ISSN 2007-9737

Depth, Spatial, and Temporal Features for Visual Odometry in Unstructured Agricultural Environments

Víctor Romero-Bautista, Leopoldo Altamirano-Robles*, Raquél Díaz-Hernández

Instituto Nacional De Astrofísica, Óptica y Electrónica,
Mexico

{victor.romero, robles, raqueld}@inaoep.mx

Abstract. Unstructured agricultural environments comprises several elements that present significant challenges to visual odometry (VO) methods. Deep learning-based VO methods have been proposed in state-of-the-art which have been demonstrated outstanding performance in structured environments even superior to conventional methods, nevertheless, these methods fail when face to unstructured environments such as the agricultural. In this work, we propose deep learning based VO model that exploits depth, spatial, and temporal features. To do this, the proposed model comprises two image processing pathways: the scene and depth pathway. The features extracted in these pathways are then fused to computes the relative pose given as R and t components. We conduct experiments evaluating the proposed model employing the agricultural Rosario dataset. Results show that the use of depth features improves significantly the performance of proposed model, obtaining consistent and coherent estimated trajectories in training and testing sequences.

Keywords. Unstructured environment, visual odometry, agriculture, deep learning.

1 Introduction

Monocular visual odometry problem consist into estimate the camera pose (R, t) in a particular space given a pair of monocular images through time, this is also know as ego-motion. In the state-of-the-art there are several works that tackled this problem which are usually developed and tested under structured environments and they are showed an outstanding performance, nevertheless, there are scarce research about monocular visual odometry in unstructured agricultural environments.

Unstructured agricultural environments generally comprises features, such as mountains, vegetation, uneven terrain, absence of man-made structures, and non-solid elements that can produce movement in the presence of winds such as plant foliage. These attributes present significant challenges to visual odometry (VO) methods.

This work addresses the problem of monocular VO in unstructured agricultural environments. Conventionally, the drawbacks presented in these type of environments have been addressed by incorporating sensor fusion such as in [6, 22], however, this involves several challenges such as the joint calibration of several sensors, as well as the handling of accumulated sensor errors, which requires a more complex system [7].

Deep learning strategies have been employed to solve problems where classical VO methods present limitations. These strategies have demonstrated certain robustness under adverse conditions such as illumination drastic changes and dynamic environments; even in some aspects these methods have surpassed classical ones [16]. However, because these methods usually are trained and tested under structured environments, when they are assessed under the unstructured environment conditions, such as in agricultural environments, these systems fail in trajectory estimation [19].

To tackle this problem, in this work we exploit the depth, spatial, and temporal image features

by building a deep learning-based monocular VO method with two image processing pathways: the scene and depth pathways. For this, we employ the DeepVO [24] framework, and take the modified version of DeepVO [18] as baseline to construct the scene pathway and monodepth2 [10] encoder to construct the depth pathway. Generally, proposed method comprises the modules: feature extractor to capture depth and spatial image information, memory unit to correlate temporal features, pathway heads to prepare features to be fused, and the estimator that computes the relative pose given as translation t and rotation R components.

Experiment results show that the depth information added in a second processing pathway helps to improve significantly the performance of proposed model. The estimated trajectories in training and testing sequences are consistent and coherent, which helps to reduce the Absolute Trajectory Error since the first 50 epochs.

This paper is organized as follows. Section 2 presents related works. Section 3 details the proposed methodology. Section 4 presents and discusses the results derived. Finally, Section 5 provides the conclusions and future work.

2 Related Work

In the literature, diverse methods of visual odometry based on deep learning have been proposed. One of the first works is PoseNet [12], which consists of the feature extraction stage by a convolutional network (or encoder) and the estimator (or decoder) composed of fully connected layers, where the global pose is estimated given an input image, this approach is highly susceptible to appearance changes. Because of this, more sophisticated approaches have been proposed that aim to exploit temporal information implicit in sequential images. One of these early approaches is DeepVO [24], which receives two sequential images and, in addition to using a convolutional network as an encoder, adds a recurrent network to capture temporal dependencies between the image pairs.

Recent works following this approach [27, 28, 26] demonstrate that the geometric information extracted by the decoder, combined with the temporal dependencies extracted by the recurrent neural network, improve the performance of a visual odometry model when faced to challenges such as velocity changes or abrupt movements, and low-texture regions, achieving similar results to classical visual odometry methods.

Additional works attempt to improve the performance of these approaches by adding optical flow [25, 14], and depth [13, 30] information. In [25, 14], the flow field image generated by an optical flow estimation model is used as input to the visual odometry model to estimate pose, which is trained using a supervised strategy. On the other hand, [13, 30] employ depth images as input for the pose estimation model and are trained in self-supervised approach.

These works have demonstrated considerable progress in the field of visual odometry. However, these models have been trained and tested on structured outdoor and indoor datasets such as KITTI [9], EuRoC [3], and TUM [23]. Therefore, when tested under unstructured conditions, these models fail [19].

Similarly to the previous works [13, 30], in this paper, we consider that the use of image depth fused with scene information can be exploited to improve the performance of a VO model under the conditions of an unstructured agricultural environment. But differently to [13, 30], we do not use the image depth directly as input to the pose estimation model or VO model, instead, we use a trained encoder based on monodepth2 to extract feature maps that contains the depth information. Specifically, the proposed model employs two image processing pathways, the scene pathway which encodes image features, such as color and texture; and depth pathway that encodes image depth features. Results shows that the use of depth information helps to improve the performance of a baseline VO model, obtaining good estimation results since the first 50 epochs.

3 Methodology

Given a sequence of monocular rgb images X_t up to time t, the proposed model computes the conditional probability of relative poses Y_t. As

Fig. 1 shows, proposed model consists of two image processing pathways: the scene pathway (above) where image appearance features are captured, and the depth pathway (bottom) that encodes image depth features; in both pathways spatio-temporal features are captured by the temporal correlation stage (memory unit). In general, each pathway is composed of the feature extractor, spatio-temporal feature correlation (memory), pathway heads, and estimator. The two pathways are fused and processed by the estimator that decodes the features into relative pose. In the following subsections, the model design and dataset details are given.

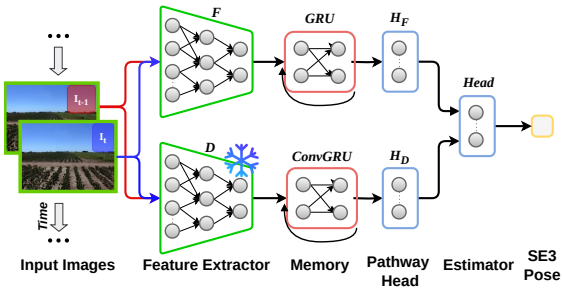


Fig. 1. Proposed model framework

3.1 Model design

The proposed model follows the DeepVO [24] framework, which is a deep learning-based visual odometry and trained in an end-to-end fashion. Specifically, this work is based on the modified version of DeepVO (M-DeepVO) for unstructured agricultural environments presented in [18].

To capture depth, space, and temporal properties of the input images, proposed model is composed by the scene and depth image processing pathways, which are then fused to estimate the relative pose, a brief description of the main components of proposed model are given below.

Scene pathway: Blocks in green depicted in Fig. 2 show to scene pathway. First, the input rgb images $I_{t-1}^{3XHXW}, I_t^{3XHXW}$ are concatenated resulting with shape of $[6, H, W]$ and passed to feature extractor F based on ResNet-18 [11] to

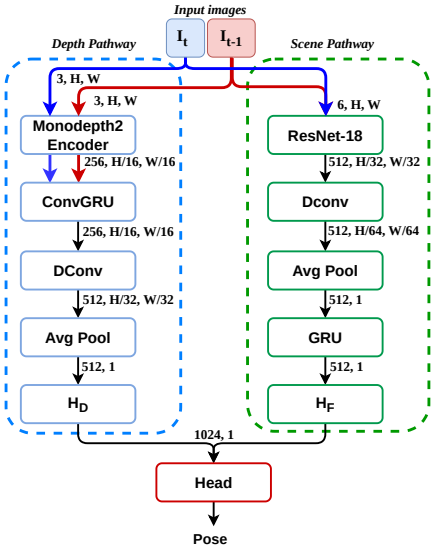


Fig. 2. Proposed model blocks. Blue blocks own to depth pathway, and green blocks correspond to spatial pathway

extract image scene features, here a $DConv$ layer was added to the output of ResNet, which employs Deformable Convolution [8] version 2 [31], where following the insights of [2] this block was set at the end of spatial image processing to act as best features selector before the adaptative average pool layer ($AvgPool$). $DConv$ layer is presented in Fig. 3, which consist of two standard convolutional operators that adds linear transformations, the main operator based on deformable convolution (DCN), and the residual connection. Then to capture temporal information between $F(I_{t-1}), F(I_t)$, Gated Recurrent Unit (GRU) [5] is employed, which was set to 512 internal units. At the end of scene pathway, the head H_F was set, which is a fully connected layer to add linear transformation, and prepare the features pathway to be fused.

Depth pathway: Blocks in blue presented in Fig. 2 comprise the depth pathway. First, encoder D based on pretrained monodepth2 receives two rgb images $I_{t-1}^{3XHXW}, I_t^{3XHXW}$ separately, the depth image features encoded are passed to the $ConvGRU$ [1] to capture temporal depth dependencies between $D(I_{t-1}), D(I_t)$. Then the feature map resulting is processed by $DConv$ layer

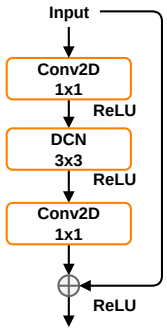


Fig. 3. DConv block. The main operator is deformable convolution (DCN) with kernel size of 3x3

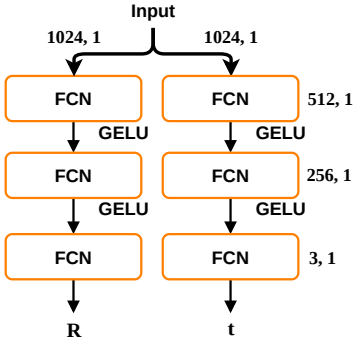


Fig. 4. Estimator stage, which is composed of two MLPs conformed by fully connected layers FCN, they contain two hidden layers, the first with 512 neurons, and second with 256 neurons. The input are concatenated features of scene and depth pathways, and the outputs correspond to rotation R and translations t

to select best features before adaptative average pool layer ($AvgPool$). Finally, the pathway head H_D prepares the depth features to be fused with the scene fatures. Here, the decoder D based on pretrained monodepth2, is freezed, so it's not trained.

Estimator: It is represent in red block ($Head$) in Fig. 2. Here, the output of depth and scene head pathways are fused, and then send to a pose net module to compute the relative pose. The pose net module is depicted in Fig. 4, which is composed of two MLPs with two hidden layers, one to estimate the rotations R and the other to estimate the translations t.

To learn the hyper-parameters θ, the Euclidean distance between the groundtruth ($\mathbf{p}_\tau, \varphi_\tau$) at time τ and its estimated one ($\hat{\mathbf{p}}_\tau, \hat{\varphi}_\tau$) are minimized. The loss function is composed of Mean Squared Error (MSE) of the translations $\mathbf{p}$ and rotations φ, which is defined as:

$$\theta^* = \underset{\theta}{argmin} \sum_{i=1}^{N} \sum_{\tau=1}^{t} \|\hat{\mathbf{p}}_\tau - \mathbf{p}_\tau\|_2^2 + \kappa \|\hat{\varphi}_\tau - \varphi_\tau\|_2^2 . \tag{1}$$

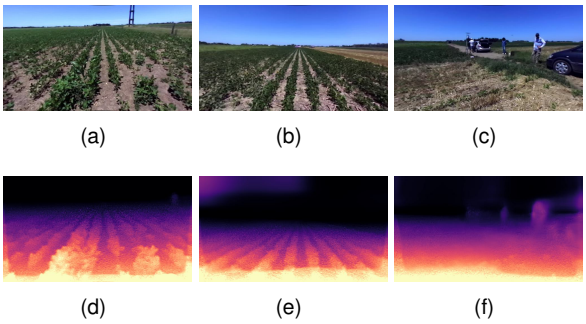
where θ^* represents the optimal parameters; $\|\cdot\|$ is 2-norm; κ is the scale factor to balance the weights of positions and orientations; and N is the number of samples.

3.2 Revisiting Monodepth2

Monodepth2 [10] is a network that estimates the depth of a monocular image. It consists of two networks, one composed of U-Net architecture to estimate depth (Depth Network) and the other corresponding to a resnet-18 architecture to estimate camera pose (Pose Network). U-Net employs a resnet-18 architecture for the encoder, and for the decoder they use an architecture that they generated themselves.

The training is self-supervised, meaning it generates its own supervision signal to estimate loss during training. It consists of predicting (estimating) the appearance of a query image based on the viewpoint of another image (view synthesis). In this way, the model is trained by minimizing the error in reconstructing an image.

In this work, only the Depth Network encoder was used, which was obtained by training monodepth2 from scratch using the Rosario dataset. Training was conducted employing the stereo and monocular approach for 50 epochs. Fig. 5 shows the estimated depth image (bottom) given a monocular input RGB image (top) obtained using the Monodepth2 model trained with Rosario dataset. Since the Rosario dataset does not include depth groundtruth data, we employed the default scale of 0 to 100 meters used for outdoor environments, thus the most intense colors are close to 0 meters, and the dark colors are close to 100 meters.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Image depth estimated by monodepth2 trained with Rosario dataset. Top row (a-c) shows the input RGB monocular image, bottom row (d-f) shows the respective image depth estimation

3.3 Rosario Dataset

Proposed in [17], the Rosario dataset is composed of six sequences taken in an agricultural environment. Each sequence includes: image data captured with RGB stereo camera with image size of 672x376 and rate of 15 Hz, IMU data with rate of 140 Hz, and the groundtruth obtained with GNSS employing frequency of 5 Hz to record data which is given in TUM format. The scenes in this dataset are composed of crops, where the camera moves in a linear fashion through the fields until it reaches a return point to rotate and continue with the linear motion, simulating the movement of a field robot. Fig. 6 shows frames of the Rosario dataset scenario.

(a) (b) (c)

Fig. 6. Frames captured in the Rosario dataset

Because of the supervised approach employed to train our model, the image data was matched with the groundtruth data, thus reducing the total number of images originally captured with rate of 15 Hz to 5 Hz based on the amount of data captured with GNSS. We also change the groundtruth given in TUM format to KITTI format. Table 1 shows the sequence information, including the number of frames resulting from the GNSS matching, as well as the length and duration of the sequence.

Table 1. Sequences description of Rosario dataset. The original sequences name given in [17] starts from 1 to 6, in this work, we change the sequences name, starting from 00 to 05

Seq. name	Num. frames	Length (m)	Duration (min.)
00	2208	611.55	9.3
01	1338	321.87	4.4
02	998	177.02	3.3
03	781	144.84	2.7
04	1385	321.87	5.2
05	2219	708.11	9.8

3.4 Metric

The Absolute Trajectory Error (ATE_{trans}) was employed to measure the performance of the proposed model.

$$ATE_{trans} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\| T_{gt,i}^{trans} - T_{est,i}^{trans} \right\|^2}. \quad (2)$$

where $T_{est,i}$ is the estimated trajectory; $T_{gt,i}$ is the groundtruth; N means the total number of poses in a trajectory; and $trans$ represents the translation of the variables.

4 Experimental results

To evaluate the proposed method of using image depth information in the modified DeepVO model, two sets of experiments were conducted. The first consists of analyzing the performance of the proposed model every 10 epochs until reaching 100 epochs, to identify the adaptability of the image depth information obtained from the pre-trained monodepth2 model with the information from the scene being trained each epoch; and then increasing the training sequence up to 500 epochs in order to identify any improvement when the model is trained for a long period, this experiment is called as depth information impact. And the

second experiment consist of analyze the general performance of the proposed model with different training and testing sequences, this is called as general performance. In both experiments, a performance comparison between modified DeepVO [18] (M-DeepVO) and the proposed model was made.

For the experiments, image frames were resized to 320x160, κ was set to 50, and learning rate was set to $5x10^{-5}$ using the Adaptative Gradient (AdaGrad) as optimizer.

The proposed model is implemented in PyTorch [15], the training sequences were deployed using a single NVIDIA RTX 2070 GPU with batch size of 32.

4.1 Depth information impact

To identify the impact of depth information obtained by incorporating the monodepth2 encoder through the depth pathway, the performance of the modified DeepVO base model [18] (M-DeepVO), which only incorporates the scene pathway, was compared with the proposed model. For this, we trained both models setting sequences $\{00, 01, 03, 04, 05\}$ (7931 frames in total) for training and sequence $\{02\}$ (998 frames in total) for testing. Training phase was made along 500 epochs, where the first 100 epochs were checked each 10 epochs, and the rest each 100 epochs.

Graph in Fig. 7 shown the ATE_{trans} obtained during the first 100 epochs in steps of 10, and the rest in steps of 100, by modified DeepVO (M-DeepVO) in red line color and the proposed model in blue line color, with the training sequence $\{00\}$ and the test sequence $\{02\}$.

Fig. 8 shown the qualitative results of baseline modified DeepVO [18] (M-DeepVO) in the training $\{00\}$ and testing $\{02\}$ sequences along the first 100 epochs with step of 10 (see (a), (b), for training sequences, and (d), (e) for testing sequences); and then for 500 epochs in steps of 100 (see (c) for training sequences, and (f) for testing sequences), where the dashed lines represent the groundtruth and the colored lines correspond to the respective model in training epoch.

In Fig. 9 the qualitative results of training the proposed model are shown, where the dashed

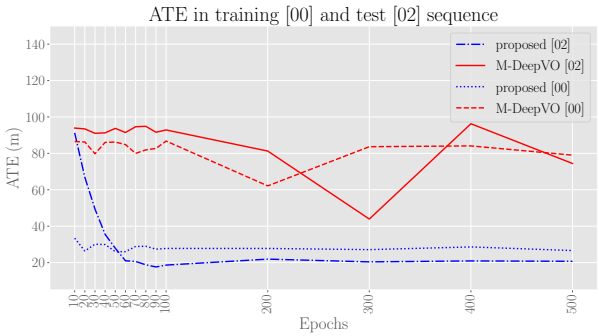


Fig. 7. Comparison of ATE_{trans} obtained by modified DeepVO [18] and the proposed along training in sequence $\{00\}$ and testing with sequence $\{02\}$

lines represent the groundtruth and the colored lines represent the estimated trajectories by the model in a respective epoch. (a) and (b) show the results of the first 100 epochs in steps of 10 with the training sequence $\{00\}$; (d) and (e) show the results of the first 100 epochs in steps of 10 with testing sequence $\{02\}$. (c) and (f) show the results for 500 epochs in steps of 100 for training $\{00\}$ and testing $\{02\}$ sequences respectively.

4.2 General performance

To identify the general performance of the proposed model, we trained the baseline M-DeepVO and the proposed model with different sets of training and testing sequences. Table 2 presents the qualitative results of this experiment, where the set of training and testing sequences can be identified. In addition, runtime measurements in sequences are reported, given in time spend in minutes along a full sequence (all frames in a sequence) and per step (using pair image as input) measured in Hertz.

Qualitative comparison results between M-DeepVO and the proposed model are depicted in Fig. 10, the dashed lines represent the groundtruth and colored lines correspond to the estimated trajectories by a respective model. (a) and (f) show the results in training sequences $\{00, 05\}$ respectively. (b) to (e) show the results in testing sequences $\{01, 02, 03, 04\}$, which are reported in Table 2.

ISSN 2007-9737

Table 2. ATE_{trans} (in meters) and runtime (in minutes and Hertz) obtained by proposed model, remarked column means the test sequence ATE_{trans}

Methods	**Sequence Name**						**Runtime in sequence**	
	Train					**Test**	**Full seq. (min)**	**Step (Hz)**
	00	02	03	04	05	01	01	
M-DeepVO [18]	94.85	**12.12**	34.51	31.51	146.2	92.76	0.533	41.812
Proposed	**27.55**	26.70	**29.52**	**25.38**	**33.13**	**77.78**	0.983	22.677
	00	01	03	04	05	02	02	
M-DeepVO [18]	79.02	42.46	**20.39**	48.9	236.6	43.9	0.4	41.583
Proposed	**26.59**	**26.95**	26.56	**28.96**	**33.20**	**17.58**	0.733	22.681
	00	01	02	04	05	03	03	
M-DeepVO [18]	89.26	**16.23**	**9.07**	**14.05**	**30.63**	35.04	0.316	41.105
Proposed	**27.42**	23.49	25.91	24.84	32.97	**24.11**	0.583	22.314
	00	01	02	03	05	04	04	
M-DeepVO [18]	85.18	74.04	81.05	95.14	104.5	79.12	0.55	41.969
Proposed	**27.09**	**24.77**	**26.89**	**31.03**	**34.34**	**41.79**	1.05	21.984

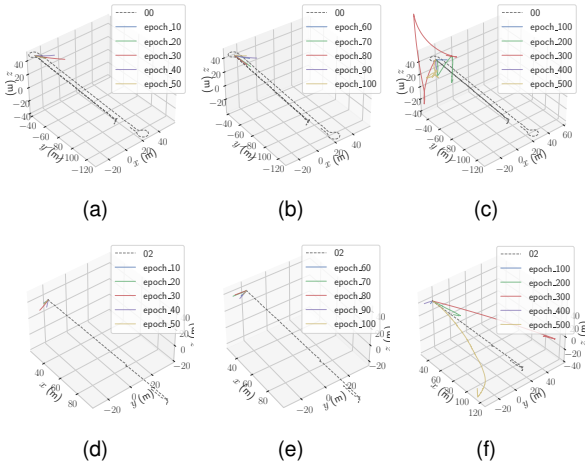


Fig. 8. Qualitative results of modified DeepVO [18] model during training. (a-c) correspond to trajectory estimation in training sequence $\{00\}$ (top graphs); (d-f) correspond to trajectory estimation in test sequence $\{02\}$ (bottom graphs)

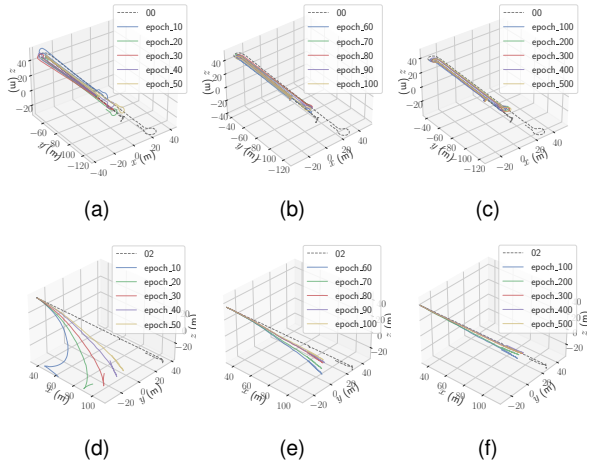


Fig. 9. Qualitative results of proposed model during training. (a-c) correspond to trajectory estimation in training sequence $\{00\}$ (top graphs); (d-f) correspond to trajectory estimation in test sequence $\{02\}$ (bottom graphs)

4.3 Comparison With State-of-the-art Methods

Table 3 presents a comparison of ATE_{trans} results obtained by the proposed model and state-of-the-art methods, which are: the baseline modified DeepVO (M-DeepVO) [18] method, which is an end-to-end model fully based on deep learning, trained with the Rosario dataset; the hybrid method proposed by Shu et al. in [20], it's based on a classical indirect monocular VSLAM system which employs a deep learning-based method to recover depth information from a monocular image; this method was developed using the Rosario dataset; MonoViT [29], which was developed using the KITTI dataset (a structured environment), to obtain the results in

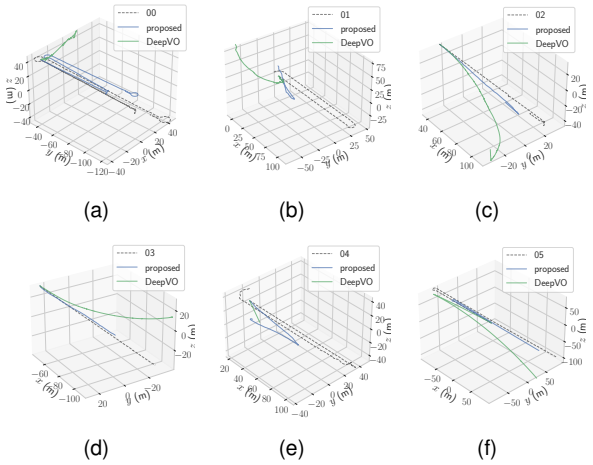


Fig. 10. Qualitative comparison between M-DeepVO and the proposed model. (a) and (f) correspond to estimated trajectories of sequences $\{00, 05\}$ employed for training; (b-e) are estimated trajectories of sequences $\{01, 02, 03, 04\}$ employed for testing, quantitative results of these test sequences are presented in Table 2

Table 3, this method was trained with Rosario dataset; and ORB-SLAM3 stereo [4], which is a classic indirect method based on feature extraction. Results of MonoViT and ORB-SLAM3 were taken from [19].

Table 3. Comparison of ATE_{trans} (m) results with state-of-the-art methods

Methods	**Sequence Name**			
	01	02	03	04
M-DeepVO [18]	99.56	7.923	**5.74**	53.39
Shu [20]	12.46	16.98	14.52	13.58
MonoViT [29]	72.17	**7.3**	8.08	24.70
ORB-SLAM3 [4]	**6.41**	10.53	7.32	**7.06**
proposed	77.78	17.58	24.11	41.79

According with Table 3, proposed model exhibits better performance than M-DeepVO in complex sequences $\{01, 04\}$ (due to approximately 180-degree turns) even when M-DeepVO is trained with 2K epochs and the proposed model with 500 epochs, which is able to recognize movement direction changes due to such turns (see (b) and (e) in Fig. 10). While in sequences $\{02, 03\}$, M-DeepVO has a lower ATE_{trans}, qualitatively, the proposed model presents a more consistent trajectory estimation (see (c) and (d) in Fig. 10).

On the other hand, in sequences $\{02, 03\}$, proposed model presents competitive performance compared to the hybrid SLAM method [20]. MonoViT presents a lower ATE_{trans} in all sequences than the proposed model; nevertheless, in sequence $\{01\}$, it presents a similar result (with a difference of 5.61 meters). Due to the stereo configuration, which allows it to accurately estimate depth and consequently the scene scale, ORB-SLAM3 stereo performs better in complex sequences $\{01, 04\}$. In sequence $\{02\}$, the proposed model presents a competitive result compared to ORB-SLAM3 stereo (with difference of approximately 7 meters).

4.4 Discussion

The incorporation of depth information add considerable improvements to the baseline modified DeepVO [18] (M-DeepVO) which is evidenced by the proposed model results.

In the depth information impact experiment, the proposed model presents the best performance compared with the baseline modified DeepVO (M-DeepVO) model, even in the first 50 epochs as (a) and (b) in Fig 9 shows. On the other hand, baseline modified DeepVO (M-DeepVO) presents a low performance in the first 100 epochs as (a), (b), (d) and (e) show in Fig. 8; and when it is trained up to 500 epochs, the performance improves but the estimated trajectories are not consistent as (c) and (f) in Fig. 8 show. This experiment reveals that depth information captured by the monodepth2 decoder can be used to boost DeepVO model, its impact can be observed since 10 epochs, where the training and testing sequence maintain consistent and coherent with the groundtruth as Fig. 9 shows. This impact is not incremental along a large number of epochs, i.e. the improvement reach a state where the estimated trajectory is not relevant improved, e.g. when the model reach 300 epochs in the results presented in (f) Fig. 9, the estimated trajectories do not present a relevant improvement, it seems that the following epochs (400 and 500) do not present any effect in comparison with the first 100 epochs, it can

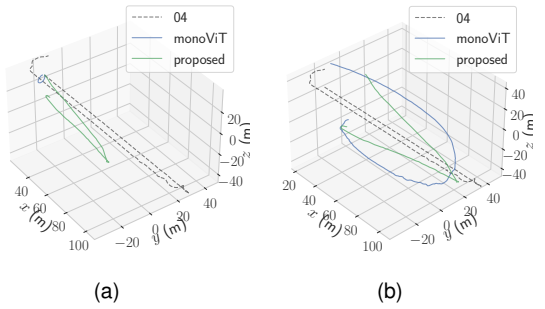
also be seen in Fig. 7, where the ATE decreased significantly in the first 100 epochs, then the ATE maintains a similar state or value. On the other hand, M-DeepVO can improve its performance with greater number of epochs as was demonstrated in [18], where it was trained along 2K epochs to obtain good results.

The general performance experiment assessed the proposed model performance by switching the training sequences to use the sequences $\{01, 02, 03, 04\}$ for testing. This is for evaluate the capacity to maintain a similar performance with different training and testing scenarios. This results are presented in Table 2, where the performance of the proposed model is better than the baseline modified M-DeepVO, because in major of training and test sequences the proposed model present lower ATE_{trans}. A peculiar case is given when sequence $\{03\}$ is used for testing, here M-DeepVO present a lower ATE_{trans} than proposed model in all of training sequences present, but in tested sequence, the proposed model performs lower ATE_{trans}, it means that this model can generalize better. Considering runtime reported in Table 2, proposed model had a longer execution time than the M-DeepVO, approximately twice as long. This performance in runtime is expected due to the incorporation of a second processing pathway in the proposed model, where the layer that requires the most execution time is $ConvGRU$, which makes the proposed model require more execution time.

Another important aspect that can be observed in Fig. 10, is the consistency of the trajectories generated by the proposed model compared with M-DeepVO; the trajectories estimated by proposed model present more consistency and coherence with the groundtruth, but when the test trajectory present curves, the performance decrease, this can be caused by the scarcity of trajectory curves samples for training.

Compared with state-of-the-art methods presented in table 3, the proposed model performs competitive with the hybrid method [20] in sequences $\{02, 03\}$, a similar performance was presented with the classic ORB-SLAM3 stereo [4] method. Considering that [20] and ORB-SLAM3 methods are based on a classical approach, they require manual calibration for correct operation, while the proposed model, due that it is based fully on a deep learning approach, does not require any calibration, making it simpler and faster to implement.

Although MonoViT's quantitative results are better to those obtained by the proposed model, these results shown in Table 3 are scaled and aligned with the groundtruth, which improves its performance as shown in (b) Fig. 11, while results presented by the proposed model are not scaled nor aligned. Qualitatively, the proposed model exhibits a better trajectory estimation consistency with respect to the groundtruth, as shown in (a) Fig. 11, where estimated trajectories are not scaled. Considering sequence $\{04\}$, ATE_{trans} of the MonoVit trajectory estimation without scaling is 91.59 meters, whereas ATE_{trans} of proposed model trajectory estimation without scaling is 41.79 meters, which means that proposed model performs better.



(a) (b)

Fig. 11. Qualitative comparison between MonoViT and proposed model. (a) Estimated trajectories without scaling; (b) Estimated trajectories with scaling

5 Conclusion

Depth information added in a second processing pathway helps to improve the performance of the baseline modified DeepVO as the proposed model shows; the geometric patterns learned by monodepth2 encoder helps to produce better training and testing trajectories estimations in less number of epochs.

The proposed model present good performance compared with baseline modified DeepVO even when this last is trained with 2K epochs. Proposed model shows susceptibility to curves presented in trajectories $\{01, 04\}$, in this sequences the performance decreased.

As future work, we consider to increase the training and testing sequences by adding the Rosario dataset version 2 [21] sequences. We expect this help to improve the performance of the proposed model in curves movements.

Acknowledgments

This work is supported by the SECIHTI doctoral scholarship granted to Víctor Romero-Bautista, with CVU number: 877984.

References

1. **Ballas, N., Yao, L., Pal, C. J., Courville, A. C. (2016).** Delving deeper into convolutional networks for learning video representations. 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, pp. 0–10.
2. **Burgos-Madrigal, A., Romero-Bautista, V., Díaz-Hernández, R., Altamirano-Robles, L. (2024).** Evaluation of deformable convolution: An investigation in image and video classification. Mathematics, Vol. 12, No. 16, pp. 0–24. DOI: 10.3390/math12162448.
3. **Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M., Siegwart, R. Y. (2016).** The euroc micro aerial vehicle datasets. The International Journal of Robotics Research, Vol. 35, pp. 1157 – 1163.
4. **Campos, C., Elvira, R., Rodr'iguez, J. J. G., Montiel, J. M. M., Tardós, J. D. (2020).** Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics, Vol. 37, pp. 1874–1890.
5. **Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y. (2014).** On the properties of neural machine translation: Encoder–decoder approaches. **Wu, D., Carpuat, M., Carreras, X., Vecchi, E. M.**, editors, Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar, pp. 103–111. DOI: 10.3115/v1/W14-4012.
6. **Cremona, J., Civera, J., Kofman, E., Pire, T. (2023).** Gnss-stereo-inertial slam for arable farming. Journal of Field Robotics, Vol. 41, pp. 2215 – 2225.
7. **Cremona, J., Comelli, R., Pire, T. (2022).** Experimental evaluation of visual-inertial odometry systems for arable farming. Journal of Field Robotics, Vol. 39, pp. 1123 – 1137.
8. **Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y. (2017).** Deformable convolutional networks. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 764–773.
9. **Geiger, A., Lenz, P., Stiller, C., Urtasun, R. (2013).** Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, Vol. 32, pp. 1231 – 1237.
10. **Godard, C., Aodha, O. M., Brostow, G. J. (2018).** Digging into self-supervised monocular depth estimation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3827–3837.
11. **He, K., Zhang, X., Ren, S., Sun, J. (2015).** Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.
12. **Kendall, A., Grimes, M. K., Cipolla, R. (2015).** Posenet: A convolutional network for real-time 6-dof camera relocalization. 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2938–2946.
13. **Li, S., Wang, X., Cao, Y., Xue, F., Yan, Z., Zha, H. (2020).** Self-supervised deep visual odometry with online adaptation.

2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6338–6347.

14. **Pandey, T., Peña, D., Byrne, J., Moloney, D. (2021).** Leveraging deep learning for visual odometry using optical flow. Sensors, Vol. 21, No. 4, pp. 0–12.

15. **Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S. (2019).** Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, Vol. 32, pp. 8024–8035.

16. **Phan, T.-D., Kim, G.-W. (2025).** Toward specialized learning-based approaches for visual odometry: A comprehensive survey. J. Intell. Robotic Syst., Vol. 111, pp. 44.

17. **Pire, T., Mujica, M., Civera, J., Kofman, E. (2018).** The rosario dataset: Multisensor data for localization and mapping in agricultural environments. The International Journal of Robotics Research, Vol. 38, pp. 633 – 641.

18. **Romero-Bautista, V., Altamirano-Robles, L., Díaz-Hernández, R. (2025).** Towards end-to-end visual odometry for unstructured agricultural environments. **López-Monroy, A. P., Rosales-Pérez, A., , Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., Olvera-López, J. A.**, editors, Mexican Conference on Pattern Recognition, Springer Nature Switzerland, Cham, Vol. 15715, pp. 245–256.

19. **Romero-Bautista, V., Altamirano-Robles, L., Díaz-Hernández, R., Zapotecas-Martínez, S., Sanchez-Medel, N. (2024).** Evaluation of visual slam algorithms in unstructured planetary-like and agricultural environments. Pattern Recognit. Lett., Vol. 186, pp. 106–112.

20. **Shu, F., Lesur, P., Xie, Y., Pagani, A., Stricker, D. (2020).** Slam in the field: An evaluation of monocular mapping and localization on challenging dynamic agricultural environment. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1760–1770.

21. **Soncini, N., Cremona, J., Vidal, E., Garc'ia, M., Castro, G., Pire, T. (0).** The rosario dataset v2: Multi-modal dataset for agricultural robotics. The International Journal of Robotics Research, Vol. 0, No. 0, pp. 02783649251368909. DOI: 10.1177/02783649251368909.

22. **Song, K., sheng Qiu, R., Yang, G., Li, J. (2022).** Monocular visual-inertial odometry for agricultural environments. IEEE Access, Vol. 10, pp. 103975–103986.

23. **Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D. (2012).** A benchmark for the evaluation of rgb-d slam systems. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 573–580.

24. **Wang, S., Clark, R., Wen, H., Trigoni, A. (2017).** Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 2043–2050.

25. **Wang, W., Hu, Y., Scherer, S. A. (2020).** Tartanvo: A generalizable learning-based vo. **Kober, J., Ramos, F., Tomlin, C.**, editors, Conference on Robot Learning, PMLR, Vol. 155, pp. 1761–1772.

26. **Xu, C., Zeng, T., Luo, Y., Song, F., Si, B. (2025).** Spatiotemporal dual-stream network for visual odometry. IEEE Robotics and Automation Letters, Vol. 10, pp. 3867–3874.

27. **Xue, F., Wang, X., Li, S., Wang, Q., Wang, J., Zha, H. (2019).** Beyond tracking: Selecting memory and refining poses for deep visual odometry. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8567–8575.

28. **Zhai, G., Liu, L., Zhang, L., Liu, Y. (2019).** Poseconvgru: A monocular approach for visual

ego-motion estimation by learning. Pattern Recognition, Vol. 102, pp. 107187. DOI: https://doi.org/10.1016/j.patcog.2019.107187.

29. **Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S. (2022).** Monovit: Self-supervised monocular depth estimation with a vision transformer. 2022 International Conference on 3D Vision (3DV), pp. 668–678.

30. **Zhou, W., Zhang, H., Yan, Z., Wang, W., Lin, L. (2023).** Decoupledposenet: Cascade decoupled pose learning for unsupervised camera ego-motion estimation. IEEE Transactions on Multimedia, Vol. 25, pp. 1636–1648.

31. **Zhu, X., Hu, H., Lin, S., Dai, J. (2018).** Deformable convnets v2: More deformable, better results. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9300–9308.

Article received on 18/09/2025; accepted on 26/11/2025.
**Corresponding author is Leopoldo Altamirano-Robles.*