demonstrating that DMG6D consistently and accurately predicts the spatial position of the target object.

TABLE I. **EVALUATION ON THE LINEMOD DATASET**. THE RESULTS ARE REPORTED USING THE ADD<0.1D METRIC. SYMMETRIC OBJECTS ARE DENOTED WITH AN ASTERISK (*), AND WE HIGHLIGHT THE OPTIMAL PERFORMANCE WITHIN EACH INPUT MODALITY GROUP IN BOLD.

INPUTS	**RGB**			**RGB-D**			**Depth Only**			
METHODS	PoseCNN[7]	PVNet[8]	DPOD[35]	DenseFusion[9]	PVN3D[10]	FFB6D[11]	CloudAAE[36]	CATRE[21]	SwinDePose[31]	OURS
ape	77.0	43.6	**87.7**	92.3	97.3	**98.4**	74.5	63.7	95.4	**96.6**
benchvise	97.5	**99.9**	98.5	93.2	99.7	**100.0**	96.6	98.6	98.2	**99.5**
camera	93.5	86.9	**96.1**	94.4	99.6	**99.9**	65.6	89.7	96.9	**99.1**
can	96.5	95.5	**99.7**	93.1	99.5	**99.8**	90.2	96.1	98.2	**99.3**
cat	82.1	79.3	**94.7**	96.5	99.8	**99.9**	90.7	84.3	**98.7**	**98.7**
driller	95.0	96.4	**98.8**	87.0	99.3	**100.0**	97.3	98.6	98.5	**100.0**
duck	77.7	52.6	**86.3**	92.3	98.2	**98.4**	50.0	63.9	92.7	**94.9**
eggbox*	97.1	99.2	**99.9**	99.8	99.8	**100.0**	99.7	99.8	**100.0**	99.9
glue*	**99.4**	95.7	96.8	**100.0**	**100.0**	**100.0**	93.5	99.4	**100.0**	**100.0**
holepuncher	52.8	82.0	**86.9**	92.1	**99.9**	99.8	57.9	93.2	93.6	**98.6**
iron	98.3	98.9	**100.0**	99.0	99.7	**99.9**	85.0	98.4	96.9	**99.5**
lamp	97.5	**99.3**	96.8	95.3	99.8	**99.9**	82.1	98.7	99.1	**99.9**
phone	87.7	92.4	**94.7**	92.8	99.5	**99.7**	94.4	97.5	98.8	**99.9**
MEAN	88.6	86.3	**95.2**	94.3	99.4	**99.7**	82.1	90.9	97.5	**98.9**

Figure 6. **Visualization of DMG6D on LineMod Objects.** We demonstrate the visual effects of DMG6D on LineMod objects. We transform the predicted 3D keypoints into the camera coordinate system and project them onto the image using the camera's intrinsic matrix. To enhance the visualization of the results, we utilize RGB images to display the outcomes, with the projected points shown in green.

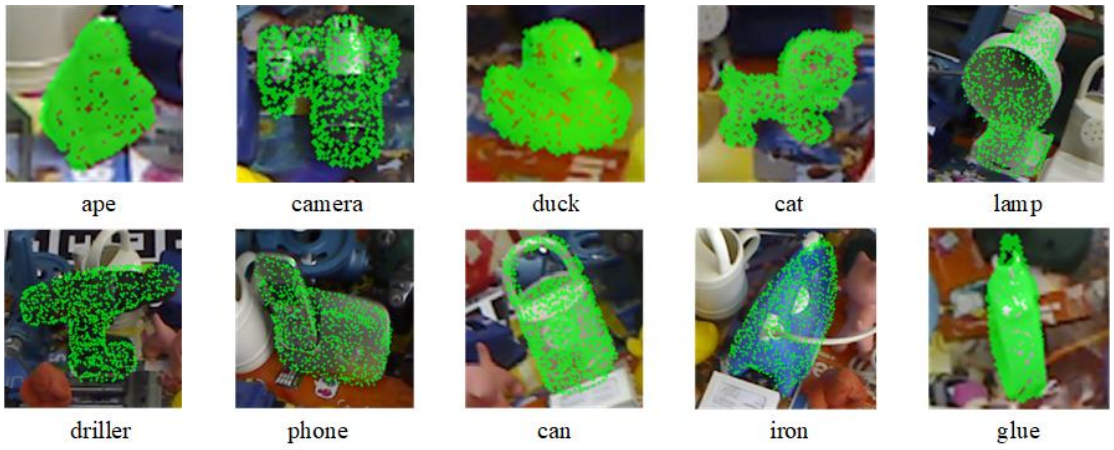


TABLE II. **EVALUATION ON THE YCB-VIDEO DATASET.** RESULTS ARE REPORTED USING THE ADD(S) METRIC, AND WE HIGHLIGHT THE OPTIMAL PERFORMANCE IN BOLD.

INPUTS	**Depth Only**		
METHODS	CloudPose (ICP)[12]	CloudAAE (ICP)[36]	Ours (ICP)
MEAN	93.0	94.0	**97.8**

Tested on the YCB-Video dataset: Table 2 presents the quantitative evaluation results of the proposed method against other state-of-the-art methods. We calculate ADD for asymmetric objects and ADD-S for symmetric objects. The results indicate a 3.8% improvement over the current state-of-the-art method CloudAAE and a 4.8% enhancement over the classical method CloudPose illustrated in Fig. 1.a, among all methods utilizing depth data as input. This underscores the robustness of the DMG6D framework in handling complex multi-object scenarios.

E. Ablation Study

In this section, we conduct an ablation study on the introduction of angular prior information for deep data feature mining effective shape and the global fusion mechanism that is most important for model performance. For accurate evaluation, we conducted tests on the YCB-Video dataset. The results shown in Table 3 are the average ADD(S) on the YCB-Video dataset.

TABLE III. **RESULTS OF ABLATION STUDY.** WE INVESTIGATED THE IMPACT OF THE GLOBAL FUSION MECHANISM ON THE 6D POSE ESTIMATION RESULTS ON THE YCB-VIDEO DATASET. "GFF" DENOTES THE GLOBAL FEATURE FUSION MECHANISM, WHILE "NV" REPRESENTS THE NORMAL VECTOR ANGLE MODALITY INFORMATION.

Aspect	**Average ADD(S)**
Full Model	97.8
w/o GFF	95.7
w/o NV	95.9

In order to verify the effectiveness of using the global fusion mechanism and normal vector angular prior information for learning deep data feature embeddings, we conducted an ablation study on the YCB-Video dataset. For comparison, we remove the global feature fusion mechanism with normal vector angle modal information at each layer of the codec process respectively, and the results in Table 3 show that the complete model with global feature fusion mechanism obtains better pose estimation results, highlighting the superiority of the global feature fusion mechanism in the face of the multi-information stream feature extraction process. Meanwhile, the network with normal vector angle prior information also performs better, indicating that the introduction of new modalities from the physical geometry perspective can effectively improve the ability of neural networks to understand image features.

V. CONCLUSION

In this paper, we present the DMG6D framework, a novel approach aimed at advancing the state-of-the-art in pose estimation using single depth images. Our framework introduces a distinctive fusion mechanism tailored for extracting depth, angular, and spatial features, thereby enhancing the richness of information obtainable from individual depth images. Notably, our methodology exhibits remarkable versatility, facilitating seamless integration of information from various modalities while mitigating the risks associated with overfitting, owing to its inherent simplicity and efficiency.

Furthermore, we propose a methodology that significantly improves multi-target keypoint localization detection by incorporating a self-attentive mechanism. This enhancement ensures the extraction of robust directional features essential for precise target keypoint fitting, thereby enhancing the accuracy and reliability of pose estimation.

The efficacy of our proposed framework is substantiated through empirical validation on benchmark datasets such as YCB-Video and LineMod. Our results demonstrate significant performance enhancements over existing techniques for 6D pose estimation from single depth images, underscoring the superiority of our approach.

Moreover, we expand the scope of depth modalities by extracting novel representations of physical geometry angles. This augmentation enables data-driven neural networks to incorporate crucial insights into physical geometry, transcending the limitations of purely data-driven methodologies. We envision that this advancement will inspire future research, facilitating the extraction of latent information from single modalities across diverse domains and fostering continuous innovation in the field.

REFERENCES

[1] Collet A, Martinez M, Srinivasa S S. The Moped Framework: Object Recognition and Pose Estimation for Manipulation. The International Journal of Robotics Research, 2011, 30(10): 1284-1306.

[2] Tremblay J, To T, Sundaralingam B, et al. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. arXiv preprint arXiv:1809.10790, 2018.

[3] Akinola I, Xu J, Song S, et al. Dynamic Grasping with Reachability and Motion Awareness. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021: 9422-9429.

[4] Marchand E, Uchiyama H, Spindler F. Pose Estimation for Augmented Reality: A Hands-on Survey. IEEE Transactions on Visualization and Computer Graphics, 2015, 22(12): 2633-2651.

[5] Wu D, Zhuang Z, Xiang C, et al. 6D-VNet: End-to-End 6-DOF Vehicle Pose Estimation from Monocular RGB Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019: 0-0.

[6] Sun P, Wang W, Chai Y, et al. RSN: Range Sparse Net for Efficient, Accurate LiDAR 3D Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 5725-5734.

[7] Xiang Y, Schmidt T, Narayanan V, et al. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. arXiv preprint arXiv:1711.00199, 2017.

[8] Peng S, Liu Y, Huang Q, et al. PVNet: Pixel-wise Voting Network for 6DOF Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4561-4570.

[9] Wang C, Xu D, Zhu Y, et al. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3343-3352.

[10] He Y, Sun W, Huang H, et al. PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DOF Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11632-11641.

[11] He Y, Huang H, Fan H, et al. FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3003-3013.

[12] Gao G, Lauri M, Wang Y, et al. 6D Object Pose Regression via Supervised Learning on Point Clouds. In: 2020 IEEE International

Conference on Robotics and Automation (ICRA). IEEE, 2020: 3643-3649.
[13] Hu Q, Yang B, Xie L, et al. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11108-11117.
[14] Boubou S, Narikiyo T, Kawanishi M. Differential Histogram of Normal Vectors for Object Recognition with Depth Sensors. In: 2016 International Conference on Autonomous Robot Systems and Competitions (ICARSC). IEEE, 2016: 162-167.
[15] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.
[16] Loewke N. Depth-Based Image Segmentation. 2015.
[17] Kumari S, Jha R R, Bhavsar A, et al. AutoDepth: Single Image Depth Map Estimation via Residual CNN Encoder-Decoder and Stacked Hourglass. In: 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 340-344.
[18] Zelener A, Stamos I. CNN-Based Object Segmentation in Urban LiDAR with Missing Points. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016: 417-425.
[19] Qi C R, Su H, Mo K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 652-660.
[20] Zhao L, Tao W. JSNet: Joint Instance and Semantic Segmentation of 3D Point Clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 12951-12958.
[21] Liu X, Wang G, Li Y, et al. CATRe: Iterative Point Clouds Alignment for Category-Level Object Pose Refinement. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 499-516.
[22] He K, Gkioxari G, Dollár P, Girshick R. "Mask R-CNN". arXiv, 24 January 2018.
[23] Xiao T, Liu Y, Zhou B, et al. Unified Perceptual Parsing for Scene Understanding. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018: 418-434.
[24] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2881-2890.
[25] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.
[26] Chen M, Xue H, Cai D. Domain Adaptation for Semantic Segmentation with Maximum Squares Loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 2090-2099.
[27] Calli B, Singh A, Walsman A, et al. The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research. In: 2015 International Conference on Advanced Robotics (ICAR). IEEE, 2015: 510-517.
[28] Hinterstoisser S, Holzer S, Cagniart C, et al. Multimodal Templates for Real-Time Detection of Texture-Less Objects in Heavily Cluttered Scenes. In: 2011 International Conference on Computer Vision. IEEE, 2011: 858-865.
[29] Wang Z, Sun X, Wei H, et al. Enhancing 6-DoF Object Pose Estimation through Multiple Modality Fusion: A Hybrid CNN Architecture with Cross-Layer and Cross-Modal Integration. Machines, 2023, 11(9): 891.
[30] Ku J, Harakeh A, Waslander S L. In Defense of Classical Image Processing: Fast Depth Completion on the CPU. In: 2018 15th Conference on Computer and Robot Vision (CRV). IEEE, 2018: 16-22.
[31] Li Z, Stamos I. Depth-Based 6DOF Object Pose Estimation Using Swin Transformer. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023: 1185-1191.
[32] Hinterstoisser S, Lepetit V, Ilic S, et al. Model-Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In: Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I. Springer Berlin Heidelberg, 2013: 548-562.
[33] Barron J T. A General and Adaptive Robust Loss Function. arXiv preprint arXiv:1701.03077, 2017.
[34] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.
[35] Zakharov S, Shugurov I, Ilic S. DPOD: 6D Pose Object Detector and Refiner. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1941-1950.
[36] Gao G, Lauri M, Hu X, et al. CloudAAE: Learning 6D Object Pose Regression with Online Data Synthesis on Point Clouds. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 11081-11087.