First Author, Second Author, Third Author *et al.* Journal of computer science and technology: Instruction for authors. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 33(1): 1–14 January 2018. DOI 10.1007/s11390-015-0000-0

6D Object Pose Estimation in Cluttered Scenes from RGB Images

Xiaolong Yang[1,2], Xiaohong Jia[1,2*], Yuan Liang[3], and Lubin Fan[3]

[1] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

[2] *University of Chinese Academy of Sciences, Beijing 100049, China*

[3] *Alibaba Group, Hangzhou 311121, China*

E-mail: yangxiaolong17@mails.ucas.ac.cn; xhjia@amss.ac.cn; liangyuan.ly@alibaba-inc.com; lubin.flb@alibaba-inc.com

Received July 15, 2018 [Month Day, Year]; accepted October 14, 2018 [Month Day, Year].

Abstract We propose a feature-fusion network for pose estimation directly from RGB images without any depth information in this study. First, we introduce a two-stream architecture that consists of segmentation stream and regression stream. The segmentation stream is used to process the spatial embedding feature and obtain the corresponding image crop. These features are further coupled with image crop in the fusion network. We use an efficient Perspective-n-Point (E-PnP) algorithm in the regression stream to extract robust spatial features between 3D and 2D keypoints. Finally, we also perform iterative refinement with end-to-end mechanism, which can further improve the estimation performance . We conduct experiments on two public datasets of YCB-Video and the challenging Occluded-LineMOD. Our method outperforms state-of-the-art approaches in both speed and accuracy.

Keywords two-stream network, 6D pose estimation, fusion feature

1 Introduction

The 6D object pose estimation has been widely used in computer vision tasks in daily life, such as robot grasping and manipulation, autonomous navigation and augmented/mixed reality, with the continuous development of sensor technology. High requirements are put forward for the following applications for various real-world needs: (1) proper handling of objects with irregular shapes , low-resolution textures, and different materials; (2) robustness to heavy occlusion, lighting changes in various environments, and potential technical noise; (3) real-time speed as possible.

Many RGB-D algorithms based on depth information can infer the pose relatively accurately when texture features of the object is blurred or the scene is blocked by other objects. The traditional method extracts features from RGB-D data, combines them with depth masks, and performs verification and inference [1, 2, 3, 4, 5, 6, 7, 8]. However, reliance on manual functions and fixed matching procedures limits the empirical performance in the case of severe occlusion and lighting changes. RGB-D-based algorithms are generally restricted applications that require the hardware of the RGB-D sensor. Hence, the majority of the existing RGB-D-based methods present difficulty in meeting

Regular Paper

This work was partially supported by the National Natural Science Foundation of China (12022117 and 61802406), the Beijing Natural Science Foundation (Z190004), the Beijing Advanced Discipline Fund (115200S001), and Alibaba Group through Alibaba Innovative Research Program.

*Corresponding Author

requirements of accurate pose estimation and fast inference [9].

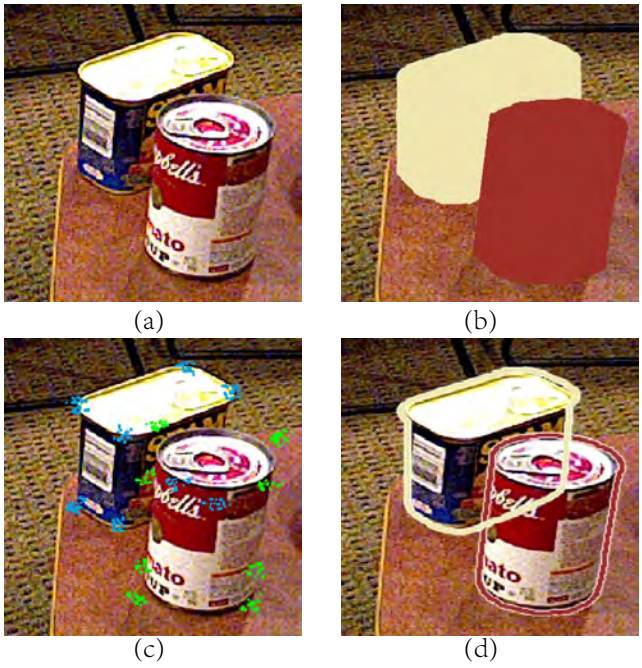


Fig.1. Example of pose estimation using our approach. (a) Input image with occluded known objects. (b) Results of object detection and segmentation. (c) Locations of keypoints. (d) Final results of the pose estimation.

We focus on methods of image-based 2D recognition of object poses due to increasing demands of source acquisition and the limitation of depth sensors. Traditional RGB-based methods rely on the correspondence between 3D and 2D keypoints. First, find unified feature points through the known 3D model and 2D pixels of the object and then calculate regression parameters of rotation and translation for further processing using perspective-n-point (PnP) algorithm. This approach is robust when the object has fruitful textures. However, it might fail if the texture becomes blurred or multiple objects block one other [12, 13, 17, 33]. Recent approaches have attempted to overcome these difficulties that usually apply deep neutral networks to return 6D poses from images directly or detect keypoints related to objects. However, both situations endow the object with the property of the global entity, which makes the procedure susceptible to heavier occlusion.

In this paper, we propose a novel method for generating high-quality real-time pose estimation of RGB images to address this occluded problem. The key idea is to build a two-stream architecture, which includes a segmentation stream to process various features of the input image and a regression stream to solve the 2D-3D correspondences via bounding corners. Different from existing methods that use image segmentation [10] and bounding boxes [11] to calculate global features separately, we further propose a fusion network for transforming and mapping RGB textures and location relationship at the per-pixel level. Abundant features ensure that our model considers both texture and geometric information of the object. Moreover, we design an iterative refinement process with end-to-end mechanism; that remarkably improves the accuracy of estimated poses. An example of pose estimation using our approach is shown in Fig. 1.

We conduct various comparisons to demonstrate the superiority of our approach over many representative competitors in terms of pose estimation performance and efficiency. Our method present superior performance on popular YCB-Video [12] and Occluded-LineMOD [13] datasets. Notably, our method still produces reliable estimation results when dealing with challenging samples in heavily cluttered scenes. Finally, we illustrate benefits of the iterative refinement process through an ablation study. This paper extends our recent ACM SIGGRAPH publication [14], and shows the following additional contributions:

- Additional qualitative and quantitative experiments to prove the superiority of our pose estimation method;
- Improved loss function and other elaborate descriptions with further advanced performance;

- Extended application: advertising replacement and wall decoration suggestion.

2 Related Work

We briefly discuss two relevant types of work based on the input, that is, RGB images with or without depth.

2.1 Pose estimation based on RGB-D images

Recent methods are typically data-driven. Song *et al* [15, 16] enhanced 3D object detection through discretization of 3D voxel spaces. The researchers proposed 3D BBox for the voxelized 3D model input and estimated poses. Although geometric information is encoded effectively, these time-consuming and space-expensive methods based on the voxel representation take nearly 20 seconds and 300M+ memory for a single model frame in [16].

Other methods directly use 3D deep learning architectures from 3D point cloud data and perform 6-DoF detection pose estimation. On the basis of the pioneering work of PointNet [17], both Frustrum [11] and VoxelNet [10] have demonstrated considerable progress in point cloud detection and surpassed a large number of competitors on the KITTI benchmark [18].

Visual-recognition deep neutral networks also perform pose estimation on the basis of RGB-D input. These methods require detailed steps of feature extraction and training to optimize 3D prior labels. For instance, PoseCNN [12] applies closest point (ICP) process iteratively, and DenseFusion [9] uses RGB-D data color and depth information of pixels.

The acquisition and widespread use of these methods are obstacles to real-world applications bacause they require many sources with depth information. By comparison, our method relies only on RGB images to achieve high-quality pose recognition and real-time performance.

2.2 Pose estimation based on RGB images

Existing methods based on RGB images are mainly divided into two categories. Traditional methods depend on keypoint detection and matching relationship with known object models [19, 20, 21, 22, 23, 24]. Other methods, which use learn mechanism to predict 2D keypoints, are presented to solve poses via the PnP algorithm [28] and address this challenge [1, 25, 26, 27]. Although excellent and efficient in some tasks that require speed performance, the robustness and accuracy of these methods reduce when faced with low-texture or low-resolution input images. Additional studies have focused on directly using CNN-based architectures for pose estimation due to deep learning [29, 30]. These studies mainly obtain different forms of orientation estimation based on RGB images. Xiang *et al* [12] learned and clustered 3D features of the object model; and then extracted the orientation information via a viewpoint-aware strategy. Mousavian *et al* [31] attempted to obtain geometric constraints and exploit 3D object parameters to restore poses While concentrating on a single-view. These methods present poor accuracy under low-texture or low-resolution input despite their important advantages in speed performance. Therefore, other technologies focus on the use of powerful deep learning based on CNN architectures to obtain target pose estimation directly from a single image [29, 30], and constantly concentrate on the prediction of orientation information. Xiang *et al* [12] attempted to learn 3D features, cluster object models, and obtain perceptual predictions on the basis of viewpoints. Mousavian *et al* [31] optimized the single view to extract potential geometric constraints, recover 3D parameters, and achieve pose estimation. Sundermeyer *et al* [32] proposed a novel encoding method to convert orientation

information into feature vectors implicitly; and then determine the optimal match for the test object frame.

Many methods focus on local information to solve the problem of occlusion and improve accuracy performance. Seg-Driven [33] crops the input image in different sizes, and the final result is weighed by combining multiple local pose estimations. SilhoNet [34] continuously obtains the direction information using contour scanning mechanism. Pix2Pose [35] recovers the correspondence between 3D and 2D from the pixel level through GAN generation technology. Although these studies have achieved excellent experimental design and acceleration network, they still exhibit poor accuracy. ACC [36] attempted to address this problem by reconstructing the mesh of the entire original model on the basis of the optimized DCNN model [37, 38]. However, this method can only be applied to a single target and demonstrates poor generalization in challenging situations.

Our method is inspired by DenseFusion [9], in which proposed a heterogeneous architecture and jointly considered of potential features of geometry and texture. However, the fusion method in our approach is used for spatial coordinates and object appearance textures without depth information. We show that our novel feature extraction and mapping scheme outperforms only PnP or image cropping-based algorithms in this study. Furthermore, we introduce a general end-to-end refinement strategy to enhance our performance.

3 Methodology

In this paper, we address challenges with high occlusion or under barren light with a heterogeneous network (Sec. 3.2) that processes texture and location information differently on the basis of DenseFusion [9]. We also design a new and completely independent coupling method (Sec. 3.3). Moreover, we build a two-stream network based on image cropping and bounding box through an encoding decoding process; that shares some similarities with Seg-Driven [33] in segmentation processing (Sec. 3.4). Finally, the estimator of poses will be further optimized with a differentiable module (Sec. 3.5). Compared the previous expensive optimization procedures [39, 12], our module accounts for only a fraction of the total reasoning time.

3.1 Architecture

Fig. 2 depicts the overall workflow of our method. The segmentation stream consists of the following steps: Conduct semantic segmentation and crop in each input frame. The existing technology can achieve excellent performance for known objects. We then feed segmentation labels for each segmented local RGB patch with the bounding box of each cropped image and the entire original image with different sizes to the fusion network, which couples spatial features after segmentation with texture features of the image itself.

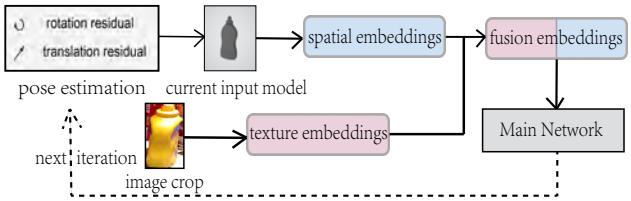


Fig. 3. Refinement procedure with the end-to-end mechanism. This additional iterative optimization directly follows the main network.

The regression stream performs an efficient PnP (EPnP) solution on spatial channels (XYZ) of the obtained fusion feature, and returned 2D-3D correspondences (rotation R and translation t) and texture channels (RGB) jointly participate in confidence calculation. We can treat these correspondences as rough pose estimates according to the magnitude of the confidence and design a subsequent refinement iteration for performance improvement, as illustrated in Fig. 3. Details

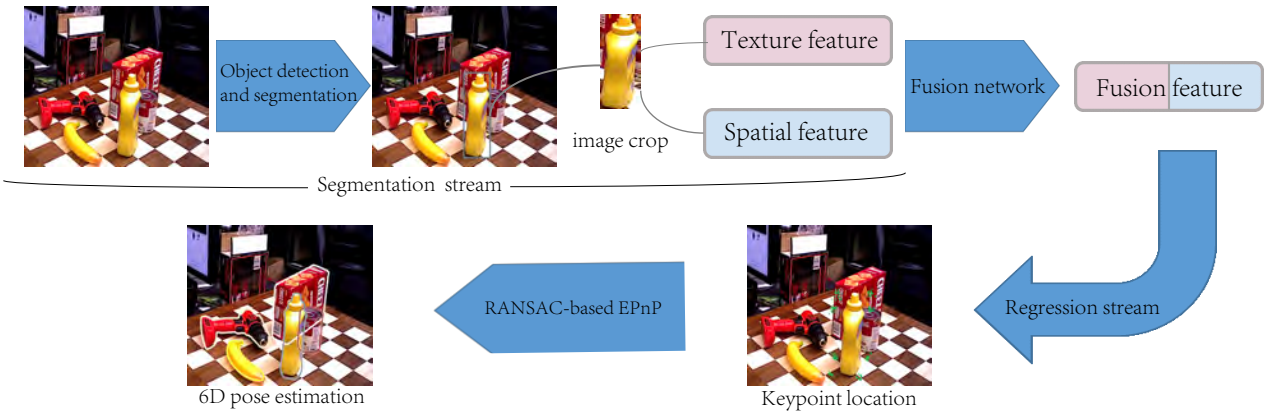


Fig.2. Proposed network architecture. A two-stream architecture network is designed for feature processing, including segmentation and regression. A fusion network can combine texture and spatial features also exists.

are described in the following section.

3.2 Segmentation Stream

We use classic backbone, which consists of two streams of segmentation and regression. We use the highly effective and efficient YOLOv3 [40] and YOLOv4 [41] for initial object detection and obtain many areas in the segmentation. We then crop each detected area on the original input image in the segmentation stream. Overlapping may occur in divided parts due to the occlusion between objects. Furthermore, we label every unit of the $n \times n$ grid with object or background information. Specifically, we can generate accurate semantic labels using the real 3D model and the corresponding depth information originally contained in datasets when training the model to greatly reduce the impact of occlusion on the image. The uneven distribution of positive and negative samples in practice is due to the significantly smaller area occupied by the investigated object than the background. Therefore, we utilize Focal Loss [42] to address negative effects of sample distribution on model training. Moreover, we apply a pixel-wise median frequency balancing technique [38] because target objects demonstrate different sizes and changing the ratio of cropped images will affect the performance of the pose estimation task.

3.3 Fusion Network

We will fuse individual multi-dimensional information for different detected objects with various sizes. The existing method DenseFusion [9] also considers the fusion of features, and links the dense depth and color features on cropped images through a multilayer perceptron (MLP) to form a new global feature. However, this method retains unnecessary errors generated by the feature extraction in the previous step, such as pixel information of other objects caused by occlusion and overlapping or background information due to segmentation accuracy. Therefore, undesigned fusion features will reduce the performance of pose estimation. A novel pixel-wise [9] strategy is used in our implementation to achieve an efficient fusion network, especially under heavy occlusion and imperfect segmentation situations.

The proposed fusion network based on PointNet [17] supplements texture features that correspond to spatial features on the processing object. Our fusion network performs local fusion and prediction on each $n \times n$ grid. Hence, we design different prediction weights according to visible and invisible parts of targets to minimize and decrease the negative influence of heavy occlusion and

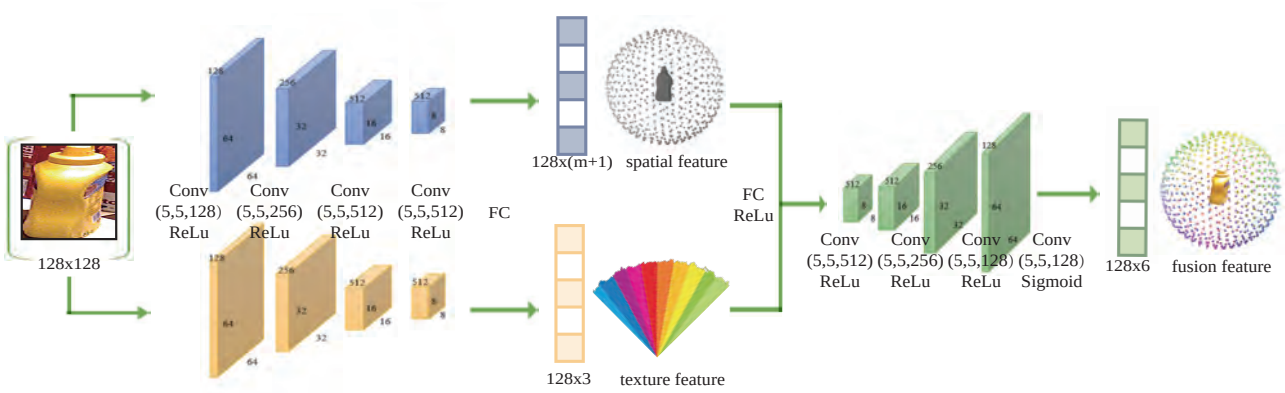


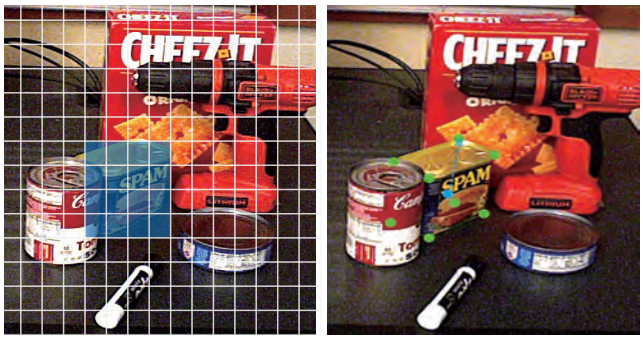
Fig.4. Architecture of the fusion network. This detailed flowchart explains the process of extracting spatial and texture information on the basis of heterogeneous CNN, and then feeds them into another one to obtain fusion features. **m+1** refers to the $m+1$-channelled semantic segmentation map in Sec.3.2 and **128** includes XYZ and RGB.

imperfect segmented noises.

We use the projection transformation to couple the spatial features and texture features on each grid, based on the real 3D model, ground-truth pose labels, and uv texture contained in original datasets combined with known internal camera parameters. We then expect to generate dense feature vectors with uniform size after passing the fusion network. Fig. 4 shows the architecture of the fusion network in detail. Two features results with different dimensions are produced when the cropped image, which is resized to 128×128, passes through a heterogeneous neural network. The upper network will obtain $m+1$-channelled spatial information based on background and occlusion whereas the lower network will obtain texture information based on texture and appearance, and then couple them to obtain 128×6 fusion features, including XYZ and RGB intrinsic properties.

3.4 Regression Stream

Regression stream plays an important role in using the efficient PnP algorithm to obtain spatial features from the feature vector and link the position information of 2D and 3D key points. Following Seg-Driven [33], we set key points to eight vertices of bboxes of each target. We focus on anchor point of the center in the bounding cube and calculate offset vector and texture information deviation of center and corner points for every vertex:



(a) (b)

Fig.5. Simple example of our seg-reg two-stream network. (a) Segmentation stream labels every grid of the cropped image. (b) Offset vector and texture deviation of the center and corner points in the regression stream.

C denotes as the 2D coordinate of each anchor center point. An offset estimation $f_i(C)$ is proposed for the i^{th} vertex. Hence, we can express the vertex as $C + f_i(C)$ with the precise 2D coordinate C_i^{GT} from the original dataset. The texture of anchor point T , a corresponding offset $f_i(T)$ and precise texture T_i^{GT} are expressed as follows:

$$\Delta_i(C) = C + f_i(C) - C_i^{GT}, \qquad \Delta_i(T) = T + f_i(T) - T_i^{GT}. \tag{1}$$

The novel loss function is expressed as follows:

$$E_{pos} = \sum_{Grid} \sum_{i=1}^{n} \|\Delta_i(C)\|_1 + \|\Delta_i(T)\|_1 . \quad (2)$$

Here, we use the L_1 loss function instead of the L_2 one, because the former is less sensitive with ill-posed samples than the latter. We set equal weights because both spatial (x, y, z) and texture (r, g, b) features present the same dimensions.

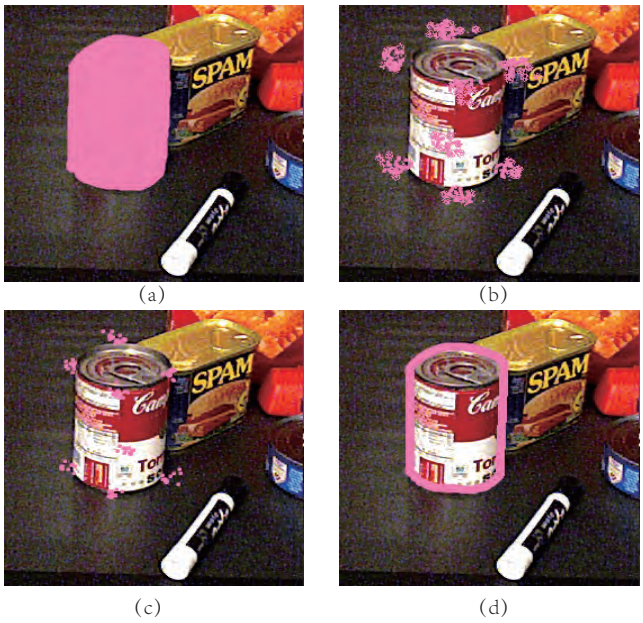


Fig.6. Process of regression stream. (a) Detected Coke can using the segmentation stream. (b) EPnP algorithm will generate many series of pose estimations, shown as pink dots. (c) We select the $n = 10$ group with the maximum confidence for RANSAC repetition. (d) Contour result after iterative optimization.

A sigmoid-based function is used as the activation function to produce the confidence Con_i for any group of EPnP algorithm results.

$$\mathcal{L}_i = \frac{1}{8} \sum_{n} \| \left(R^{GT} p_n + t^{GT}\right) - \left(R_i p_n + t_i\right) \|, \quad (3)$$

where p_n denotes the selected n-th 3D point of corners from the real 3D model. However, we want to decide which pose estimation may be the optimal hypothesis because the PnP algorithm may lead to multiple sets of approximate solutions. Thus, we need to use RANSAC to balance the confidence among each prediction. The loss function to minimize becomes:

$$\mathcal{L} = \frac{1}{N} \sum_{i} \left(\mathcal{L}_i Con_i - w \log\left(c_i\right)\right), \quad (4)$$

where Con_i represents the score of approximation between our pose estimation and the true three-dimensional transformation. We then design a new loss penalty term as follows:

$$E_{pro} = \sum_{Grid} \sum_{i=1}^{n} \| -exp(-\Theta \left\|\delta_i(C) + \delta_i(T)\right\|_2) + Con_i \|_1 , \quad (5)$$

where Θ is a trade-off parameter. Finally, the regression loss becomes:

$$E = \omega_{pos} E_{pos} + \omega_{pro} E_{pro}, \quad (6)$$

where ω_{pos} and ω_{pro} weight the effects of these two losses.

The large number of result candidates generated by iterative algorithm for every target will consume unnecessary computing resources. Hence, we use the EPnP algorithm with RANSAC mechanism ($n = 10$) [43] to obtain the 6D pose with consideration for efficiency. Fig. 6 illustrates the 2D-3D corresponding procedure between the image and the 3D model.

3.5 Iterative refinement

We propose an iterative optimal procedure with network architecture that evidently advances the pose estimation performance in a robust and effective manner.

Optimization adjustment aims to reduce the noise caused by calculation and trade-off as extensively as possible through the iterative form and improve the performance of pose estimation. New predictions from the beginning are unnecessary because we obtained a satisfactory result. However, we add a module behind the main network to polish the previous results. Therefore, we use the result of the main network as the initial value, perform refinement iterations, and then utilize

the iteration result as the new input and repeat this step. The important operation of iterative optimization aims to combine the 3D model from the original dataset with the obtained pose estimation, which presents the maximum confidence, and conducts a projection transformation to output the corresponding image. The image continues to iterate in the main network due to the segmentation stream to improve the accuracy of performance.

This module is illustrated in Fig. 3. We obtain the pose estimation after K ($K = 2$ in our experiment) iterations as follows:

$$Pos = M_{R_K,t_K} \cdot M_{R_{K-1},t_{K-1}} \cdots M_{R_1,t_1} \cdot M_{R_0,t_0}, \quad (7)$$

where M_{R_K,t_K} represents the $K-th$ Euclidean transformation, including rotation and translation. Fig. 9 intuitively shows the effect with and without iterative optimization ($K = 1$).

4 Experiment Results

We introduce the evaluation on the two popular datasets of Occluded-LINEMOD [13] and YCB-Video [12] that outperforms those of other methods. The YCB dataset consists of various models with distinguishable textures under different surroundings while presenting many rich labels. Hence, this information is a commonly used 6oF datasets in many existing methods. The Occluded-LINEMOD dataset is widely used in challenging situations with serious occlusion and messy background information. The performance of existing SOTA on this dataset is unsatisfactory.

We compare our proposed approach with the following recent advanced methods: PoseCNN [12], BB8 [46], Tekin [26], Heatmaps [47], Pix2Pose [35], SilhoNet [34] , Seg-Driven [33] , PVnet [48] and CDPN [52]. The commonly used metrci ADD-S [12] proposed by the author of the YCB-Video dataset was used in this study. We set ADD-0.1d in our expriments to adjust whether a pose estimation will be positive under the condition that the metric should be below 10% of the model diameter.

4.1 Ablation Study on Different Architectures

We investigate the effectiveness of different architectures in this section. The architecture we proposed can differ in the following aspects: 1) fusion feature, and 2) iterative refinement. We compare different features in the first part and omit the second item $\Delta_i(c)$ related to texture information in Eq. (2) (Sec. 3.4). The module that only use spatial features is similar to the Seg-Driven [33] method. Both approaches rely only on segmenting images and use a regression algorithm (e.g. EPnP) to calculate the 2D-3D correspondence. The experimental results of different modules are reported in Table 1. The fusion feature (third row) strongly promotes the improvement of performance based on only spatial features (first row). The symmetric Eggbox model with unclear texture may cause a slight drop in performance, but the average result of the fusion feature presents an increasing of more than 5% compared with the use of spatial feature alone. The iterative refinement part clearly showed that the results with iterative refinement (even rows) are constantly better than those without (odd rows). The ablation study proves the superiority of fusion feature and iterative refinement modules.

4.2 Quantitative Evaluation on Occluded-LINEMOD

We produce deep-fake images on random PASCAL VOC [51] images to find suitable training and testing objects. Fig. 7 illustrates some sampled results of our algorithm. Table 2 presents the quantitative evaluations on Occluded-LINEMOD. Our method demon-

Fig.7. Results on the Occluded-LINEMOD dataset. The entire scene is largely obscured and prensents challenging camera viewpoints.

Table 1. Ablation study on Occluded-LINEMOD. The metric is ADD-0.1d. Symmetric objects have bold-name. SF: spatial feature, TF:texture feature, IR:iterative refinement.

Methods	SF	TF	IR	Ape	Can	Cat	Driller	Duck	**Eggbox**	**Glue**	Holepun	Average
XYZ only	✓			14.9	60.1	16.5	48.2	25.1	35.6	44.4	35.9	35.1
XYZ+Refine	✓		✓	17.5	63.1	16.6	48.3	25.4	**35.8**	44.6	38.6	36.2
Fusion	✓	✓		37.4	68.6	26.0	48.5	25.0	30.3	45.2	45.1	40.8
Fusion+Refine	✓	✓	✓	**39.1**	**69.8**	**26.9**	**49.0**	**25.6**	31.9	**46.5**	**53.1**	**42.7**

strates superior performance with global inference technologies, such as PoseCNN, Tekin, BB8, and Pix2Pose. Compared with other methods, the proposed method also slightly outperforms Seg-Driven, Heatmaps, and PVnet, which use deep information in the data-build stage. Furthermore, we compare the time efficiency in Table 3. Our method is approximately five times faster than other methods due to its succinct network architecture and EPnP-RANSAC mechanism. Moreover, our approach only needs less than 30 ms for each image. Notably, our proposed approach shows high accuracy and efficiency even with large occlusions.

Table 2. Quantitative evaluation on the Occluded-LINEMOD dataset. Symmetric objects are presented in bold font. Red and blue labels denote the best and second best results, respectively.

Object	PoseCNN	Tekin	BB8	Pix2Pose	Heatmap	Seg-Driven	PVnet	CDPN	Ours
Ape	9.6	7.0	28.5	8.3	16.5	12.1	15.81	28.92	39.1
Can	45.2	1.2	11.2	12.1	42.5	39.9	63.31	55.98	69.8
Cat	0.9	3.6	9.6	9.3	2.8	8.2	16.68	13.24	26.9
Driller	41.4	1.4	0.2	10.9	47.1	45.2	65.65	51.37	49.0
Duck	19.6	5.1	6.8	6.3	11.0	17.2	25.24	22.97	25.6
Eggbox	22.0	9.6	4.0	13.8	24.7	22.1	50.17	35.98	31.9
Glue	38.5	6.5	4.7	11.3	39.5	35.8	49.62	39.68	46.5
Holepun	22.1	8.3	8.1	10.7	21.9	36	39.67	51.06	53.1
Average	24.9	5.3	9.1	10.3	25.8	27.0	40.77	37.4	42.7

Table 3. Efficiency comparison on the Occluded-LINEMOD dataset. All are conducted in the same environment.

Method	PoseCNN	Tekin	BB8	Pix2Pose	Heatmap	Seg-Driven	PVnet	Ours
FPS	4	40	3	-	4	22	8	30

Fig.8. YCB-Video results. Every two rows is a pair of images, representing different perspectives of the same scene. Our method generates accurate pose estimations, even in the presence of large occlusions. Furthermore, it can process multiple objects in real time.

4.3 quantitative Evaluation on YCB-Video

Fig. 8 presents additional test results in a large number of scenarios. Our method exhibits satisfactory estimation results in all tests. We compare our method with the baseline PoseCNN and three latest approaches in Table 4. Our method outperforms both PoseCNN and SilhoNet. Furthermore, our method is only slightly better than Heatmap and Seg-Driven because symmetrical objects have similar appearance and are textureless.

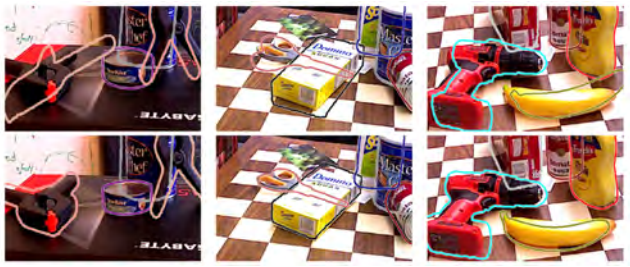


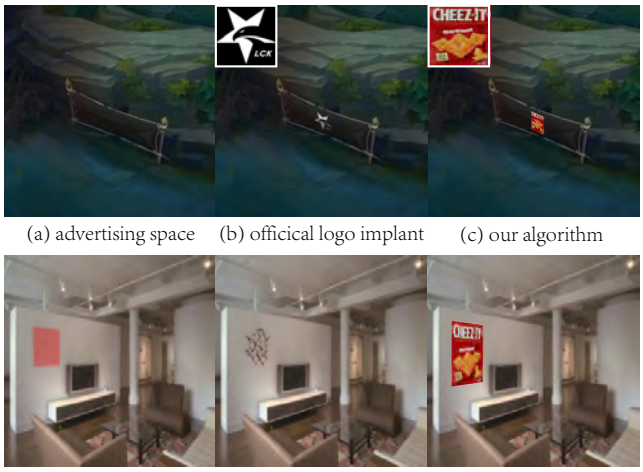
Fig. 9. Comparison of before (top) and after(bottom) the iterative refinement module. Benefits of our end-to-end procedure, particularly in the presence of large occlusions, are presented.

Table 4. Quantitative evaluation on the YCB-Video dataset. The metric is ADD-0.1d. Symmetric objects are presented in bold font. Red and blue labels denote the best and second best results, respectively.

Object	PoseCNN	SilhoNet	Heatmap	Seg-Driven	CDPN	Ours
master_chef_can	3.6	23.8	32.9	33.0	35.5	37.1
cracker_box	25.1	20.1	62.6	44.6	45.6	45.8
sugar_box	40.3	48.5	44.5	75.6	71.5	69.4
tomato_soup_can	25.5	25.1	31.1	40.8	49.6	52.3
mustard_bottle	61.9	60.8	42.0	70.6	74.8	78.2
tuna_fish_can	11.4	25.3	6.8	18.1	25.2	24.1
pudding_box	14.5	17.0	58.4	12.2	48.4	32.6
gelatin_box	12.1	26.2	42.5	59.4	57.8	46.9
potted_meat_can	18.9	22.2	37.7	33.3	37.6	40.1
banana	30.3	32.8	16.8	16.6	25.1	27.5
pitcher_base	15.6	25.9	57.2	90.0	80.8	82.0
bleach_cleanser	21.2	20.8	65.3	70.9	81.9	82.1
bowl	12.1	22.5	25.6	30.5	22.6	23.0
mug	5.2	12.3	11.6	40.7	51.2	55.6
power_drill	29.9	26.0	46.1	63.5	57.4	59.8
wood_block	10.7	18.7	34.3	27.7	25.1	27.5
scissors	2.2	3.4	0.1	17.1	11.6	12.4
large_marker	3.4	3.0	3.2	4.8	4.5	6.1
large_clamp	28.5	29.7	10.8	25.6	25.9	27.7
extra_large_clamp	19.6	20.4	29.6	8.8	13.4	16.4
foam_brick	54.5	42.0	51.7	34.7	40.7	44.8
Average	21.3	25.1	33.6	39.0	42.2	42.4

4.4 Ablation study of iterative refinement

Fig. 9 illustrates some sampled predictions and comparisons. The results showed that our refinement module can improve the performance of the original main network.



(a) advertising space (b) officical logo implant (c) our algorithm

(d) advertising space (e) recommended decoration (f) our algorithm

Fig. 10. Examples of object placement. The top row is found in virtual media whereas the bottom row is located in the real environment.

4.5 Applications

6D object pose estimation is a key technology for real-world applications, as mentioned in Sec. 1. We apply our method in object placement. We skin known textures at a specified location and pretend that the object is in the dataset because our method targets known objects. We translate and rotate the target object according to the pose, and then place it in the specified position after estimating the pose by our algorithm. As shown in the first row of Fig. 10. The image on the left is the advertising space of the most popular e-sports game in the world, and all viewers will see it during the game. The middle image shows that the official game logo is placed on the advertising space. The image on the right uses our algorithm with the texture of the cracker box in the YCB dataset. Notably, many recommended decoration algorithms can also be viewed as an implementation of object placement. Liang *et al* [53] determined the optimal decoration recommendation on the wall. Meanwhile, we also applied our proposed method to achieve the same performance in the real environment.

5 Conclusions

We proposed a novel method for 6-DoF real-time pose estimation of targets only from images without depth information to build a two-stream architecture that includes a segmentation stream for processing various features of the input image and a regression stream for solving 2D-3D correspondences via bounding corners. We also proposed a fusion network to transform and map textures and location relationships at the pixel level. Abundant features ensure that our model considers both the appearance and geometric information of the object. Moreover, we design an iterative procedure to improve the accuracy of results.

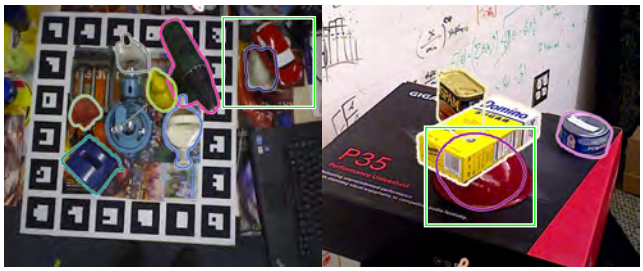


Fig. 11. Unsatisfactory examples (highlighted in the green box). Left: heavily occluded symmetrical Eggbox model in the Occluded-LINEMOD dataset. Right: texture-less symmetrical Bowl model in the YCB-Video dataset.

However, our method may produce unsatisfactory results on objects with symmetry or texture-less models. Fig. 11 shows two examples of this scenario. If an excessive number of occlusions exists in the scene, our method fails to obtain sufficient information to estimate accurate poses. This issue will be our main concern in a future investigation.

References

[1] Brachmann E, Krull A, Michel F, Gumhold S, Shotton J, Rother C. Learning 6D Object Pose Estimation Using 3D Object Coordinates. In *European Conference on Computer Vision (ECCV)*, 2014, pp.536-551.

[2] Hinterstoisser S, Holzer S, Cagniart C, Ilic S, Konolige K, Navab N, Lepetit V. Multimodal Templates for Real-

time Detection of Texture-less Objects in Heavily Cluttered Scenes. In *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp.858-865.

[3] Hinterstoisser S, Lepetit V, Ilic S, Holzer S, Bradski G, Konolige K, Navab N. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In *Asian Conference on Computer Vision (ACCV)*, 2012, pp.548-562.

[4] Kehl W, Milletari F, Tombari F, Ilic S, Navab N. Deep Learning of Local RGB-D Patches for 3D Object Detection and 6D Pose Estimation. In *European Conference on Computer Vision (ECCV)*, 2016, pp.205-220.

[5] Rios C R, Tuytelaars T. Discriminatively Trained Templates for 3D Object Detection: A Real Time Scalable Approach. In *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp.2048-2055.

[6] Tejani A, Tang D, Kouskouridas R, Kim T K. Latent-Class Hough Forests for 3D Object Detection and Pose Estimation. In *European Conference on Computer Vision (ECCV)*, 2014, pp.462-477.

[7] Wohlhart P, Lepetit V. Learning Descriptors for Object Recognition and 3D Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp.3109-3118.

[8] Cao Y, Ju T, Xu J, Hu S. Extracting Sharp Features from RGB-D Images. *Computer Graphics Forum*, 2017, 36(8): 138-152.

[9] Wang C, Xu D, Zhu Y, Martin R, Lu C, Li F, Savarese S. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp.3343-3352.

[10] Xu D, Anguelov D, Jain A. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp.244-253.

[11] Qi C R, Liu W, Wu C, Su H, Guibas L J. Frustum PointNets for 3D Object Detection From RGB-D Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp.7918-7927.

[12] Xiang Y, Schmidt T, Narayanan V, Fox D. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *Robotics: Science and Systems*, 2018.

[13] Krull A, Brachmann E, Michel F, Yang M Y, Gumhold S, Rother C. Learning Analysis-by-Synthesis for 6D Pose Estimation in RGB-D Images. In *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp.954-962.

[14] Xiaolong Yang , Xiaohong Jia. 6D Pose Estimation with Two-stream Net. In ACM SIGGRAPH Posters, 2020.

[15] Song S, Xiao J. Sliding Shapes for 3D Object Detection in Depth Images. In *European Conference on Computer Vision (ECCV)*, 2014, pp.634-651.

[16] Song S, Xiao J. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp.808-816.

[17] Qi C R, Su H, Mo K, Guibas L J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp.652-660.

[18] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp.3354-3361.

[19] Aubry M, Maturana D, Efros A A, Russell B C, Sivic J. Seeing 3D Chairs: Exemplar Part-based 2D-3D Alignment using a Large Dataset of CAD Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp.3762-3769.

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

[21] Ferrari V, Tuytelaars T, Gool L V. Simultaneous Object Recognition and Segmentation from Single or Multiple Model Views. *International Journal of Computer Vision*, 2006, 67(2): 159-188.

[22] Rothganger F, Lazebnik S, Schmid C, Ponce J. 3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints. *International Journal of Computer Vision*, 2006, 66(3): 231-259.

[23] Zhu M, Derpanis K G, Yang Y, Brahmbhatt S, Zhang M, Phillips C, Lecce M, Daniilidis K. Single Image 3D Object Detection and Pose Estimation for Grasping. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp.3936-3943.

[24] Nakajima Y, Saito H . Robust Camera Pose Estimation by Viewpoint Classification Using Deep Learning. Computational Visual Media, 2017, 3(2): 189-198.

[25] Suwajanakorn S, Snavely N, Tompson J J, Norouzi M. Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning. In *Advances in Neural Information Processing Systems (NIPS)*, 2018, pp.2059-2070.

[26] Tekin B, Sinha S N, Fua P. Real-Time Seamless Single Shot 6D Object Pose Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp.292-301.

[27] Tremblay J, To T, Sundaralingam B, Xiang Y, Fox D, Birchfield S. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. In *Conference on Robot Learning (CoRL)*, 2018, pp.292-301.

[28] Fischler M A, Bolles R C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 1981, 24(6): 381-395.

[29] Schwarz M, Schulz H, Behnke S. RGB-D Object Recognition and Pose Estimation based on Pre-trained Convolutional Neural Network features. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp.1329-1335.

[30] Tulsiani S, Malik J. Viewpoints and Keypoints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp.1510-1519.

[31] Mousavian A, Anguelov D, Flynn J, Kosecka J. 3D Bounding Box Estimation Using Deep Learning and Geometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp.7074-7082.

[32] Sundermeyer M, Marton Z C, Durner M, Brucker M, Triebel R. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In *European Conference on Computer Vision (ECCV)*, 2018, pp.699-715.

[33] Hu Y, Hugonot J, Fua P, Salzmann M. Segmentation-Driven 6D Object Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp.3385-3394.

[34] Billings G, Roberson M J. SilhoNet: An RGB Method for 6D Object Pose Estimation. *IEEE Robotics and Automation Letters*, 2019, 4(4): 3727-3734.

[35] Park K, Patten T, Vincze M. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp.7668-7677.

[36] Castro P, Armagan A, Kim T K. Accurate 6D Object Pose Estimation by Pose Conditioned Mesh Reconstruction. arXiv:1910.10653, 2019.

[37] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences. arXiv:1404.2188, 2014.

[38] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence(T-PAMI)*, 2017, 39(12): 2481-2495.

[39] Li C, Bai J, Hager G D. A Unified Framework for Multi-View Multi-Class Object Pose Estimation. In *European Conference on Computer Vision (ECCV)*, 2018, pp.254-269.

[40] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. arXiv:1804.02767, 2018.

[41] Bochkovskiy A, Wang C, Liao H M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934, 2020.

[42] Lin T, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp.2980-2988.

[43] Lepetit V, Moreno N F, Fua P. EPnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision*, 2009, 81(2): 155.

[44] Besl P J, McKay N D. Method for Registration of 3D Shapes. *International Society for Optics and Photonics*, 1992, 1611(1): 586-606.

[45] Kehl W, Manhardt F, Tombari F, Ilic S, Navab N. SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again. In *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp.1521-1529.

[46] Rad M, Lepetit V. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth. In *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp.3828-3836.

[47] Oberweger M, Rad M, Lepetit V. Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation. In *European Conference on Computer Vision (ECCV)*, 2018, pp.119-134.

[48] Peng S, Liu Y, Huang Q, Zhou X, Bao H. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp.4561-4570.

[49] Li Y, Wang G, Ji X, Xiang Y, Fox D. DeepIM: Deep Iterative Matching for 6D Pose Estimation. In *European Conference on Computer Vision (ECCV)*, 2018, pp.683-698.

[50] Muja M, Lowe D G. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In *VISAPP of IEEE International Conference on Computer Vision (ICCV)*, 2009, pp.331-340.

[51] Everingham M, Van Gool L. , Williams C. K. I. , Winn J. , Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 2010, 88(2): 303–338.

[52] Li Z, Wang G, Ji X. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp.7678-7687.

[53] Liang Y, Fan L, Ren P, Xie X, Hua X. DecorIn: An Automatic Method for Plane-based Decorating. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2020.

Xiaolong Yang received his B.S. degree in Information and Computing Science from Northwestern Polytechnical University in 2017 and is currently pursuing his M.S. and Ph.D. degrees at Key Laboratory of Mathematics Mechanization, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences(CAS). His research interest is on computer graphics and computer vision.

Xiaohong Jia is an associate professor at Key Laboratory of Mathematics Mechanization, Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS). She received her Ph.D. and Bachelor's degree from the University of Science and Technology of China in 2009 and 2004, respectively. Her research interests include computer graphics, computer aided geometric design and computational algebraic geometry.

Yuan Liang is an algorithm engineer at Alibaba DAMO Academy. He received his BSc and PhD degree from Tsinghua University, China, in 2014 and 2019, respectively. His research interests include interactive multimedia analysis and geometric processing from visual media.

Lubin Fan is a senior algorithm engineer at Alibaba Damo Acadamy. He received the BSc and PhD degree in Mathematics from Wuhan University in 2009 and Zhejiang University in 2014, respectively. From 2014 to 2017, he was a postdoctoral fellow at the Visual Computing Center, King Abdullah University of Science and Technology (KAUST). He joined the Alibaba Group in 2017. His research interests include image and video processing, geometric processing and procedural modeling.