IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 September 2023, accepted 10 October 2023, date of publication 18 October 2023, date of current version 27 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325661

METHODS

Deep Photo-Geometric Loss for Relative Camera Pose Estimation

YONGJU CHO[1], SEUNGHO EUM[2], JUNESEOK IM[2], ZAHID ALI[2], HYON-GON CHOO[1], AND UNSANG PARK[2]

[1]Electronics and Telecommunications Research Institute, Daejeon 34129, South Korea
[2]Department of Computer Science and Engineering, Sogang University, Seoul 04170, South Korea

Corresponding author: Unsang Park (unsangpark@sogang.ac.kr)

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean Government (The research of the fundamental media.contents technologies for hyper-realistic media space) under Grant 23ZH1220.

ABSTRACT CNN-based absolute camera pose estimation methods lack scene generalizability as the network is trained with scene-specific parameters. In this paper, we aim to solve the scene generalizability problem in 6-DoF camera pose estimation using a novel deep photo-geometric loss. We train a CNN-based relative pose estimation network end-to-end, by jointly optimizing the proposed deep photo-geometric loss along with the pose regression loss. Most traditional pose estimation methods use local keypoints to find 2D-2D correspondences, which fails under occlusion, textureless surfaces, motion blur, or repetitive structures. Given camera intrinsics, poses and depth, our method generates uniform 2D-2D photometric correspondence pairs via epipolar geometry during the training process with constraints to avoid textureless surfaces and occlusion, without the need of manually annotated keypoints information. The network is then trained with the correspondences information in such a way that not only the network learns from auxiliary photometric consistency information but also efficiently leverages scene geometry, consequently, we call it photo-geometric loss. The input to the photo-geometric loss layer is taken from the activation maps of the deep network, which contains much more information than a simple 2D-2D correspondence, and thus alleviating the need to choose a robust pose regression loss and its hyperparameters. With extensive experiments on three public datasets, we show that the proposed method significantly outperforms state-of-the-art relative pose estimation methods. The presented method also depicts state-of-the-art results on these datasets under cross-database evaluation settings, which proves its significance in terms of scene generalization.

INDEX TERMS Camera calibration, camera pose estimation (CPE), structure from motion (SfM), multi-view stereo (MVS), perspective-n-point (PnP), 6-DoF camera, photometric information, geometric consistency information.

I. INTRODUCTION

Camera pose estimation (CPE) is a vital task for many computer vision problems, such as Structure from Motion (SfM) [40], Simultaneous Localization And Mapping (SLAM) [3], and depth inference from multi-view stereo [42]. Before the advent of convolutional neural networks (CNN), the task of CPE was mainly accomplished by extracting sparse keypoints like SIFT, SURF, and then establishing 2D-2D correspondences between the matching keypoints, followed by the estimation of the camera pose by solving Perspective-n-Point (PnP) problem [11]. This traditional method of CPE, however, suffers from two main problems, i) noisy features due to repetitive structures and non-Lambertian surfaces, ii) an overall slow process of computing the features from multiple images and then solving for correspondence. On the other hand, after the success of CNNs on a variety of computer vision tasks, such as image classification, object detection, image retrieval, semantic segmentation, etc., researchers proposed CNN-based solutions to predict camera

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

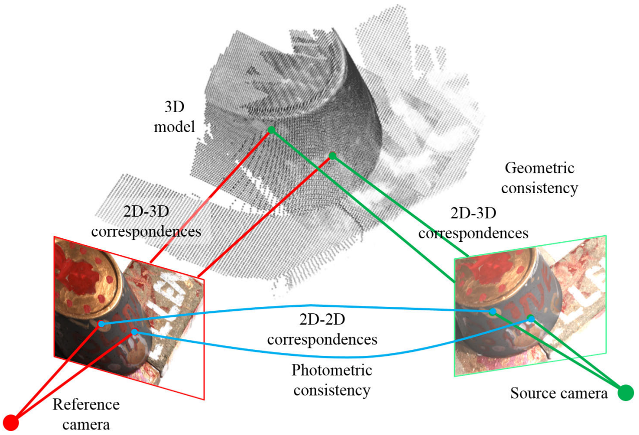

FIGURE 1. Visualization of 2D-2D correspondences (cyan) used to train the proposed end-to-end RTNet. The 2D-2D correspondences are only used at training time and not required at test time.

pose. In contrast with the hand-crafted feature-based CPE methods, the main benefit CNN-based counterparts offer is the single step pose estimation (PE) with much faster processing. Some researchers have successfully demonstrated the use of deep learning (DL) modules to replace only parts of the traditional PE pipeline, such as keypoint extraction [10], [12], [29], keypoint matching [35], [43], and RANSAC [4], [30].

The mitigation of classical PE schemes to DL techniques is also inspired by the aim to eliminate manual engineering for feature selection. The classical two-stage PE pipeline does not take full benefit from the global photometric and geometrical constraints. On the other hand, CNN architectures that directly regress the 6-DoF camera pose [13], [19] looks at the complete image at once, and simultaneously regress the pose after utilizing the global context information from the input image. However, they too fail to capture the structural or geometric information of the given scene. Therefore, the researchers focused on utilizing alternative informations in the form of 2D-to-3D correspondence (reprojection loss) [18], temporal cues [9], global positioning [5], reinforcement learning by using probabilistic selection of deterministic hypothesis [4]. Nevertheless, accurate CPE using an end-to-end trainable deep network is a challenging problem due to difficulties in modeling the photometric and geometric constraints simultaneously.

In this paper, we propose to extract the photometric and geometric consistency information from the ground truth (GT) depth and camera parameters, rather than using off-the-shelf keypoint extractors. We uniformly sample the 2D locations of textured regions of the reference image, use camera parameters and GT depth to inverse-project the 2D locations to 3D space, and then project the 3D points on the source image. We apply geometric constraints to reject occluded points and pixel neighborhood similarity to avoid textureless regions. The remaining 2D locations, not only represent 2D-2D correspondence (photometric consistency) between the reference and source images, but also follow geometric consistency (epipolar geometry) due to the use of depth and camera information during the projection process. The extracted 2D-2D correspondence information is then used to train a Siamese network in such a way that the network is forced to generate photo-geometric consistent feature maps at the final convolutional layer. The proposed network called RTNet (rotation and translation estimation network) is jointly trained with pose regression (PR) loss to finally infer pose. We observed that, when trained in this manner, the proposed RTNet outperforms state-of-the-art (SOTA) relative pose estimation (RPE) methods on DTU and 7Scenes datasets. Moreover, under the cross-database scenario, RTNet evenly outperforms SOTA relative and absolute pose estimation (APE) methods

We uniformly sample the 2D locations of textured regions of the reference image, use camera parameters and GT depth to inverse-project the 2D locations to 3D space, and then project the 3D points on the source image. We apply geometric constraints to reject occluded points and pixel neighborhood similarity to avoid textureless regions. The remaining 2D locations, not only represent 2D-2D correspondence (photometric consistency) between the reference and source images, but also follow geometric consistency (epipolar geometry) due to the use of depth and camera information during the projection process. The extracted 2D-2D correspondence information is then used to train a Siamese network in such a way that the network is forced to generate photo geometric consistent feature maps at the final convolutional layer. The proposed network called RTNet (rotation and translation estimation network) is jointly trained with pose regression (PR) loss to finally infer pose. We observed that, when trained in this manner, the proposed RTNet outperforms state-of-the-art (SOTA) relative pose estimation (RPE) methods on DTU and 7Scenes datasets. Moreover, under the crossdatabase scenario, RTNet evenly outperforms SOTA relative and absolute pose estimation (APE) methods.

Our work is inspired by [21] in the context of end-to-end RPE, as well as CNN-based feature extraction and correspondence finding methods such as SuperPoint, SuperGlue [10], [35]. Our work is also influenced by the idea of using loss map from the deep layers as [16]. Compared to these prior works, our approach makes the following contributions:

Most prior CNNs for PE are trained using independent single images, labeled with their corresponding absolute camera pose. Hence, their performance is closely bounded within the parameters of the dataset used for training. In RTNet, we leverage the geometric as well as photometric constraints between pairs of images to predict relative pose, independent of the global pose information. This not only improves the performance among other RPE methods, but also show better generalization ability under cross-database evaluation.

Unlike SuperPoint, SuperGlue and HF-Net [34], RTNet does not require pre-training the feature descriptor network with a separate feature point correspondence dataset.

Unlike SuperGlue and HF-Net, that are multi-stage (hierarchical) PE methods, RTNet is an end-to-end trainable RPE method. Furthermore, in contrast with SuperPoint, SuperGlue, HF-Net, and LF-Net [29], where main focus of these methods is to generate discriminative keypoint descriptors for feature matching purpose, the sole purpose of using correspondences information by RTNet is to leverage photo-geometric consistency on PE.

Unlike LF-Net that updates the two branches of the network in iterative manner, RTNet updates the entire network at once durnig backpropagation.

Unlike other CNN-based RPE methods, RTNet utilizes 2D-to-2D correspondence loss along with the PR loss to optimize the prediction network. These 2D-to-2D correspondences are generated via epipolar geometry, and thus, are not only photometrically consists, but also geometrically consistent.

Unlike traditional approaches, at the test time, RTNet directly predicts the relative camera pose from input images without any need to generate 2D-to-2D correspondences.

Similar to [16], we compute the loss from the deep layers as it contains much more information than a simple 2D-3D correspondence, and thus alleviating the need for choosing a robust loss and its hyperparameters. However, [16] simply add up the dense layers to compute loss, while RTNet minimizes the Euclidean distance between the 2D-2D correspondences on the feature map to leverage photometric consistency.

We confine RTNet to use all the GT information which is readily available for the multi-view stereo-based depth estimation networks, such as RMVSNet [42]. Consequently, RTNet can directly be plugged into RMVSNet pipeline for the homographic warping process.

The proposed method shows SOTA results on Microsoft 7Scenes dataset [17] and DTU Robot Image dataset [1] in comparison with the previous RPE methods.

II. RELATED WORK

Traditionally, CPE has been approached by computing the pose from 2D-3D correspondences between 2D pixels in the query image and 3D points in the scene model [6], which are determined through handcrafted feature descriptor matching [36], [37]. This assumes that the scene is represented by a 3D structure-from-motion (SfM) model. The full 6 degree-of-freedom (DoF) pose of a query image can be estimated very precisely. However, these methods require a large database of features and efficient retrieval methods. Their performance is affected by the changing environmental conditions. Furthermore, they are computationally expensive and often do not scale well.

The evolution of CPE methods using DL techniques is not very old. The first research that successfully implemented an end-to-end trainable CNN-based 6DoF CPE was PoseNet [19] in 2015. Published in 2015, PoseNet utilized GoogLeNet [14] pretrained on ImageNet [33] as base network for feature extraction, to leverage deep features learnt on image classfication task to handle the complicated out-of-plane regression problem using transfer learning. However, PoseNet over-fitted its training data and failed to generalize on new scenes. Therefore, to incoperate scene geometry, Kendall et al. proposed a novel loss function based on scene reprojection error [18] and showed its efficiency in appearance-based localizations. Following PoseNet, several researchers also proposed modifications to the PoseNet architecture, with major focus being the feature extractor and the regressor part. Melekhov et al. [24] replaced the feature extractor with an hourglass style encoder decoder architecture with a base network consisting of ResNet34 layers. SVS-Pose [27], suggested a VGG16 feature extractor. BranchNet [41] reduced the number of convolutional layers from the GoogLeNet backbone in PoseNet and proposed a two separate FC layer branches for regressing translation and rotation independently.

In contrast with end-to-end DL based methods, local learning methods focus on local and related problems by imposing a less tight coupling between the input images and the output poses. This way, the local learning models are combined with structure-based pipelines in a more generalized way. However, these approaches b29,b42 assumes stricter initialization settings and sometimes contain fine-tuning steps [10], [35] with separate training datasets.

Nevertheless, the idea that DL methods can be used for regressing the pose with an end-to-end learning, without the need to manually engineer the features, gave rise to more research in this field. Additionally, the DL methods offered robustness against lighting conditions and viewpoint changes, constant runtime at inference and low memory footprint. MapNet [5] used relative pose loss along with the APE loss, as well as an additional GPS-based localization data to constraint the loss. VLocNet [39], on the other hand, proposed to jointly learn absolute and relative pose using three separate network branches, one for APE and the other two for RPE. To further extract scene information, VLocNet++ [31] added semantic segmentation as another auxilary learning appoarch to the exhisting VLocNet architecutre. However, this requires additional segmentation labels of the scenes, as well as more memory requirements to accomodate another 5 network for semantic segmenation. Glocker et al. [16] presented Neural Reprojection Error (NRE) as a subtitute to the reprojection error used by previous APE methods [18], where the NRE is computed from dense loss map. However, with a theoretical model for APE algorithms, Valada et al. [38] concluded that the APE techniques are not guaranteed to generalize from the training data in practical scenarios. Sattler et al. also commented that the APE is more closely related to image retrieval approaches than to methods that accurately estimate camera pose via 3D geometry.

In contrast with the APE, where the underlying models establish a one-to-one relationship between images and their absolute pose with respect to a global reference, RPE tries to find relative pose between two images by modeling the relationship between the visual features on the two images. In particular, RPE provides means for relation and representation learning for previously unseen scenes and objects [25]. However, the RPE methods differ from each other based on the underlying task, broadly classified into three categories, i) image retreval (IR), ii) visual odometry (VO), and iii) pose regression (PR). IR-based PE methods decouple feature learning and PE. Given a query image, RelPoseNet [21] estimates its pose to be the pose of its nearest neighbour (NN) in a reference dataset utilizing the visual similarity between the query and the reference images. RelPoseNet trains a relative pose regressor with pair of images, and at test time, uses the deep features of the network to calculate the NN in the feature space. Similarly, RelocNet [2] retrieves a relevant image from a database, which presents high camera frustum overlap with an unseen query. Subsequently, RelocNet uses the pose of the images stored in the database, to compute the pose of the query by applying a transformation produced by CNN that is trained with a camera frustum overlap loss.

VO-based RPE methods assume that the given pair of images are sequential images. Consequently, VO-based RPE methods estimate the incremental motion between the pair of images. Zheng et al. [44] proposed a CNN-Recurrent Neural Network (RNN) architecture called DeepVO, in which the features from CNN are fed to an RNN for learning the dynamics and relations among the sequence of images. The final pose between the pair of images is regressed with a Mean Square Error (MSE) loss layer. VidLoc [9] exploited the temporal smoothness constraint between short sequences of consecutive frames by training a RNN for video-clip localization. Laskar et al. [20] proposed an end-to-end architecture for learning ego-motion from a sequence of RGB-D images using a prior set of discretized velocities and directions. However, these methods are applicable to image sequences.

In comparison to the VO-based and IR-based RPE methods, the relative PRbased methods assume that the scene graph information is given, and thus the relative pose can be regressed directly from the GT pose. Moulon et al. [25] presented a Siamese architecture with spatial pyramid pooling layer to regress relative pose, where the pairwise scene overlap information is estimated from an open source tool [26], demonstrated improved performance compared to the local feature-based approaches that utilizes SIFT and ORB features. Szegedy et al. [13] replaced the base network with a GoogLeNet, and experimented with different 6 combinations at the FC layer and loss layers. Rather than directly regressing pose, a recent RPE method DirectionNet [7] estimates a discrete distribution over keypoint locations by factorizing poses as a set of 3D direction vectors. DirectionNet shows a near 50% reduction in error over direct regression methods on synthetically generate dataset. However, some researchers [18], [23] argued that the 3D pose is continous and must be solved in a regression framework instead.

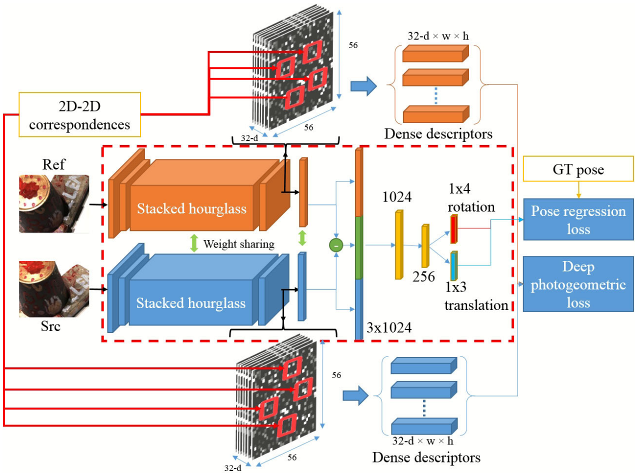

FIGURE 2. Overall architecture of the proposed RTNet using deep photo-geometric and pose regression loss.

The proposed method partly adopts features from many of these previously discussed techniques. However, contrary to the previous end-to-end trainable APE and RPE methods that trains individual models with PR loss, reprojection loss, or location-aware loss, our method is the first to jointly train an end-to-end CNN architecture that simultaneously regresses the 6-DoF relative pose by imposing photometric and geometric consistency using 2D-2D correspondences that follows epipolar geometry. By jointly learning both tasks, our approach is robust to unseen environments, thereby combining the advantages of both local feature and DL-based localization methods. The presented method does not require separate handcrafted feature extractor to generate 2D-2D correspondences. Rather, it uniformly samples 2D-2D correspondences from the GT depth and camera parameters, which is discussed in the following sections.

III. PROPOSED APPROACH

The overall architecture of the proposed RPE method is shown in Fig. 2. The inputs to RTNet are a pair of images (a reference and a source), GT relative pose, and the 2D-2D correspondences. The pair of images are passed to the Siamese network and their deep features are extracted. We used stacked hourglass [28] structure for feature extraction with ResNet as the base network. As shown in Fig. 2, the deep descriptors of the 2D-2D correspondence pairs are taken from the final convolutional layer, thus we name them as 2D-2D deep correspondences. Given correspondences, the deep photo-geometric loss layer minimizes the L2-distance between these corresponding deep features. Each feature descriptor have a size of 32 × w × h, where w × h represent the spatial window around the corresponding points. We experimented with different sizes of w × h and reported the results in the results section. The deep features from the two branches are then flattened, and concatenated. Inspired by the research on object detection [32], where feature sharing between proposal generation

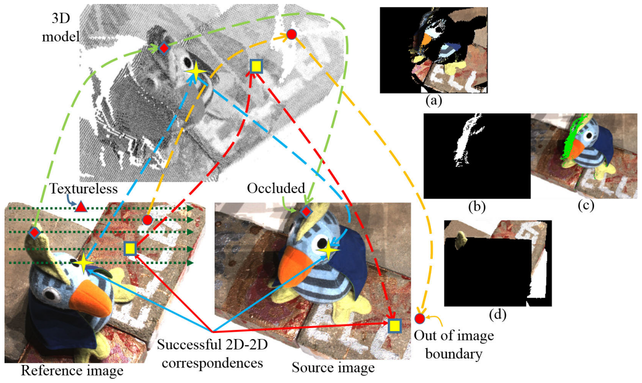

FIGURE 3. **Process of finding 2D-2D correspondences between a reference and a source image. (a) Shows the pixels that can be successfully projected on src. image from ref., and thus their 2D locations are used as correspondences, (b) and (c) represent the regions where ref. image pixels are occluded, and (d) shows the pixels of ref. image that are mapped outside the boundaries of src. image. (Please note that (a)-(d) are only for concept visualization. We do not map pixels. Rather, we only make use of their location).**

part and object detection part has shown both speed and performance gains, we also share the deep features between deep correspondence and the PR part of the RTNet. In the following subsections, we describe the deep photo-geometric loss, the correspondences generation method and the PR part in more detail.

A. CORRESPONDENCE GENERATION AND DEEP PHOTO-GEOMETRIC LOSS

Given GT pose, scene graph, 3D model, and intrinsics, we generate the 2D-2D correspondence pairs. Figure 3 shows this process, where the points are projected from reference image to source image using the 3D model. We assume that for RPE purpose, as far as the two pixels on the pair of images follow a 2D-2D correspondence, they need not to be local keypoints. Therefore, we sample uniform 2D-2D correspondence pairs from the pair of images. First, the reference image is scanned row-wise (dark green arrows on reference image in Fig. 3), and uniform points are sampled. Second, we apply the texturness constraint on the pixels indexed by the sampled 2D points, and reject the 2D points that falls on textureless regions (such as the point represented by red triangle in Fig. 3). Third, remaining 2D points are inverse projected to the 3D model using 2D-3D projection, and then projected to the source image using 3D-2D mapping (represented by large dashed arrows in Fig. 3). It is worth mentioning that due to these 2D-3D and 3D-2D projections, the candidate 2D points strictly follow the scene geometry, consequently contributing to the learning process with geometric information. Fourth, 2D points that are occluded in source view (such as the point represented by red diamond in Fig. 3) or mapped outside the source image boundaries (point represented by the red circle in Fig. 3) are also rejected. Remaining 2D points that are successfully mapped from reference image to source image are saved as the 2D-2D corre- sponding pairs for that particular image pair. Example of 2D-2D corresponding pairs are illustrated with the yellow star and the yellow square in Fig. 3. Although, the 2D-2D correspondences correspond to photometric consistency of the input RGB images, we scale the correspondences on the respective deep feature maps. Consequently, we call these correspondences as 2D-2D deep correspondences. Moreover, it can be observed that these 2D-2D corresponding pairs not only follow the geometric constraints, but also follow the photometric constraints as they represent the same regions on the two views. Thus, we name the loss between these deep correspondences as deep photo-geometric loss. As shown in Fig. 2, we minimize the L2-distance between the deep features from the corresponding 2D-2D locations of the final convolution layers of the Siamese network. Let $(x_i^{ref}, y_i^{ref}) \in \{(x_1^{ref}, y_1^{ref}), (x_2^{ref}, y_2^{ref}), \ldots, (x_n^{ref}, y_n^{ref})\}$ denote the n uniformly sampled candidate 2D points on reference image I_{ref}, then the deep photo-geometric consistency loss is given by:

$$L_{PG} = \sum_{i=1}^{n} \left\| F(x_i^{ref}, y_i^{ref}) - F(p(x_i^{ref}, y_i^{ref})) \right\| \tag{1}$$

where F is a 32×w×h deep feature map from the last convolutional layer of the Siamese network and p represents the projection function that projects 2D points from reference to the source.

1) AVOIDING TEXTURELESS SURFACE

The proposed method finds corresponding pairs from the two images based on epipolar geometry, and thus, does not force the corresponding pairs to be keypoints. This relieves the need to engineer feature descriptors or use a separate fine-tuning stage with an annotated dataset for feature extractor as in [10] and [35]. However, the proposed method can end up minimizing the distance between the corresponding pairs that are sampled from textureless regions (non-Lambertian surfaces). Thus, it is vital to reject correspondences from textureless regions. We adopt a fast and simple, yet effective approach to filter out textureless regions. We compare the intensity of the pixel indexed by the candidate 2D point with its immediate 8 neighbors, and if the intensities of the 5 out of 8 neighbors differs above a specified threshold with that of the candidate 2D point, then the candidate 2D point is added to the pool of corresponding pairs. Mathematically denoted as:

$$T(I_{ref}(x_i^{ref}, y_i^{ref})) = \sum_{m,n} s(I_{ref}(m,n) - I_{ref}(x_i^{ref}, y_i^{ref})) \geqq 5,$$
$$s(x) = \begin{cases} 1 & x \geqq threshold_{intensity} \\ 0 & otherwise \end{cases} \tag{2}$$

where m, n represent the indexes of the neighboring pixels, T represent the texturness of the pixel indexed by (x_i^{ref}, y_i^{ref}) on reference image I_{ref}. The red triangle in Fig. 3 represents the textureless surface, and thus, it is rejected during the sampling of corresponding pairs.

2) OCCLUSION AND OUT-OF-IMAGE-BOUNDARY HANDLING

Handcrafted features simply fail to find corresponding keypoints in occluded regions. The proposed method rejects candidate 2D correspondence point on reference image if it is occluded in source image. We define an occluded region as a region on the source image, such that more than one points from reference image map to that region. Two 2D points (x_i^{ref}, y_i^{ref}) and (x_j^{ref}, y_j^{ref}) on reference image are treated as occluded in source image if:

$$p(x_i^{ref}, y_i^{ref}) = p(x_j^{ref}, y_j^{ref}) \tag{3}$$

here p defines the projection from reference to the source image. In the case of occlusion, both points are rejected for simplicity. Similarly, candidate 2D points that are projected outside the boundaries of the source image, are also rejected. The red circle in Fig. 3 represents the 2D point that is mapped outside the source image boundary, while the red diamond represents the 2D point that is occluded in the source image, and thus, they are both rejected during the 2D-2D correspondence sampling process. Although, the correspondences can be generated on the fly during training, for speed gains we generated the 2D-2D correspondences offline and loaded them during the training time.

B. RELATIVE POSE REGRESSOR

The pose regressor network which consists of a series of fully-connected layers, is fed with the deep photo-geometric consistent features from the Siamese network. Following [18], we use the 1×4 unit quaternion vector representation for the rotation ΔR and 1×3 vector representation for the translation Δt. The PR loss is given by:

$$L_{PR} = \left\| \Delta\hat{t} - \Delta t \right\| + \beta \left\| \Delta\hat{R} - \Delta R \right\| \tag{4}$$

where β represents the balancing parameter between rotation and translation estimations. The behavior of the β on training loss led us to a few interesting observations, as described below.

1) PR LOSS ONLY

When training the network with only PR loss, the prediction accuracies show high dependency on the value of β. In general, a proportionally higher value of β as the scale of translation values shows better results. This requires recomputing the value of β every time for an unseen dataset. To remove the β dependency, we found that convergence can be achieved by training the network with a slightly high value of β (100-200) first, and then fine-tune the network with mean absolute percentage error (MAPE). Equation 4 can be converted to MAPE ($L_{\%}$) as:

$$L_{\%} = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{\left\|\hat{R}^i_3 - R^i_3\right\|}{\hat{R^i_3}+c} + \frac{\left\|\hat{R}^i_2 - R^i_2\right\|}{\hat{R^i_2}+c} + \frac{\left\|\hat{R}^i_1 - R^i_1\right\|}{\hat{R^i_1}+c} + \frac{\left\|\hat{R}^i_0 - R^i_0\right\|}{\hat{R^i_0}+c} + \frac{\left\|\hat{t}^i_2 - t^i_2\right\|}{\hat{t^i_2}+c} + \frac{\left\|\hat{t}^i_1 - t^i_1\right\|}{\hat{t^i_1}+c} + \frac{\left\|\hat{t^i_0} - t^i_0\right\|}{\hat{t^i_0}+c}\Big) \tag{5}$$

where N is the batch size and $c = 0.0001$ is a non-zero constant for numerical stability. Note that the asymmetric nature of the MAPE positively affects the learning process in the presence of +ve and −ve GT values.

2) JOINT OPTIMIZATION OF DEEP PHOTO-GEOMETRIC AND PR LOSS

However, when training RTNet by jointly optimizing the deep photo-geometric error and PR error as given in Eq. 6, we found that the value of $\beta = 1$ in Eq.4 works well for all the three datasets, and thus can be removed.

$$L_{joint} = L_{PG} + L_{PR} \tag{6}$$

Please note that RTNet is trained with L_{joint}. $L_{\%}$ was only used to study the affects of β when training the network without L_{PG}. Unlike the reprojection error [18], the photo-geometric error did not exhibit divergence in training when used as an additional loss along with PR loss.

IV. EXPERIMENTS

A. DATASETS

We evaluated the proposed RTNet on three datasets and compared it with SOTA methods on prediction accuracy and generalization capability against unknown scenes.

1) DTU DATASET [1]

The DTU dataset is captured using a calibrated camera, mounted on a robotic arm. The camera positions and light positions are controlled via a computer program, to collect large amounts of high-quality image data with different lighting conditions. A structured light scanner is also used to capture the 3D surface geometry of the viewed object. DTU dataset is helpful when evaluating image matching algorithms as the image correspondences can be determined from the known camera and scene geometry. The dataset includes 124 scenes containing different number of camera positions.

2) 7SCENES DATASET [17]

Captured in an indoor office environment, this dataset is comprised of RGB-D images collected from seven different scenes (rooms), where each scene consists of multiple sequences. The images are captured with a handheld Kinect RGB-D camera (640 × 480 resolution), and the GT poses are extracted using KinectFusion. The dataset is difficult for relocalization and tracking tasks, due to the different camera motions in the presence of motion blur, perceptual aliasing, and textureless features in the room.

3) UNIVERSITY DATASET [21]

Similar to the 7Scenes dataset, University dataset contains 5 similar indoor scenes. However, in contrast to 7Scenes, all the scenes in the University dataset are registered to a common global coordinate frame, eliminating the need to train and test the models scene-wise. The dataset contains 9,694 training and 5,068 test images. However, we used all

TABLE 1. Performance comparison of RPE methods on DTU dataset. (translation (m), rotation (^{o})).

Method	Median ERROR
spp-net [25]	0.2026m, 21.27^o
RPNet [13]	0.1251m, 13.62^o
RelPoseNet [21]	0.3700m, 11.35^o
Ours(Squeeznet)(w=1,h=1)	0.0347m, 2.81^o
Ours(Res32)(w=1,h=1)	0.0027m, 0.21^o
Ours(Res50)(w=1,h=1)	0.0013m, 0.15^o
Ours+DPE(Res50)(w=1,h=1)	**0.0010m, 0.08^o**
Ours+DPE(Res50)(w=3,h=3)	0.0046m, 0.41^o

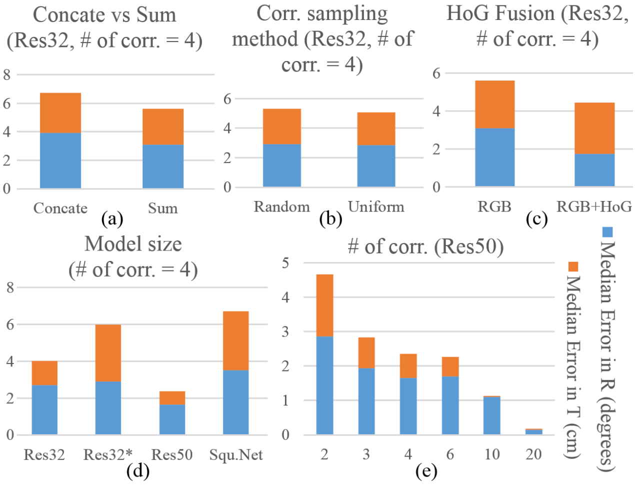

FIGURE 4. Effect of different parameters on the PE.

the images for training and evaluated the models on DTU and 7Scenes datasets.

B. EVALUATIONS

In this section, we quantitatively demonstrate the performance of the proposed method on the DTU and 7Scenes datasets, as well as compare the cross-database performance on these datasets using the University Dataset.

1) EVALUATION ON DTU DATASET

To the best of our knowledge, [25] used the DTU dataset for the first time for comparing the performance of CNN-based PE method and handcrafted feature-based method. We chose DTU dataset as the main dataset for evaluating the proposed RTNet, as the DTU dataset provides accurate and reliable orientation and translation information in comparison with 7Scenes or University dataset, where the GT poses are dependent upon the success of underlying PE methods. We compare our results with the current SOTA RPE methods in Table 1. For performance comparison, we used the metrics in [18]: Positional error (m) and angular error (degrees). The proposed method is evaluated with three different base networks, i.e., SqueezeNet, Res50 and Res32. Additionally, we found that preprocessing the images with deep photo enhancer (DPE) [8] results in slightly smaller PE errors as compared with using RGB images. Moreover, we found that increasing the spatial window (w × h) around the 2D correspondences increases the PE errors. This may be caused by the fact that the neighboring deep features around the 2D correspondence pairs, which represent the compressed representation of the input image, are not guaranteed to follow the three constraints defined in section III-A.

In Fig.4 we present the effect of varying different parameters of the RTNet structures on its performance. Results show that using sum instead of concatenate (Fig.4(a)) and Res50 instead of Res32 or SqueezeNet (Fig.4(d)) resulted in reduced PE errors. Res32* in Fig.4(d) denotes a modified version of RTNet, where we increased the depth of PR part of the RTNet by adding 3 conv. layers for successive feature dimension reduction, however, that negatively affected the performance. Sampling policy shows a very slight effect on the performance (Fig.4(b)), however, using too fewer correspondences leads to bigger PE errors (Fig.4(e)). Finally, Fig.4(c) shows the comparison between using RGB images vs using RGB+Histogram of Gradients (HoG) features. We found a small reduction in rotation error but increased translation error when fusing HoG features with the RGB images.

2) EVALUATION ON 7SCENES DATASET

The 7Scenes dataset does not provide calibrated camera parameters and thus it is impossible to find 2D-2D correspondences using epipolar geometry. In order to evaluate RTNet on the 7Scenes dataset, we confined to use keypoint correspondence information from the VisualSfM pipeline. Please note that the use of off-the-shelf keypoint information when experimenting with the 7Scenes dataset is due to the unavailability of the calibrated camera parameters. The experimental results are compared with the SOTA 6-DoF CPE methods in Table 2. The proposed RTNet (bold values) not only outperforms SOTA RPE methods (RelPoseNet and RelocNet), but also outperform APE methods, except the VLocNet++ (underline values). The VLocNet++ requires extra segmentation labels to train the network. It is also worth mentioning that the RPE methods in Table 2 are evaluated on a 5× bigger test set as compared with the test set used by APE schemes. Following the convention in [21], we pair every reference image in the test set with top-5 nearest neighboring source images and predict 5 poses. In contrast, APE methods estimate only one pose for each test image.

There is an interesting observation from Table 3. When trained on DTU dataset, the proposed RTNet shows a very small error on the 7Scenes dataset (2nd col., last row), but when the roles of these two datasets are switched, the performance drops very much (3rd col., last row). We believe, that this is due to the reason that when training RTNet with the DTU dataset, we use accurate GT

TABLE 2. Performance comparison on 7Scenes dataset with representative end-to-end camera pose estimation methods.

Method	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs
PoseNet	0.32m, 8.12^o	0.47m, 14.40^o	0.29m, 12.00^o	0.48m, 7.68^o	0.47m, 8.42^o	0.59m, 8.64^o	0.47m, 13.80^o
Hourglass-Pose	0.15m, 6.53^o	0.27m, 10.84^o	0.19m, 11.63^o	0.21m, 8.48^o	0.25m, 7.01^o	0.27m, 10.84^o	0.29m, 12.46^o
BranchNet	0.18m, 5.17^o	0.34m, 8.99^o	0.20m, 14.15^o	0.30m, 7.05^o	0.27m, 5.10^o	0.33m, 7.40^o	0.38m, 10.26^o
PoseNet+	0.13m, 4.48^o	0.27m, 11.30^o	0.17m, 13.00^o	0.19m, 5.55^o	0.26m, 4.75^o	0.23m, 5.35^o	0.35m, 12.40^o
MapNet	0.09m, 3.24^o	0.20m, 9.29^o	0.12m, 8.45^o	0.19m, 5.45^o	0.19m, 3.96^o	0.20m, 4.94^o	0.27m, 10.57^o
VLocNet++	0.023m, 1.44^o	0.018m, 1.39^o	0.016m, 0.99^o	0.024m, 1.14^o	0.024m, 1.45^o	0.025m, 2.27^o	0.021m, 1.08^o
RelPoseNet∗	0.13m, 6.46^o	0.26m, 12.72^o	0.14m, 12.34^o	0.21m, 7.35^o	0.24m, 6.35^o	0.24m, 8.03^o	0.27m, 11.80^o
RelocNet∗	0.12m, 4.14^o	0.26m, 10.40^o	0.14m, 10.50^o	0.18m, 5.32^o	0.26m, 4.17^o	0.23m, 5.08	0.28m, 7.53^o
Ours∗	**0.047m, 1.82^o**	**0.0485m, 1.83^o**	**0.026m, 1.65^o**	**0.0513m, 1.81^o**	**0.038m, 1.41^o**	**0.0464m, 1.47^o**	**0.0356m, 0.86^o**

∗ denotes schemes that compute relative camera pose. All other methods are APE methods. Underline → best among all, **bold** → best among RPE methods

TABLE 3. Performance comparison under cross-database settings.

Method	Train DTU Test 7Scenes	Train 7Scenes Test DTU	Train Univ. Test DTU	Train Univ. Test 7Scenes
HSC-Net[21]∗	0.39m, 18.77^o	0.23m, **20.23^o**	0.60m, 21.25^o	0.39m, 18.96^o
Pose-Net[22]	9.99m, 122.06^o	6.26m, 126.65^o	6.30m, 129.00^o	3.48m, 111.27^o
Ours∗	**0.03m, 5.66^o**	**0.18m**, 20.27^o	**0.19m, 21.11^o**	**0.03m, 1.36^o**

∗ denotes schemes that compute relative camera pose

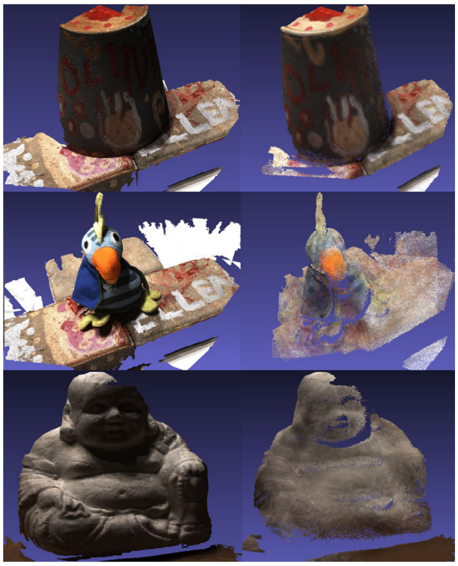

FIGURE 5. Comparison of 3D reconstruction results using GT (left) and predicted pose (right).

information to leverage photo-geometric consistency, which consequently shows good performance on 7Scenes. However, the correspondence information for 7Scenes is computed from the SFM pipeline (using SIFT), which is less accurate than the GT correspondences, resulting in relatively poor RPE performance. The same behavior is observed when we used University dataset as the training set and the DTU dataset as the test set. In fact, the errors in the 3rd and 4th column seem very similar due to the fact that the nature of the two datasets, i.e., capturing method, camera movement, and capturing environment, is very similar. This is further clarified by the small errors in the 5th column. Comparing the cross-database performance in Table 3 and the performance given in Table 2, it can be observed that the proposed method outperforms the SOTA RPE, and even outperforms most of the APE methods. This proves the claim that using geometric consistent 2D-2D correspondence information can help learning more generalized features that can be applied to unseen scenarios.

3) 3D RECONSTRUCTION WITH GROUND TRUTH DEPTH

In Fig. 5 we present 3D reconstruction results on some of the test scans of the DTU dataset, using the predicted pose from the RTNet. We used Fusibile [14] for the 3D reconstruction and GT depths. Even though, RTNet exhibits less completeness as compared with the GT results, it still shows promising results towards 3D reconstruction using deep learning based CPE.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced a deep photo-geometric loss to overcome the lack of scene generalization problem posed to the absolute camera pose estimation methods. The basis of photo-geometric consistency is laid upon 2D-2D photometric correspondences computed using epipolar geometry between pairs of images. We presented a CNN-based end-to-end relative camera pose estimation architecture RTNet, that jointly optimizes the proposed photo-geometric loss and the pose regression loss. In this way, the proposed method not only leverages the photometric consistency but also forces the geometric consistency on the network. With experiments on three different public datasets, we showed that the photo-geometric loss helps the deep network to learn more generalized features, consequently outperforming the state-of-the-art pose estimation methods on previously unseen scenes. Currently, the proposed method is tightly coupled with finding the accurate 2D-2D correspondences from camera parameters and depth at training time. We plan to invest future efforts towards relaxing the need to find accurate correspondences by utilizing the generalization power of the network. Currently, the proposed method is tightly coupled with finding the accurate 2D-2D correspondences from camera parameters and depth at training time. However, the camera pose estimation method using deep learning shows limited performance in terms of accuracy compared to traditional methods [11]. Therefore, we plan to invest future efforts towards relaxing the need to find accurate correspondences and pose information, by utilizing the generalization power of the network.

REFERENCES

[1] H. Aans, K. S. Pedersen, and A. L. Dahl, "On recall rate of interest point detectors," in *Proc. 3DPVT*, May 2010.

[2] V. Balntas, S. Li, and V. A. Prisacariu, "RelocNet: Continuous metric learning relocalisation using neural nets," in *Proc. ECCV*, Sep. 2018, pp. 751–767.

[3] A. J. Ben Ali, M. Kouroshli, S. Semenova, Z. S. Hashemifar, S. Y. Ko, and K. Dantu, "Edge-SLAM: Edge-assisted visual simultaneous localization and mapping," *ACM Trans. Embedded Comput. Syst.*, vol. 22, no. 1, pp. 1–31, Jan. 2023.

[4] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "DSAC—Differentiable RANSAC for camera localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2492–2500.

[5] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2616–2625.

[6] F. Camposeco, A. Cohen, M. Pollefeys, and T. Sattler, "Hybrid camera pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 136–144.

[7] K. Chen, N. Snavely, and A. Makadia, "Wide-baseline relative camera pose estimation with directional learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3257–3267.

[8] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6306–6314.

[9] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "VidLoc: A deep spatio-temporal model for 6-DoF video-clip relocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2652–2660.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1–10.

[11] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, (CVPR)*, Jun. 2019, pp. 8084–8093.

[12] S. En, A. Lechervy, and F. E. Jurie, "RPNet: An end-to-end network for relative camera pose estimation," in *Proc. ECCV Workshops*, 2018.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[14] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.

[15] H. Germain, V. Lepetit, and G. Bourmaud, "Neural reprojection error: Merging feature learning and camera pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 414–423.

[16] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time RGB-D camera relocalization," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2013, pp. 173–179.

[17] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6555–6564.

[18] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. ICCV*, 2015, pp. 2938–2946.

[19] K. Konda and R. Memisevic, "Learning visual odometry with a convolutional network," in *Proc. 10th Int. Conf. Comput. Vis. Theory Appl.*, 2015, pp. 486–490.

[20] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 920–929.

[21] X. Li, S. Wang, Y. Zhao, J. Verbeek, and J. Kannala, "Hierarchical scene coordinate classification and regression for visual localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11980–11989.

[22] S. Mahendran, H. Ali, and R. Vidal, "3D pose regression using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 494–495.

[23] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Image-based localization using hourglass networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 870–877.

[24] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Relative camera pose estimation using convolutional neural networks," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.*, 2017, pp. 675–687.

[25] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open multiple view geometry," in *Proc. Int. Workshop Reproducible Res. Pattern Recognit.* Cham, Switzerland: Springer, 2016, pp. 60–74.

[26] T. Naseer and W. Burgard, "Deep regression for monocular camera-based 6-DoF global localization in outdoor environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 1525–1530.

[27] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.

[28] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "LF-Net: Learning local features from images," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA, Curran Associates, 2018, pp. 6237–6247.

[29] T. Probst, D. P. Paudel, A. Chhatkuli, and L. Van Gool, "Unsupervised learning of consensus maximization for 3D vision problems," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 929–938.

[30] N. Radwan, A. Valada, and W. Burgard, "VLocNet++: Deep multitask learning for semantic visual localization and odometry," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4407–4414, Oct. 2018.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1137–1149.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," in *Proc. IJCV*, 2015, pp. 211–252.

[33] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12708–12717.

[34] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4937–4946.

[35] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search," in *Proc. ECCV*. Cham, Switzerland: Springer, 2012, pp. 752–765.

[36] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1744–1756, Sep. 2017.

[37] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixé, "Understanding the limitations of CNN-based absolute camera pose regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3297–3307.

[38] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 6939–6946.

[39] X. Wei, Y. Zhang, Z. Li, Y. Fu, and X. Xue, "DeepSFM: Structure from motion via deep bundle adjustment," in *Proc. ECCV*, 2020, pp. 230–247.

[40] J. Wu, L. Ma, and X. Hu, "Delving deeper into convolutional neural networks for camera relocalization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 5644–5651.

[41] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5520–5529.

[42] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2666–2674.

[43] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 340–349.

[44] Y. Zheng, Y. Kuang, S. Sugimoto, K. Åström, and M. Okutomi, "Revisiting the PnP problem: A fast, general and optimal solution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2344–2351.

YONGJU CHO received the B.S. and M.S. degrees in electrical and computer engineering from Iowa State University, in 1997 and 1999, respectively, and the Ph.D. degree in electrical and computer engineering from Michigan State University, in 2009. He joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, in 2001. He has been involved in the development of a data broadcasting systems, multimedia networking, and immersive media. His research interests include digital signal processing, adaptive wireless video communications, and computer vision.

SEUNGHO EUM received the bachelor's degree in computer science and engineering from Kookmin University, in 2021. He is currently pursuing the master's degree in computer vision with the Graduate School of Computer Science and Engineering, Sogang University.

JUNESEOK IM received the bachelor's degree in computer engineering from Seokyeong University, in 2021. He is currently pursuing the master's degree in computer vision with the Graduate School of Computer Science and Engineering, Sogang University. He received the AI Grand Challenge Hosted from the Korea Institute of Science and Technology Information and Communications.

ZAHID ALI received the B.E. degree in electronic engineering from the NED University of Engineering and Technology, in 2009, the M.E. degree in computer engineering from Chosun University, in 2013, and the Ph.D. degree in computer from Sogang University, in 2020. His research interest includes the intersection of computer vision and deep learning. He received the Global IT Scholarship from NIIED, South Korea, in 2013.

HYON-GON CHOO received the B.S. and M.S. degrees in electronic engineering, in 1998 and 2000, respectively, and the Ph.D. degree in electronic communication engineering from Hanyang University, South Korea, in 2005. He is currently a Principal Researcher with the Electronics and Telecommunications Research Institute, Daejeon, South Korea. He is the Director of the Immersive Media Research Section. His research interests include video coding for machines, holography, multimedia protection, and 3D broadcasting technologies.

UNSANG PARK received the B.S. and M.S. degrees in materials science from Hanyang University, Seoul, South Korea, in 1998 and 2000, respectively, and the M.S. and Ph.D. degrees in computer science from Michigan State University, MI, USA, in 2004 and 2009, respectively. He is currently a Professor with the Department of Computer Science and Engineering, Sogang University. His research interests include pattern recognition, image processing, computer vision, and machine learning.