3D Reconstruction in Dynamic Environments

Chi Zhang
Shanghai Jiaotong University
zhangciiiii@sjtu.edu.cn

Bo Yue
Shanghai Jiaotong University
yuebo2017JD@sjtu.edu.cn

Abstract

In this project, we build our model on a RGB-D sensor, Realsense D435, that is able to consistently inpaint and map scenes containing multiple dynamic elements. Our model consists of two sub models: the image inpainting one and the 3D reconstruction one. The image inpainting part aims to detect and eliminate all dynamic objects, and inpaint the occluded background with plausible imagery, which is approached by two networks: mask net and pose net. Mask net generates masks for dynamic objects to erase them from the static background while pose net produces pose transformation between two consecutive frames to obtain the pose of sensor in world coordinate. Afterwards, the 3D reconstruction part employs information of the pose of sensor to project static pixels in RGB image and corresponding depth image into 3D world coordinate as point clouds. This serves as a panorama of the static scene, and is achieved by means of truncated signed distance function(TSDF).

In real-world application, a sequence of frames, or a video is recorded by D435 as input of our model while a map with point clouds and a sequence of frames containing only static background serve as output.

1. Introduction

Localization and mapping are basic capabilities of robotic systems operating in real-world environments. The majority of simultaneous localization and mapping(SLAM) methods focus on static environments, but the deployment in real-world situations requires them to handle dynamic objects, since dynamic objects degrade the performance of vision-based robotic pose-estimation or localization tasks. SLAM method which is still robust in dynamic environments, i.e. DynaSLAM is a quite challenging task today. This is because a dynamic object may disguise itself as a part of the static background and moves after a long duration and therefore the robot needs to map consistently with static background only. DynaSLAM, though difficult to implement, can be quite useful in application such as virtual and augmented reality for tidy background.

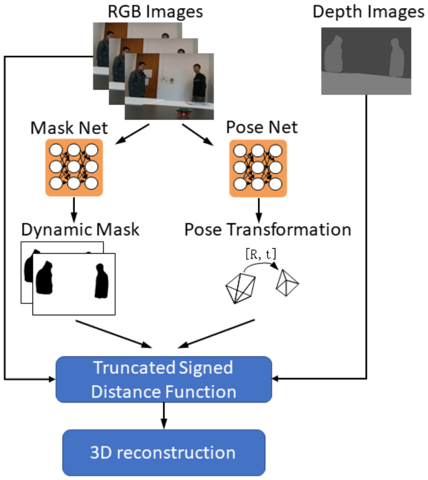


Figure 1. The whole process of our project.

In this paper, we propose two networks in image inpaint part, shown in Fig.1: the aforementioned maks net and pose net. The two networks are based on deep neural networks to detect, eliminate and inpaint static background. In terms of mask net, the traditional method relies on geometry and thus heavily depends on the precision of camera sensor which generates RGB-D images. It does not make sense for all camera sensors to meet the requirements of a relatively-high precision level, however. Additionally, our approach are not semantic-based such as the renowned Mask R-CNN [6], since it is time-consumed if applied to generate masks. Instead, we use the result of semantic segmentation to produce labelled masks to train our mask net for fast application in other scenes. We design three loss functions for mask net, namely consensus loss, explanatory loss and smoothness loss. Consensus loss takes into account the difference between our generated mask and the ground truth; explanatory loss assumes most of the pixels in a certain frame to be static; smoothness loss takes into consideration that the masked region should tend to be large blocks, rather than multiple discrete scatters. Based on upon these, the mask net takes three consecutive images of frames as input and output masks for the first and the third frame.

As for pose net, we employ residuals, namely the difference between two consecutive frames to generate pose transformation of our sensor for mapping in 3D reconstruction part. Pose net also take three consecutive frames as input while output two pose transformations from the first frame to the second frame and from the third frame to the second frames. Notably, in both net, we feature three consecutive frames and employ the middle frame as out goal frame for simpler code realization.

For 3D reconstruction part, we feature TSDF method to restore the static background with point clouds. The TSDF encodes the distance from the voxel to the closest surface, and is updated at every frame. The TSDF is efficiently represented using voxel hashing.

Our experimental datasets entail TUM RGB-D dataset [21] and Bonn RGB-D[1] [16]. Also, we apply D435 in generating videos for our dataset.

2. Related Work

Semantic Segmantation Deep learning method provides new ideas for the SLAM task. Random sample consensus (RANSAC) is used to select feature points for estimation, which will fail if the points are on dynamic objects or sky. [9] propose exclude feature points using a mask produced by semantic segmentation. This idea is also used in [31], which applies semantic segmentation to update the static probability distribution of the pixels. Segmentation networks get developed recently. [15] can have a better segmentation performance. Depth information can be estimated with the semantic segmentation together in [14]. Semantic segmentation networks dedicated to outdoor scene location tasks have also been proposed in [20]. Encoding and decoding ideas are used to improve network performance in [26]. In addition to monocular segmentation, 3D segmentation based on binocular has also been studied [11, 10, 29].

However, there are many disadvantages to dealing with SLAM using only semantic segmentation, and use other methods to assist it can achieve better results. [18] takes full advantage of using instance-level semantic segmentation, which does not require known models of the objects, and the mask for objects can be completed from depth and surface normal cues. [1] adopts the consistency of consecutive frames, which firstly use Mask R-CNN to roughly select the potential dynamic objects and use the residual to filter. [28] do the segmentation firstly and then use information of consecutive frames to label the dynamic objects. Instead, [13] abandons the consistency of consecutive frames and process each frame individually. [13] proposes the tracking idea to generate the mask in the whole scenario with the help of the labeled mask for the first frame. Another paper proposes to use k-means on depth images to perform clusters for semantic refinement [3]. This paper use end-to-end deep learning framework, which employ gray scaled figures and mGAN to produce gray scaled static background [2].

In addition to the idea of semantic segmentation, other methods are also used in this task [7]. These methods not merely assist semantic segmentation, but also use far more advanced techniques. These methods are a combination of both traditional and learning-based techniques. adds optical flow to infer the position of the mask at the next frame, and the output semantic-level mask will be input into the network, forming a cyclic structure. [23] uses optical flow to calculate the probability of pixels to be dynamic, and a novel memory module is proposed for the probability updating. [24] proposed object flow based on optical flow and segmentation, and a scoring function based on Gaussian Mixture Model and Euclidean Distance is designed for objects location. Scene flow can provide more information than optical flow, and [4] use deep learning to estimate scene flow and depth map, then the motion of segmented objects can be inferred.

Image Inpainting After we have eliminated the dynamics from the frame, we have to inpaint the holes. Here, we assume the holes are often blocks not are not big, which is formulated in three loss functions for mask net. Among the non learning-based approaches to image inpainting, propagating appearance information from neighboring pixels to the target to-be-eliminated region is the usual procedure [22]. Accordingly, these methods often succeed in dealing with narrow holes, where color and texture vary smoothly, but fail to handle big holes, resulting in oversmoothing or lack of details, in other words. Differently, patch-based methods [5] operate by iteratively searching for relevant patches from the image static, namely non-hole regions. However, due to the number of images or frames and the large number of pixels in each frame, these approaches are often computationally expensive and therefore not fast enough for real-time applications. Moreover, they do not make semantically aware patch selections.

For learning-based approach, deep learning based methods usually initialize the image holes with a constant value, and further pass it through a CNN. Context Encoders [17] were among the first ones to successfully use a standard pixel-wise reconstruction loss, as well as an adversarial loss for image inpainting tasks. Due to the different reflecting nature of different semantic part, Yang *et al.* [27] take the result from Context Encoders as input and then propagates the texture information from non-hole regions to fill the hole regions as post-processing. Taking advantage

[1] available at URL: `https://vision.in.tum.de/data/datasets/rgbd-dataset/download` & `http://www.ipb.uni-bonn.de/data/rgbd-dynamic-dataset/`

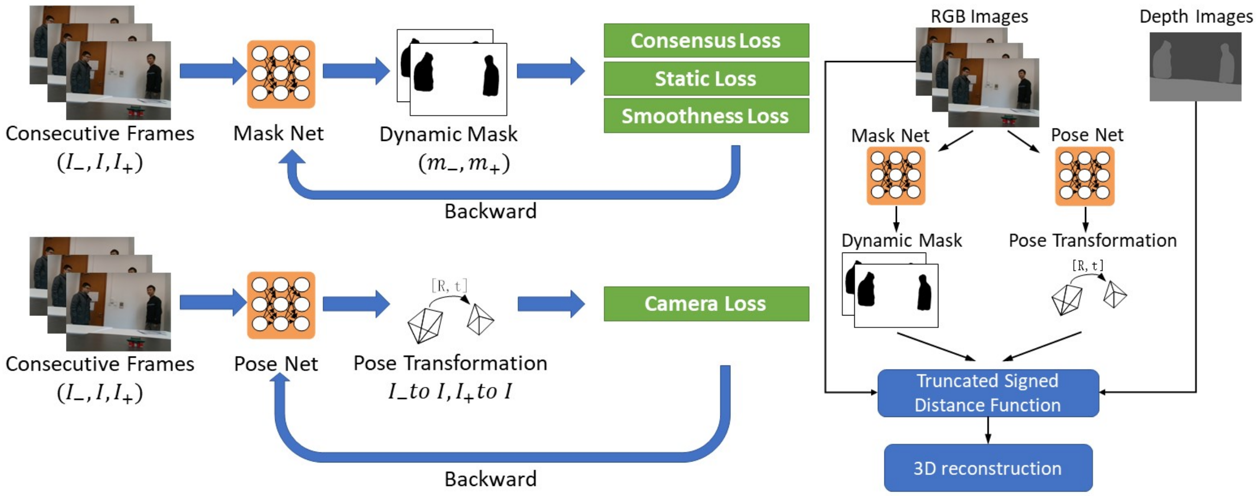


Figure 2. The detailed structure of mask net and pose net.

of aforementioned patch-based methods, Song *et al.* [19] present a refinement network in which a blurry initial hole-filling result is used as the input, then iteratively replaced with patches from the closest non-hole regions in the feature space. Iizuka *et al.* [8] extend Content Encoders by defining both global and local discriminators, then apply a post-processing. Following this work, Yu *et al.* [30] replaced the post-processing with a refinement network powered by the contextual attention layers. The recent work of Liu *et al.* [12] obtains amazing inpainting results by using partial convolutions. In contrast, the work by Ulyanov*et al.* [25] proves that there is no need for external dataset training. The generative network itself can rely on its structure to complete the corrupted image. However, this approach usually applies several iterations (∼50000) to get good and detailed results.

3. Our Approach

In this section, we present our specific approach to build our model, particularly the two networks of the image inpainting part. We make two innovations here. First, we use a mask net instead of residuals to generate dynamic masks. Second, we use a pose net instead of ORB-SLAM2 to provide pose estimations. The details of two networks are shown in Fig.2.

3.1. Mask Net and Pose Net

We use C_ψ, M_χ to represent the pose net and the mask net respectively. The subscript ψ, χ represent the parameters of the network. Consecutive frames are denoted by I_-, I, I_+, where I represents the target frame and I_-, I_+ represent the neighboring reference frames.

With the consecutive frames and the pose net, the pose transformation can be estimated as:

$$e_-, e_+ = C_\psi(I_-, I, I_+) \tag{1}$$

where e_-, e_+ represent the camera motion from reference frame I_-, I_+ to the target frames I. With the consecutive frames and the mask net, the segmentation mask can be estimated as:

$$m_-, m_+ = M_\chi(I_-, I, I_+) \tag{2}$$

where $m_-, m_+ \in [0, 1]$ represent the segmentation mask, the larger number signifies a higher probability for the pixel to be static.

3.2. Loss function

The parameters ψ, χ are updated by reducing the loss

$$L = \lambda_C L_C + \lambda_M L_M \tag{3}$$

where λ_C, λ_M are the weights for camera loss and mask loss. The camera loss which helps the training of pose net is given by

$$L_C = \sum_{s \in \{+,-\}} \sum_{\Omega} \rho(I, w_c(I_s, e_s, d_s)) \cdot m_s \tag{4}$$

where w_c represents the image warp process with the depth d_s in the whole pixel domain Ω. The robust error $\rho(x, y)$ is computed as

$$\rho(x, y) = \lambda_\rho \sqrt{(x-y)^2 + \epsilon} + (1 - \lambda_\rho)\mu(x, y) \tag{5}$$

$$\mu(x, y) = 1 - \frac{(2\mu_x\mu_y + c_1)(x\mu_x y + c2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)} \tag{6}$$

where the second equation is called as structure similarity loss [**?**]. The μ_x, μ_y are the local mean over the pixel neighborhood and the σ_x, σ_y are the variance over the pixel neighborhood. λ_ρ, c_1, c_2 are fixed constants.

The loss for mask net L_M is formulated by

$$L_M = L_{con} + L_{sta} + L_{smo} \tag{7}$$

Consensus loss C_{con} aims to help mask net identify dynamic objects in the frame. The Mask_RCNN is involved to provide segmentation for frame (I_-, I_+), and the ground truth of the dynamic mask $m_{-,gt}, m_{+,gt}$ can be generated. The $m_{s,gt}$ supervise the training of mask net by

$$L_{con} = \sum_{s\in\{+,-\}} \sum_{\Omega} H(m_s, m_{s,gt}) \tag{8}$$

where the $H(x, y)$ is cross entropy loss.

Static loss L_{sta} aims to make the whole scene tends to be static. The loss is given by

$$L_{sta} = \sum_{s\in\{+,-\}} \sum_{\Omega} H(m_s, \mathbf{1}) \tag{9}$$

Smoothness loss L_{smo} aims to make the mask for the objects to be large blocks and make the edges to be smooth. The loss is given by

$$L_{smo} = \|\lambda_e \nabla m_-\|^2 + \|\lambda_e \nabla m_+\|^2 \tag{10}$$

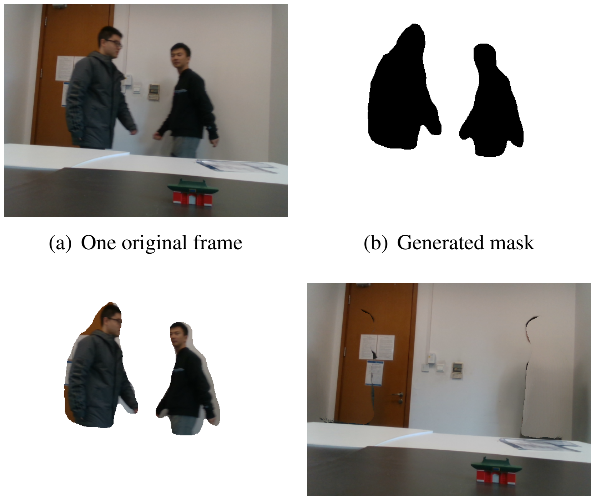
where $\lambda_e = e^{\nabla I}$ and ∇ is the is the gradient of the mask.

4. Results

In this section, we present the results of our approach. In image inpainting part, we successfully restore a static background video. Due to the fact that a static paper cannot include a live video, we present several inpainted pictures to showcase our results. As to 3D reconstruction part, we manage to build 3D static background model in world coordinate with voxels. Again, due to the fact that the paper is 2D while the static background model is 3D, we present two pictures of the model from two aspects.

4.1. Image Inpainting Part

As mentioned above, the image inpainting part eliminates dynamics in each frame from a certain video sequence and incorporates these static frames together to produce a final synthetic video sequence only with static background. A detailed process of inpainting a frame is depicted in Fig.3. The original video is taken by us at a studyroom in our school's main library. It is clear that we succeed in inpainting each frame, thus successfully inpainting the whole video sequence.

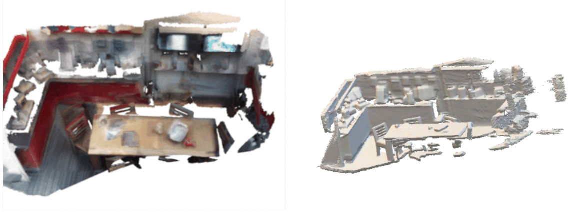


(a) One original frame
(b) Generated mask
(c) Elimianted dynamics
(d) Inpainted frame

Figure 3. The up-left figure is one original frame from a video sequence recorded by our sensor D435 while the up-right figure is the generated mask from our mask net. The down-left figure is the eliminated dynamic objects by means of generated masks while the down-right figure is the inpainted frame for the original frame, which is used to produce video sequence of static background only.

4.2. 3D Reconstruction Part

As mentioned above, we employ TSDF method to reconstruct 3D static background model in this part. By taking RGB-D images as input, we succeed in projecting only static pixels from inpainted frames into the world coordinate to form a 3D static background model, shown in Fig.4.



(a) Mapping of the static scene
(b) 3D reconstructed model

Figure 4. The left figure is the mapping of static scene using inpainted RGB-D pictures while the right figure is the 3D reconstructed static background model using TSDF method.

5. Conclusion

We presented how we conduct our project: 3D reconstruction in dynamic environments, which is composed of two parts. In the first part, we succeed in inpainting a video recorded by us. In the second part, we manage to project pixels in RGB-D images into world coordinate as point clouds by means of TSDF method. We test our approach on the popular TUM RGB-D dataset, as well as on the Bonn RGB-D dynamic dataset and also videos recoreded by our sensor D435. We meet the requirements in the project.pdf.

Future work might include researching more on semantic SLAM methods, adding the thought of tracking into our model and taking into account the texture of dynamic objects. We wish to combine traditional method along with learning-based method in semantic, namely image inpainting part. Also, pose estimation of dynamic objects is expected to be included to better eliminate dynamics. Finally, as far as a certain secsor is concerned, the texture of object influences the pixels representing these objects in RGB-D images, thanks to their different reflecting properties.

Acknowledgement

We appreciate Prof. Gao and teaching assistant Wu for the assistance and guidance they give us when conducting this project. Furthermore, we thank them for the server they provide for us to implement our algorithm and run our dataset.

References

[1] Berta Bescos, José M Fácil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 3(4):4076–4083, 2018.

[2] Berta Bescós, José Neira, Roland Siegwart, and Cesar Cadena. Empty cities: Image inpainting for a dynamic-object-invariant space. *CoRR*, abs/1809.10239, 2018.

[3] Mihai Bujanca, Mikel Luján, and Barry Lennox. Fullfusion: A framework for semantic reconstruction of dynamic scenes. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 10 2019.

[4] Zhe Cao, Abhishek Kar, Christian Hane, and Jitendra Malik. Learning independent object motion from unlabelled stereoscopic videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5594–5603, 2019.

[5] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *Proceedings of SIGGRAPH 2001*, pages 341–346, August 2001.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[7] Yuan-Ting Hu, Jia-Bin Huang, and Alexander Schwing. Maskrnn: Instance level video object segmentation. In *Advances in Neural Information Processing Systems*, pages 325–334, 2017.

[8] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, 36(4):107:1–107:14, 2017.

[9] Masaya Kaneko, Kazuya Iwami, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Mask-slam: Robust feature-based monocular slam by masking using semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 258–266, 2018.

[10] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019.

[11] Peiliang Li, Tong Qin, et al. Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–661, 2018.

[12] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. *CoRR*, abs/1804.07723, 2018.

[13] Keviskokitsi Maninis, Sergi Caelles, Yuhua Chen, Jordi Ponttuset, Laura Lealtaixe, Daniel Cremers, and L Van Gool. Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1515–1530, 2019.

[14] Vladimir Nekrasov, Thanuja Dharmasiri, Andrew Spek, Tom Drummond, Chunhua Shen, and Ian Reid. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7101–7107. IEEE, 2019.

[15] Vladimir Nekrasov, Chunhua Shen, and Ian Reid. Light-weight refinenet for real-time semantic segmentation. *arXiv preprint arXiv:1810.03272*, 2018.

[16] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. 2019.

[17] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *CoRR*, abs/1604.07379, 2016.

[18] Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20. IEEE, 2018.

[19] Yuhang Song, Chao Yang, Zhe Lin, Hao Li, Qin Huang, and C.-C. Jay Kuo. Image inpainting using multi-scale feature image translation. *CoRR*, abs/1711.08590, 2017.

[20] Erik Stenborg, Carl Toft, and Lars Hammarstrand. Long-term visual localization using semantically segmented images. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6484–6490. IEEE, 2018.

[21] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[22] Alexandru Telea. An image inpainting technique based on the fast marching method. *J. Graphics, GPU, Game Tools*, 9:23–34, 2004.

[23] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[24] Yihsuan Tsai, Minghsuan Yang, and Michael J Black. Video segmentation via object flow. pages 3899–3908, 2016.

[25] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Deep image prior. *CoRR*, abs/1711.10925, 2017.

[26] Xide Xia and Brian Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[27] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. *CoRR*, abs/1611.09969, 2016.

[28] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1168–1174. IEEE, 2018.

[29] HW Yu and Beom Hee Lee. A variational feature encoding method of 3d object for probabilistic semantic slam. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3605–3612. IEEE, 2018.

[30] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. *CoRR*, abs/1801.07892, 2018.

[31] Fangwei Zhong, Sheng Wang, Ziqi Zhang, and Yizhou Wang. Detect-slam: Making object detection and slam mutually beneficial. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1001–1010. IEEE, 2018.