International Journal of Innovative Computing, Information and Control
Volume **18**, Number **5**, October **2022**

ICIC International ©2022 ISSN 1349-4198
pp. 1571–1586

A CNN-BASED MULTI-MODEL ENSEMBLE METHOD FOR INDOOR AND OUTDOOR MULTI-VIEW STEREO RECONSTRUCTION

Bhattarabhorn Wattanacheep and Orachat Chitsobhuk

School of Engineering
King Mongkut's Institute of Technology Ladkrabang
Chalongkrung Road, Ladkrabang, Bangkok 10520, Thailand
{ bhattarabhorn.wa; orachat.ch }@kmitl.ac.th

Received April 2022; revised July 2022

Abstract. *Camera poses estimation is a critical process that ensures the success of Three-Dimensional (3D) modelling. We present a Convolutional Neural Network (CNN)-based multi-model ensemble method for indoor and outdoor multi-view stereo reconstruction capable of learning across multiple domains, including images from both indoor and outdoor environments. Each domain's images have distinct properties and shooting viewpoints, which leads to difficulty in efficient learning such a large difference and requires large amount of computational resources. In order to reduce complexity of the end-to-end single model, the proposed model is divided into multiple learning agents consisting of domain-specific agents and domain relationship agent. The domain-specific agent is trained independently on its own set of unique image characteristics, for example, one for indoor datasets and another for outdoor datasets. The domain relationship agent then ensembles and analyzes the multiple domain features and finalizes the estimation. In terms of average root mean square error, we compare the performance of the combined domain single model with the suggested ensemble CNN model. The experimental results indicate that the proposed model outperforms the others, with rotation and translation prediction errors of 0.112012266.*

Keywords: 3D reconstruction, Convolutional neural network, Deep learning, Transfer learning, Ensemble CNN

1. **Introduction.** Nowadays, 3D modeling techniques are widely adopted in a variety of fields, for example, robotics, aircraft, Unmanned Aerial Vehicles (UAVs) [1-3], navigating autonomous vehicles [4], mobile robotics and augmented reality [5], virtual application skeleton-based action recognition [6] simulation and various components of large-scale localization. The basic techniques commonly used in finding structures for 3D modeling are Structure from Motion (SfM) and Simultaneous Localization and Mapping (SLAM) [7]. These techniques require feature extraction from two-dimensional images, for example, determining the angle of the object within the image and approximating geometric structure and camera poses to help estimate 3D model structural parameters. The ability of a feature detector such as Speeded Up Robust Features (SURF) [8], Oriented FAST and Rotated BRIEF (ORB) [9] or SIFT [10] methods used to discover corresponding feature points is critical. To detect feature points and match images, initial images are iteratively compared to find rotations and translations, within each image pair. The prediction time is proportional to the image size. Typically, Random Sample Consensus (RANSAC) is employed to assist the camera pose estimation algorithm in selecting feature point pairs. If the pairing discovered in the first stage is accurate, the camera pose will be estimated correctly. As a result, when the predetermined number of matches is satisfied, prioritized

DOI: 10.24507/ijicic.18.05.1571

search strategies [11-13] will complete the procedure. Situations such as motion blurring, light variation, and errors from different image order may all lead to feature matching failure. Moreover, if the estimation of the feature points is insufficient or their positions are errors, this makes it incapable of finding sufficient corresponding features between pairs of images, resulting in degraded rotation and translation prediction.

3D objects are widely used in computer vision applications ranging from human-machine interactions to autonomous vehicles and robotics [14]. Deep Learning (DL) has achieved impressive success in 2D fields [15-17,20,39] with various applications such as face recognition and image classification [18,19,41], pattern recognition, image enhancement, object detection [20,21,54], and semantic segmentation [22,23]. Since everything we perceive in the real world is in 3D space, 3D data can help improve the performance of computer vision-based applications [24].

In the recent years, several 3D databases have been made available to the public [25-27]. These advancements have enabled computer vision researchers to work with real-world objects, and DL-based 3D shape analysis research, including 3D classification, segmentation, retrieval, and reconstruction [55]. However, unlike the regular sampled 2D images, 3D shapes are irregular triangle meshes or point clouds; it is a challenging task for DL to extract distinguishing features [28] that can characterize the shapes and parts of a 3D object. In addition, knowledge from previously trained networks can be transferred to train on new problems. Conventional machine learning and deep learning algorithms, so far, have generally been designed to operate on solving specific tasks. When the feature space changes, the models must be rebuilt from scratch. Most models require a large amount of information to learn, especially, when solving complex problems. This leads to difficulty in labeling the ground truth for each image in these large training datasets. ImageNet's dataset, for example, consists of over a million images divided into several categories. Transfer learning attempts to derive the based learning model and apply the knowledge obtained for one task to solving a relevant one.

Transfer learning research begins with learning a target job and then modifies CNNs to predict camera poses from input images and extract valuable features that are robust to motion blur and illumination for localization problems. Kendall et al. [29] transferred learning from PoseNet and demonstrated that it was ineffective to predict camera poses using the high-dimensional output of fully connected layers, as they cannot produce the best results due to overfitting with the PoseNet training data.

[30] proposed a sorting algorithm that took advantage of scene semantics to create consistency between indoor and outdoor models. The research detected building windows and used as a key in reconstructing the three-dimension scenes since they were visible both inside and outside. The detected windows were then classified as indoor or outdoor using semantic classifiers and imported to Patch-based Multi-View Stereo (PM-VS) [31]. The results were compared to those of SfMs and they illustrated the efficiency of PMVS even in the case of noisy windows and misaligned indoor and outdoor position.

3D reassembly is an innovative and practical application which integrates indoor and outdoor 3D reconstruction algorithms into a single application. However, dense geometry of window detection is necessary since it affects the assembling of the 3D images. In [32], the authors adopted deep learning technology to automatically learn specific area patterns from a single input image. The dimension of the features was reduced after transferred learning from fully connected layer of VGG19. Finally, a regression estimation method was employed. Based on the Support Vector Regression (SVR) principles, this resulted in rotation and translation estimates with lower prediction error and independence from object geometry. The SVR's kernel functions were used to transform the original dataset

(linear/nonlinear) into higher dimensional space so that the data became linearly separable and make hyperplane decision boundary among classes. The larger the dimensional space, the longer the processing time. This resulted in a very long prediction time when dealing with a huge dataset. Furthermore, testing using pictures taken with various imaging characteristics (for example, learning with direct shooting to the objects but testing with parallel shooting to them) may lead to unsatisfactory results.

Later, there was a study on transfer learning using the neural network instead of SVR to learn on different image characteristics especially with shooting point of views [33]. Pooling was adopted to reduce the number of features and the neural network parameters were adjusted during training time on multiple epochs until reaching the minimum prediction errors. Parallel processes from neuron nodes of one layer to another layer allowed faster decision and offered higher accuracy.

Nevertheless, it was quite a complicated task to understand a wide range of image characteristics. If the model was previously trained on certain picture attributes, it will need to be retrained in order to learn about new aspects. Adjusting the model and dealing with all of the data volatility would be challenging. In this case, the model must be trained on substantial samples of all possible aspects from such as both indoor and outdoor datasets to recognize all desired cases. Another difficulty would be dealing with imbalanced datasets since we could not get sufficient looks of the underlying classes, resulting in poor accuracy for classes with less observation. Even if a deep learning model is fine-tuned to surpass the competition, the model may still have flaws in certain situations. Consequently, it is reasonable to assume that a strategy that makes use of a variety of deep learning techniques will deliver superior performance. Several researches have been published in the literature to this objective, with the goal of increasing the accuracy of prediction by integrating models into one another. As a result, there is a study that presents the concepts of ensembling [50,51,53], which is the process of integrating various learning algorithms in order to gain their collective performance, i.e., to improve the performance of existing models by combining several models into a single reliable model. Models are stacked together to improve their performance and obtain a single final prediction to ensure the most stable and accurate prediction possible. To reduce training complexity while increasing model accuracy, we proposed a multi-domain model in which each sub-model was trained independently on its own set of unique image characteristics, such as one for indoor and one for outdoor datasets, followed by an aggregate model to refine the final solution. The suggested model's hierarchical structure enabled it to estimate rotation and translation more accurately for indoor and outdoor camera pose estimation.

We demonstrated that estimation for rotation and translation can be calculated concurrently for all images without the need for an initial image for iteratively computing pairwise estimation, resulting in faster predictions and the ability to understand the image characteristics well regardless of the change in brightness, thereby alleviating the constraint associated with shooting methods. This led to a better understanding of more versatile and efficient shooting. Additionally, our study was shown to be applicable to images with less dense geometry and a wide variety of challenging datasets. We illustrated the effectiveness of our approach compared to the SIFT-based methodology especially for handling large texture region and repeating structure on the same datasets [32,33].

The remainder of this paper is organized as follows. Section 2 describes framework for 3D reconstruction parameter estimation for combined domain single model and the proposed ensemble of CNN models for indoor and outdoor multi-view stereo reconstruction. The simulation experiments are discussed in Section 3. Finally, Section 4 concludes this paper.

2. **Methodology.** In this research, we proposed an estimation of rotation and translation parameters of 2D pictures used to reconstruct 3D images based on transfer learning from a CNN model. However, training a CNN for multiple domains containing a variety of image attributes using a single model is a challenging task since the final model may have a deep and complicated architecture, which not only requires high computational cost but also can damage the accuracy and validation of the model [34]. Multiple domains can contain significantly different aspects of the shot, and the perspective and factors related to image acquisition resulted in a need for a large training dataset in order to cope with such a wide range of local characteristics. Moreover, introducing a new domain data requires the model to be retrained from scratch resulting in a lengthy development time. As a result, we introduced the CNN-based multi-model ensemble method for indoor and outdoor multi-view stereo reconstruction. We divided our agents into domain-specific agents and domain relationship agents to offer flexibility in learning on multiple domains of picture data to evaluate and estimate the required parameters. The efficacy of the based-line end-to-end single model and the recommended ensemble CNN model were explored and compared for indoor and outdoor image domains with highly diverse image acquisition formats and image properties.

2.1. **Proposed framework for 3D reconstruction parameter estimation.** The suggested system was divided into two parts: the first part is a preprocessing data section, which extracted significant properties from the image and sent them to the second part (ensemble CNN model), for learning and estimating by the learning agent network model.

For the preprocessing process, we transferred knowledge from the pre-trained CNN model of the prototype VGG19 trained with ImageNet [35] dataset and extracted the features from the fully connected layer 7 with feature dimension of 4096. Since our work did not require classification of data, 4096 features were dropped from the fully connected layer 7 of VGG19, which was delivered to our camera pose estimation system. However, the resulting features had a very high dimension, which might lead to data fragmentation (sparsity) and made data analysis more complex, slow, and inefficient. Therefore, it was necessary to reduce feature dimension that might not be highly relevant information for prediction. The problem was minimized by reducing the data dimension using Latent Semantic Analysis (LSA) [52], which was based on a mathematical technique called Singular Value Decomposition (SVD) [36], reducing feature dimension while maximizing the signal energy into as few coefficients as possible. The compact features were then sent to the ensemble model to estimate the camera pose parameters.

In this study, we compared the performance between the proposed ensemble CNN model (see in Figure 2) and a single model trained with multiple domains (see in Figure 1). The preprocessing of the data was handled in exactly the same way by both models; the only difference was in the design of the deep learning model.

2.1.1. *Combined domain single model.* Figure 1 showed a combined domain single model, which consisted of two single CNN models, one for estimating rotation parameters ($\hat{r}_{1234}$) and the other for estimating translation parameters ($\hat{t}_{123}$), which were used to learn transferred features from indoor and outdoor data. Each CNN contained four convolutional layers to extract spatial characteristics. There were max pooling layers after the second and fourth layers of each CNN followed by flatten layers to approximate the required parameters.

2.1.2. *Ensemble CNN model (A CNN-based multi-model ensemble method for indoor and outdoor multi-view stereo reconstruction).* The ensemble CNN model was separated into

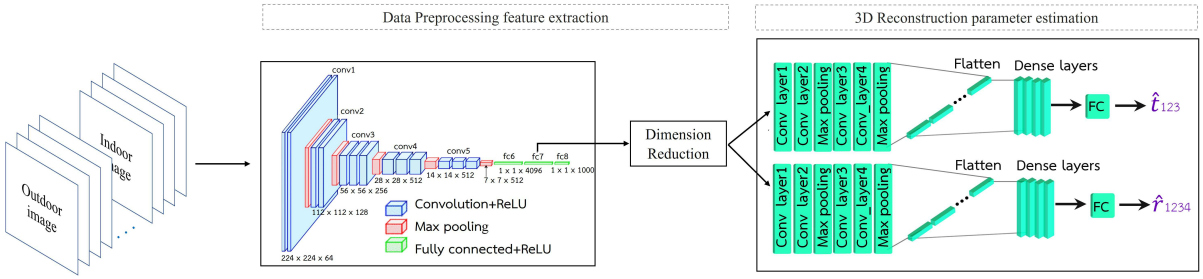


FIGURE 1. Combined domain single model

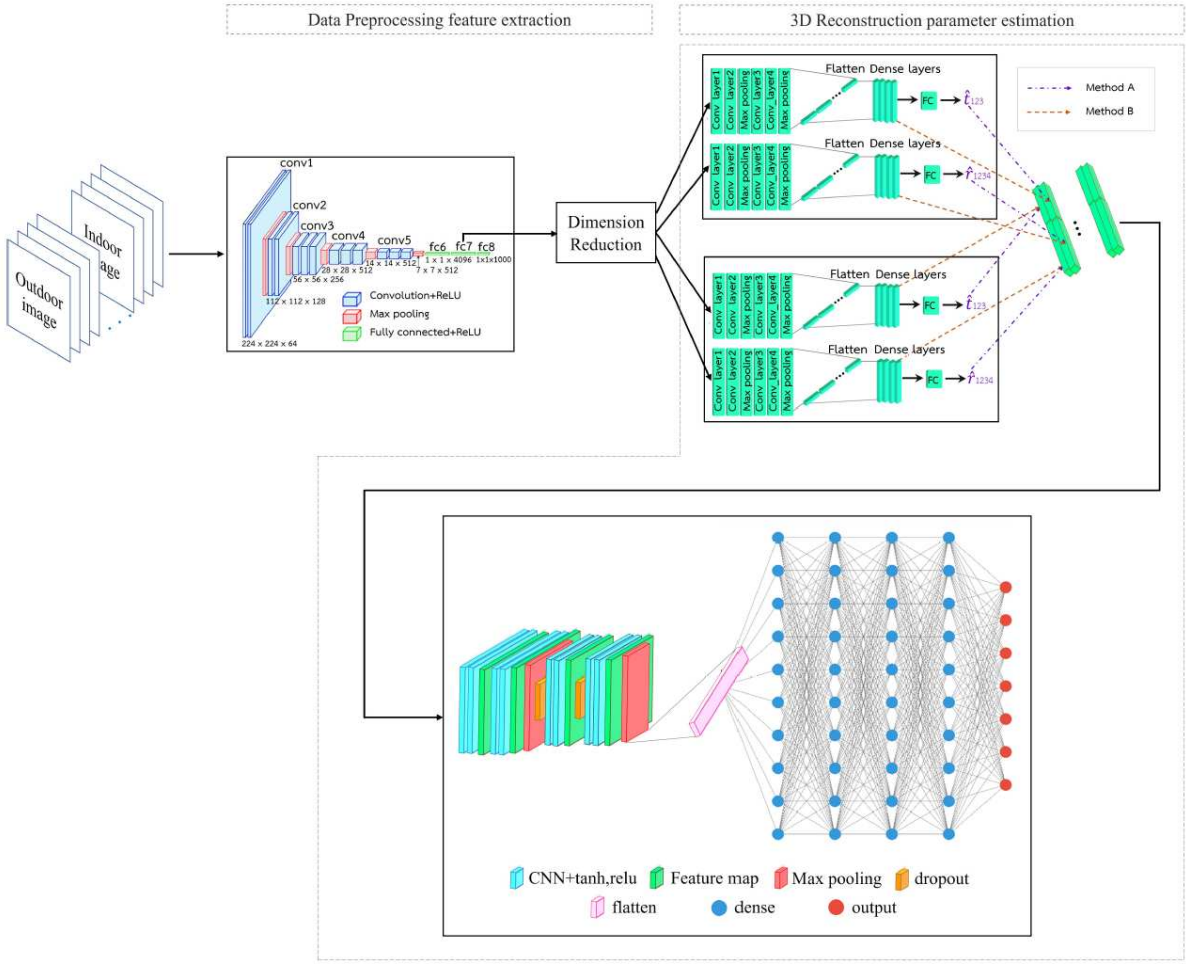


FIGURE 2. A CNN-based multi-model ensemble method for indoor and outdoor multi-view stereo reconstruction

domain-specific agents and domain relationship agents to enable the flexibility in learning a variety of multiple domain image dataset.

For a given dataset, a single algorithm may not be able to make the perfect prediction. Machine learning algorithms have limitations and achieving a high-accuracy model is challenging. The overall accuracy of the model could be improved if we build and combine multiple models. The combination can be implemented by combining the outputs from each model with two goals in mind: minimizing model error while maintaining generalization.

Domain-specific agents had the same structure as a single model's architecture but were trained the model with domain-specific data. Two learning agents were used to test the prototype: one was an indoor leaning agent, and the other was an outdoor learning agent. The parameter estimation results from all the domain-specific learning agents were sent to the domain relationship agent, which was used to assess the relationship between domains and predict the final rotation and translation parameters. We had investigated two methods of the feature arrangement from each domain-specific agent as illustrated in Figure 2 Method A (dash-dotted line) extracting features of the output layer of each domain-specific agent and arranged them into a structure of 7×2 features (4 rotation ($\hat{r}_{1234}$) and 3 translation parameters ($\hat{t}_{123}$)) while Method B (dashed line) combine features from the layer before the output layer of domain-specific agent and constructed 128×2 features (64 rotation and 64 translation features) from indoor and outdoor domain-specific agents.

The domain relationship agent consisted of a 2D CNN with a set of 2 convolutional layers and 2 max pooling layers followed by dense layers to derive inferences of regression prediction of the final camera pose parameters.

2.2. **Model performance measurement.** After predicting the rotation and translation of the test set, the model accuracy is computed using Root Mean Square Error (RMSE) [56] measurement.

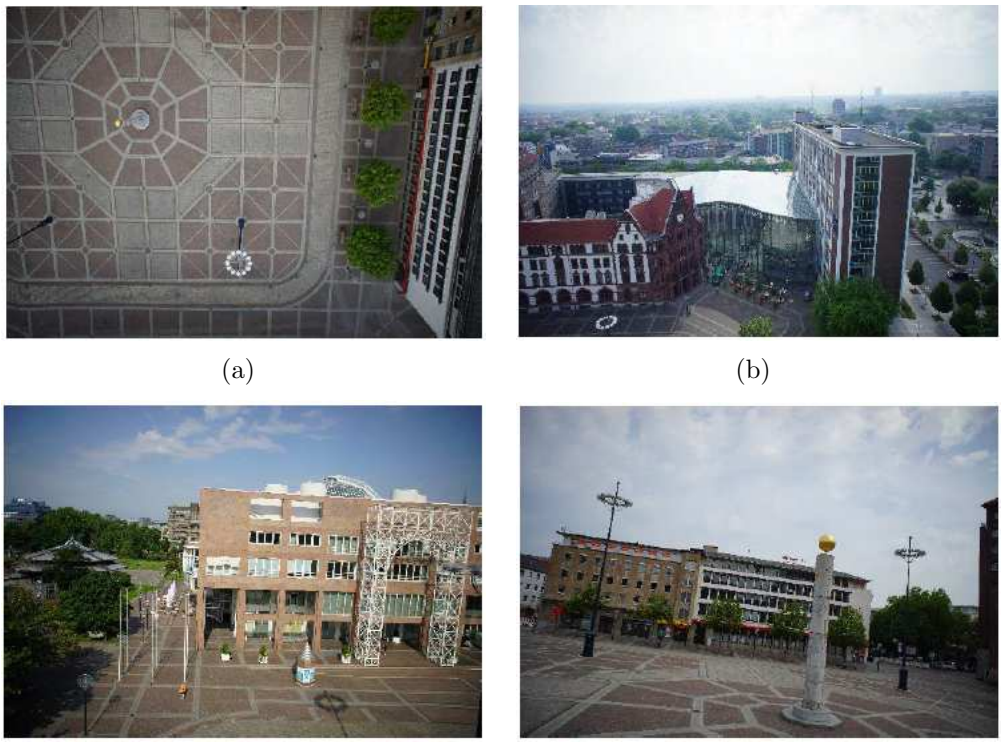
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{1}$$

where N = number of samples; x_i = observed values; $\hat{x}_i$ = predictions.

3. **Experimental Results.** In this research, datasets from three sources were used. The first one was Dome dataset, indoor photography from the Dome in the CMU Panoptic Studio [37]. The 209,934 images were taken by shooting around the object, with a variety of rotation and translations. The dataset was divided into 146,957 training and 62,977 test images. Each image came with the extrinsic camera parameters; a set of 480 possible patterns of different individual or group movements, for example, moving, walking around and baseball swings, as shown in Figure 3. The second source was Map dataset, outdoor aerial photography [38] taken from a drone camera, which flew above four different locations – nadir square, Stadthaus, Rathaus, and obelisk oblique square. There were large translations along the X and Y axes, while the Z-axis varied slightly. It comprised of 1500 training images and 578 testing images, as illustrated in Figure 4. The third source was the outdoor Cambridge dataset [29] which had six sub-datasets: GreatCourt, KingsCollege, OldHospital, ShopFacade, StMarysChurch and Street, as show in Figure 5. GreatCourt contains 1532 training images and 760 testing images. KingsCollege has the largest spatial extent of 5000 m^2 amongst all the datasets. It consists of 1220 training images and 346 testing images. OldHospital has a spatial extent of 2000 m^2, and contains 895 training and 182 testing images. ShopFacade contains 230 training images and 103 testing images. StMarysChurch contains 1487 training images and 530 testing images and Street contains 3015 training images and 2923 testing images. For the outdoor Cambridge dataset, we only used the test set to predict rotation and translation using in Section 2.1.2 Method B for benchmarks.

A variety of criteria were explored to determine the most effective model structure, including filter analysis, CNN layers, activation functions, optimizer type, batch size (epoch), and the appropriate number of dense layers. Filter size [2, 2] and [1, 2], number of filters [16, 32, 64, 128], output activation [TANH, RELU, LINEAR], Optimizer [Adam

FIGURE 3. Dome dataset, indoor photography from the Dome in the CMU Panoptic Studio [37]



(a) (b)

(c) (d)

FIGURE 4. Outdoor dataset of aerial photographs [38]: (a) Nadir square; (b) Stadthaus; (c) Rathaus; (d) Obelisk oblique square

and Adadelta], batch size [32, 50, 100], number of epochs [200, 500], and number of density layers [2 (decrease density), 4] were among the parameters. The experiments were divided into two parts: the first part was a parameter analysis of the combined domain single model described in Section 3.1 and that of the ensemble CNN model detailed in Section

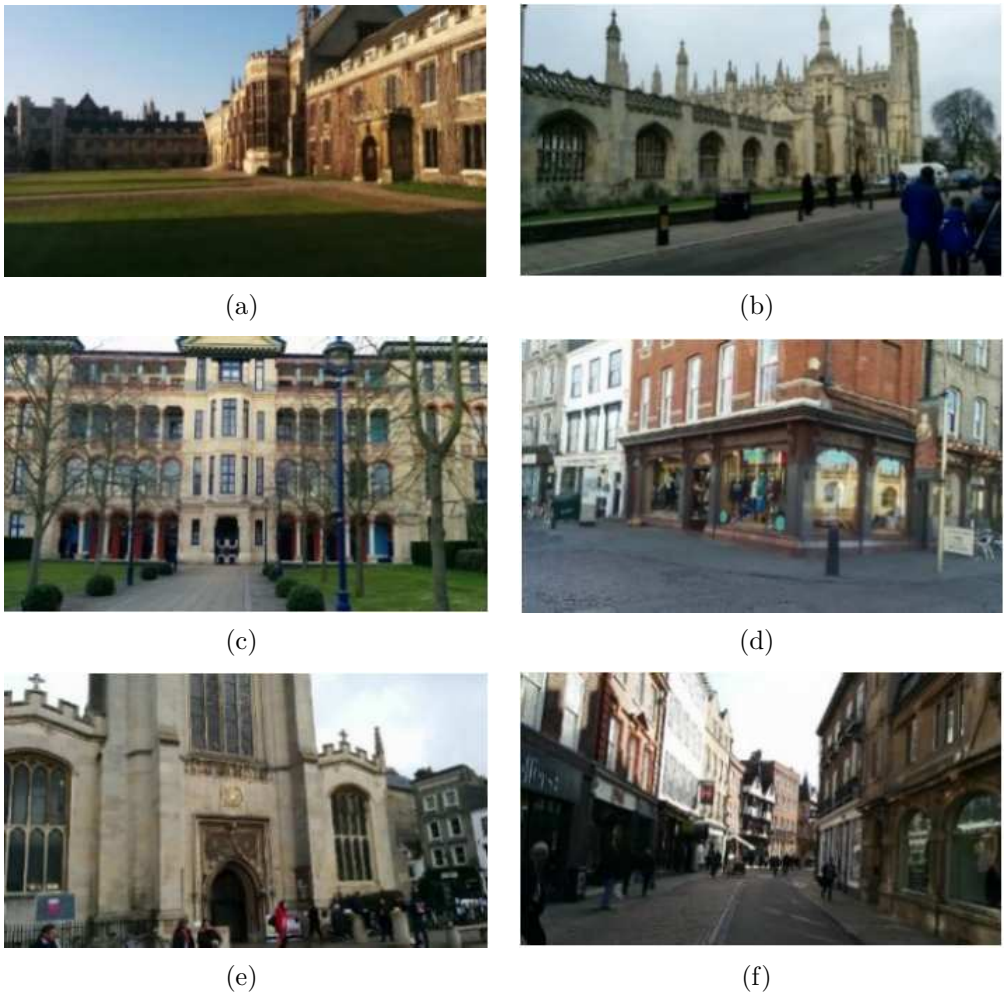


FIGURE 5. Outdoor dataset of Cambridge dataset [29]: (a) GreatCourt; (b) KingsCollege; (c) OldHospital; (d) ShopFacade; (e) StMarysChurch; (f) Street

3.2, and the second part was performance comparison among the research references [32], the combined domain single model and our proposed ensemble CNN model using the Root Mean Square Error (RMSE) as the performance metric illustrated in Section 3.3.

3.1. **Parameter analysis of a combined domain single model.** In the real world, we were occasionally confronted with unbalanced datasets, where the amount of data from several domains was unequally distributed. The indoor dataset was several times larger than the outdoor dataset approximately $1 : 100$ for this experiment, which could impact the system's overall performance. When the data from multiple domains are combined for the purpose of training a model, the proportions of features from larger domains tend to dominate the results. This means that predictions for domains with large volumes of data tend to be more accurate, as opposed to domain data with a limited volume, which results in poor predictive performance. According to the results of the experiment, we discovered that this effect starts to appear when the proportion of distinct data for each domain is different by 3 or more times. Consequently, we create a model trained on both the same $(1 : 1)$ and distinct quantities $(1 : 3, 3 : 1)$. The expectation is that, when evaluated with our approach, the approximations will provide comparable results even for domain datasets with unequal distribution. To evaluate the effects of an unbalanced dataset, we

conducted three trials with different ratios of training images from the two datasets: 1 : 1, 3 : 1, and 1 : 3 of the number of training images from the indoor dataset relative to the number of training images from the outdoor dataset, respectively. For performance evaluation, an equal number of test images were distributed across two domains. The average RMSE of rotation and translation across various optimizers, activations, epochs, and other configurations was shown in Table 1.

TABLE 1. The average RMSE values of rotation and translation predictions from the combined domain single model

Ratio of training images from two domain dataset	Activation	Minimum RMSE of rotation and translation
Set 1 – 1 : 1 (number of training indoor images = the number of training outdoor images = 1000)	Tanh+linear	0.616
	Tanh+linear (Reduce Dense)	0.655
	Tanh+relu+linear	**0.474**
	Tanh+relu+linear (Reduce Dense)	0.602
Set 2 – 3 : 1 (number of training indoor images = 1500, the number of training outdoor images = 500)	Tanh+linear	0.675
	Tanh+linear (Reduce Dense)	**0.548**
	Tanh+relu+linear	0.551
	Tanh+relu+linear (Reduce Dense)	0.697
Set 3 – 1 : 3 (number of training indoor images = 500, the number of training outdoor images = 1500)	Tanh+linear	0.620
	Tanh+linear (Reduce Dense)	0.684
	Tanh+relu+linear	**0.495**
	Tanh+relu+linear (Reduce Dense)	0.548

The average RMSE of rotation and translation predictions from training model with three sets of indoor and outdoor datasets and testing with 578 images each from indoor and outdoor test set was shown in Table 1. The minimum average RMSE of the training set 1 was 0.47409921 from the model parameters: 4 CNN layers with the number of filters (128, 32, 32, 16) for rotation model and (128, 16, 64, 16) for translation model, of each layer respectively, adadelta optimizer, 500 epochs, activation function (tanh+relu+linear), and 4 dense layers whereas the minimum average RMSE of the training set 2 was 0.548575566 from the model parameters: 4 CNN layers with the number of filters (64, 32, 32, 16) for rotation and (128, 16, 32, 16) for translation, of each layer respectively, adadelta optimizer, 500 epochs for rotation and 200 epochs for translation, activation function (tanh+linear (Reduce Dense)), and 2 dense layers. The minimum average RMSE of the third training set was 0.495257541 from the model parameters: 4 CNN layers with the number of filters (128, 128, 128, 16) for rotation and (128, 32, 32, 16) for translation, of each layer respectively, adadelta optimizer for rotation and adam optimizer for translation, 500 epochs, activation function (tanh+relu+linear).

3.2. **Parameter analysis of ensemble CNN method.** In this section, we trained two domain-specific agents, one from an indoor (Dome) dataset and the other from an outdoor (Map) dataset, with an integration of rotation and translation prediction into the same model. The experimental results with different feature arrangements from each domain-specific agent as mentioned in Section 2.1.2 named as Methods A and B from the training sets are shown in Figure 6.

It can be seen in Figure 6 that the minimum average RMSE of training set 1 was 0.131917386317026 with best model parameters as 4 CNN layers with the number of filters (64, 128, 128, 128), adam optimizer, 500 epochs, and 4 dense layers for indoor

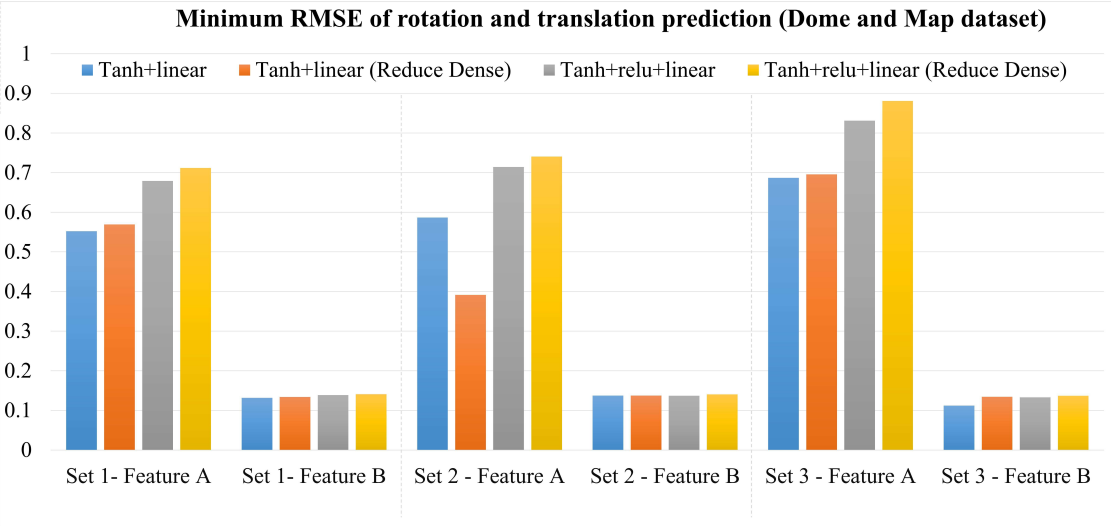


FIGURE 6. The minimum average RMSE of rotation and translation predictions for training sets 1, 2, and 3 with different ratios of training images from indoor (Dome) and outdoor (Map) datasets

Dome dataset and the number of filters (128, 128, 64, 16), adam optimizer, 500 epochs, and 4 dense layers for outdoor Map dataset, and activation functions (tanh+linear) for both datasets.

The minimum average RMSE of training set 2 is 0.137139154124717 with best model parameters as 4 CNN layers, the number of filters (128, 128, 128, 64) and 500 epochs for indoor Dome dataset and the number of filters (128, 128, 128, 128) and 200 epochs for outdoor Map dataset and activation functions (tanh+relu+linear) and adam optimizer for both datasets.

Accordingly, the minimum average RMSE of training set 3 was 0.11201226608203 with best model parameters as 4 CNN layers, the number of filters (128, 128, 128, 32) with adadelta optimizer and 200 epochs for indoor Dome dataset, and the number of filters (128, 128, 16, 16) with adam optimizer and 500 epochs for outdoor Map dataset, and activation functions (tanh+linear) for both datasets.

From the experiment, we noticed that the feature arrangement influenced the accuracy of the rotation and translation predictions. A feature arrangement Method B that combined features from the layer before the output layer of a domain-specific agents and constructed 128×2 features (64 rotation features and 64 translation features) from both indoor and outdoor domain-specific agents yielded the best results. The model trained on feature arrangement Method B was tested on the Cambridge dataset and the achieving median localization errors for both position and orientation were presented in Figure 7.

Comparing the localization accuracy across the datasets, as shown by a cumulative histogram, can reveal the differences in relative errors between them. It can be shown that the Street in the Cambridge Landmarks' outdoor dataset appears to be the most difficult to analyze. The same observation is pointed out by Walch et al. [40] concerning this unique behavior on this dataset. This dataset comprises videos recorded in opposite compass directions with similar spatial positions resulting in large angular deviations at similar global position. Although, OldHospital has a smaller spatial extent, it has relatively lower localization accuracy than KingsCollege because of large spatial camera movements. The cumulative distributions errors show that datasets with large angular deviations (ShopFacade) resulted in higher orientation errors than scenarios where the camera did not undergo severe rotations.

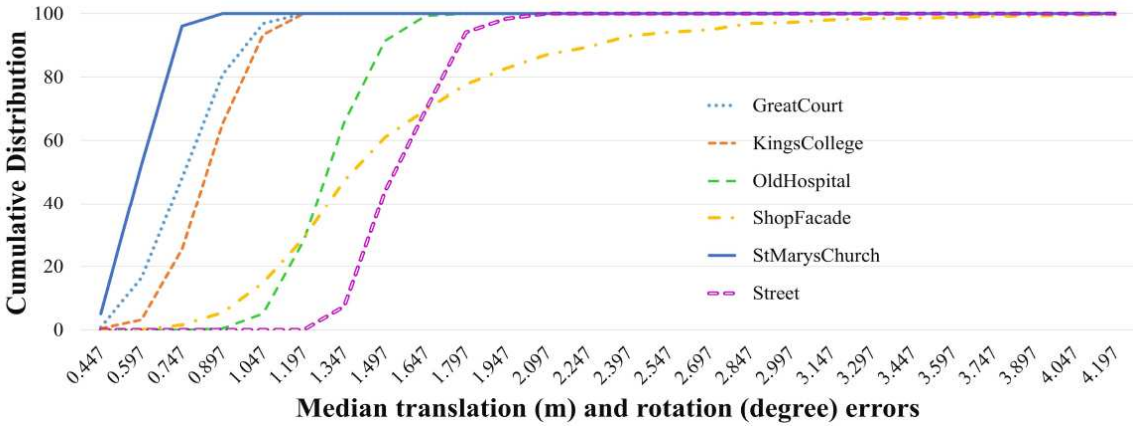


FIGURE 7. The median of rotation and translation predictions for Cambridge dataset

3.3. **The performance comparison of the reference model, the combined domain single model, and the proposed CNN-based multi-model ensemble method.** In this section, we illustrated the performance comparison among the reference technique [33], the implemented combined domain single model (2.1.1), and our proposed learning agent network model with different methods of feature arrangements (Methods A and B) on 3 training sets with various ratios of training images from Dome and Map datasets and several activation functions as shown in Table 2.

TABLE 2. The minimum average RMSE of the proposed ensemble CNN model and combined domain single model for multiple data domains

The number of training images divided	[33]	Combined domain single model	Ensemble CNN (Method A)	Ensemble CNN (Method B)
Set 1	2.501	0.474	0.495	0.131
Set 2		0.548	0.391	0.137
Set 3		0.495	0.686	0.112

Table 2 shows that the proposed ensemble CNN model with feature arrangement Method B provides the best performance for all training sets and activation function options with best minimum average RMSE of 0.112012266 which is 0.362086944 less prediction errors than the combined domain single model and 0.27951777 less prediction errors than the proposed ensemble CNN model with feature arrangement Method A, respectively. Moreover, when compared to the reference technique [33], the proposed ensemble CNN with feature organization Method B achieves lower prediction errors by 0.342181824, 0.411436412, and 0.460477478 for the three training sets and lower average RMSE by 2.389914218.

Predictions from a variety of reliable models can be combined to improve accuracy. A good model has skill, which means it can make better predictions than likelihood. Another important consideration is that the models must be good in a variety of ways, with a range of prediction errors. The reason that model averaging is effective is that different models do not always make the same mistakes on the test set when they are compared. The reason that model averaging is effective is that different models do not always make the same mistakes on the test set. When multiple neural networks' predictions are combined, a bias is introduced, which counteracts the variance of a single trained neural network

model. The result is predictions that are less sensitive to training data specifics, training scheme selection, and the serendipity of a single training run.

Table 3 illustrates the implementation of the Cambridge dataset, which is often used to evaluate the efficacy of camera pose estimation algorithms. A different shooting style was used in this dataset compared to Dome (indoor with differing view angles to the object) and Map (outdoor with the downward view angles from a drone flying horizontally). There are six subsets, the first five of which were locales, and the sixth of which was the street. The last one was the most difficult combination of images since the images were generally textureless. The results in Table 3 showed that the results for the first 10 methods were based on Shavit and Ferens [49], where LearnLess (DSAC++) [48] provided the least median error for subsets 1 to 5. However, no report for the last subset was found by Brachmann and Rother [48]. Despite this, our technique showed significantly lower median rotation and translation errors than the others. For subsets 1 to 5, the total of median errors from our approach was fewer than Brachmann and Rother [48] by 0.51° and 0.84 m, respectively, and less than Active Search (SfM) [49] by 0.89° and 2.2 m for subsets 2 to 6. Furthermore, for subsets 1 to 6, the methods in [33,34] reported the median errors, in which our method provides less median error by 32.21 m, 0.37 m and 39.58°, 1°, respectively.

TABLE 3. Median translation (m) and rotation errors for different pose estimators – using the Cambridge dataset

Algorithm	Great-Court	Kings-College	Old-Hospital	Shop-Facade	StMarys-Church	Street
PoseNet [29]	NA	1.97 m, 5.40°	2.31 m, 5.38°	1.46 m, 8.08°	2.65 m, 8.48°	3.67 m, 6.50°
Dense PoseNet [29]	NA	1.66 m, 4.86°	2.57 m, 5.14°	1.41 m, 7.18°	2.45 m, 7.96°	2.96 m, 6.00°
Bayesian PoseNet [42]	NA	1.74 m, 4.06°	2.57 m, 5.14°	1.25 m, 7.54°	2.11 m, 8.38°	2.14 m, 4.96°
LSTM-Pose [43]	NA	0.99 m, 3.65°	1.51 m, 4.29°	1.18 m, 7.44°	1.52 m, 6.68°	NA
SVS-Pose [44]	NA	1.06 m, 2.81°	1.50 m, 4.03°	0.63 m, 5.73°	2.11 m, 8.11°	NA
PoseNet+Reprojection error pose loss [45]	7.00 m, 3.7°	0.99 m, 1.1°	2.17 m, 2.9°	1.05 m, 4.0°	1.49 m, 3.40°	20.7 m, 25.7°
VLocNet [46]	NA	0.83 m, 1.42°	1.07 m, 2.41°	0.59 m, 3.53°	0.63 m, 3.91°	NA
DSAC [47]	2.80 m, 1.5°	0.30 m, 0.5°	0.33 m, 0.6°	0.09 m, 0.40°	0.55 m, 16°	NA
LearnLess (DSAC++) [48]	0.4 m, 0.2°	0.18 m, 0.3°	0.20 m, 0.3°	0.06 m, 0.30°	0.13 m, 0.4°	NA
Active Search [49]	NA	0.42 m, 0.6°	0.44 m, 1.0°	0.12 m, 0.40°	0.19 m, 0.5°	0.85 m, 0.8°
VGG19+CNN [33]	0.16 m, 0.18°	0.20 m, 0.21°	0.22 m, 0.48°	0.11 m, 0.20°	0.10 m, 0.39°	0.77 m, 0.76°
Our new system	**0.06 m, 0.12°**	**0.18 m, 0.05°**	**0.09 m, 0.46°**	**0.09 m, 0.13°**	**0.09 m, 0.08°**	**0.68 m, 0.38°**

4. **Conclusion.** In this research, we presented the estimation of the rotation and translation parameters of two-dimensional images used to reconstruct the 3D images using the proposed ensemble CNN model called a CNN-based multi-model ensemble method. To ensure flexibility in learning a variety of domains, the ensemble CNN model was composed of domain-specific agents and the domain relationship agent. The arrangement of features acquired from domain-specific agents and to be learned by the domain relationship agent was quite significant and had a large influence on the accuracy of the proposed ensemble CNN model. The feature arrangement that combined features from the layer before the output layer of both indoor and outdoor domain-specific agents and constructed 128×2 features (64 rotation features and 64 translation features) produced the best results, according to the experiments. Different ratios of training images and several test sets from indoor and outdoor domain datasets from various locations and shooting perspectives were evaluated in the experiments. The ensemble CNN model showed the highest predictive performance compared to the other algorithms. The results revealed that outdoor prediction has great potential for improvement. Since the Map dataset received by the drone's camera produces a large misalignment estimation error, we would like to train a specialized CNN that learns on additional depth parameters to assist with feature estimation prior to final aggregate estimation. We hope to expand our technique into these areas in the future.

REFERENCES

[1] C. Forster, M. Pizzoli and D. Scaramuzza, SVO: Fast semi-direct monocular visual odometry, *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp.15-22, doi: 10.1109/ICRA.2014.6906584, 2014.

[2] J. Engel, J. Sturm and D. Cremers, Camera-based navigation of a low-cost quadrocopter, *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.2815-2821, doi: 10.1109/IROS.2012.6385458, 2012.

[3] M. Achtelik, M. Achtelik, S. Weiss and R. Siegwart, Onboard IMU and monocular vision based control for MAVs in unknown in- and outdoor environments, *2011 IEEE International Conference on Robotics and Automation*, pp.3056-3063, doi: 10.1109/ICRA.2011.5980343, 2011.

[4] H. Lim, S. N. Sinha, M. F. Cohen and M. Uyttendaele, Real-time image-based 6-DOF localization in large-scale environments, *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1043-1050, doi: 10.1109/CVPR.2012.6247782, 2012.

[5] S. Lynen, T. Sattler, M. Bosse, J. Hesch, M. Pollefeys and R. Siegwart, Get out of my lab: Large-scale, real-time visual-inertial localization, *Computer Vision and Pattern Recognition*, doi: 10.15607/RSS.2015.XI.037, 2019.

[6] C. Ding, K. Liu, F. Cheng and E. Belyaev, Spatio-temporal attention on manifold space for 3D human action recognition, *Appl. Intell.*, pp.560-570, 2021.

[7] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd Edition, Cambridge University Press, 2004.

[8] 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, 2008.

[9] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, ORB: An efficient alternative to SIFT or SURF, *2011 International Conference on Computer Vision*, pp.2564-2571, doi: 10.1109/ICCV.2011.6126544, 2011.

[10] E. N. Mortensen, H. Deng and L. Shapiro, A SIFT descriptor with global context, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol.1, pp.184-190, doi: 10.1109/CVPR.2005.45, 2005.

[11] S. Choudhary and P. J. Narayanan, Visibility probability structure from SfM datasets and applications, *European Conference on Computer Vision (ECCV)*, 2012.

[12] Y. Li, N. Snavely and D. P. Huttenlocher, Location recognition using prioritized feature matching, *European Conference on Computer Vision (ECCV)*, vol.6312, pp.791-804, 2010.

[13] T. Sattler, B. Leibe and L. Kobbelt, Efficient & effective prioritized matching for large-scale image-based localization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.9, pp.1744-1756, doi: 10.1109/TPAMI.2016.2611662, 2017.

[14] E. Ahmed, A. Saint, A. E. R. Shabayek, K. Cherenkova and D. Aouada, Deep learning advances on different 3D data representations: A survey, *http://arxiv.org/abs/1808.01462v2*, 2019.

[15] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (eds.), Red Hook, NY, USA, Curran Associates, Inc., 2012.

[16] S. Tural, R. Samet, S. Aydin and M. Traore, Deep learning based classification of military cartridge cases and defect segmentation, *IEEE Access*, vol.10, pp.74961-74976, 2022.

[17] W. Wang, C. Tang, X. Wang, Y. Luo, Y. Hu and J. Li, Image object recognition via deep feature-based adaptive joint sparse representation, *Computational Intelligence and Neuroscience*, vol.2019, Article ID: 8258275, 2019.

[18] S. Lee and K. Jo, Person browser system based on named entity recognition for broadcast news interview videos, *Int. J. Control Autom. Syst.*, vol.19, pp.186-199, https://doi.org/10.1007/s12555-019-0391-z, 2021.

[19] M. Tan and Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, *International Conference on Machine Learning*, pp.6105-6114, arXiv Preprint, arXiv: 1905.11946, 2019.

[20] C. Szegedy et al., Going deeper with convolutions, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp.1-9, doi: 10.1109/CVPR.2015.7298594, 2015.

[21] N. Westlake, H. Cai and P. Hall, Detecting people in artwork with CNNs, *European Conference on Computer Vision (ECCV)*, pp.825-841, 2016.

[22] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao and B. Catanzaro, Improving semantic segmentation via video propagation and label relaxation, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.8856-8865, 2019.

[23] V. Badrinarayanan, A. Kendall and R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.12, pp.2481-2495, doi: 10.1109/TPAMI.2016.2644615, 2017.

[24] A. Voulodimos, N. Doulamis, A. Doulamis and E. Protopapadakis, Deep learning for computer vision: A brief review, *Computational Intelligence and Neuroscience*, vol.2018, Article ID: 7068349, 2018.

[25] A. X. Chang, T. Funkhouser, L. Guibas et al., ShapeNet: An information-rich 3D model repository, *http://arxiv.org/abs/1512.03012*, 2019.

[26] S. Choi, Q.-Y. Zhou, S. Miller and V. Koltun, A large dataset of object scans, *http://arxiv.org/abs/1602.02481*, 2019.

[27] S. Song, S. P. Lichtenberg and J. Xiao, SUN RGB-D: A RGB-D scene understanding benchmark suite, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp.567-576, 2015.

[28] I. K. Kazmi, L. You and J. J. Zhang, A survey of 2D and 3D shape descriptors, *Proc. of the 2013 10th International Conference Computer Graphics, Imaging and Visualization*, Los Alamitos, CA, USA, pp.1-10, 2013.

[29] A. Kendall, M. Grimes and R. Cipolla, PoseNet: A convolutional network for real-time 6-DoF camera relocalization, *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[30] A. Cohen, J. L. Schönberger, P. Speciale, T. Sattler, J.-M. Frahm and M. Pollefeys, Indoor-outdoor 3D reconstruction alignment, in *Computer Vision – ECCV 2016, Lecture Notes in Computer Science*, B. Leibe, J. Matas, N. Sebe and M. Welling (eds.), vol.9907, Springer, https://doi.org/10.1007/978-3-319-46487-9_18, 2016.

[31] Y. Furukawa and J. Ponce, Accurate, dense, and robust multiview stereopsis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.32, no.8, pp.1362-1376, doi: 10.1109/TPAMI.2009.161, 2010.

[32] B. Wattanacheep and O. Chitsobhuk, Prediction of 3D rotation and translation from 2D images, *The 7th International Conference on Computer and Communications Management (ICCCM2019)*, Association for Computing Machinery, New York, NY, USA, pp.49-52, https://doi.org/10.1145/3348445.3348485, 2019.

[33] B. Wattanacheep and O. Chitsobhuk, Camera pose estimation using CNN, *2020 the 3rd International Conference on Control and Computer Vision (ICCCV'20)*, Association for Computing Machinery, New York, NY, USA, pp.84-88, doi: https://doi.org/10.1145/3425577.3425593, 2020.

[34] O. Lantang, G. Terdik, A. Hajdú and A. Tiba, Comparison of single and ensemble-based convolutional neural networks for cancerous image classification, *Annales Mathematicae et Informaticae (54.)*, pp.45-56, http://dx.doi.org/10.33039/ami.2021.03.013, 2021.

[35] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *The 3rd International Conference on Learning Representations (ICLR2015)*, https://arxiv.org/abs/1409.1556, 2015.

[36] R. A. Sadek, SVD based image processing applications: State of the art, contributions and research challenges, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.3, no.7, pp.26-34, https://arxiv.org/ftp/arxiv/papers/1211/1211.7102.pdf, 2012.

[37] H. Joo, H. S. Park and Y. Sheikh, MAP visibility estimation for large-scale dynamic 3D reconstruction, *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, pp.1122-1129, doi: 10.1109/CVPR.2014.147, 2014.

[38] F. C. Nex, M. Gerke, F. Remondino, H. J. Przybilla, M. Baumker and A. Zurhorst, ISPRS benchmark for multi-platform photogrammetry, *Annals of the Photogrammetry, Remote Sensing and Spatial Information Science*, vol.II-3/W4, Munich, Germany, pp.135-142, 2015.

[39] S. M. Noe, T. T. Zin, P. Tin and I. Kobayashi, Automatic detection and tracking of mounting behavior in cattle using a deep learning-based instance segmentation model, *International Journal of Innovative Computing, Information and Control*, vol.18, no.1, pp.211-220, 2022.

[40] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck and D. Cremers, Image-based localization using LSTMs for structured feature correlation, *Proc. of the IEEE International Conference on Computer Vision*, pp.627-637, 2017.

[41] A. F. Siregar and T. Mauritsius, Ulos fabric classification using android-based convolutional neural network, *International Journal of Innovative Computing, Information and Control*, vol.17, no.3, pp.753-766, 2021.

[42] K. Alex and R. Cipolla, Modelling uncertainty in deep learning for camera delocalization, *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp.4762-4769, arXiv Preprint, arXiv: 1509.05909, 2016.

[43] S. Bell, C. L. Zitnick, K. Bala and R. Girshick, Inside-outside Net: Detecting objects in context with skip pooling and recurrent neural networks, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp.2874-2883, doi: 10.1109/CVPR.2016.314, 2016.

[44] T. Naseer and W. Burgard, Deep regression for monocular camera-based 6-DoF global localization in outdoor environments, *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.1525-1530, doi: 10.1109/IROS.2017.8205957, 2017.

[45] A. Kendall and R. Cipolla, Geometric loss functions for camera pose regression with deep learning, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp.6555-6564, 2017.

[46] V. Abhinav, N. Radwan and W. Burgard, Deep auxiliary learning for visual localization and odometry, *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp.6939-6946, 2018.

[47] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold and C. Rother, DSAC – Differentiable RANSAC for camera localization, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp.6684-6692, https://github.com/cvlab-dresden/DSAC, 2017.

[48] E. Brachmann and C. Rother, Learning less is more – 6D camera localization via 3D surface regression, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp.4654-4662, https://github.com/vislearn/LessMore, 2018.

[49] Y. Shavit and R. Ferens, Introduction to camera pose estimation with deep learning, *arXiv Preprint*, arXiv: 1907.05272, 2019.

[50] J. Zhang, L. Chen, J. Tian et al., Breast cancer diagnosis using cluster-based undersampling and boosted C5.0 algorithm, *Int. J. Control Autom. Syst.*, vol.19, pp.1998-2008, https://doi.org/10.1007/s12555-019-1061-x, 2021.

[51] A. Manna, R. Kundu, D. Kaplun et al., A fuzzy rank-based ensemble of CNN models for classification of cervical cytology, *Sci. Rep.*, vol.11, 14538, https://doi.org/10.1038/s41598-021-93783-8, 2021.

[52] C.-H. Liao, S.-M. Chen, B.-C. Kuo and K.-C. Pai, A Chinese vocabulary learning system: Latent semantic analysis approach, *International Journal of Innovative Computing, Information and Control*, vol.10, no.6, pp.2179-2191, 2014.

[53] A. Y. Yousif, S. M. Younis, S. A. Hussein and N. M. G. Al-Saidi, An intelligent computing for diagnosing COVID-19 using available blood tests, *International Journal of Innovative Computing, Information and Control*, vol.18, no.1, pp.57-72, 2022.

[54] S. Lata and O. Surinta, An end-to-end Thai fingerspelling recognition framework with deep convolutional neural networks, *ICIC Express Letters*, vol.16, no.5, pp.529-536, doi: 10.24507/icicel.16.05.529, 2022.
[55] J. Wietrzykowski and D. Belter, Stereo plane R-CNN: Accurate scene geometry reconstruction using planar segments and camera-agnostic representation, *IEEE Robotics and Automation Letters*, vol.7, no.2, pp.4345-4352, doi: 10.1109/LRA.2022.3150841, 2022.
[56] A. G. Barnston, Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score, *Weather and Forecasting*, vol.7, no.4, pp.699-709, 1992.

Author Biography

Bhattarabhorn Wattanacheep received the B.E. degree in Computer Engineering from King Mongkut's Institute of Technology, Ladkrabang, Thailand, in 2012, the M.S. degree in Computer Engineering from King Mongkut's Institute of Technology, Ladkrabang, Thailand, in 2014. She is currently a student in Ph.D. program (Electrical Engineering at King Mongkut's Institute of Technology Ladkrabang, Thailand). Her research interests include robotics, image processing and optimization of technologies using machine learning.

Orachat Chitsobhuk received the B.E. degree in Electronics Engineering from King Mongkut's Institute of Technology, Ladkrabang, Thailand, in 1992, the M.S. degree in Computer Engineering from Arizona State University, AZ, in 1997, and the Ph.D. degree in Electrical Engineering from University of Texas, Arlington, US, in 2001. She is currently an associate professor and a lecturer at King Mongkut's Institute of Technology Ladkrabang, Thailand. Her research interests include image and scene analysis, machine learning and pattern recognition, and hardware design for image processing applications.