

GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks

Yasin Almalioglu¹, Muhamad Risqi U. Saputra¹, Pedro P. B. de Gusmo¹, Andrew Markham¹, and Niki Trigoni¹

Abstract—In the last decade, supervised deep learning approaches have been extensively employed in visual odometry (VO) applications, which is not feasible in environments where labelled data is not abundant. On the other hand, unsupervised deep learning approaches for localization and mapping in unknown environments from unlabelled data have received comparatively less attention in VO research. In this study, we propose a generative unsupervised learning framework that predicts 6-DoF pose camera motion and monocular depth map of the scene from unlabelled RGB image sequences, using deep convolutional Generative Adversarial Networks (GANs). We create a supervisory signal by warping view sequences and assigning the re-projection minimization to the objective loss function that is adopted in multi-view pose estimation and single-view depth generation network. Detailed quantitative and qualitative evaluations of the proposed framework on the KITTI [1] and Cityscapes [2] datasets show that the proposed method outperforms both existing traditional and unsupervised deep VO methods providing better results for both pose estimation and depth recovery.

I. INTRODUCTION

Visual odometry (VO) and depth recovery are essential modules of simultaneous localization and mapping (SLAM) applications. In the last few decades, VO systems have attracted a substantial amount of attention, enabling robust localization and accurate depth map reconstruction. Monocular VO is confronted with numerous challenges such as large scale drift, the need for hand-crafted mathematical features and strict parameter tuning [3], [4]. Supervised deep learning based VO and depth recovery techniques have showed good performance in challenging environments and successfully alleviated issues such as scale drift, need for feature extraction and parameter finetuning [5]–[8]. VO as a regression problem in supervised deep learning exploits the capability of convolutional neural network (CNN) and recurrent neural network (RNN) to estimate camera motion, to calculate optical flow, and to extract efficient feature representations from raw RGB input [5]–[7], [9]. In recent years, unsupervised deep learning approaches have achieved remarkable results in various domains eliminating the need for labelled data [10], [11].

Years of research in visual SLAM have been inspired by human navigation that easily locates obstacles even in

¹Yasin Almalioglu, Muhamad Risqi U. Saputra, Pedro P. B. de Gusmo, Andrew Markham, and Niki Trigoni are with the Computer Science Department, The University of Oxford, UK {yasin.almalioglu, muhamad.saputra, pedro.gusmao, andrew.markham, niki.trigoni}@cs.ox.ac.uk

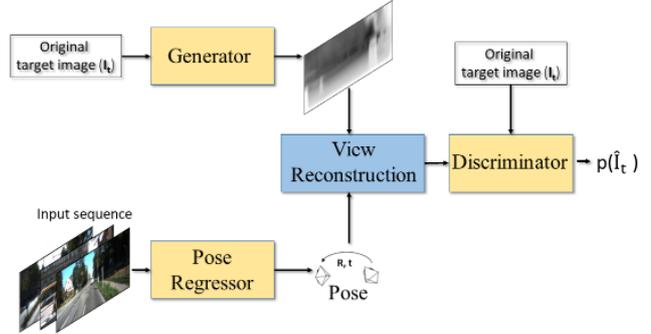


Fig. 1: Architecture overview. The unsupervised deep learning approach consists of depth generation, multi-view pose estimation, view reconstruction, and target discrimination modules. Unlabelled image sequences from different temporal points are given to the networks to establish a supervision signal. The networks estimate relative translation and rotation between consecutive frames from different perspectives parametrized as 6-DoF motion, and depth image as a disparity map for a given view.

unknown environments. A neuroscientific insight is that human brain saves a structural perception of the world, which makes it capable of real and imaginary scene reconstruction through vast previous experiences [12], [13]. In this study, we propose a novel real-time localization and generative depth estimation approach that mimics the remarkable ego-motion estimation and re-constructive scene representation capabilities of human beings by training an unsupervised deep neural network. The proposed network consists of a pose regressor and depth generator network. The former regresses 6 degree-of-freedom (DoF) pose values using CNN-RNN modules, and the latter generates depth maps using deep convolutional generative adversarial network (GAN) [14]. The model takes a sequence of monocular images to estimate 6-DoF camera motion and depth map that is sampled in the same input data distribution, which is trained in an end-to-end and unsupervised fashion directly from raw input pixels. A view reconstruction approach is utilized as part of the objective function for the training in a similar way to prior works [15]–[18]. The entire pose estimation and depth map reconstruction pipeline is a consistent and systematic learning framework which continually improves its performance by collecting unlabelled monocular video

data from numerous environments. This way, we want to mimic and transfer a continuous learning functionality from humans into VO domain.

In summary, the main contributions of our method are as follows:

- To the best of our knowledge, this is the first monocular VO method in literature, which uses adversarial and recurrent unsupervised learning approaches for joint pose and depth map estimation.
- We propose a novel adversarial technique for GANs to generate depth images without any need for depth groundtruth information.
- No strict parameter tuning is necessary for pose and depth estimation, contrary to traditional VO approaches.

Evaluations made on the KITTI [1] and Cityscapes [2] datasets prove the success of our pose estimation and depth map reconstruction approach. As the outline of this paper, the previous work in VO is discussed in section II. Section III gives an overview of the proposed approach named GANVO. Section IV describes the proposed unsupervised deep learning architecture and its mathematical background in detail. Section V shows our quantitative and qualitative results with comparisons to the existing VO methods. Finally, section VI concludes the study with some interesting future directions.

II. RELATED WORK

Camera motion and depth map estimations are well studied problems in machine vision domain. Many traditional techniques have been proposed in the last decade with state-of-the-art results. However, existing traditional techniques require accurate image correspondence between consecutive frames, which is frequently violated in challenging environments with low texture, complex scene structure, and occlusions [19]–[22]. Deep learning approaches are adopted in several components of the traditional techniques to solve problems such as feature extraction and matching, pose estimation, and optical flow calculation. An external supervised training has played a key role in deep learning approaches to solve these aforementioned issues.

Deep VO techniques involve 6-DoF camera pose estimation, depth map reconstruction, object segmentation, optical flow extraction, and sensor fusion approaches [5], [9], [23]–[29]. Deep 6-DoF pose regression from raw RGB images was first proposed using CNNs, which was also extended to raw RGB-D images for challenging environments [30]. Using RNNs that captures the temporal dynamics boosted the accuracy of 6-DoF pose estimation in deep VO approaches, resulting in a competitive performance against model-based VO methods [5]. The unsupervised deep learning studies on simultaneous estimation of the camera motion, image depth, surface normal and optical flow to learn structure from motion indicate that the joint training of VO components provides an implicit supervisory signal to the network [16], [17], [31]. One critical issue of these unsupervised studies is the fact that they use auto encoder-decoder based traditional depth estimators with a tendency to generate overly smooth

images [32]. To solve this, we apply GANs that provide shaper and more accurate depth maps. The second issue of the aforementioned unsupervised techniques is the fact that they only employ CNNs that only analyse just-in-moment information to estimate camera pose [5], [7]. We address this issue by employing a CNN-RNN architecture to capture temporal relations across frames.

III. ARCHITECTURE OVERVIEW

As shown in Fig. 1, the raw RGB sequences consisting of a target view and source views are stacked together to form an input batch to the multi-view pose estimation module. The module regresses the relative 6-DoF pose values of the source views with respect to the target view. In parallel, the depth generation module generates a depth map of the target view. The view reconstruction module synthesizes the target image using the generated depth map, estimated 6-DoF camera pose and nearby colour values from source images. The view reconstruction constraint that provides a supervision signal forces the neural network to synthesize a target image from multiple source images acquired from different camera poses. The view discriminator module tries to distinguish the synthesized target image from the original target image.

In the proposed adversarial scheme for a GAN, the objective of the generator is to trick the discriminator, i.e., to generate depth map for the target view reconstruction such that the discriminator cannot distinguish the reconstruction from the original. As opposed to the typical GANs, the output image of the generator is mapped to the color space of the image and the discriminator distinguishes this mapping from the original rather than a direct comparison of the generator output. The proposed scheme enables us to generate depth maps in an unsupervised manner.

IV. UNSUPERVISED DEPTH AND POSE ESTIMATION WITH GENERATIVE ADVERSARIAL NETWORK

The network architecture of the proposed method is shown in Fig. 2. The details of the architecture are explained in the following sections.

A. Depth Generator

The first part of the architecture is a depth network that generates a single-view depth map of the target frame. The depth network is based on a GAN design that tries to learn the probability distribution of the input images $p(\mathbf{I}_t)$, consisting of a generator network, G , a discriminator network D , and an encoder network E . The encoder E maps the input target image \mathbf{I}_t to a feature vector \mathbf{z} , i.e. $E(\mathbf{I}_t) = \mathbf{z}$. G maps the feature vector \mathbf{z} to the depth image space and D maps an input RGB image to an image likelihood.

B. Pose Regressor

The second network shown in the bottom of Fig. 2 tries to estimate relative pose $\mathbf{p} \in \mathbb{SE}(3)$ introduced by motion and temporal dynamics across frames. The convolution part of the pose estimation network extracts features from the input

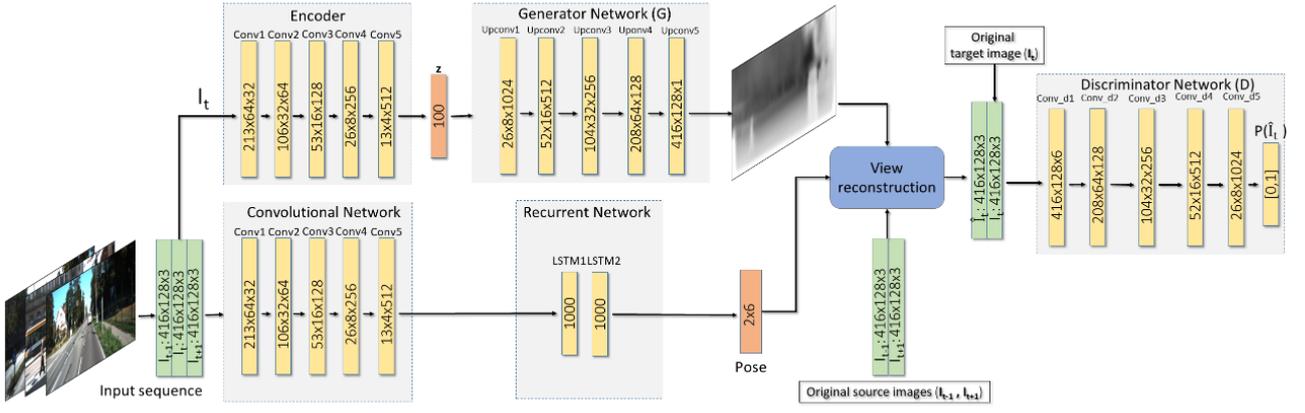


Fig. 2: The proposed architecture for pose estimation and depth map generation. The spatial dimensions of layers and output channels show the tensor shapes that flow through the network. Generator network G maps a feature vector \mathbf{z} generated by the encoder network E to the depth image space. The pose estimation network consists of a convolutional network that extracts VO related features, and a recurrent network that captures temporal relations among the input frame sequences. The pose results are collected after the recurrent network, which has 6 $(N - 1)$ output channels for 6-DoF motion parameters, where N is the length of the sequence. The view reconstruction algorithm is followed by a discriminator network D that maps the reconstructed RGB image to a likelihood of the target image. D decides whether it is shown the reconstruction or the original image.

frame sequences and propagates them to the RNN part. The LSTM modules output 6 $(N - 1)$ channels for 6-DoF pose values consisting translation and rotation parameters, where N is the length of the sequence. Although one LSTM is able to capture the sequential relations, a second LSTM improves the learning capacity of the network, resulting in a more accurate pose regression.

C. View Reconstruction

A sequence of 3 consecutive frames is given to the pose network as input. An input sequence is denoted by $\langle I_{t-1}, I_t, I_{t+1} \rangle$ where $t > 0$ is the time index, I_t is the target view, and the other frames are source views $I_s = \langle I_{t-1}, I_{t+1} \rangle$ that are used to render the target image according to the objective function:

$$\mathcal{L}_g = \sum_s \sum_p |I_t(p) - \hat{I}_s(p)| \quad (1)$$

where p is the pixel coordinate index, and \hat{I}_s is the projected image of the source view I_s onto the target coordinate frame using a depth image based rendering module. The rendering is based on the estimated depth image \hat{D}_t , the 4×4 camera transformation matrix \hat{T}_t and the source view I_s [33]. We denote the homogeneous coordinates of a pixel in the target view as p_t , and the camera intrinsics matrix as K . Coordinates of p_t are projected onto the source view p_s with:

$$p_s = K \hat{T}_t^{-1} \hat{D}_t(p_t) K^{-1} p_t. \quad (2)$$

Since the value of p_s is not discrete, an interpolation is required to find the expected intensity value at that position. To do that, we use bilinear interpolation using the 4 discrete

neighbours of p_s [34]. The mean intensity value for projected pixel is estimated as follows:

$$\hat{I}_s(p_t) = I_s(p_s) = \sum_{i \in \{top, bottom\}, j \in \{left, right\}} w^{ij} I_s(p_s^{ij}) \quad (3)$$

where w^{ij} is the proximity value between projected and neighbouring pixels, which sums up to 1.

D. View Discriminator

A realistic image is synthesized by the view reconstruction algorithm using the depth image generated by G and estimated pose values. D discriminates between the reconstructed image and the real image sampled from the target data distribution p_{data} , playing an adversarial role. These networks are trained by optimizing the objective loss function:

$$\mathcal{L}_d = \min_G \max_D V(G, D) = \mathbb{E}_{\mathbf{I} \sim p_{data}(\mathbf{I})} [\log(D(\mathbf{I}))] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (4)$$

where \mathbf{I} is the sample from the p_{data} distribution and \mathbf{z} is a feature encoding on the latent space.

E. The Adversarial Training

In contrast to the original GAN [14], we remove fully connected hidden layers for deeper architectures and use batchnorm in G and D networks. Pooling layers are replaced by strided convolutions and LeakyReLU activation is used for all layers in D . In G network, pooling layers are replaced by fractional-strided convolutions and ReLU activation is used for all layers except for the output layer that uses tanh non-linearity. The GAN with these modifications generates non-blurry images and resolves the convergence

problem during the training [35]. The final objective for the optimization of weights in the architecture is:

$$\mathcal{L}_{final} = \mathcal{L}_g + \beta \mathcal{L}_d \quad (5)$$

where β is the balance factor. The optimal β is experimentally found to be the ratio between the expected values \mathcal{L}_g and \mathcal{L}_d at the end of the training.

V. EXPERIMENTS AND RESULTS

We implemented the architecture with the publicly available Tensorflow framework [36]. Batch normalization is employed for all of the layers except for the output layers. The weights of the network are optimized with Adam optimization to increase the convergence rate, with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 0.1 and mini-batch size of 8. For training purposes, the input tensors of the model are assigned to sequential images of size 128×416 , whereas they are not limited to any specific image size at test time. Three consecutive images are stacked together to form the input batch. We use the KITTI dataset [37] for benchmarking and the Cityscapes dataset [2] for evaluating cross-dataset generalization ability in the experiments. The model is trained on a NVIDIA TITAN V model GPU. We compare the proposed method with standard training/test splits on the KITTI dataset for the odometry and monocular depth map estimation tasks.

A. Pose estimation benchmark

We have evaluated the pose estimation performance of our GANVO on the standard KITTI visual odometry split. The dataset contains 11 driving sequences with groundtruth odometry obtained through the IMU/GPS sensors, where the sequences 00 – 08 are used for training and 09 – 10 for testing without any use of the pose and depth groundtruth during the training session. The network regresses the pose predictions as 6-DoF relative motion (Euclidean coordinates for translation and rotation) between sequences. We compare the pose estimation accuracy with the existing unsupervised deep learning approaches with the same sequence length of 5, and monocular ORB SLAM. The results are evaluated using Absolute Trajectory Error (ATE) [38] for 5 consecutive input frames with an optimized scaling factor to resolve scale ambiguity, which is reported to be the best sequence length for the compared methods [17], [39]. As shown in Table I and Fig. 3, the proposed method outperforms all the competing unsupervised and traditional baselines, without any need of global optimization steps such as loop closure detection, bundle adjustment and re-localization, revealing that GANVO captures long-term high level odometry details in addition to short-term low level odometry features.

B. Single-view depth evaluation

We evaluate the performance of the proposed depth estimation approach on a benchmark split of the KITTI dataset [40] to compare with the existing learning-based and traditional depth estimation approaches. In a total of 44,000 frames, we use 40,000 frames for training and 4,000

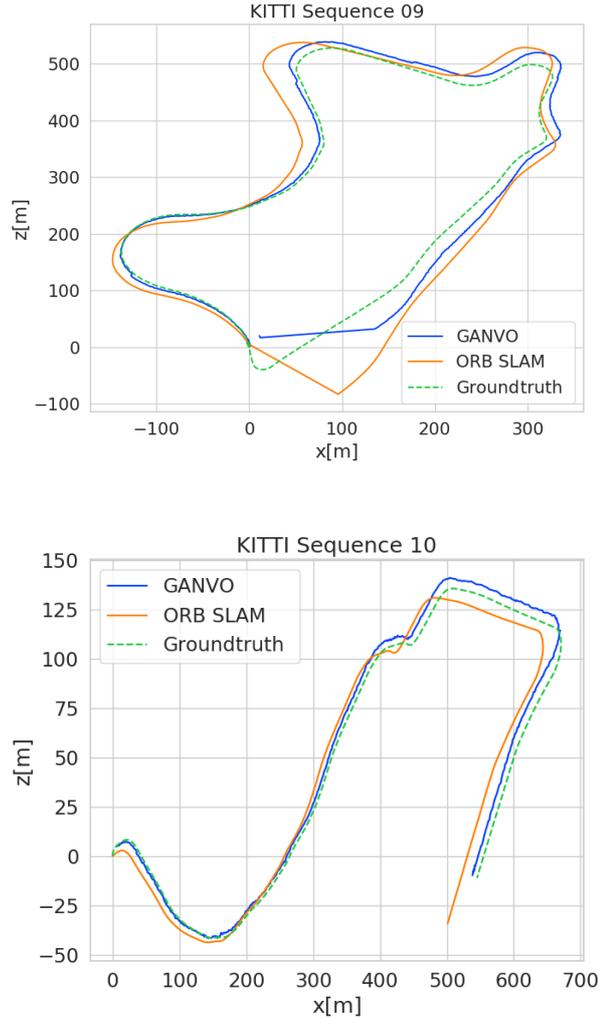


Fig. 3: Sample trajectories comparing the proposed unsupervised learning method GANVO with ORB SLAM, and the ground truth in meter scale. GANVO shows a better odometry estimation in terms of both rotational and translational motions.

frames for validation. The sequence length of the input data is set to be 3 frames during the training session to have the same evaluation setup with the compared methods, where the central frame is the target view for the depth estimation. The groundtruth captured by Light Detection and Ranging (LiDAR) sensor is projected into the image plane for the evaluation in terms of error and accuracy metrics. The predicted depth map, D_p , is multiplied by a scaling factor, \hat{s} , that matches the median with the groundtruth depth map, D_g , to solve the scale ambiguity issue, i.e. $\hat{s} = \text{median}(D_g) / \text{median}(D_p)$. Moreover, we test the adaptability of the proposed approach training on the Cityscapes dataset and finetuning on the KITTI dataset.

Figure 4 shows example depth map results predicted by

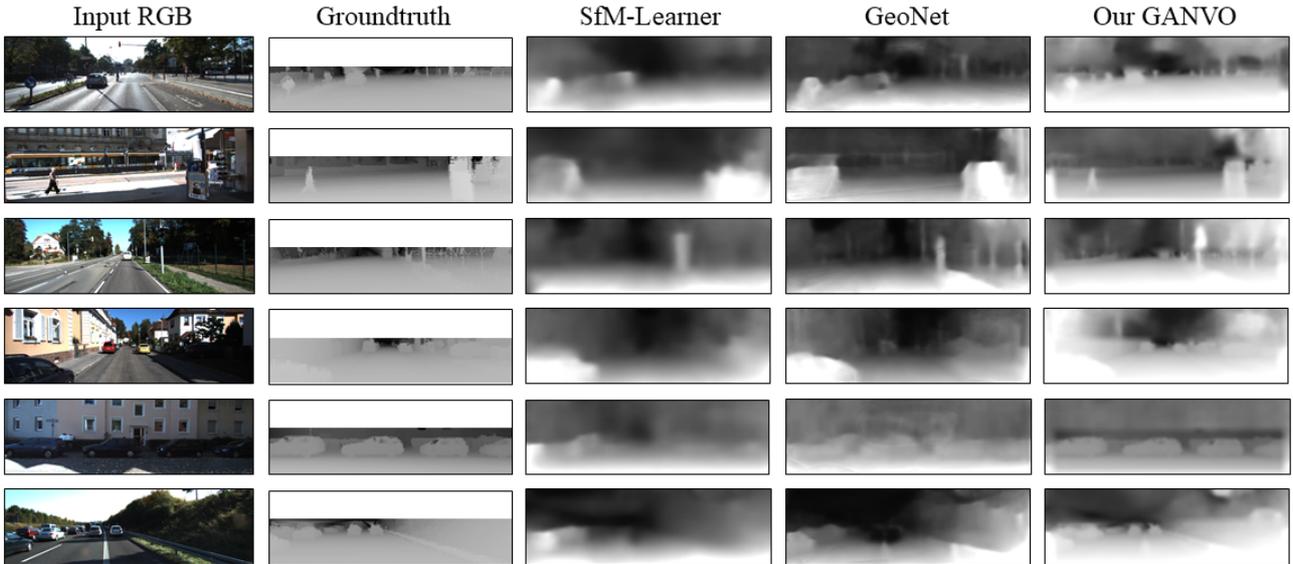


Fig. 4: Comparison of unsupervised monocular depth estimation between SfM-Learner, GeoNet and the proposed GanVO. The groundtruth captured by Light Detection and Ranging (LiDAR) is interpolated for visualization purpose. GanVO captures details in challenging scenes containing low textured areas, shaded regions, and uneven road lines, preserving sharp, accurate and detailed depth map predictions both in close and distant regions.

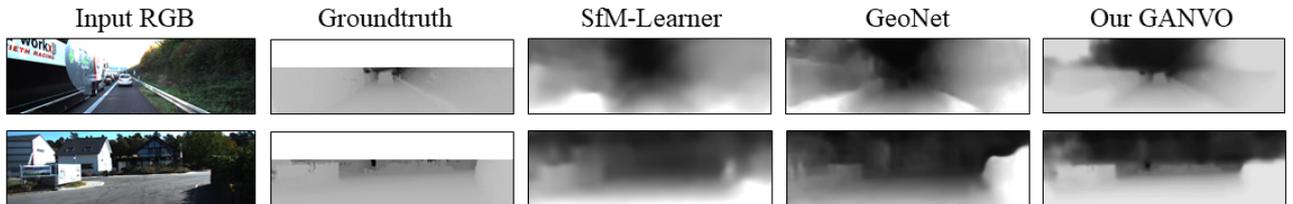


Fig. 5: Typical failure cases of our model. Sometimes, all of the compared methods struggle in vast open rural scenes and huge objects occluding the camera view.

Method	Seq.09	Seq.10
ORB-SLAM	0.014 ± 0.008	0.012 ± 0.011
SfM-Learner [17]	0.016 ± 0.009	0.013 ± 0.009
GeoNet [39]	0.012 ± 0.007	0.012 ± 0.009
Our GANVO	0.009 ± 0.005	0.010 ± 0.013

TABLE I: Absolute Trajectory Error (ATE) on KITTI odometry dataset. We also report the results of the other methods for comparison that are taken from [17], [39]. Our method outperforms all of the other methods.

the proposed method, SfM-Learner [17] and GeoNet [39]. It is clearly seen that GANVO outputs sharper and more accurate depth maps compared to the other methods that fundamentally use an encoder-decoder network with various implementations. An explanation for this result is that adversarial training using the convolutional domain-related feature set of the discriminator allows to obtain less blurry results [32]. Furthermore, it is also seen in Fig. 4 that the depth maps predicted by the proposed GANVO captures the small objects in the scene whereas the other methods tend to ignore

them. A loss function in image space leads to averaging all likely locations of details, whereas an adversarial loss function in feature space with a natural image prior makes the proposed GANVO more sensitive to the likely positions of the details in the scene [32]. The proposed GANVO also accurately predicts the depth values of the objects in low-textured areas caused by the shading inconsistencies in a scene. In Fig. 5, we demonstrate typical failure cases in the compared unsupervised methods, which are caused by challenges such as poor road signs in rural areas and huge objects covering the most of the visual input. Even in these cases, GANVO performs slightly better than the existing methods.

As shown in Table II quantitatively, our unsupervised approach significantly outperforms both the existing unsupervised methods [17], [39], [42] and even supervised methods [40], [41]. Furthermore, in the benchmark where the approaches are compared in terms of their adaptability to different environments, the approaches are trained on the Cityscapes dataset and finetuned on the KITTI dataset. In this benchmark, the proposed method results in clearly better

Method	Supervision	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	$\bar{\delta} < 1.25$	$\bar{\delta} < 1.25^2$	$\bar{\delta} < 1.25^3$
Eigen [40] Coarse	Depth	K	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen [40] Fine	Depth	K	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu [41]	Depth	K	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Monodepth [31]	Pose	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
SfM-Learner [17]	No	K	0.183	1.595	6.709	0.270	0.734	0.902	0.959
GeoNet [39]	No	K	0.155	1.296	5.857	0.233	0.793	0.931	0.973
GANVO	No	K	0.150	1.141	5.448	0.216	0.808	0.939	0.975
Garg et al. [42] cap 50m	Pose	K	0.169	1.080	5.104	0.273	0.740	0.904	0.962
SfM-Learner [17] cap 50m	No	K	0.201	1.391	5.181	0.264	0.696	0.900	0.966
GeoNet [39] cap 50m	No	K	0.147	0.936	4.348	0.218	0.810	0.941	0.977
GANVO cap 50m	No	K	0.137	0.892	3.671	0.201	0.867	0.970	0.983
Monodepth [31]	Pose	CS + K	0.124	1.076	5.311	0.219	0.847	0.942	0.973
SfM-Learner [17]	No	CS + K	0.198	1.836	6.565	0.275	0.718	0.901	0.960
GeoNet [39]	No	CS + K	0.153	1.328	5.737	0.232	0.802	0.934	0.972
GANVO	No	CS + K	0.138	1.155	4.412	0.232	0.820	0.939	0.976

TABLE II: Monocular depth estimation results on the KITTI dataset [1] by a benchmark split [40]. The use of the KITTI dataset in the training is shown with the letter K, and the Cityscapes dataset [2] with CS. We report the error and accuracy values for the other methods for comparison taken from [17], [31], [39]. The best results are shown in bold. Garg et al. [42] report 50m cap and we list them in a separate row for comparison.

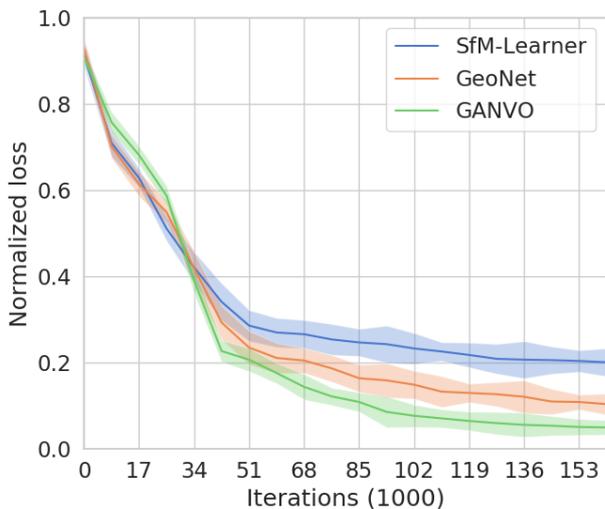


Fig. 6: Normalized loss values of GANVO compared to the unsupervised methods SfM-Learner [17] and GeoNet [39] for multiple training experiments with various set of hyperparameters. GANVO is less sensitive to set of hyperparameters with lower mean and variance of the normalized loss values.

error and accuracy compared to the existing unsupervised methods. Moreover, GANVO obtains much closer results to Monodepth [31] which is supervised by left-right image consistency, i.e. pose.

Figure 6 displays a robustness analysis of the proposed GANVO and the existing unsupervised approaches in terms of a parameter sensitivity analysis. We demonstrate that the proposed architecture is robust to hyper-parameter tunings such as different initialization of the network weights, different split of the dataset, and change of optimization parameters.

VI. CONCLUSIONS

In this study, we proposed an unsupervised generative deep learning method for pose and depth map estimation for a monocular video sequences, demonstrating the effectiveness of adversarial learning in these tasks. The proposed method outperforms all the competing unsupervised and traditional baselines in terms of pose estimation, and captures more detailed, sharper and more accurate depth maps of the scenes. In a future work, we would like to explicitly address scene dynamics and investigate a learning representation for a full 3D volumetric modelling with semantic segmentation.

Acknowledgments: We would like to thank the authors referred in this study for sharing their code. This work is funded by the NIST grant 70NANB17H185. YA would like to thank the Ministry of National Education in Turkey for their funding and support.

REFERENCES

- [1] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3061–3070.
- [2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
- [3] J. Engel, T. Schöps, and D. Cremers, “Lsd-slam: Large-scale direct monocular slam,” in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.
- [4] R. Mur-Artal and J. D. Tardós, “Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras,” *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [5] S. Wang, R. Clark, H. Wen, and N. Trigoni, “Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks,” in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2043–2050.
- [6] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, “Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem,” in *AAAI*, 2017, pp. 3995–4001.
- [7] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, “Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots,” *Neurocomputing*, vol. 275, pp. 1861–1870, 2018.

- [8] M. Turan, Y. Almalioglu, E. Konukoglu, and M. Sitti, "A deep learning based 6 degree-of-freedom localization method for endoscopic capsule robots," *arXiv preprint arXiv:1705.05435*, 2017.
- [9] P. Muller and A. Savakis, "Flowdometry: An optical flow and deep learning based approach to visual odometry," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 624–631.
- [10] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," *arXiv preprint arXiv:1710.11041*, 2017.
- [11] M. Turan, E. P. Ornek, N. Ibrahimli, C. Giracoglu, Y. Almalioglu, M. F. Yanik, and M. Sitti, "Unsupervised odometry and depth learning for endoscopic capsule robots," *arXiv preprint arXiv:1803.01047*, 2018.
- [12] T. I. Brown, V. A. Carr, K. F. LaRocque, S. E. Favila, A. M. Gordon, B. Bowles, J. N. Bailenson, and A. D. Wagner, "Prospective representation of navigational goals in the human hippocampus," *Science*, vol. 352, no. 6291, pp. 1323–1326, 2016.
- [13] E. J. Wamsley, M. A. Tucker, J. D. Payne, and R. Stickgold, "A brief nap is beneficial for human route-learning: The role of navigation experience and eeg spectral power," *Learning & Memory*, vol. 17, no. 7, pp. 332–336, 2010.
- [14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [15] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 781–788.
- [16] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," *arXiv preprint arXiv:1709.06841*, 2017.
- [17] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, vol. 2, no. 6, 2017, p. 7.
- [18] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5515–5524.
- [19] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtm: Dense tracking and mapping in real-time," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2320–2327.
- [20] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 2007, pp. 225–234.
- [21] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1434–1441.
- [22] M. Turan, Y. Y. Pilavci, I. Ganiyusufoglu, H. Araujo, E. Konukoglu, and M. Sitti, "Sparse-then-dense alignment-based 3d map reconstruction method for endoscopic capsule robots," *Machine Vision and Applications*, vol. 29, no. 2, pp. 345–359, 2018.
- [23] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *European conference on computer vision*. Springer, 1992, pp. 237–252.
- [24] M. Turan, Y. Almalioglu, H. Araujo, T. Cemgil, and M. Sitti, "Endosensorfusion: Particle filtering-based multi-sensory data fusion with switching state-space model for endoscopic capsule robots using recurrent neural network kinematics," *arXiv preprint arXiv:1709.03401*, 2017.
- [25] C. Zhao, L. Sun, P. Purkait, T. Duckett, and R. Stolkin, "Learning monocular visual odometry with dense 3d mapping from dense 3d flow," *arXiv preprint arXiv:1803.02286*, 2018.
- [26] M. R. U. Saputra, A. Markham, and N. Trigoni, "Visual slam and structure from motion in dynamic environments: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, p. 37, 2018.
- [27] M. Turan, Y. Almalioglu, H. Gilbert, A. E. Sari, U. Soylu, and M. Sitti, "Endo-vmfusenet: deep visual-magnetic sensor fusion approach for uncalibrated, unsynchronized and asymmetric endoscopic capsule robot localization data," *arXiv preprint arXiv:1709.06041*, 2017.
- [28] M. Turan, J. Shabbir, H. Araujo, E. Konukoglu, and M. Sitti, "A deep learning based fusion of rgb camera information and magnetic localization information for endoscopic capsule robots," *International journal of intelligent robotics and applications*, vol. 1, no. 4, pp. 442–450, 2017.
- [29] M. Turan, Y. Almalioglu, E. P. Ornek, H. Araujo, M. F. Yanik, and M. Sitti, "Magnetic-visual sensor fusion-based dense 3d reconstruction and localization for endoscopic capsule robots," *arXiv preprint arXiv:1803.01048*, 2018.
- [30] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for rgb-d based object recognition," in *Experimental Robotics*. Springer, 2013, pp. 387–402.
- [31] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, vol. 2, no. 6, 2017, p. 7.
- [32] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666.
- [33] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," in *Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291. International Society for Optics and Photonics, 2004, pp. 93–105.
- [34] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *European conference on computer vision*. Springer, 2016, pp. 286–301.
- [35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning," in *OSDI*, vol. 16, 2016, pp. 265–283.
- [37] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.
- [38] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [39] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2018.
- [40] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.
- [41] F. Liu, C. Shen, G. Lin, and I. D. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, 2016.
- [42] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.