2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

Attention-guided RGB-D Fusion Network for Category-level 6D Object Pose Estimation

Hao Wang[1], Weiming Li[1], Jiyeon Kim[2] and Qiang Wang[1]

Abstract— **This work focuses on estimating 6D poses and sizes of category-level objects from a single RGB-D image. How to exploit the complementary RGB and depth features plays an important role in this task yet remains an open question. Due to the large intra-category texture and shape variations, an object instance in test may have different RGB and depth features from those of the object instances in training, which poses challenges to previous RGB-D fusion methods. To deal with such problem, an Attention-guided RGB-D Fusion Network (ARF-Net) is proposed in this work. Our key design is an ARF module that learns to adaptively fuse RGB and depth features with guidance from both structure-aware attention and relation-aware attention. Specifically, the structure-aware attention captures spatial relationship among object parts and the relation-aware attention captures the RGB-to-depth correlations between the appearance and geometric features. Our ARF-Net directly establishes canonical correspondences with a compact decoder based on the multi-modal features from our ARF module. Extensive experiments show that our method can effectively fuse RGB features to various popular point cloud encoders and provide consistent performance improvement. In particular, without reconstructing instance 3D models, our method with its relatively compact architecture outperforms all state-of-the-art models on CAMERA25 and REAL275 benchmarks by a large margin.**

I. INTRODUCTION

Estimating six degree-of-freedom (6D) pose for three-dimensional object is important in many computer vision applications such as robotic grasping and manipulation [1], [2], virtual reality [3] and augmented reality [3], [4]. 6D object pose estimation [5] has been extensively studied in the past few years and there are two task settings defined for instance-level objects [6], [7] and category-level objects [8], [9] respectively. For instance-level pose estimation, the object instance's CAD model is known in training and assumed to not change in test. Different from this, category-level pose estimation predicts 6D poses and sizes for a variety of object instances belonging to the same category without knowing their exact 3D models. Due to intra-category variations of textures and shapes, an object instance in test may be unseen in training and may look quite different from the training samples. This makes the category-level pose estimation task much more challenging than its instance-level counterpart.

With the dramatic growth of RGB-D sensors, many works [10]–[13] take advantage of the complementary RGB and depth data and achieve dominant performance in 3D computer vision tasks. In instance-level pose estimation task, several recent works such as DenseFusion [14] and FFB6D [15] make in-depth studies on fusion mechanisms of RGB and Depth features. DenseFusion proposes a dense fusion strategy instead of naive concatenation operation on high-level features. FFB6D proposes a full flow bidirectional fusion network that performs fusion on each of the encoding and decoding layers as communication bridges between the two modalities. Although these works achieved state-of-the-art performance in several instance-level pose estimation benchmarks, the benefits of fusing RGB features or RGB-D fusion features for category-level pose estimation are still limited, which motivates this work.

In this paper, we focus on the research question: how to optimally fuse information from the RGB and depth modalities for category-level 6D pose estimation? To answer the question, firstly, we observe that some RGB-depth correlations are instance-specific and difficult to generalize. For example, the logo printed on a bottle instance may have strong correlation to the bottle's depth to infer its 6D pose. However, such correlation becomes not applicable for another bottle instance with a different logo. This indicates that direct RGB-D fusion may introduce noise in the fused features, especially for unseen objects. To avoid this, fusion is supposed to learn to distinguish instance-specific and category-shared RGB-depth correlations and exploit the latter. Secondly, object parts play different roles in estimating 6D pose and the parts that only belong to some specific instances are not generalizable for unseen instances without these parts. The above observations inspire that attention guidance that can discriminate such category-shared and instance-specific correlations is essential to make proper RGB-D fusion for the category-level pose estimation task.

Inspired by the recent success of transformers in representation learning of sequential and visual data, we propose an Attention-guided RGB-D Fusion Network (ARF-Net). The key design component is the ARF module that learns to adaptively fuse RGB and depth features with guidance of attention mechanisms. Considering that 6D pose is closely related to 3D representation, our ARF module aims to fuse RGB and depth features into point cloud feature representation via structure-aware attention and relation-aware attention. Specifically, in the ARF module, a structure-aware attention is firstly employed to extract spatial relationship among object points. Then, a relation-aware attention is utilized to infer the correlation between the RGB features

[1] Hao Wang, Weiming Li and Qiang Wang are with SAIT China Lab, Samsung Research Center, Beijing, China {hao1.wang, weiming.li, qiang.w}@samsung.com

[2] Jiyeon Kim with Samsung Advanced Institute of Technology(SAIT), South Korea jiyeon31.kim@samsung.com

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9981242

and depth features, which enables the fusion module to selectively fuse RGB features to corresponding depth features. We build our ARF-Net with multiple ARF modules to make the fusion progressively, which is proved to be effective. With the structure-enhanced and relation-enhanced fused features, our ARF-Net directly establishes canonical correspondences instead of deforming categorical shape priors to reconstruct instance 3D models which is commonly used in top-performing methods [16]–[18]. Finally, we leverage a point cloud alignment method [19] to estimate object 6D pose and size. Experimental results show that our method outperforms all state-of-the-art models on the CAMERA25 and REAL275 [8] benchmarks by a large margin. To summarize, our main contributions are:

1) We propose an Attention-guided RGB-D Fusion (ARF) module with structure-aware attention and relation-aware attention designs that effectively address the issues in extracting optimal RGB-D fused representations for unseen object instances.
2) Based on the ARF module, we build an ARF-Net for category-level 6D object size and pose estimation. With the fused 3D representations by ARF module, the ARF-Net consists of a simple structure, yet achieves significant performance improvement.
3) We conduct ablation studies and extensive experiments on the public CAMERA25 and REAL275 datasets, which show that our method is effective and outperforms all state-of-the-art methods by a large margin.

II. Related Work

A. Instance-level 6D pose estimation

Object instances in instance-level 6D pose estimation are provided with known 3D models and they are the same during training and testing. Recent methods follow data-driven methodology and can be divided into three classes: direct pose regression methods, keypoint-based methods and differentiable rendering-based methods. Direct pose regression methods such as PoseCNN [7], DenseFusion [14] and GDR-Net [20], directly regress pose parameters from feature embedding. Keypoint-based methods such as PVNet [21], Pix2Pose [22] and PVN3D [23], first predict sparse keypoints or dense object coordinates and then use a PnP or a pose fitting method to recover object 6D pose. The third class of methods leverage differentiable rendering techniques for end-to-end pose optimization via rendering latent 3D representations [24] or by a self-supervised learning that alleviates dependencies on labeled data [25]–[27]. Besides these one-step prediction methods, iterative refinement methods for pose estimation also receive great research interests [28]–[30]. However, the iterative refinements are time-consuming and the two-step pipelines are not end-to-end optimizable.

Instead of using RGB and depth data separately, several RGB-depth fusion mechanisms are derived. DenseFusion [14] proposes a dense fusion strategy instead of naive concatenation operation with two networks that extract and fuse appearance and geometry features respectively. Since the two networks are separate, they are not able to communicate and share information and thus limit the expression ability of the learned representation. Therefore, FFB6D [15] proposes a full flow bidirectional fusion modules as communication bridges between the two networks. FFB6D achieve state-of-the-art performance in several instance-level pose estimation benchmark. CMA [31] uses a self-attention on concatenated RGB and depth features for subsequent pose predictor. Different from these methods, our work focuses on a more general setting where the object CAD models are not available and object instance may be unseen by the model in its training.

B. Category-level 6D Pose and Size Estimation

Category-level 6D pose and size estimation is formally introduced in [8]. To deal with category-level pose estimation without using known 3D models, Wang et al. [8] propose a canonical shape representation called normalized object coordinate space (NOCS) to represent different object instances within a category in a unified manner. They first establish NOCS map with RGB input and then align NOCS coordinates with the observed object points to obtain 6D pose and size. Recently, several top-bottom methods are implemented which first use 2D object detection or instance segmentation with RGB image to detect objects and then use RGB-D data to predict 6D object poses. Chen et al. [32] train a variational auto-encoder (VAE) to capture a pose-independent feature, which is fused with a pose-dependent encoder to directly estimate object pose. Tian et al. [16] reconstruct coordinates in NOCS by explicitly modeling the deformation field from categorical shape priors. To use both RGB and Depth data, the above methods usually fuse features with the DenseFusion approach [14]. However, few works specifically discuss what appearance features from RGB images are beneficial and how to fuse them with depth features for category-level pose estimation.

Recently, Wang, et al. [17] introduced a simple instance relation network that captures the relation of the input RGB image and point cloud to extract representative feature embedding for each instance, which does not show much enhancement on feature representation ability. Chen et al. [18] introduce a global structure similarity between observed point features and prior shape features in order to adaptively inject semantic features into the prior. Different from these methods that leverage multi-source features by a late fusion mechanism, Lin et al. [33] use a module of Spherical Fusion to learn embedding of pose sensitive features from the appearance and shape observations of each scale. Different from these prior works, we build our fusion module with two attention modules that not only modeling RGB-to-Depth relationship but also instance structure cues and utilize them in a progressive way which can boost network performance for category-level 6D pose estimation task.

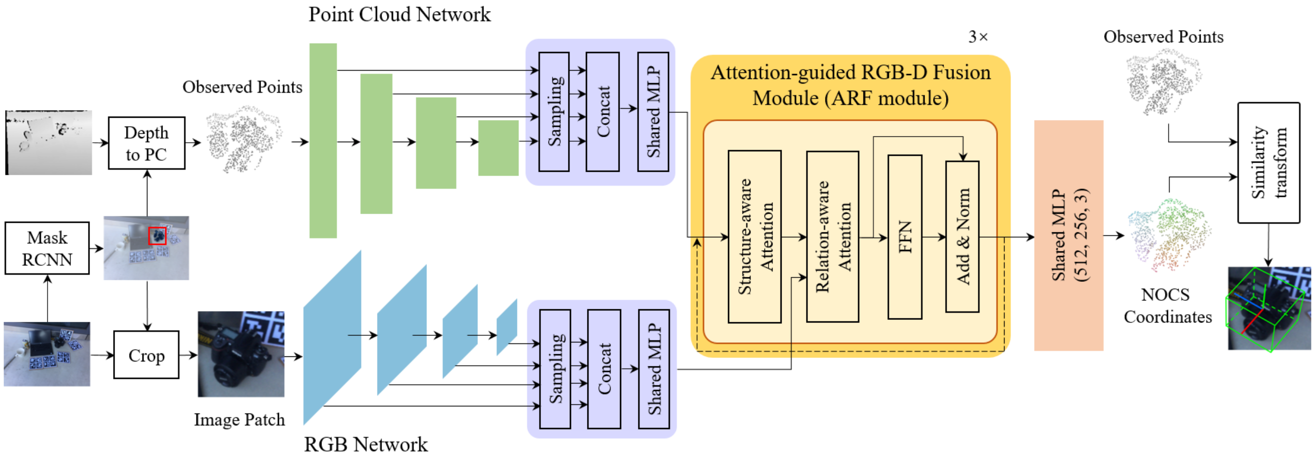


Fig. 1. Illustration of our proposed attention guided RGB-D fusion network. A key component is the Attention-guided RGB-D fusion(ARF) module. In the ARF module, the structure-aware attention captures structural relationship among object parts and the relation-aware attention captures RGB-to-depth correlations between the appearance and geometric features. The RGB-D fused representation is fed to an MLP to regress NOCS coordinates for object pose and size estimation.

III. Proposed Method

A. Overview

Given an aligned RGBD image, we first employ an instance segmentation method to detect and segment object instances. We then lift the depth image to point cloud with the camera's intrinsic matrix. Object RGB image is cropped by object bounding box and object point cloud is cropped by object mask. As shown in Fig. 1, our ARF-Net takes a pair of object point cloud and object image as input and predicts corresponding points in the NOCS. The NOCS is proposed in [8] and is a shared canonical representation for all possible object instances within a category. 6D object pose and size is obtained by aligning between the observed point cloud of each object instance and its corresponding points in the NOCS.

In our ARF-Net, a CNN network is used to extract appearance features from the object RGB image, and a point cloud network (PCN) is used to extract geometry features from the observed object instance point clouds. The appearance features and geometry features are fed into the ARF module to fuse the two modality features. Based on the fusion module, object appearance features can be adaptively propagated to geometry features and geometry features can be adaptively enhanced. Features fused by ARF modules are adopted for predicting dense correspondences in NOCS by a Multi-Layer Perceptron (MLP) decoder. Next, we will describe the proposed structure-aware attention, relation-aware attention and pose generation method in detail.

B. Attention-guided RGB-D Fusion

Since geometry information provides more relevant cues to determine object pose and size, this work fuses representations from the RGB modality to enhance 3D representations for pose estimation. We extract the initial pointwise geometry features from the observed point cloud with a PCN backbone network. These initial features only capture local point cloud features without awareness of relationships to other object parts nor complemented information from RGB features. Therefore, we propose to fuse the two-modal representations by our ARF module.

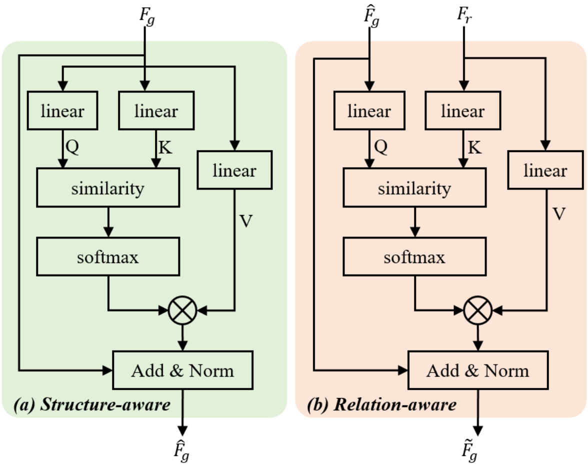


Fig. 2. An illustration of our (a) structure-aware attention and (b) relation-aware attention module.

The design of our ARF module is inspired by the transformer architecture, whose multi-head attention mechanism recently is proved to have a strong expressive ability for capturing long-term dependencies from sequence data. We leverage this advantage to model relationships not only between appearance features and geometry features but also between geometry features, and based on the relation modeling to select the significant cross-modal cues and spatial structure cues. In the fusion module, we first obtain structure-aware geometry features and then relation-aware cross-modal features.

Structure-aware Attention. We first use an attention module to build long-term dependencies among points. To gather multi-scale point features, we up-sample multiple resolution point features to the same resolution and concatenate these point features, where we use F_g to represent point features. A shared MLP is used to compress the feature dimension of F_g into a fixed one. As depicted in Fig. 2(a), the attention module takes the F_g as input, which are projected with linear operations to produce query, key and value of the multi-head attention module, i.e. $Q_g^i = F_g \cdot W_{Q_g}^i, K_g^i = F_g \cdot W_{K_g}^i, V_g^i = F_g \cdot W_{V_g}^i$, where i is the index of attention heads, $W_{Q_g}^i, {K_g}^i$ and $W_{V_g}^i$ are all $\in R^{d \times d}$.

$$\hat{F}_g^i = softmax(Q_g^i \cdot (K_g^i)^T / \sqrt{d}) \cdot V_g^i \tag{1}$$

$$\hat{F}_g = F_g + concat(\hat{F}_g^1, \hat{F}_g^2, \ldots, \hat{F}_g^M) \tag{2}$$

Here M is the number of attention heads. In the multi-head attention module, the attention operation is calculated in parallel in multiple heads. In each head, the attention map is computed between every local feature pair in the projected embedding space. Features of each point are aggregated by weighting features of all points through the attention weights. Such operation can provide more structure information for the object instance. After obtaining all the point features from each attention head, we concatenate them and add them to the input point features to obtain the structure aware geometry features.

Relation-aware Attention. In order to model correlation between RGB features and point features, we use another attention module to explore relations between the two representation spaces. Since the RGBD images are well-aligned, we can easily retrieve pixel-wise RGB features for each 3D point. For each point, we first gather its corresponding RGB features from each scale and concatenate them into multi-scale RGB features. Then we also apply a shared MLP to compress it to the same feature dimension as the point features. We use F_r to represent the RGB features. As depicted in Fig. 2(b), the attention operation takes F_r as the key and value, and $\hat{F}_g$ enhanced point features by structure-aware attention as the query. The multi-modal fused features $\widetilde{F}_g$ are computed similar as (1) and (2).

The learned attentions indicate relation between the appearance features and the geometry features. Intuitively, higher correlation means more contribution of the corresponding appearance feature to a certain point. Therefore, we take the learned relevance as a guidance to highlight beneficial appearance information for fusion. The relation-aware attention module uses an addition operation to perform fusion of the adaptively aggregated appearance features to the former geometry features to obtain the multi-modal fused features $\widetilde{F}_g$. Then, the feedforward network (FFN) composed of one-layer MLP and one linear layer is employed to get the complete multi-modal features $\widetilde{F}_g = \widetilde{F}_g + FFN(\widetilde{F}_g)$.

In our implementation, we resort to two multi-attention modules in our attention-guided RGB-D fusion (ARF) module to extract robust 3D features from both point and RGB features. Through this RGBD fusion mechanism, we not only enhance the geometry features with rich semantic appearance features, but also explore global structure information. In this way, the network can leverage local and global multi-modal information to improve geometry representation learning for accurate pose estimation. Moreover, with sequential multiple ARF modules, point features can be enhanced in a progressive way.

C. Correspondence based Pose Estimation

So far we have obtained dense multi-modal features from fusion module. The fused features are fed into a dense correspondences decoder which aims to recover the 3D canonical coordinates corresponding to input instance points. After that, dense 3D-3D correspondences are built for each instance. The point decoder module is a shared MLP module for predicting point-wise correspondences. Finally, RANSAC-based Umeyama algorithm [19] can be adopted to recover object 6D pose and size based on the built 3D-3D dense correspondences.

IV. Experiments

A. Datasets

We evaluate our method on the benchmark of NOCS dataset [8] for category-level 6D object pose and size estimation. The dataset contains 6 table-scale object categories including bottle, bowl, can, camera, laptop and mug. Two subsets CAMERA25 and REAL275 are available. CAMERA25 is a synthetic dataset generated by a context-aware mixed reality approach which includes 300,000 composite images of 1,085 object instances, among which 25,000 images of 184 instances are used for evaluation. REAL275 is a more challenging real-world dataset captured with clutter, occlusion and various lighting conditions whose training set contains 4,300 images of 7 scenes and the test set contains 2,750 images of 6 scenes. For each of the training and testing sets, 6 unique object instances per category are used. It is one of the most comprehensive datasets for category-level 6D pose and size estimation.

B. Evaluation Metrics

For category-level pose estimation, we report mean Average Precision (mAP) of 3D intersection over union (IoU$_x$) at different thresholds for 3D object detection and mAP at different rotation and translation thresholds (n°m cm) for 6D pose estimation following the same metrics as [8].

IoU$_x$ It computes the overlap between predicted 3D bounding boxes and the ground truth 3D bounding boxes respectively. If the ratio of overlapping is larger than a specified ratio, the prediction is judged to be correct. We report mAP of 3D IoU at 50 and 75 ratios.

***n*° *m* cm** This metric computes the rotation and translation errors between the predicted pose and the ground truth pose. If the rotation error is smaller than the angle threshold n degree and the translation error is smaller than a distance threshold m cm, the prediction is judged to be correct. We use 5 and 10 for angle thresholds and 2 and 5 for translation thresholds for this metric.

TABLE I. Comparison of our method with state-of-the-art methods on CAMERA25 and REAL275 benchmarks. '-' denotes no results are reported under this metric.

Method	CAMERA25						REAL275					
	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
NOCS	83.9	69.5	32.3	40.9	48.2	64.6	78.0	30.1	7.2	10.0	13.8	25.1
CASS	-	-	-	-	-	-	77.7	-	-	23.5	-	58.0
SPD	93.2	83.1	54.3	59.0	73.3	81.5	77.3	53.2	19.3	21.4	43.2	54.1
FS-Net	-	-	-	-	-	-	92.2	63.5	-	28.2	-	60.8
DualPoseNet	92.4	86.4	64.7	70.7	77.2	84.7	79.8	62.2	29.3	35.9	50.0	66.8
CFN	93.8	88.0	72.0	76.4	81.0	87.7	79.3	55.9	27.8	34.3	47.2	60.8
SGPA	93.2	88.1	70.7	74.5	82.7	88.4	80.1	61.9	35.9	39.6	61.3	70.7
Ours	**93.8**	**90.4**	**77.1**	**80.9**	**85.2**	**90.4**	**83.2**	**73.1**	**40.9**	**45.6**	**66.2**	**75.7**

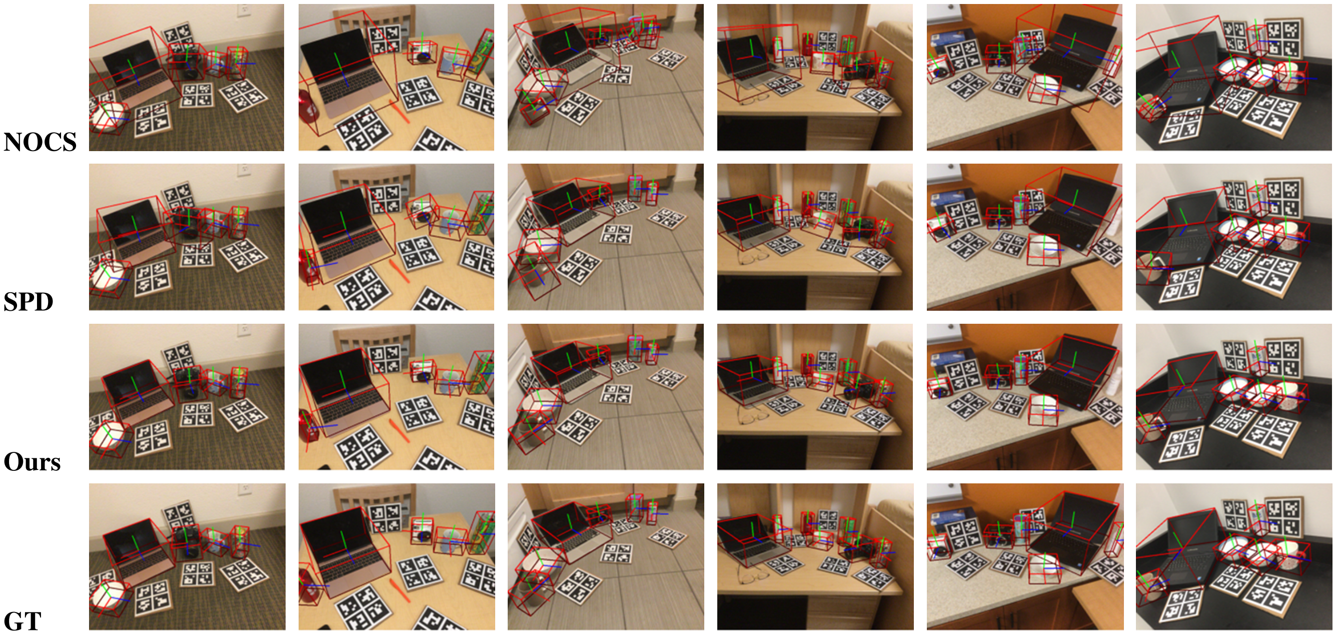


Fig. 3. Qualitative comparison of our method with NOCS, SPD on REAL275 dataset.

C. Implementation Details

Most previous works of category-level pose estimation decoupled instance segmentation and object pose estimation. Following this scheme, we use the target object masks estimated from Mask R-CNN [34] to crop RGB-D image as SPD [16], and recover instance point cloud data using camera intrinsic parameters. For each detected object, its RGB crop is scaled to 192x192 and its point cloud is sampled to 1024 points. RGB encoder is a ResNet18 [35] pretrained with ImageNet [36] whose network structure is similar with SPD [16]. In order to make a fair comparison with existing methods, we choose PointNet++ [37] as the point encoder. For the ARF modules, we use 4 heads and take 512-d input for each attention, and adopt a three-layer MLP with layer sizes 512, 256 and 3 for dense prediction. We employ L1 loss to train the predicted coordinates.

D. Comparison with State-of-the-Arts

We compare our method with representative state-of-the-art methods, including NOCS [8], CASS [32], SPD [16], FS-Net [38], DualPoseNet [33], CFN [17] and SGPA [18].

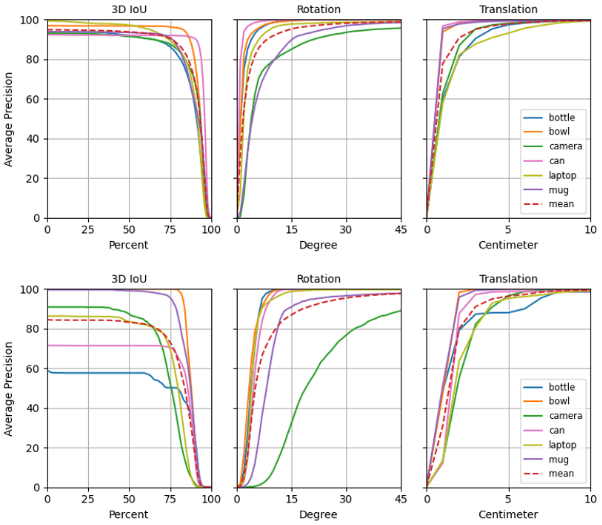


Fig. 4. Average precision results of different thresholds with 3D IoU, rotation error, and translation error on CAMERA25 (1^{st} row) and REAL275 dataset (2^{nd} row).

TABLE II. Evaluation of our fusion module with three 3d representation exactors of point cloud, including pointnet++, dgcnn and 3dgcn.

Method	5°2cm	5°5cm	10°5cm
Baseline(Pointnet++)	13.9	16.8	47.7
ARF-Net(Pointnet++)	40.5	45.6	75.7
Baseline(DGCNN)	14.0	16.9	48.3
ARF-Net(DGCNN)	45.0	50.3	76.9
Baseline(3DGCN)	35.5	42.8	72.2
ARF-Net(3DGCN)	48.2	53.9	78.0

TABLE III. Evaluation of the advantage of our ARF modules on REAL275 dataset.

Method	5°2cm	5°5cm	10°5cm
Baseline (BL)	13.9	16.8	47.7
Concatenation	23.7	25.2	56.4
DenseFusion	28.5	30.7	65.9
1x ARF	31.2	35.2	68.1
2x ARFs	35.7	40.0	72.4
3x ARFs	40.9	45.6	75.7
4x ARFs	41.0	46.1	75.5

Among these methods, DualPoseNet, FS-Net and CASS directly regress 6D pose and size, while the others first predict NOCS coordinates and later estimate poses and size by similarity transformation. Table I lists the comparative results on the CAMERA25 and REAL275 datasets. Quantitative results in Table I show the superiority of our proposed network on both datasets. The mAP of 10°5cm on REAL275 dataset reaches 75.7% which is 5% higher than the SOTA method SGPA. We also achieve 45.6% on 5°5cm, 40.9% on 5°2cm and 66.2% on 10°2cm which are all improved nearly 5%. In terms of IoU_{75}, ARF-Net outperforms SGPA by 11.2%. Notice that our ARF-Net can also gain a large improvement over CFN[17] which also uses an attention-based fusion. One reason should be that our model not only captures multi-modal cues, but also leverages spatial information among object parts which are essential for determining object poses. Therefore, the superiority of our method is obvious on the REAL275 dataset. Specifically, both the performance gains of the 5°2cm and 5°5cm metrics on CAMERA25 dataset are more than 4%. These comparative results demonstrate that our ARF-Net outperforms previous methods by large margins under all the evaluation metrics. Fig. 3 presents a qualitative comparison of NOCS, SPD and our method on REAL275 dataset. Visual examples in the figure demonstrate that the accuracy improvements of our method are consistent with those in Table I. Fig. 4 further illustrates detailed error evaluation results for each category on the two datasets, which shows that we can obtain very accurate pose results for most categories. Compared with previous methods, the performance on camera is also well improved.

E. Ablation Studies

In this section, we conduct the following ablation studies to justify the design of our ARF-Net. The studies are conducted on REAL275 dataset.

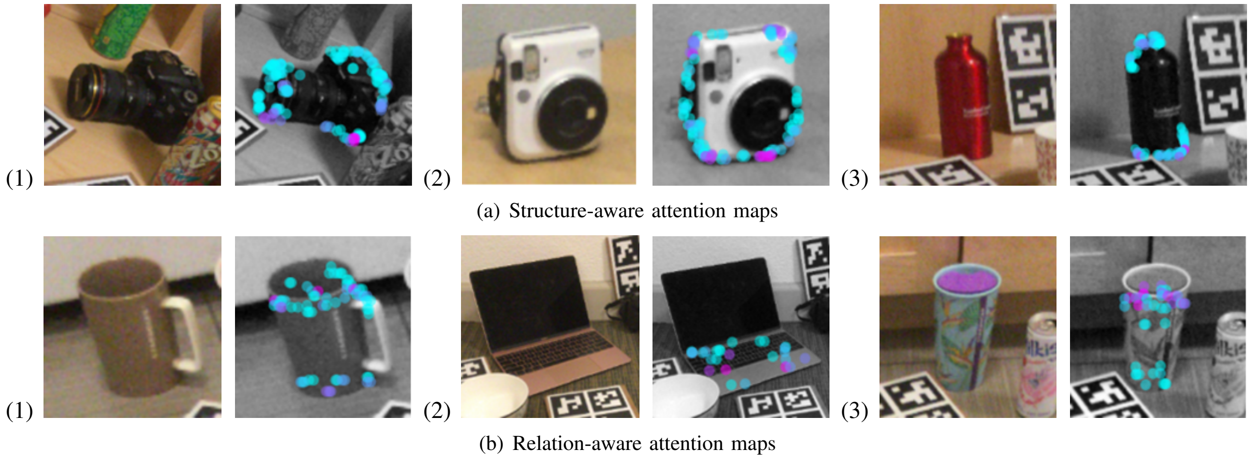
TABLE IV. Evaluation of different attention designs for fusion on REAL275 dataset.

Method	5°2cm	5°5cm	10°5cm
Baseline (BL)	13.9	16.8	47.7
Only SA	37.7	42.1	68.7
Only RA	23.4	27.5	62.3
RA+SA	39.1	43.3	73.9
SA+RA(ours)	40.9	45.6	75.7

To illustrate the effectiveness and reliability of the fusion method, we compare the performance of three commonly used 3D point cloud feature extractors under our fusion framework, including Pointnet++ [37], DGCNN [39], and 3DGCN [40]. Our baseline framework is a single modal network with the point cloud encoder and the shape decoder. The ARF-Net adds the RGB feature encoder and the ARF modules into the baseline for prediction. We report the experimental results in Table II. Compared with the baselines, different point cloud extractors with our ARF fusion operations can achieve consistent performance improvements on REAL275 dataset. Specifically, the ARF-Net built on PointNet++ improves the mAP of 5°2cm, 5°5cm and 10°5cm from 13.9%, 16.8% and 47.7% to 40.5%, 45.6% and 75.7%. Both the ARF-Nets built on DGCNN and 3DGCN achieve higher than 50% and 75% on 5°5cm and 10°5cm, respectively, which are much better than previous methods and also have large performance gains compared with their baselines. The experimental results demonstrate that through the guidance of attentions for fusing RGB and point features, the network can leverage distinctive multi-modal features which lead to higher pose accuracies.

To demonstrate the effectiveness and superiority of our method, we compare the performance under several fusion methods including direct concatenation of multi-scale RGB features and multi-scale point features, dense fusion method also used in CASS [32] and our ARF modules. We also investigate the effect of different choices of the ARF number on the pose accuracy. In Table III, we gradually increase the number of ARF modules from 1 to 4. Compared with the baseline, fusing RGB-D features can obviously obtain better prediction accuracies. The performance of one ARF module is better than the commonly used fusion mechanisms and outperforms many previous methods. Moreover, progressively utilizing more ARF modules will bring better fusion features. Increasing the ARF numbers to three can bring continuous performance improvement of all the metrics while four ARF modules could not bring accuracy improvement on some pose metrics. Therefore, three ARF modules are enough to learn a good embedding from multi-modal representations.

In Table IV, we quantitatively evaluate different choices of the two attentions to further illustrate the effectiveness of our attention design components. The result of ‘Only SA’ is implemented by adding three ARF modules without relation-aware attention to the baseline. The ‘Only RA’ is a fusion network without structure-aware attention. The ‘RA+SA’ first uses relation-aware attention and then structure-aware atten-



(a) Structure-aware attention maps

(b) Relation-aware attention maps

Fig. 5. Visualization of the learned structure-aware and relation-aware attention maps by our ARF module. The color hue varies from blue to red that corresponds to attention value from small to large (best viewed in color print). The structure-aware attention map indicates the significance of different object parts for the task. The relation-aware attention map indicates the correlations between the RGB features and depth features that are used to guide RGB-depth fusion. As shown in subfigure (b), the relation-aware attentions successfully avoid the instance-specific logos and textures printed on the cans, and locate on the category-shared regions on shape contours.

tion. Among different attention designs, our ARF modules can achieve best performance on each pose metrics.

The experimental results not only reflect the advanced nature of our method, but also illustrate the importance of the multimodal information fusion. Moreover, with a careful design, networks can directly reconstruct an accurate canonical shape without any shape prior information.

Visualization of Attention maps. In Fig. 5, we visualize the structure-aware and relation-aware attention maps learned by our ARF module. The attention maps are obtained by projecting each point's attention values to the image for intuitive visualization.

In the structure-aware attention map, the attention value is computed by aggregating the positions of top-5 ranked retrieved keys for all the queries in the structure-aware attention module. In this way, the map indicates the significance of contributions from different object parts. As shown in Fig. 5, the attention value is color-coded increasingly from blue to red and the top value attentions tend to be located at regions of 3D shape variations that are characteristic to provide useful cues to infer object pose. It is notable that these parts are mostly common parts shared through different instances of the category such as the camera body and circular rim of lens. In the second image in Fig. 5(a), since the camera's 3D shape does not include a long lens, the structure-aware attention focuses on points in the camera's body part that is shared by all instances in the camera category.

Computed in a similar way as the structure-aware attention map, the relation-aware attention map indicates the correlations between the RGB features and depth features. It can be seen that the top value relation-aware attentions tend to be located at regions where RGB and depth features have correlations and such corrections are category-shared. As shown in the third image in Fig. 5(b), the attentions successfully avoid the colorful logos and textures printed on the can, which only belong to this can and not shared by the category.

The above visualization demonstrates that our ARF modules learn meaningful attentions that are useful for the category-level object pose estimation task, in particular, with category-shared properties that can generalize to unseen object instances.

V. CONCLUSIONS

In this work, we present a novel attention-guided RGB-D fusion Network for category-level 6D object pose estimation. We propose ARF modules that learn to adaptively fuse RGB and depth features with guidance from both structure-aware attentions and relation-aware attentions. Based on the attention guidance, the network is capable of capturing spatial relationship among object parts and RGB-to-depth correlations between the appearance and geometric features. In particular, our ARF-Net directly establishes canonical correspondences with a compact decoder based on the multimodal features learned from the ARF modules. Extensive experiments on a challenging public benchmark demonstrate that our method is superior to previous methods. Experiments also show that our method can effectively fuse RGB features to various popular point cloud encoders.

Although we have shown improved performance in this task, there still exist some issues for future works. For example, severe partiality of objects caused by heavy occlusion or low quality images caused by motion blur or over-exposing may bring challengs to existed RGBD methods. We will continue to investigate efficient and robust methods for real-world scenarios.

References

[1] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 627–12 637.

[2] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.

[3] P. Cipresso, I. A. C. Giglioli, M. A. Raya, and G. Riva, "The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature," *Frontiers in psychology*, p. 2086, 2018.

[4] M. Gattullo, G. W. Scurati, M. Fiorentino, A. E. Uva, F. Ferrise, and M. Bordegoni, "Towards augmented reality manuals for industry 4.0: A methodology," *Robotics and Computer-Integrated Manufacturing*, vol. 56, pp. 276–286, 2019.

[5] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, and J. He, "Deep learning on monocular object pose detection and tracking: A comprehensive overview," *arXiv preprint arXiv:2105.14291*, 2021.

[6] A. Krull, E. Brachmann, F. Michel, M. Y. Yang, S. Gumhold, and C. Rother, "Learning analysis-by-synthesis for 6d pose estimation in rgb-d images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 954–962.

[7] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[8] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.

[9] A. Ahmadyan, L. Zhang, A. Ablavatski, J. Wei, and M. Grundmann, "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7822–7831.

[10] K. Chen, Y.-K. Lai, and S.-M. Hu, "3d indoor scene modeling from rgb-d data: a survey," *Computational Visual Media*, vol. 1, no. 4, pp. 267–278, 2015.

[11] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[12] Y. Wang, C. Wang, P. Long, Y. Gu, and W. Li, "Recent advances in 3d object detection based on rgb-d: A survey," *Displays*, vol. 70, p. 102077, 2021.

[13] Y. Hu, Z. Chen, and W. Lin, "Rgb-d semantic segmentation: a review," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 1–6.

[14] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.

[15] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "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, pp. 3003–3013.

[16] M. Tian, M. H. Ang, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 530–546.

[17] J. Wang, K. Chen, and Q. Dou, "Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4807–4814.

[18] K. Chen and Q. Dou, "Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2773–2782.

[19] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.

[20] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 611–16 621.

[21] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4561–4570.

[22] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7668–7677.

[23] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "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, pp. 11 632–11 641.

[24] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 710–10 719.

[25] G. Wang, F. Manhardt, J. Shao, X. Ji, N. Navab, and F. Tombari, "Self6d: Self-supervised monocular 6d object pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 108–125.

[26] Z. Yang, X. Yu, and Y. Yang, "Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3907–3916.

[27] F. Manhardt, G. Wang, B. Busam, M. Nickel, S. Meier, L. Minciullo, X. Ji, and N. Navab, "Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning," *arXiv preprint arXiv:2003.05848*, 2020.

[28] S. Iwase, X. Liu, R. Khirodkar, R. Yokota, and K. M. Kitani, "Repose: Fast 6d object pose refinement via deep texture rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3303–3312.

[29] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 683–698.

[30] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 574–591.

[31] L. Zou, Z. Huang, F. Wang, Z. Yang, and G. Wang, "Cma: Cross-modal attention for 6d object pose estimation," *Computers & Graphics*, vol. 97, pp. 139–147, 2021.

[32] D. Chen, J. Li, Z. Wang, and K. Xu, "Learning canonical shape space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 973–11 982.

[33] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3560–3569.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[37] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[38] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis, "Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1581–1590.

[39] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[40] Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, "Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1800–1809.