QaQ: Robust 6D Pose Estimation via Quality-Assessed RGB-D Fusion

Théo Petitjean, Zongwei Wu, Olivier Laligant, Cédric Demonceaux

QaQ: Robust 6D Pose Estimation via Quality-Assessed RGB-D Fusion

Théo Petitjean[1,2], Zongwei Wu[3], Olivier Laligant[1], Cédric Demonceaux[3]
[1] ImViA UR 7535, University of Burgundy [2]SPIE
[3]ICB UMR CNRS 6303, University of Burgundy
theo.petitjean@spie.com

Abstract

RGB-D 6D pose estimation has recently drawn great research attention thanks to the complementary depth information. Whereas, the depth and the color image are often noisy in real industrial scenarios. Therefore, it becomes challenging for many existing methods that fuse equally RGB and depth features. In this paper, we present a novel fusion design to adaptively merge RGB-D cues. Specifically, we created a Quality-assessment block that estimates the global quality of the input modalities. This quality represented as an α parameter is then used to reinforce the fusion. We have thus found a simple and effective way to improve the robustness to low-quality inputs in terms of Depth and RGB. Extensive experiments on 6D pose estimation demonstrate the efficiency of our method, especially when noise is present in the input.

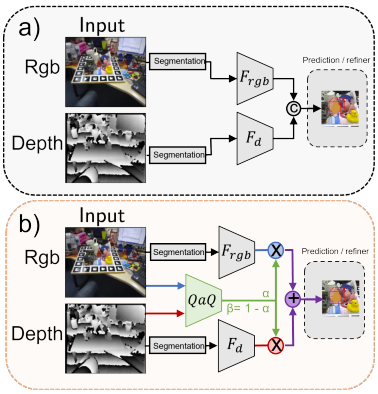


Figure 1. a) The baseline of DenseFusion [6] pose estimation. b) Our proposed Quality-assessment Querying, termed QaQ, enables adaptive fusion in order to amplify the contribution of the most truthful modality and minimize the other.

1 Introduction

The use of computer vision for position estimation in industrial environments has gained increasing interest in recent years thanks to its potential to improve efficiency and accuracy in manufacturing processes. Accurate position estimation is crucial for the successful execution of robotic tasks such as assembly, packaging, and inspection. However, achieving accurate position estimation in such environments presents several challenges, particularly when the captured images may be of low quality.

In fact, in industrial environments, the quality of images can be affected by various factors such as low lighting conditions, dust, and vibrations. Furthermore, using denoising algorithms to improve image quality can slow down the process, which is often unacceptable in industrial settings, especially for embedded systems. Therefore, a position estimation algorithm that can provide accurate results even dealing with low-quality images is highly demanded.

In this paper, we propose a novel approach for position estimation in industrial environments that leverages a learning-based module to estimate the quality of input images. Specifically, our approach deeply investigate the correlation/consistency between input modalities, i.e., RGB and Depth, to create adaptive weights to automatically control the contribution of each modality. In such a manner, the modality with better quality is prioritized, leading to better robustness and more stable performance in the industrial context.

We validate the effectiveness of our approach through extensive experiments. We show our method can serve as a plug-in-and-play module to improve the baseline performance. To simulate the industrial context, we additionally simulate white noise of different levels on top of input images. In such cases, our method constantly performs better compared to our baseline, especially while an important noise is present.

2 Related Work

6D pose estimation involves predicting the matrix that transforms object coordinates into camera coordinates. Traditional methods rely on feature extraction and matching techniques [51, 52], while recent deep learning-based approaches have shown great potential [44, 47]. However, accurately predicting 3D poses remains a challenge. RGB-based pose estimation methods have emerged, leveraging point of interest detection techniques [1, 3, 53], and deep learning networks that directly predict pose from RGB images [45, 46]. Never-

theless, achieving high accuracy for 6D pose estimation using RGB-only data remains problematic, especially under challenging scenarios such as occlusion, reduced visibility, mixed foreground-background, etc. 3D point cloud-based pose estimation methods [48, 49], employing deep learning networks adapted to 3D point clouds [7, 36], have shown promise. However, due to the sparsity and lack of texture in point clouds, their 6D pose estimation accuracy remains limited.

Recently, pose estimation using RGB-D data has gained popularity. Classical RGB-D methods [26, 28, 27, 29, 30] utilize two-step algorithms, falling into categories like holistic, dense correspondence [21, 22], and keypoint-based methods. KeyPoint-based approaches, such as PVN3D [35], have demonstrated state-of-the-art results. Specifically, to improve the robustness and efficiency of pose estimation networks, several methods use RGB-D images and merge features extracted from both point clouds and RGB images. Late fusion strategies, as seen in DenseFusion [6], PointFusion [37], and MoreFusion [31], have shown effectiveness in better understanding input data. However, concatenation-based strategies may propagate noise and poorly optimized fusions can significantly impact the network's performance, as demonstrated by FFB6D [41]. To address this, we propose a new fusion method, balanced by a parameterized coefficient output from a quality-assessment module. In such a manner, our method takes the depth quality into account, improving the fusion design with more effectiveness.

3 Overview

The overall architecture of our method is presented in Figure 1. On top of a baseline model (in gray), we propose a **Q**uality **A**ssessment **Q**uerying module, denoted as **QaQ** block, for permanent control/weighting of the contribution of each modality based on its quality. Specifically, the QaQ module takes the paired RGB-D images as inputs and deeply analyzes the correlation, consistency, and alignment between them. Our intuition comes from the observation that, despite the modality-specific features, there exist modality-shared clues which can be exploited for Quality-assessment. In such a manner, while one modality is degraded with undesirable noise, the other one can be used to query and quantify such degradation. Therefore, the presence of noise in one modality can be transferred as a weighting parameter to control its contribution through the deep network. Our QaQ block can be served as a plug-in-and-play module that can be easily adapted to any existing network. Without modifying the architecture of the baseline, by fusing the multi-modal clues more intelligently, our QaQ module leads to a simple yet efficient manner to improve the network performance and its robustness against noise in the industrial context.

3.1 Quality-assessment Querying module

The QaQ module is a deep neural network model that takes as input a pair of images (RGB-D), and outputs α and β, where $\beta = 1 - \alpha$, referring to the quality of the input modalities, respectively. Specifically, as shown in Figure 2, our QaQ module consists of three components: an early fusion, a spatial feature extraction layer, and a classifier.

We first fuse the RGB and depth feature from the input side of our QaQ module. Specifically, they are concatenated along the channel dimension and passed through a 2D convolutional layer. The resulting feature map is then fed into the encoder network which consists of convolutional and pooling layers. Such a design leads to gradually reduced resolution during the feature extraction, hence enabling a large receptive field to better cover contextualized clues.

Lastly, the extracted output is flattened and fed into a classifier to generate the final adjusting weights. Our classifier consists of a dropout layer, a fully connected layer, and a sigmoid activation function. The output weights are two normalized scalar values between 0 and 1, referring to the quality score of the input modality.

3.2 Plug-In-And-Play

Our Quality-assessment block can be served as a plug-in-and-play module that can easily be adapted into any existing RGB-D baseline. In our application, we choose the well-acknowledged DenseFusion [6] as our baseline.

Specifically, the original DenseFusion extracts feature via dual encoders F_{rgb} and F_d and extracted two level outputs. Then DenseFusion fuses RGB and depth in an equal without explicitly modeling their quality. Formally, let R_i be the extracted visual feature and D_i be the extracted depth feature, the original DenseFusion outputs the fused feature f_i by:

$$f_i = MLP(Concat(R_i, D_i)) \tag{1}$$

where $Concat$ stands for the channel-wise concatenation. Differently, with the help of our QaQ block, we obtain the weighting parameters α to automatically control the modal contribution. Therefore, we can replace the Eq. 1 by follow:

$$f_i = MLP(\alpha \times R_i + (1 - \alpha) \times D_i) \tag{2}$$

4 Experiments

4.1 Experimental Setups

Dataset: We evaluate the performance on two widely-used benchmarks, i.e., **LineMod** [9] and **YCB** [8]. We follow the same train/test data split as previous works by [10, 6, 31]. While comparing the performance, for

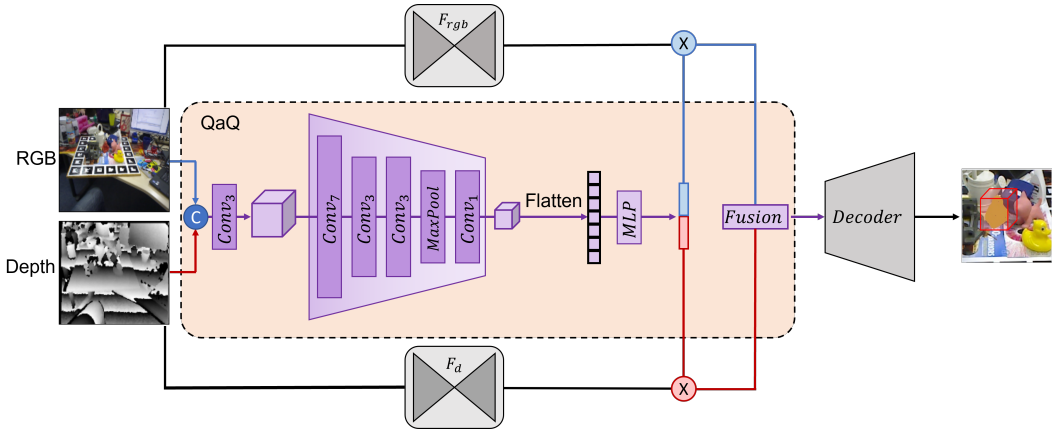
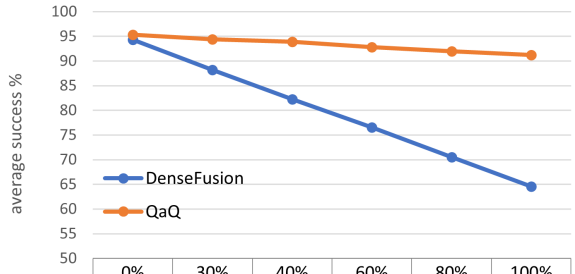


Figure 2. QaQ is an artificial intelligence block that estimates the quality of the input paired modalities, i.e., RGB and Depth. We weight the contribution of each modality by two learned parameters α and β where $\beta = 1 - \alpha$ to achieve a quality-aware RGB-D fusion. Therefore, the plugged-in adaptive fusion block can boost the baseline performance and reinforce the robustness of 6D pose estimation.

Table 1. Quantitative result on ADD metric for LineMOD dataset. Symmetric objects are denoted in *italics*. **Bold** denotes the best result.

Comparison	Without Noise		30% Inputs Noised	
	DF [6]	Ours	DF [6]	Ours
ape	**92.3**	90.3	85.1	**88.1**
bench vi. 2	93.2	**94.3**	89.0	**93.2**
camera	94.4	**96.8**	77.3	**95.3**
can	93.1	**95.6**	90.2	**96.0**
cat	**96.5**	95.8	92.3	**95.6**
driller	87.0	**90.0**	84.4	**89.7**
duck	**92.3**	92.1	85.3	**89.4**
eggbox	98.8	**100**	**100**	**100**
glue	**100**	**100**	99.2	**99.7**
hole p.	92.1	**92.8**	68.6	**91.4**
iron	97.0	**98.1**	93.4	**98.3**
lamp	95.3	**96.9**	92.5	**96.2**
phone	92.8	**96.5**	87.1	**95.0**
AVERAGE	94.3	**95.3**	88.2	**94.4**



	0%	30%	40%	60%	80%	100%
DenseFusion	94,3	88,2	82,3	76,6	70,5	64,6
QaQ	95,3	94,4	93,9	92,8	92	91,2

Figure 3. Average performance vis-à-vis the proportion of noised input samples. We maintain the same level of SNR, around 50 dB, while gradually noise more inputs. Our method consistently performance better than our baseline, validating our effectiveness.

a fair comparison, we apply the same noise to both baseline and our method.

Metrics: We follow previous works [10, 27] to evaluate the performance with the conventional ADD/ADD-S metric for non-symmetric and symmetric objects, respectively. These metrics are defined as the average distance between the predicted pose (R_i and t) and the ground truth ($\hat{R}_i$ and $\hat{t}$) for each vertex i in object o, as shown in Equation 3:

$$\begin{aligned} ADD &= \frac{1}{m}\sum_{x\in o}\left\|(Rx+t)-(\hat{R}x+\hat{t})\right\|, \\ ADD\text{-}S &= \frac{1}{m}\sum_{i\in o}\min_{j\in o}\left\|(Ri+t)-(\hat{R}j+\hat{t})\right\|. \end{aligned} \tag{3}$$

To evaluate the performance of our approach on the YCB dataset, we follow previous works [10, 31, 35] and report the area under the ADD-S curve (AUC) with a maximum threshold of 0.1m, as well as the percentage of ADD values smaller than 2cm.

Simulated Noise: To analyze the model robustness, we simulate different levels of noise. Specifically, we simulate noise over a proportion of input samples to make the benchmarks more challenging and more realistic, i.e., industrial settings. To quantify the level of noise in our experiments, we use the signal-to-noise ratio (SNR). The mean SNR is around 45(dB). The proportion of noised samples gradually augments from 30% to 100%. Among these samples, 1/3 of them are noised on RGB images, 1/3 of them are noised on depth

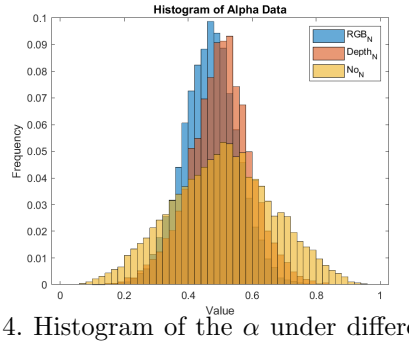


Figure 4. Histogram of the α under different infereior settings. **Blue** histogram stands for the case where the noise is applied to RGB samples (RGB_N), **orange** histogram for noise applied to Depth ($Depth_N$), and **yellow** histogram for oracle cases, i.e., no simulated noise (No_N).

Table 2. Quantitative evaluation on the YCB data with AUC and ADD(2cm) metrics. Here all the testing samples are noised.

	DF [6]		Ours	
	AUC	*ADD*	AUC	*ADD*
Oracle	93.1	96.8	**94.1**	**97.1**
$SNR_D = 39$ $SNR_{rgb} = 6$	57.4	64.2	**64.5**	**67.1**
$SNR_D = 33$ $SNR_{rgb} = 3$	37.6	40.1	**58.6**	**61.0**

images, and 1/3 of them are noised on both modalities. For the RGB image, we introduce additive Gaussian white noise (GWN) with a standard deviation σ_{noise}, while the depth information is corrupted by multiplicative GWN to avoid generating spurious data on the 0 values produced by the physical sensor.

4.2 Noise analyses

Linemod: Table 1 presents the quantitative performance under original settings, i.e., without noise, and in degraded settings, i.e., 30% samples are noised. We can see that our QaQ module can effectively improve the baseline accuracy in almost all cases. Without noise, our QaQ module enables +1% absolute gain. When dealing with noisy samples, the absolute gain becomes more significant and achieves around +6%.

We further noise more samples from the input side and report the performance in Figure 3. It can be seen that our module consistently improve the baseline robustness against noise and achieve more stable performance, validating the effectiveness of our module.

YCB: For this dataset, we noise all the input samples and variate the noise degree. Specifically, we consider three scenarios: no noise, lower-level noise on each modality, and stronger-level noise on each modality. Table 2 shows that in all three scenarios, the QaQ module improves the average success rate, especially in high-noise settings where our module shows better robustness against noise with +20% absolute gain.

4.3 Discussion and Key Takeaways

To better understand our approach, we provide in Figure 4 the α histogram under different degraded settings, i.e., without noise or adding noise on one modality. By comparing the cases without noise and with noise, we can observe that the histogram becomes more concentrated. In other words, there are less extreme cases where α is near 0 or 1. In fact, in general settings without noise, we can simply use one modality and achieve relatively great performance, i.e., the cases where α is near 0 or 1 in the **yellow** histogram. However, when there exist noise in input modalities, it is no more the case and we need to leverage multi-modal clues for accurate prediction, i.e., there is less α near 0 or 1 in **orange** or **blue** histograms. This difference highlights the necessity and interest in realizing RGB-D fusion in challenging cases.

Moreover, we find out that the α value decreases when the noise is concentrated on RGB and increases when the noise is on depth. Hence, we can conclude that the QaQ module seeks to balance the weight of each modality when noise is present, with lower α values implying a lower weight of RGB in the fusion and vice versa. The variable α value also validates our concept of adaptive fusion.

5 Conclusion

In this paper, we highlight a limitation of existing 6D pose estimation methods, including the widely used DenseFusion, in dealing with noise. Our analysis suggests that the fusion design, which equally merges RGB and Depth features, may be the main performance bottleneck. To address this issue, we propose a novel adaptive fusion approach that effectively aggregates multi-modal clues. Our method can be easily integrated into traditional networks and demonstrates significant improvements in robustness against noise. Through our module's in-depth analysis, we gain insights into the role of adaptive multi-modal fusion. We believe that our work can inspire future research in designing more robust methods tailored for industrial environments.

6 Acknowledgement

We gratefully acknowledge the support of the French government's Plan France Relance : ImViA - Wasoria, which provided funding for this project.

References

[1] Collet, A., Martinez, M. & Srinivasa, S. The moped framework: Object recognition and pose estimation for

manipulation, The International Journal of Robotics Research. (2011)

[2] Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D. & Birchfield, S. Deep object pose estimation for semantic robotic grasping of household objects. *ArXiv Preprint ArXiv:1809*. pp. 10790 (2018)

[3] Zhu, M., Derpanis, K., Yang, Y., Brahmbhatt, S., Zhang, M., Phillips, C., Lecce, M. & Daniilidis, K. Single image 3d object detection and pose estimation for grasping. *2014 IEEE International Conference On Robotics And Automation (ICRA)*. **1** pp. 3936-3943 (2014)

[4] Marchand, E., Uchiyama, H. & Spindler, F. Pose estimation for augmented reality: A hands-on survey, " IEEE transactions on visualization and computer graphics. (2016)

[5] Marder-Eppstein, E. Project tango. (ACM,2016)

[6] Wang, C., Xu, D., Zhu, Y., Martin-Martin, R., Lu, C., Fei-Fei, L. & Savarese, S. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. *Computer Vision And Pattern Recognition (CVPR)*. (2019)

[7] Qi, C., Su, H., Mo, K. & Guibas, L. Pointnet: Deep learning on point sets for 3d classification and segmentation. *ArXiv Preprint ArXiv:1612*. pp. 00593 (2016)

[8] Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P. & Dollar, A. The ycb object and model set: Towards common benchmarks for manipulation research, 2015 International Conference on Advanced Robotics (ICAR). (2015)

[9] 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, Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2011)

[10] Xiang, Y., Schmidt, T., Narayanan, V. & Fox, D. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *ArXiv Preprint ArXiv:1711*. pp. 00199 (2017)

[11] Li, C., Bai, J. & Hager, G. "A unified framework for multi-view multi-class object pose estimation. " *ArXiv Preprint ArXiv:1803*. pp. 08103 (2018)

[12] Peng, S., Liu, Y., Huang, Q., Zhou, X. & Bao, H. Pvnet: Pixel-wise voting network for 6dof pose estimation. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 4561-4570 (2019)

[13] Chen, W., Jia, X., Chang, H., Duan, J. & Leonardis, A. G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4233-4242 (2020)

[14] Zakharov, S., Shugurov, I. & Ilic, S. Dpod: 6d pose object detector and refiner. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 1941-1950 (2019)

[15] Li, Z., Wang, G. & Ji, X. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 7678-7687 (2019)

[16] Buch, A., Kiforenko, L. & Kraft, D. "Rotational subgroup voting and pose clustering for robust 3d object recognition, " in Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. (2017)

[17] Drost, B., Ulrich, M., Navab, N. & Ilic, S. "Model globally, match locally: Efficient and robust 3d object recognition, " in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, Ieee. (2010)

[18] Vidal, J. & Y., C. Lin, and R. (Mart´ı, "6d pose estimation using an improved method based on point pair features, " in 2018 4th International Conference on Control, Automation,2018)

[19] Li, Y., Wang, G., Ji, X., Xiang, Y. & Fox, D. "Deepim: Deep iterative matching for 6d pose estimation. " *ArXiv Preprint ArXiv:1804*. pp. 00175 (2018)

[20] Tekin, B., Sinha, S. & Fua, P. "Real-Time Seamless Single Shot 6D Object Pose Prediction, " in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). (2018)

[21] Liebelt, J., Schmid, C. & Schertler, K. independent object class detection using 3d feature maps. *2008 IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1/8 (2008)

[22] Glasner, D.,Galun, M.,Alpert, S.,Basri, R.& Shakhnarovich, G. aware object detection and pose estimation. *2011 International Conference On Computer Vision*. **2** pp. 1275-1282 (2011)

[23] Doumanoglou, A., Kouskouridas, R., Malassiotis, S. & Kim, T. Recovering 6d object pose and predicting next-bestview in the crowd. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 3583- (2016)

[24] Park, K., Patten, T. & Vincze, M. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 7668-7677 (2019)

[25] Xu, Z., Chen, K. & Jia, K. W-PoseNet: Dense Correspondence Regularized Pixel Pair Pose Regression. *CoRR*. **abs/1912.11888** (2019), http://arxiv.org/abs/1912.11888

[26] 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, Springer. (2014)

[27] 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, Springer. (2012)

[28] 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, Springer. (2016)

[29] Rios-Cabrera, R. & Tuytelaars, T. "Discriminatively trained templates for 3d object detection: A real time

scalable approach, " in Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2055)
[30] Wohlhart, P. & Lepetit, V. "Learning descriptors for object recognition and 3d pose estimation, " in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). (2015)
[31] Wada, K., Sucar, E., James, S., Lenton, D. & Davison, A. MoreFusion: Multi-object Reasoning for 6D Pose Estimation from Volumetric Fusion. *CoRR*. (2020), https://arxiv.org/abs/2004.04336
[32] Pavlakos, G., Zhou, X., Chan, A., Derpanis, K. & Daniilidis, K. 6-dof object pose from semantic keypoints. *2017 IEEE International Conference On Robotics And Automation (ICRA)*. pp. 2011-2018 (2017)
[33] Rad, M. & Lepetit, V. Bb8: A scalable, accurate. (robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth,2017)
[34] Tekin, B., Sinha, S. & Fua, P. Real-time seamless single shot 6d object pose prediction. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 292-301 (2018)
[35] He, Y., Sun, W., Huang, H., Liu, J., Fan, H. & Sun, J. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. (2020,6)
[36] Zhou, Y. & Tuzel, O. "Voxelnet: End-to-end learning for point cloud based 3d object detection. " *ArXiv Preprint ArXiv:1711*. pp. 06396 (2017)
[37] Xu, D., Anguelov, D. & Jain, A. Pointfusion: Deep sensor fusion for 3d bounding box estimation. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 244-253 (2018)
[38] Qi, C., Liu, W., Wu, C., Su, H. & Guibas, L. Frustum pointnets for 3d object detection from rgb-d data. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 918-927 (2018)
[39] Qi, C., Litany, O., He, K. & Guibas, L. Deep hough voting for 3d object detection in point clouds. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 9277-9286 (2019)
[40] Zhou, G., Yan, Y., Wang, D. & Chen, Q. A novel depth and color feature fusion framework for 6d object pose estimation. *IEEE Transactions On Multimedia*. **2** (2020)
[41] He, Y., Huang, H., Fan, H., Chen, Q. & Sun, J. FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation. *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. (2021,6)
[42] Lowe, D. Object recognition from local scale-invariant features. *Proceedings Of The Seventh IEEE International Conference On Computer Vision*. **2** pp. 1150-1157 (1999)
[43] Wohlhart, P. & Lepetit, V. "Learning descriptors for object recognition and 3d pose estimation, " in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). (2015)
[44] Mousavian, A., Anguelov, D., Flynn, J. & Kosecka, J. "3d bounding box estimation using deep learning and geometry, " in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). (2017)
[45] Tulsiani, S. & Malik, J. "Viewpoints and keypoints, " in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). (2015)
[46] Schwarz, M., Schulz, H. & Behnke, S. "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features, " in Robotics and Automation (ICRA), 2015 IEEE International Conference on, IEEE. (2015)
[47] Xiang, Y., Choi, W., Lin, Y. & Savarese, S. "Data-driven 3d voxel patterns for object category recognition, " in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). (2015)
[48] Song, S. & Xiao, J. "Sliding shapes for 3d object detection in depth images, " in European conference on computer vision, Springer. (2014)
[49] Song, S. & Xiao, J. "Deep sliding shapes for amodal 3d object detection in rgb-d images. (" in Proceedings of the IEEE Computer Vision,2016)
[50] Fischler, M. & Bolles, R. "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, " Communications of the ACM. (1981)
[51] Aubry, M., Maturana, D., Efros, A., Russell, B. & Sivic, J. "Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of cad models, " in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR). (2014)
[52] Ferrari, V., Tuytelaars, T. & Gool, L. "Simultaneous object recognition and segmentation from single or multiple model views, " International Journal of Computer Vision. (2006)
[53] 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)
[54] Athar, S., Wang, Z. & Wang, Z. Deep Neural Networks for Blind Image Quality-assessment: Addressing the Data Challenge. (2021)
[55] Zhang, R., Isola, P., Efros, A., Shechtman, E. & Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CVPR*. (2018)
[56] Prashnani, E., Cai, H., Mostofi, Y. & Sen, P. PieAPP: Perceptual Image-Error Assessment through Pairwise Preference. (2018)
[57] Gu, J., Cai, H., Chen, H., Ye, X., Ren, J. & Dong, C. PIPAL: a Large-Scale Image Quality-assessment Dataset for Perceptual Image Restoration. (2020)
[58] Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image Quality-assessment: from error visibility to structural similarity. *IEEE Transactions On Image Processing*. **13**, 600-612 (2004)
[59] Zhang, W., Ma, K., Yan, J., Deng, D. & Wang, Z. Blind Image Quality-assessment Using a Deep Bilinear

Convolutional Neural Network. *IEEE Transactions On Circuits And Systems For Video Technology.* **30**, 36-47 (2020,1)

[60] Esfandarani, H. & Milanfar, P. NIMA: Neural Image Assessment. *CoRR.* **abs/1709.05424** (2017)

[61] Wagner, M., Lin, H., Li, S. & Saupe, D. Algorithm Selection for Image Quality-assessment. (2019)

[62] Jiang, X., Shen, L., Feng, G., Yu, L. & An, P. Deep Optimization model for Screen Content Image Quality-assessment using Neural Networks. (2019)