6 Supplementary

To make our model fully reproducible, we present complete implementation details in Section 6.1. Besides, our code library will be released upon acceptance. We report more comparisons between our QVM module and the 2D matching/relation techniques [1, 5, 43] in Section 6.2 to demonstrate the superiority of QVM in instance-level 3D matching. For clear reference, we display samples from the newly proposed RoboTools benchmark in Section 6.3. In Section 6.4 and Section 6.5, we present more detection qualitative results and voxel visualizations, respectively. Finally, we provide extended related works discussions in Section 6.6, where we exhaustively compare VoxDet with the existing instance-level tasks, including visual tracking, instance pose estimation, and instance retrieval.

6.1 Implementation Details

Model Structure: We adopt ResNet50 [44] with feature pyramid network [25] as our feature extractor $\psi(\cdot)$. The default multi-scale ROIAlign in [25] is leveraged to obtain the 2D proposal features, where the dimensions are $N = 500, C = 256, w = 7$. In our 2D-3D mapping, we set $C/d = 32, d = 8$, which results in the voxel feature dimension $C_v = 256, D = 16, L = 14$. All the 3D convolutions in TVA and QVM take kernel size as 3 and the padding equals to 1, so that the dimension of the voxels remains the same throughout the two modules. For the $\mathrm{Rot}(\cdot, \cdot)$ function, we have followed [10] to use `torch.nn.functional.affine_grid()` and `torch.nn.functional.grid_sample()` functionalities. Though the 2D-3D mapping can learn the rotations in the physical world, it sacrifices some semantics information in the feature channels when reshaping. Therefore, in QVM, we have a global matching branch to retrieve the lost semantic information. To be more specific, we apply global average pooling on the support features to get a support vector $\mathbf{k} \in \mathbb{R}^{1\times C\times 1\times 1}$. Then we adopt depth-wise convolution between $\mathbf{k}$ and $\mathbf{F}^{\mathrm{Q}}$ to get a correlation map. Note that this correlation map preserved all the semantic channels from the backbone $\psi(\mathrm{cot})$, so that the lost information in the 2D-3D mapping. The map is added to the voxel relation output $\mathcal{R}_{\mathrm{v}}(\mathbf{V}^{\mathrm{S}}, \mathrm{Rot}(\mathbf{V}^{\mathrm{Q}}, \hat{\mathbf{R}}^{\mathrm{Q}}))$ for the final score.

Training Details: In the first reconstruction stage, we set the loss weights as $w_{\mathrm{recon}} = 10.0, w_{\mathrm{gan}} = 0.01, w_{\mathrm{percep}} = 1.0$. The model is trained for 16 epoch on the 9600 instances from OWID datasets. We leveraged Adam optimizer [45] with a base learning rate of 5×10^{-5} during training. In the second detection stage, we initialize the 2D-3D mapping modules in TVA and QVM with the reconstruction pre-trained weights. VoxDet first only learns the detection task, without learning the rotation estimation, *i.e.*, the loss weights are set as $w_1 = w_2 = w_3 = w_4 = w_5 = 1.0, w_6 = 0$ in the first 10 epochs, where SGD is leveraged as an optimizer with 0.02 base learning rate. Note that in this stage, the 2D-3D mapping part only takes $\frac{1}{10}$ of the base learning rate. Then in the final epoch, VoxDet learns the rotation estimation with the detection part fixed, *i.e.*, $w_1 = w_2 = w_3 = w_4 = w_5 = 0.0, w_6 = 1.0$.

6.2 More Matching Module Comparison

We compare QVM with more matching techniques in Table 4, where the averaged results on the cluttered LM-O [16] and RoboTools benchmark are reported. We first ablate the Voxel Relation module in QVM, which results in QVM†. Specifically, all the Voxel Relation in QVM† are replaced by a simple depth-wise convolution, *i.e.*, we first apply global average pooling on the template voxel to get a feature vector, which is then taken as the convolution kernel to calculate the correlation voxel from the queries. We can see such a naive design will result in a performance drop.

Table 4: Comparison with different types of matching module. We compare QVM with the correlation in [5], class-level relation proposed in [1], and the class distance defined in FSDet [43].

Method	mAR	AR50	AR75
QVM (Ours)	**21.70**	**25.40**	**19.05**
QVM†	20.80	22.45	18.35
2D Corr. [5]	18.30	20.95	16.00
2D Relation [1]	19.70	20.65	18.55
FSDet [43]	16.05	19.05	14.10
Local Matching [46, 47]	10.60	9.600	7.850

For all the rest methods, we used the same open-world detector to obtain the universal proposals, which are then matched with the template images using different matching techniques. To be more specific, 2D Corr. [5] constructs support vectors from every reference image. Then, depth-wise convolution is conducted between each support vector and the proposal patch. The resulting correlation maps are sent to an MLP for classification score. In 2D Relation [1], we substitute the simple depth-wise convolution in 2D Corr. with the spatial and channel relation proposed in [1]. In

Figure 8: The instances and test scenes in the newly built RoboTools benchmark. The 20 unique instances are recorded as multi-view videos, where the relative camera poses between frames are provided. RoboTools consists of various challenging scenarios, where the desired instance could be under severe occlusion or in different orientation.

FSDet [43], the depth-wise convolution in 2D Corr is replaced by the distance defined in [43]. Since they are geometry-unaware, we find all the 2D matching/relation techniques worse than our QVM module.

Additionally, we designed a Local Matching baseline [47, 46]. In Local Matching, we first extract local key points from the reference images and proposals using SuperPoint [47]. Then the points descriptors are matched by SuperGlue [46]. We take the mean matching score of all the points in the proposal as their classification score. We find such an implementation, though geometry-invariant, falls short in our task since it lacks semantic representation of the whole instance.

6.3 Datasets Examples

The 20 instances and 24 scenes in the newly built RoboTools benchmarks are presented in Fig. 8. Compared with existing benchmarks [16, 17], RoboTools is much more challenging with more cluttered backgrounds and more severe pose variation.

6.4 More Detection Visualizations

We present more detection qualitative comparisons in Fig. 9. VoxDet, in red, is compared with three baselines, DTOID [6], Gen6D [5], and $\mathrm{OLN}_{\mathrm{DINO}}$. Compared with previous instance detectors [6, 5], VoxDet is more robust under orientation variation and severe occlusion by virtue of the learned geometric knowledge. For example, in the LM-O benchmark, second column, when the duck is partially occluded and the egg box is in different orientations, VoxDet can still find them while Gen6D fails. Compared with similarity matching [9], VoxDet can better distinguish similar instances via

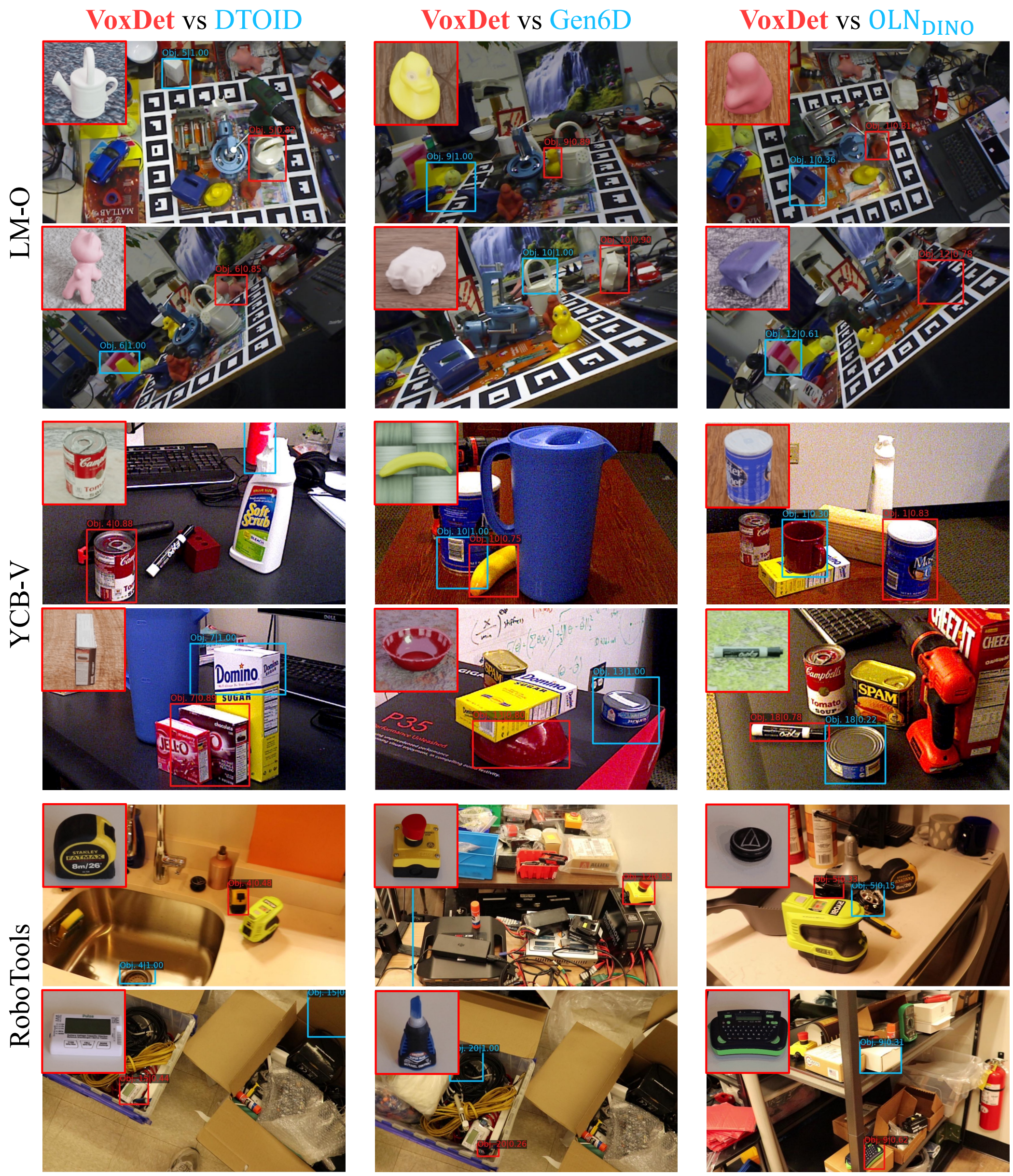

Figure 9: Detection qualitative results comparison between VoxDet and 2D baselines, DTOID [6], Gen6D [5], OLN$_{\text{DINO}}$ [13, 9] on the three benchmarks. VoxDet shows better robustness under pose variance and occlusion. These qualitative comparisons can be better visualized in our supplementary video.

the QVM module. For instance, in the RoboTools benchmark, the third column, the desired instance could be distracted by the motor, which has similar appearances but different geometry. Our VoxDet can discover such geometric differences and make correct classification, while the similarity matching falls short even if the feature from DINO [9] is stronger than ResNet50 [44].

6.5 More Voxel Visualizations

We display more voxel activation visualization in Fig. 10. In these experiments, we backpropagate the final proposal's classification scores and visualize each grid's activation value in the template voxel, following GradCAM [48]. For better visualization effects, we set a threshold and only display the volumes with high activation values with the rest nearly transparent. We find that as the query

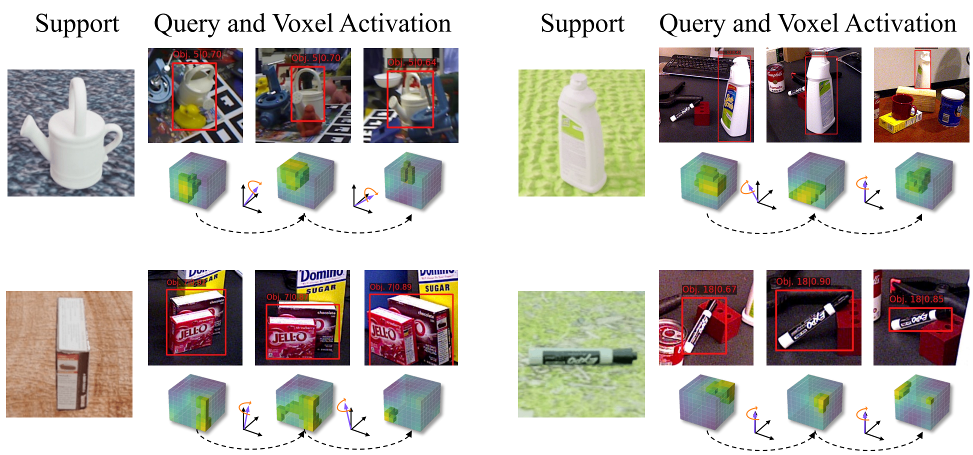

Figure 10: Visualization of the high activation grids during matching. When query instance rotates along a certain axis, the location of the high-activated grids also roughly rotates in the corresponding direction. The rotation axises are displayed for better understanding.

rotates along a certain axis, the location of the high-activated grids also rotates along corresponding axes. We attribute these results to the learned geometric knowledge in our voxel representation.

6.6 Extended Related Works

Visual Object Tracking aims to localize a general target instance in a video, given its initial state in the first frame. Early methods adopt discriminative correlation filters [49–51], where the calculation in the frequency domain is so efficient that real-time speed can be achieved on a single CPU. More recently, methods are developed on Siamese Network [52] and Transformers [53–55]. Unlike detection, object tracking has a strong temporal consistency assumption, *i.e.*, the location and appearance of the instance in the next frame do not largely vary from the previous frame. So that they only conduct detection/matching in the small search region with a single 2D template, which can't work for our whole image detection setting.

Instance Pose Estimation is developed to estimate the 6 DoF pose of an unseen instance. Some of them [56, 57] match the local point features and resort to RANSAC to optimize the relative pose. While others [5, 58] first selects the closest template frame and then conducts pose refinement on the known template poses. Most of these methods usually assume the instance detection is perfect, *i.e.*, they crop the instance out of the query image with the ground truth box and estimate the pose on the small object-centered patch. Our VoxDet can serve as their front-end, which is robust to cluttered environments, thus making the detection-pose estimation framework more reliable.

Instance Retrieval hopes to retrieve a specific instance from a large database with a single reference image [59–64]. Some early work extracts local point features from template and query patch for image matching [60, 47], which may suffer from poor discriminative capability. More recent work resorts to the deep neural network for a global representation of the instance [61–64], which is compared with the features from query images. However, most of them construct 2D template features from the reference, so that their representation is unaware of the 3D geometry of the instance, which may not be robust under severe pose variation. Besides, instance retrieval methods usually require high-resolution query images for the discriminative features, while the instance in our cluttered query image could be in low-resolution, which sets additional barriers to these approaches.

References

[1] B. Li, C. Wang, P. Reddy, S. Kim, and S. Scherer, "AirDet: Few-Shot Detection without Fine-tuning for Autonomous Exploration," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 427–444.

[2] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense Relation Distillation With Context-Aware Aggregation for Few-Shot Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 10 185–10 194.

[3] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring Plain Vision Transformer Backbones for Object Detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 280–296.

[4] I. Misra, R. Girdhar, and A. Joulin, "An End-to-End Transformer Model for 3D Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2906–2917.

[5] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang, "Gen6D: Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images," in *Proceedings of the European Conference on Computer Vision(ECCV)*, 2022, pp. 298–315.

[6] J.-P. Mercier, M. Garon, P. Giguere, and J.-F. Lalonde, "Deep Template-based Object Instance Detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1507–1516.

[7] A. Osokin, D. Sumin, and V. Lomakin, "OS2D: One-Stage One-Shot Object Detection by Matching Anchor Features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 635–652.

[8] P. Ammirato, C.-Y. Fu, M. Shvets, J. Kosecka, and A. C. Berg, "Target Driven Instance Detection," *arXiv preprint arXiv:1803.04610*, 2018.

[9] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," 2023.

[10] Z. Lai, S. Liu, A. A. Efros, and X. Wang, "Video Autoencoder: Self-Supervised Disentanglement of Static 3D Structure and Motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9730–9740.

[11] H.-Y. F. Tung, R. Cheng, and K. Fragkiadaki, "Learning Spatial Common Sense with Geometry-Aware Recurrent Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2595–2603.

[12] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang, "HoloGAN: Unsupervised Learning of 3D Representations from Natural Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7588–7597.

[13] D. Kim, T.-Y. Lin, A. Angelova, I. S. Kweon, and W. Kuo, "Learning Open-World Object Proposals without Learning to Classify," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5453–5460, 2022.

[14] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "ShapeNet: An Information-Rich 3D Model Repository," *arXiv preprint arXiv:1512.03012*, 2015.

[15] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora *et al.*, "Abo: Dataset and Benchmarks for Real-World 3D Object Understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 126–21 136.

[16] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D Object Pose Estimation using 3D Object Coordinates," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 536–551.

[17] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research," in *Proceedings of the International Conference on Advanced Robotics (ICAR)*, 2015, pp. 510–517.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, 2014, pp. 580–587.

[20] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1440–1448.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-time Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7263–7271.

[24] ——, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, pp. 1–6, 2018.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.

[26] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8681–8690.

[27] X. Wang, T. Huang, J. Gonzalez, T. Darrell, and F. Yu, "Frustratingly Simple Few-Shot Object Detection," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 9919–9928.

[28] Y. Xiao, V. Lepetit, and R. Marlet, "Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3090–3106, 2022.

[29] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards Open World Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5830–5840.

[30] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, "OW-DETR: Open-World Detection Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9235–9244.

[31] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment Anything," *arXiv preprint arXiv:2304.02643*, 2023.

[32] J. L. Schonberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[33] A. W. Harley, S. K. Lakshmikanth, F. Li, X. Zhou, H.-Y. F. Tung, and K. Fragkiadaki, "Learning from Unlabelled Videos Using Contrastive Predictive Neural 3D Mapping," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019, pp. 1–19.

[34] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer, "Deepvoxels: Learning Persistent 3D Feature Embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2437–2446.

[35] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NerF: Representing Scenes as Neural Radiance Fields for View Synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[36] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[37] R. Yang, G. Yang, and X. Wang, "Neural Volumetric Memory for Visual Locomotion Control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1–12.

[38] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A Generalizable Learning-based VO," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2021, pp. 1761–1772.

[39] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the Continuity of Rotation Representations in Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5745–5753.

[40] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[41] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv preprint arXiv:1911.01911*, 2019.

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[43] Y. Xiao and R. Marlet, "Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 192–210.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[45] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching with Graph Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947.

[47] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-Supervised Interest Point Detection and Description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2018, pp. 224–236.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[49] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[50] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "AutoTrack: Towards High-Performance Visual Tracking for UAV with Automatic Spatio-Temporal Regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 920–11 929.

[51] B. Li, C. Fu, F. Ding, J. Ye, and F. Lin, "ADTrack: Target-Aware Dual Filter Learning for Real-Time Anti-Dark UAV Tracking," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 496–502.

[52] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4277–4286.

[53] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "HiFT: Hierarchical Feature Transformer for Aerial Tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 457–15 466.

[54] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 341–357.

[55] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-End Tracking with Iterative Mixed Attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 608–13 618.

[56] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-Shot Object Pose Estimation without CAD Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6825–6834.

[57] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen, "FS6D: Few-Shot 6D Pose Estimation of Novel Objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6814–6824.

[58] Q. Gu, B. Okorn, and D. Held, "OSSID: Online Self-Supervised Instance Detection by (and for) Pose Estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3022–3029, 2022.

[59] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep Learning for Instance Retrieval: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[60] A. Babenko and V. Lempitsky, "Aggregating Local Deep Features for Image Retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 1269–1277.

[61] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307.

[62] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk, "Solar: Second-order loss and attention for image retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 253–270.

[63] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep Image Rretrieval: Learning Global Representations for Image Search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 241–257.

[64] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang, "Dolg: Single-Stage Image Retrieval with Deep Orthogonal Fusion of Local and Global Features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 772–11 781.