

Depth-DensePose: an efficient densely connected deep learning model for camera-based localization

Amr Abozeid^{1,2}, Hesham Farouk³, Samia Mashali³

¹Department of Computer Science, College of Science and Arts, Jouf University, Sakaka, Saudi Arabia

²Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt

³Department of Computers and Systems, Electronics Research Institute, Cairo, Egypt

Article Info

Article history:

Received Mar 13, 2021

Revised Dec 31, 2021

Accepted Jan 19, 2022

Keywords:

Augmented reality
Camera-based localization
Computer vision
Deep learning
Depthwise separable
convolution

ABSTRACT

Camera/image-based localization is important for many emerging applications such as augmented reality (AR), mixed reality, robotics, and self-driving. Camera localization is the problem of estimating both camera position and orientation with respect to an object. Use cases for camera localization depend on two key factors: accuracy and speed (latency). Therefore, this paper proposes Depth-DensePose, an efficient deep learning model for 6-degrees-of-freedom (6-DoF) camera-based localization. The Depth-DensePose utilizes the advantages of both DenseNets and adapted depthwise separable convolution (DS-Conv) to build a deeper and more efficient network. The proposed model consists of iterative depth-dense blocks. Each depth dense block contains two adapted DS-Conv with two kernel sizes 3 and 5, which are useful to retain both low-level as well as high-level features. We evaluate the proposed Depth-DensePose on the Cambridge Landmarks dataset, which shows that the Depth-DensePose outperforms the performance of related deep learning models for camera based localization. Furthermore, extensive experiments were conducted which proven the adapted DS-Conv is more efficient than the standard convolution. Especially, in terms of memory and processing time which is important to real-time and mobile applications.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Amr Abozeid

Department of Computer Science, College of Science and Arts, Jouf University

Saudi Arabia

Department of Mathematics, Faculty of Science, Al-Azhar University

Cairo, Egypt

Email: aabozeid@azhar.edu.eg, aaabozeid@ju.edu.sa

1. INTRODUCTION

Nowadays, visual (camera) localization is an active research field due to the rapid development and technological advances in computer vision and mobile computing. Visual localization is to estimate the pose (position and orientation) of a camera. Camera localization is an important step in emerging applications such as augmented reality (AR), mixed reality, robotics, and self-driving [1].

Generally, there are two categories of traditional visual localization methods: image-based and structure-based methods [2], [3]. In Image-based methods, the 6-degrees-of-freedom (6-DoF) camera poses are estimated by retrieving the nearest image with known ground truth poses in a reference database. Matches between images are performed by searching through the global descriptor space [4]. This descriptor may be a hand-crafted feature (such as scale invariant feature transform (SIFT) ([5], dominate SIFT [6], speeded-up robust features (SURF) [7], or oriented fast and rotated BRIEF (ORB) [8]) or learned features (such as

SuperPoint [9], ASLFeat [10], or SeqNet [11]). Structure-based methods predict the 6-DoF camera poses by matches local descriptors of 2D points from the query image and a reconstructed 3D point-cloud of the whole scene. Local descriptors are similar to global descriptors which may be hand-crafted or learned. Although the hand-crafted feature methods are accurate and robust in many situations (especially in outdoor environments). The indoor feature-based localization is still an issue since indoor scenes often have less texture, repetitive structures, and insufficient local/global features for matching [12], [13].

Recently, the convolutional neural networks (CNN) gained salient success to solve many computer vision problems [14]–[17]. Accordingly, CNN has been used to guess the camera pose for an input image. The camera localization problem was defined as a regression problem to estimate the absolute pose by using CNN [18]. Kendall *et al.* [19] developed the PoseNet which is a deep learning model for regressing the 6-DoF camera pose directly from the input image. As a result of PoseNet success, many deep learning models were developed for camera pose estimation.

In this paper, we propose, the Depth-DensePoseNet, an efficient end-to-end deep learning model for camera/image-based localization. The architecture of Depth-DensePose has many advantages: i) adopt the DenseNets [20] architecture, which lead to developing a deeper model, fast training, and produce more accurate results; ii) develop the depthwise separable convolution (DS-Conv) with two kernel sizes 3 and 5 instead of the standard convolution. The kernel size 3 is useful to obtain the low-level features while the kernel with size 5 is useful to obtain the high-level features. As a result, the DS-Conv requires memory less than the standard convolution with similar performance [21]; and iii) reduce the average pooling layers by adopting a convolution with stride=2, which preserves the resolution and requires less graphics processing unit (GPU) memory.

We evaluate the proposed Depth-DensePose on Cambridge Landmarks dataset [19], [22]. The results of these experiments show that, the Depth-DensePose model achieves lower mean squared errors (MSE) of the camera position to 0.74 m and orientation to 2.14 degree on the entire Cambridge dataset. From the results, we can conclude that the Depth-DensePose outperforms the related models for camera-based localization tasks. Other extensive experiments were conducted which proven the adapted DS-Conv is more efficient than the standard convolution. The Depth-DensePose using DS-Conv requires a 40% lower memory size and 20% less training time than standard convolution. These results are important to real-time and mobile applications. It is worth mention that, the total number of layers of the Depth-DensePose using DS-Conv is larger. Because the DS-Conv uses two different kernel sizes.

This paper is organized: section 2 presents some important related work. Section 3 discusses the proposed system and method. Section 4 discusses the Depth-DensePose implementation, experiments and results. Finally, section 5 presents the conclusion.

2. RELATED WORK

In traditional methods, the extracted handcraft features of the input image are used to search for the best pose (localization) which matches the stored features. Deep learning methods directly learn good representations (encoding) of the input images at several levels of details. The representation could be unknown features, depth, or even 3D motion between two images for localization problems. Generally, the most popular three architectures of deep neural networks (DNNs) have been applied to solve the localization problem. The three architectures are convolutional neural network (CNN), encoder-decoder network, and recurrent neural network (RNN) [23].

Kendall *et al.* [19] suggested using a modified architecture from the well-known DNN architecture for classification, GoogLeNet [24], as a backbone for creating a pose regression network. The suggested architecture, named PoseNet [19]. GoogLeNet consists of six inception modules plus two intermediate classifiers that form a total of 22 layers. Each inception module is consisting of a stack of 3×3 filters, 5×5 filters, and a max-pooling layer which are useful to build a powerful model.

Nevertheless, however, the PoseNet accuracy is less than the accuracy of some traditional methods. This is due to the large size (2048D) of the input feature vector to the regressor layer. Also, increasing the size of the feature vector causes the overfitting problem, which leads to increased testing errors when working on test images that differ from the training images. Generalization is another important problem that not fully addressed in the original PoseNet model.

As a result of PoseNet success in encoding the localization problem as a regression task, many deep learning models were developed as enhancements on the PoseNet. Walch *et al.* [25] suggested using the long short term memory (LSTM) to reduce the feature vector size which addresses the overfitting problem. In this architecture, four LSTM units after used to reduce the size of the feature vector. LSTM is a type of RNN that consists of some hidden states that collect or omit relevant contextual features. Recently, combine CNN with LSTM is useful to solve some computer vision problems [25].

Hourglass-Pose [26] is another architecture based on the ResNet34 which is an encoder-decoder model [27]. The ResNet consists of four residual blocks, each one consists of convolution, activation, and batch normalization layers. There are direct (skip) connections between the corresponding encoder and decoder blocks. These connections are useful to preserve low-level details which help to solve the vanishing gradient problem.

In SVS-Pose [28], the VGG16 model [29] was utilized instead of the GoogLeNet model. The SVS-Pose uses a 3x3 filter throughout the network. The VGG16 contains three Fully Connected (FC) layers after the convolutional layers. The VGG16 model was cut out after the first FC into two branches, one to estimate the camera position and the other to estimate the orientation separately. There are two add FC layers at the end at the estimate the final position and orientation. Instead of using one single shared localizer model, BranchNet [10] cuts the PoseNet architecture after the 5th inception module as a shared encoder. The reset layers are duplicated to form two branches that estimate position and orientation separately.

Instead of returning a single estimated camera pose for the input image, the MDPoseNet returns multiple estimated values. The main MD-PoseNet idea is that the network returns the distribution of all potential camera poses instead of the most potential camera pose [30]. Unlike single-image localization, some works have extended PoseNet to the time domain to address temporal localization. VidLoc [31] performs localization for a small video-clips by bidirectional recurrent neural networks (BLSTM) [32]. VLocNet [33] and VLocNet++ [18] proposed a network to jointly learn pose regression and visual odometry. The Kalman filtering (KFNet) model [34] utilized the Kalman filtering [35] to improve the temporal localization for the online cameras.

In deep learning research, increasing the number of layers leads to build a more accurate model [36], [37]. Nevertheless, the deeper model may cause pop/vanish features problems that negatively affect the training results. To address these issues, Huang *et al.* [20] developed a densely connected network named, DenseNets. In the DenseNets architecture, there are dense connections between the consecutive blocks, which address the popping/vanishing problem and produce more accurate results. Moreover, dense connections between blocks are useful to overcome the over-fitting problem that occurs in training on a smaller dataset. DenseNets achieve high performances for many computer vision problems [37]–[41].

3. THE PROPOSED SYSTEM AND METHOD

According to the success of DenseNets in image classification and segmentation tasks [36], [37], we utilize its structure to build an efficient Depth-DensePose model for image pose regression. However, the use of DenseNets requires taking the following challenges into account: i) the standard DenseNets [20] was initially proposed for image classification tasks and not for regression tasks, ii) the structure of DenseNets actually consists of many max-pooling layers that decrease the resolution of features, iii) the DenseNets complexity is very high and requires high GPU memory. This limits some important factors that required to achieve high performance such as image size, number of layers, kernel size (usually 1×1 and 3×3), and the training dataset size [42].

Based on the above challenges, we proposed an efficient Depth-DensePose model for camera-based localization. The Depth-DensePose structure has the following advantages:

- Adopt the DenseNets [20], [37] architecture, which lead to developing a deeper model, fast training, and produce more accurate results.
- Develop the depthwise separable convolution (DS-Conv) with two kernel sizes 3 and 5 instead of the standard convolution. The kernel size 3 is useful to obtain the low-level features while the kernel with size 5 is useful to obtain the high-level features. As a result, the DS-Conv requires memory less than the standard convolution with similar performance [21].
- Reduce the average pooling layers by adopting a convolution with stride=2, which preserves the resolution and requires less GPU memory.

As show in Figure 1, DS-Conv contains a stack of depthwise and pointwise convolutions. In depthwise, the convolution operation is applied separately on each input channel. After that, a pointwise convolution is applied to merge the output channels from the depthwise convolution. The DS-Conv dramatically decreases the memory consumption and the computational complexity. Compared with standard convolution, the convolution process doesn't perform across all channels. That means the number of connections is fewer and the model is lighter. In this paper, we adapted the DS-Conv with two kernel sizes 3 and 5. The kernel size 3 is useful to obtain the low-level features while the kernel with size 5 is useful to obtain the high-level features.

The Depth-DensePose model consists of iterative depth-dense blocks, as in Figure 2 (in Appendix). Each depth-dense block consists of two DS-Conv with different kernel sizes (exactly, 3 and 5). There are links between the depth-densely blocks which maintains the resolution and enhances the training results [2].

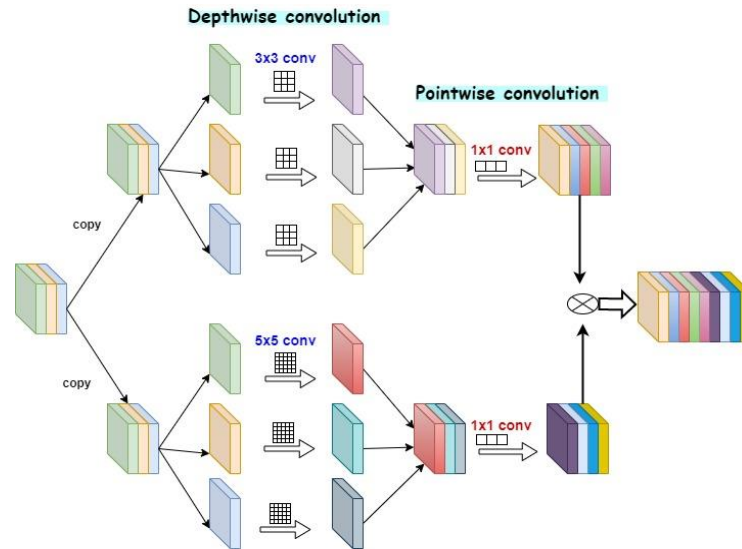


Figure 1. A proposed depthwise separable convolution with two kernel sizes (3 and 5)

After each convolution, a batch normalization (BN) and activation layers were applied. The BN is useful to create wider and deeper models. We implement the rectified linear unit (ReLU) as an activation layer. The ReLU is applied to optimize the output channels and enhanced the performance [43]. After each depth-dense block, a transition block is implemented to decrease the feature vector size. Each transition block consists of a 1×1 convolution layer followed by a BN, ReLU, and finally, convolution layer with stride 2 instated of the max-pooling layer. Like the PoseNet, the proposed structure has three regressor blocks that are applied after the transition layer 1, 2, and the final depth-dense blocks 4. Each regressor block consists of: i) an average pooling layer with appropriate kernel size as shown in Figure 2; ii) a 1×1 convolution with 128 kernels for dimension reduction followed by a Flatten layer. Flattening is converting the input two-dimensional vector into a one-dimensional vector; iii) FC layer with 512 outputs, then a ReLU activation layer, and iv) a fully connected regressor with 7 output values that represent the estimated pose.

The proposed model depth is scalable based on the iterative number of each depth-dense block (i.e. the values of $[N_1, N_2, N_3, N_4]$). The scalability is very useful to update the depth based on the size of the dataset. It's better to use a deeper model when there is a large-scale dataset. Based on experiments, the number of depth-dense blocks $[6, 8, 12, 6]$ gives the best results for the Cambridge Landmarks [19], [22] dataset.

4. RESULTS AND DISCUSSION

4.1. Implementation and training scheme

The proposed model was implemented by PyTorch [44]. PyTorch is an open-source library developed by Facebook's AI Research lab, used for machine/deep learning and computer vision applications. Table 1, lists the important implementation and training parameters. The implementation and training are done on a device equipped with NVIDIA RTX 2060 with 6 GB memory and 16 GB of RAM.

Transfer learning (TL) is very important to boost performance [45], [46]. In machine learning (ML), transfer learning is a methodology that concentrates on storing knowledge gained while solving one problem and utilizing it to a different but related problem. Therefore, we adopted the methodology of cascaded training [37], which leads to fast convergence and good performance. In this methodology, the training will conduct as iterative steps. In the first iteration, the model will train for 1,000 epochs. In the second iteration, the best training weights gained in the first iteration are utilizing, and the model will train for additional 1,000 epochs and so on. When there are no enhanced results, the cascaded training will stop.

Table 1. The important implementation and training parameters

Parameters	Values
Learning rate	0.0001
Regressor weights (w_1, w_2, w_3)	(0.3, 0.3, 1)
Batch sizes	16, 32..... 128 (based on the GPU and size of inputs)
Depth dense blocks $[N_1, N_2, N_3, N_4]$	[6, 8, 12, 6]

4.2. Experiments

Three experiments were conducted to evaluate the proposed Depth-DensePose: i) the first experiment is to evaluate the effectiveness of the proposed model compared to the related work; ii) the second experiment is to evaluate the effectiveness of using the DS-Conv compared to the standard convolution; and iii) the final experiment is to evaluate the efficiency of using the DS-Conv compared to the standard convolution.

4.2.1. Dataset

To evaluate the effectiveness, the depth-DensePose was trained and evaluated on the Cambridge Landmarks [19], [22] dataset. The Cambridge dataset is a large urban localization dataset consists of 6 scenes around the Cambridge University. Each scene is split into a set of training and testing images. The Cambridge dataset were captured using a phone camera and contains about 12,000 images labeled with their full 6-DoF camera pose. The resolution of each image is 1920×1080 px. For complexity reduction, many temporal frames were removed. Therefore, accurate temporal information is not available. The dataset labels are produced using the visual structure from motion (visual SfM) [47]. We applied several random scaling and transformations on the datasets, which lead to reduce the complexity and augment the dataset. The input images are randomly resized to 256×341 and randomly cropped to 224×224.

4.2.2. Effectiveness evaluation compared to the related work

The proposed Depth-DensePose was evaluated against the related work, especially PoseNet [19], DensePoseNet [19], LSTM-Pose [25], SVS-Pose [28], and VLocNet [33]. The MSE metric was utilized in these experiments of translation (in meters) and rotation. Table 2 summarizes the experimental MSE of position (in meters) and rotation (in degrees). According to these experiments, several results can be drawn. First, the proposed Depth-DensePose achieves a lower MSE and clearly outperforms the other related approaches. In general, the Depth-DensePose reduces the average MSE of the camera position to 0.74 and orientation to 2.14 for the entire Cambridge dataset. Second, the Depth-DensePose is even better than VLocNet [33] which failed to run on a large dataset. Furthermore, the scalability of the Depth-DensePose makes it works well on large as well as small datasets. The results in Figure 3, prove that the proposed architecture is beneficial to solve the camera-based localization problem.

Table 2. The Depth-DensePose MSE compared with other related model on the cambridge landmarks dataset

	K. College		Old Hospital		Shop Façade		Street		St M. Church		Average	
	P	R	P	R	P	R	P	R	P	R	P	R
PoseNet	1.92 m	5.40°	2.31 m	5.38°	1.46 m	8.08°	3.67 m	6.50°	2.65 m	8.48°	2.40 m	6.77°
DensePoseNet	1.66 m	4.86°	2.57 m	5.14°	1.41 m	7.18°	2.96 m	6.00°	2.45 m	7.96°	2.21 m	6.23°
LSTM-Pose	0.99 m	3.65°	1.51 m	4.29°	1.18 m	7.44°	NA	NA	1.52 m	6.68°	1.3 m	5.52°
SVS-Pose	1.06 m	2.81°	1.50 m	4.03°	0.63 m	5.73°	NA	NA	2.11 m	8.11°	1.33 m	5.17°
VLocNet	0.84 m	1.42°	1.07 m	2.41°	0.59 m	3.53°	NA	NA	0.63 m	3.91°	0.78 m	2.82°
Depth-DensePose	0.69 m	1.23°	0.60 m	0.82°	0.38 m	3.04°	1.24 m	3.34°	0.78 m	2.26°	0.74 m	2.12°

Note: P → Position, R → Rotation

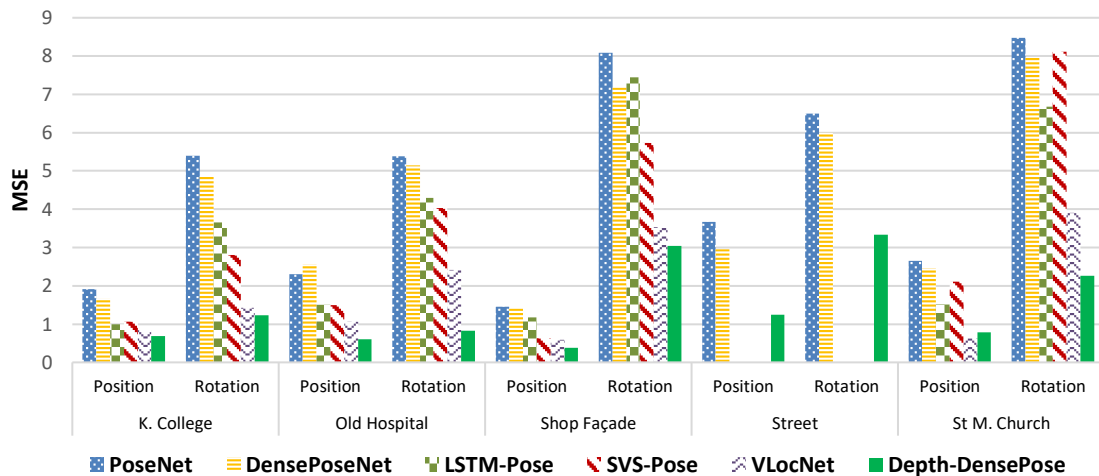


Figure 3. Evaluation of the Depth-DensePose model compared to the related models

4.2.3. Effectiveness of the adopted DS-Conv compared to the standard convolution.

Figures 4(a)-(f) show the results of experiments were conducted to evaluate the effectiveness of the adopted DS-Conv compared to the standard convolution. The proposed model was trained on the Shop Façade scene using DS-Conv and standard convolution. Training results were monitored and compared for different training epochs (especially, 100, 200, 300, and 400). From the results we can draw that, the effectiveness of the DS-Conv is comparable with the standard convolution. Furthermore, the training using DD-Conv archives lower training loss.

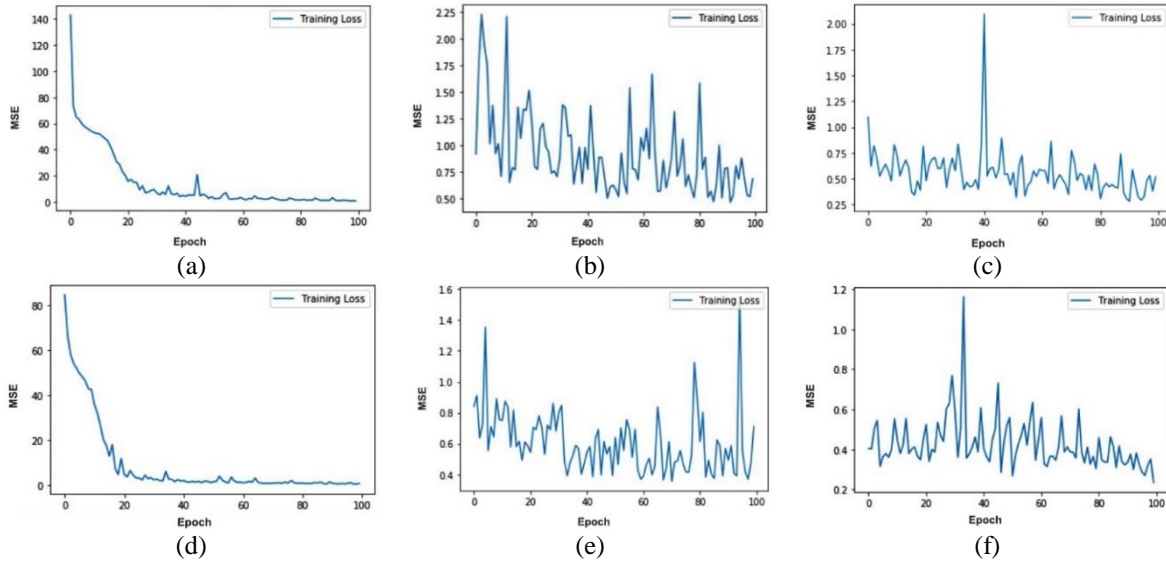


Figure 4. The Depth-DensePose training losses for different training epochs conducted on the shop façade scene (a) first 100 epochs using standard conv, (b) second 100 epochs using standard conv, (c) third 100 epochs using standard conv, (d) first 100 epochs using DS-Conv, (e) second 100 epochs using DS-Conv, and (f) third 100 epochs using DS-Conv

4.2.4. Efficiency of the adopted DS-Conv compared to the standard convolution

From Table 3 several conclusions can be drawn. First, the number of (trainable and total) parameters of the Depth-DensePose using the DS-Conv is notably lower than the Depth-DensePose using the standard convolution. Second, the Depth-DensePose using DS-Conv requires a lower memory size than standard convolution by 40%. This advantage is important to real-time and mobile application where memory size is limited. Other experiments were conducted to estimate the training time of the Depth-DensePose using DS-Conv compared to using standard convolution. Table 4, summarizes the results of the training time of one epoch for different patch sizes (especially 4, 16, and 32). As a result, the training time of the Depth-DensePose using DS-Conv is 20% lower than using standard convolution. It is worth mention that, the total number of layers of the Depth-DensePose using DS-Conv is larger. Because the DS-Conv uses two different kernel sizes.

Table 3. Params size of the Depth-DensePose using DS-Conv compared to standard convolution

	The Depth-DensePose using DS-Conv	The Depth-DensePose using standard Conv
Total params	5,571,797	9,065,685
Trainable params	5,571,797	9,065,685
Params size (MB)	21.25	34.58

Table 4. Training time (in second) of the Depth-DensePose using DS-Conv compared to standard Conv on the Cambridge dataset

	The Depth-DensePose using DS-Conv			The Depth-DensePose using standard Conv		
	4	16	32	4	16	32
K. College	70.13	38.32	28.15	86.99	42.37	35.41
Shop Façade	14.99	7.63	7.18	19.20	9.43	8.12
Street	288.63	127.89	107.47	344.38	186.57	157.47

5. CONCLUSION

In this paper, we propose, the Depth-DensePose, an efficient deep learning model for camera-based localization. The proposed Depth-DensePose utilizes both advantages of DenseNets and adapted DS-Conv to build a deeper and more efficient network. The results of the experiments demonstrated that, the Depth-DensePose achieves a lower mean squared error and clearly outperforms the other related approaches. Moreover, the proposed DS-Conv with two kernel sizes has a greater positive impact on the performance.

APPENDIX

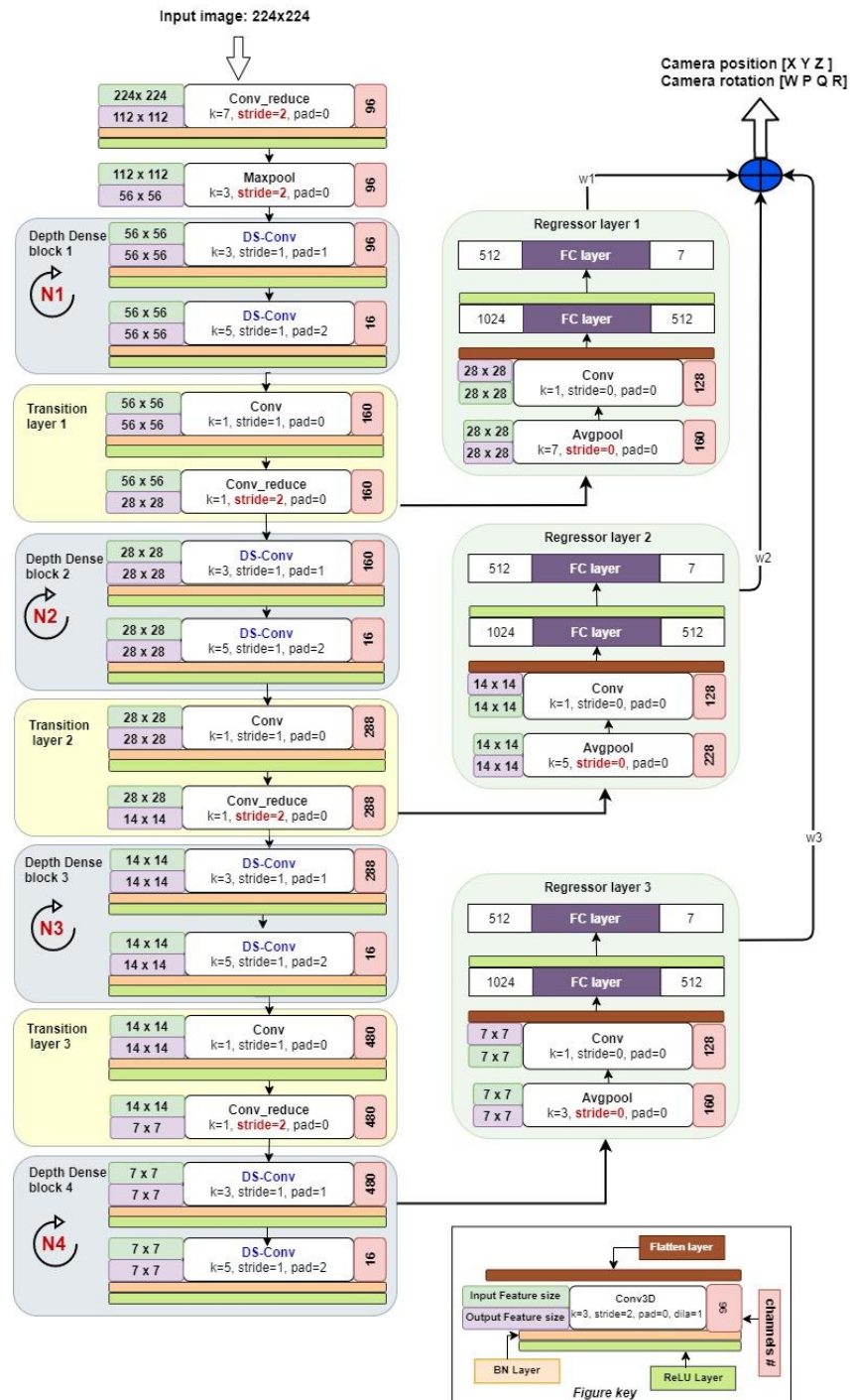


Figure 2. The proposed Depth-DensePose model for camera-based localization, where N1=6, N2=8, N3=12, and N4=6

ACKNOWLEDGEMENTS

This work was supported by the Egyptian Information Technology Industry Development Agency (ITIDA) under ITAC Program CFP # 164.




REFERENCES

- [1] J. Li, C. Wang, X. Kang, and Q. Zhao, "Camera localization for augmented reality and indoor positioning: a vision-based 3D feature database approach," *International Journal of Digital Earth*, vol. 13, no. 6, pp. 727–741, Jun. 2020, doi: 10.1080/17538947.2018.1564379.
- [2] H. Germain, G. Bourmaud, and V. Lepetit, "Sparse-to-dense hypercolumn matching for long-term visual localization," in *2019 International Conference on 3D Vision (3DV)*, Sep. 2019, pp. 513–523, doi: 10.1109/3DV.2019.00063.
- [3] S. J. Lee, D. Kim, S. S. Hwang, and D. Lee, "Local to global: efficient visual localization for a monocular camera," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2021, pp. 2230–2239, doi: 10.1109/WACV48630.2021.00228.
- [4] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: a survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, Jan. 2021, doi: 10.1007/s11263-020-01359-2.
- [5] E. N. Mortensen, Hongli Deng, and L. Shapiro, "A SIFT descriptor with global context," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 184–190, doi: 10.1109/CVPR.2005.45.
- [6] K. Eldahshan, H. Farouk, A. Abozeid, and M. H. Eissa, "Global dominant SIFT for video indexing and retrieval," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 19, pp. 5023–5035, 2019.
- [7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008, doi: 10.1016/j.cviu.2007.09.014.
- [8] M. Bansal, M. Kumar, and M. Kumar, "2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors," *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18839–18857, May 2021, doi: 10.1007/s11042-021-10646-0.
- [9] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: self-supervised interest point detection and description," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 337–33712, doi: 10.1109/CVPRW.2018.00060.
- [10] Z. Luo *et al.*, "ASLFEaT: learning local features of accurate shape and localization," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 6588–6597, doi: 10.1109/CVPR42600.2020.00662.
- [11] S. Garg and M. Milford, "SeqNet: learning descriptors for sequence-based hierarchical place recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4305–4312, Jul. 2021, doi: 10.1109/LRA.2021.3067633.
- [12] L. Wang, R. Li, J. Sun, H. Soon Seah, C. K. Quah, and L. Zhao, "Feature-based and convolutional neural network fusion method for visual relocalization," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Nov. 2018, pp. 1489–1495, doi: 10.1109/ICARCV.2018.8581204.
- [13] P. Roy and C. Chowdhury, "A survey of machine learning techniques for indoor localization and navigation systems," *Journal of Intelligent and Robotic Systems*, vol. 101, no. 3, Mar. 2021, doi: 10.1007/s10846-021-01327-z.
- [14] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, Mar. 2021, doi: 10.1016/j.isprsjprs.2020.12.010.
- [15] D. R. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," *Evolutionary Intelligence*, Jan. 2021, doi: 10.1007/s12065-020-00540-3.
- [16] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, Jun. 2020, doi: 10.1007/s13748-019-00203-0.
- [17] D. T. Mane and U. V. Kulkarni, "A survey on supervised convolutional neural network and its major applications," in *Deep Learning and Neural Networks*, IGI Global, 2020, pp. 1058–1071.
- [18] N. Radwan, A. Valada, and W. Burgard, "VLocNet++: deep multitask learning for semantic visual localization and odometry," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4407–4414, Oct. 2018, doi: 10.1109/LRA.2018.2869640.
- [19] A. Kendall, M. Grimes, and R. Cipola, "PoseNet: a convolutional network for real-time 6-dof camera relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2938–2946, doi: 10.1109/ICCV.2015.336.
- [20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
- [21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, Springer International Publishing, 2018, pp. 833–851.
- [22] A. Kendall, M. Grimes, and R. Cipola, "Cambridge Landmarks dataset," 2015. <http://mi.eng.cam.ac.uk/projects/relocalisation/> (accessed Oct. 12, 2021).
- [23] R. Li, S. Wang, and D. Gu, "Ongoing evolution of visual SLAM from geometry to deep learning: challenges and opportunities," *Cognitive Computation*, vol. 10, no. 6, pp. 875–889, Dec. 2018, doi: 10.1007/s12559-018-9591-8.
- [24] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
- [25] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 627–637, doi: 10.1109/ICCV.2017.75.
- [26] I. Melekhov, Y. Ilioinas, J. Kannala, and E. Rahtu, "Image-based localization using hourglass networks," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2017, vol. 2018-Janua, pp. 870–877, doi: 10.1109/ICCVW.2017.107.
- [27] W. Liu *et al.*, "SSD: single shot multibox detector," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 21–37, Dec. 2015, doi: 10.1007/978-3-319-46448-0_2.
- [28] T. Naseer and W. Burgard, "Deep regression for monocular camera-based 6-DoF global localization in outdoor environments," in *IEEE International Conference on Intelligent Robots and Systems*, Sep. 2017, vol. 2017-September, pp. 1525–1530, doi: 10.1109/IROS.2017.8205957.
- [29] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE*




- Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1943–1955, Oct. 2016, doi: 10.1109/TPAMI.2015.2502579.
- [30] H. Jo, W. Lee, and E. Kim, “Mixture density-PoseNet and its application to monocular camera-based global localization,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 388–397, Jan. 2021, doi: 10.1109/TII.2020.2986086.
- [31] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, “VidLoc: a deep spatio-temporal model for 6-DoF video-clip relocalization,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2652–2660, doi: 10.1109/CVPR.2017.284.
- [32] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.
- [33] A. Valada, N. Radwan, and W. Burgard, “Deep auxiliary learning for visual localization and odometry,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 6939–6946, doi: 10.1109/ICRA.2018.8462979.
- [34] L. Zhou *et al.*, “KFNet: learning temporal camera relocalization using kalman filtering,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 4918–4927, doi: 10.1109/CVPR42600.2020.00497.
- [35] R. J. Meinhold and N. D. Singpurwalla, “Understanding the kalman filter,” *The American Statistician*, vol. 37, no. 2, pp. 123–127, May 1983, doi: 10.1080/00031305.1983.10482723.
- [36] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. Ben Ayed, “HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1116–1126, May 2019, doi: 10.1109/TMI.2018.2878669.
- [37] N. Alalwan, A. Abozeid, A. A. ElHabshy, and A. Alzahrani, “Efficient 3D deep learning model for medical image semantic segmentation,” *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1231–1239, Feb. 2021, doi: 10.1016/j.aej.2020.10.046.
- [38] M. Huber, G. Schindler, C. Schorkhuber, W. Roth, F. Pernkopf, and H. Froning, “Towards real-time single-channel singing-voice separation with pruned multi-scaled densenets,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, vol. 2020-May, pp. 806–810, doi: 10.1109/ICASSP40776.2020.9053542.
- [39] H. Hu, Z. Li, L. Li, H. Yang, and H. Zhu, “Classification of very high-resolution remote sensing imagery using a fully convolutional network with global and local context information enhancements,” *IEEE Access*, vol. 8, pp. 14606–14619, 2020, doi: 10.1109/ACCESS.2020.2964760.
- [40] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K. K. Paliwal, and F. Shang, “Deep residual-dense lattice network for speech enhancement,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 8552–8559, Apr. 2020, doi: 10.1609/aaai.v34i05.6377.
- [41] R. Kandhari, M. Negi, P. Bhatnagar, and P. Mangipudi, “Use of deep learning models to detect COVID-19 from chest X-Rays,” in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Jan. 2021, pp. 1–5, doi: 10.1109/ICCCI50826.2021.9402545.
- [42] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Sep. 2014, *arXiv preprint arXiv:1409.1556*.
- [43] G. Zhao, Z. Zhang, H. Guan, P. Tang, and J. Wang, “Rethinking ReLU to train better CNNs,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 2018, vol. 2018-August, pp. 603–608, doi: 10.1109/ICPR.2018.8545612.
- [44] “An open source machine learning framework that accelerates the path from research prototyping to production deployment,” *Pytorch*, 2021. <https://pytorch.org/> (accessed Oct. 12, 2021).
- [45] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, “Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Medical Image Analysis*, vol. 54, pp. 280–296, May 2019, doi: 10.1016/j.media.2019.03.009.
- [46] N. Tajbakhsh *et al.*, “Convolutional neural networks for medical image analysis: full training or fine tuning?,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.
- [47] C. Wu, “VisualSFM: A Visual Structure from Motion System.” CCWU. <http://ccwu.me/vsfm/doc.html> (accessed Oct. 12, 2021).

BIOGRAPHIES OF AUTHORS







Amr Abozeid    is an assistant professor of computer science at Computer Science Department, College of Science & Arts (Gurayat), Jouf University, Saudi Arabia. He also worked as assistant professor of Computer Science at the Mathematics and Computer Science department, Faculty of Science, Al-Azhar University. He received B.Sc. degree in computer science and mathematics from Al-Azhar University, Cairo, Egypt, in 2005. He received the M.Sc. in computer science from the department of mathematics and computer science, faculty of science, Ain Shams University, in 2012. He received the Ph.D. in computer science from the department of mathematics and computer science, faculty of science, Al-Azhar University. His fields of research include video processing, computer vision, deep learning, and mobile computing. He can be contacted at email: aabozeid@azhar.edu.eg.



Hesham Farouk    is an associate Prof. since 2012. He joined the Electronics Research Institute, Egypt, in 1993. His fields of research are signal processing, mobile systems, Neural Networks, image compression, video processing, video compression, video indexing and retrieval, video on demand, pattern recognition and machine vision. Dr. Farouk received his Ph.D. at 2001 from Electronics & Communications Dept., Faculty of Engineering, Cairo Univ. and his M.Sc. at 1996 from Electronics & Communications Dept., Faculty of Engineering, Cairo Univ. Acting Manager Mobile, Social and Cloud Network Competence Center (MSCC) Ministry of Communication and Information Technology. Dr. Farouk joined Ministry of communications and Information Technology of Egypt in 2002 till June 2012. He was eContent Department manager, Information and infrastructure sector. Dr. Hesham participated in many national projects in MCIT developed based on portals and

digital libraries. He also participated in some strategic studies as mobile for development. Since June 2012 he is an Acting Manager Mobile, Social and Cloud Network Competence Center (MSCC), Technology Innovation and Entrepreneurship Center. Till March 2014. He was responsible about technology support to incubated companies and initiating new projects. For example, generating a Mobile applications national competition in 2012. Managing TIEC mobile lab. And help in establishing TIEC cloud lab. Then he is a technology consultant in ITI. 2012 Meanwhile, In ERI he is managing Technology Transfer office since June 2013 and the ERI technical office He has some professional activities as he is vice president to Mobile task force group running under EITESAL. Dr Farouk is Cisco certified for CCNA and as Instructor and certified from Improve academy as Innovation guide. Dr. Farouk Help in developing many National strategies as Mobile application, and EG-cloud Dr. Farouk is a lecturer at American University in Cairo, AUC, in the field of networking administration, web mastering and developing and he also Cisco certified teacher since 2003. He can be contacted at email: heshamali68@hotmail.com.



Samia Mashali     head of Digital Signal Processing group, Computers and Systems Department, Electronics Research Institute, Ministry of the higher education and Scientific Research, PhD in Electronics and Communications Engineering, Cairo University 1985. She was awarded a Badge of Honor, first degree, from Mr President Hosny Mubarak, a State Award in Engineering Science 1996. She was a technical consultant at the Ministry of Communications and Information Technology (MCIT) from 2004 till 2012. During this period, she was an education program director, the Deputy of the Egyptian Education Initiative program director, and ICDL program director. She is the Principal Investigator of six projects. She is a reviewer of a number of Electronics and Communication periodicals, a reviewer for the STDF, Academy of Scientific Research. She is the supervisor & examiner of more than 70 PhD. & MSc. Theses. She published more than 70 papers in international journals and conferences. Her main research technology topics are image processing and pattern recognition, speech processing, digital signal processing, neural networks, multimedia, computer visions, medical imaging, bioinformatics, information technology, IoT and augmented reality. She can be contacted at email: samiamashaly@gmail.com.