Learning to Reduce Scale Differences for Large-Scale Invariant Image Matching

Yujie Fu\textsuperscript{1}, Graduate Student Member, IEEE, Pengju Zhang, Bingxi Liu, Zheng Rong\textsuperscript{1}, and Yihong Wu\textsuperscript{1}

Abstract—Most image matching methods perform poorly when encountering large scale changes in images. To solve this problem, we propose a Scale-Difference-Aware Image Matching method (SDAIM) that reduces image scale differences before local feature extraction, via resizing both images of an image pair according to an estimated scale ratio. In order to accurately estimate the scale ratio for the proposed SDAIM, we propose a Covisibility-Attention-Reinforced Matching module (CVARM) and then design a novel neural network, termed as Scale-Net, based on CVARM. The proposed CVARM can lay more stress on covisible areas within the image pair and suppress the distraction from those areas visible in only one image. Quantitative and qualitative experiments confirm that the proposed Scale-Net has higher scale ratio estimation accuracy and much better generalization ability compared with all the existing scale ratio estimation methods. Further experiments on image matching and relative pose estimation tasks demonstrate that our SDAIM and Scale-Net are able to greatly boost the performance of representative local features and state-of-the-art local feature matching methods.

Index Terms—Image matching, large scale changes, scale difference reduction, scale ratio estimation, covisibility-attention-reinforced matching module.

I. INTRODUCTION

ESTABLISHING pixel-level correspondences between two images is an essential basis for a wide range of computer vision tasks such as visual localization [1], [2], 3D scene reconstruction [3] and simultaneous localization and mapping (SLAM) [4]. Such correspondences are usually estimated by sparse local feature extraction and matching [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. A local feature consists of a keypoint and a descriptor. But the scale invariance of both existing keypoint detectors and descriptors is not enough to deal with large scale changes [17]. Few inlier correspondences can be established by matching local features under the circumstances of large scale changes in images, which is called as the scale problem of local features in this paper. If the scale difference between two images is small, we call that the two images are at related scale levels in scale space [18], [19].

To alleviate the scale problem of local features, the Multi-Scale Feature Extraction method is widely used [5], [10], [12], [20], [21]. Given an image, this method extracts local features from several neighbouring scale levels of original image scale level. It improves the robustness of local features to scale changes. However, if the scale difference is too large, it still only establishes very few inlier correspondences, as depicted in Figure 1(a). The reasons are as follows. As observed by the literature [17], [22], [23], given an inlier correspondence, the two local features linked by this correspondence are probably extracted from two related scale levels. If a correspondence consists of two local features extracted from two unrelated scale levels, this correspondence is probably wrong. Given an image pair, the scale difference of the image pair is unknown. The Multi-Scale Feature Extraction method only samples several scale levels near original image scale levels. Thus, it is not able to ensure that there exist related scale levels between the two images of the image pair. When the scale difference between the two images is not very large, most of the sampled scale levels of one image are related with those of another

Fig. 1. Given a challenging image pair with large scale change, two thousand SIFT keypoints are extracted from every image. Correspondences established by SIFT (a) and our method (b) are displayed. Only those correspondences conforming to the ground truth epipolar geometry are drawn.

1051-8215 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
image so that many inlier correspondences can be established. However, if the scale difference is too large, few or even no sampled scale levels of one image are related with those of another image so that few inlier correspondences will be established.

In a word, the scale problem of local features can be ascribed to too few related scale levels to a large extent. To deal with this problem, we propose a Scale-Difference-Aware Image Matching method (SDAIM), as shown in Figure 2(a). Given an image pair, firstly, its scale difference is estimated. Then both images are resized according to the estimated scale difference so that the scale difference of the image pair is greatly reduced before multi-scale local feature extraction. This method guarantees that most sampled scale levels of two images are related and restricts most local features to be extracted from the related scale levels. As shown in Figure 1(b), after enhanced by our SDAIM, SIFT can establish much more inlier correspondences. In this paper, a scale ratio is used to depict the scale difference (Section III-A). An accurate scale ratio estimation method is needed in our SDAIM.

So far, image scale ratio estimation problem has drawn little attention. Zhou et al. used a Bag-of-Features model to encode the image pyramids of an image pair and exhaustively compared the representation vectors of sampled scale levels to estimate the scale ratio [17]. A box embedding for images is proposed to estimate the scale ratio by Rau et al. [24]. The scale ratio of an image pair depends heavily on the visual overlaps within the image pair. But both the above methods do not consider visual overlap information fully during image encoding processes. To make the most of visual overlap information, we propose a Covisibility-Attention-Reinforced Matching module (CVARM). Given two images and their dense feature maps computed by a convolutional neural network (CNN), it takes dense feature maps of both images as input, exhaustively compares the patches of the image pair and obtains a correlation map. Then it max-pools the correlation map in one branch and max-pools the correlation map along the channel axis in another branch to compute the covisibility attention maps. At last, it multiplies the correlation map by the covisibility attention maps to emphasize covisible areas and suppresses the distraction from non-covisible areas. Best viewed in color with 200% zoom in.

In summary, our contributions are listed as follows:

1) We analyze the reasons for the scale problem of local features. And we propose a Scale-Difference-Aware
Image Matching method, termed as SDAIM, that has an intrinsic invariance to scale changes.

2) We present a novel neural network, termed as Scale-Net, based on our proposed Covisibility-Attention-Reinforced Matching Module (CVARM), which can emphasize the visual overlaps of an image pair and accurately estimate the scale ratio of the image pair. Scale-Net can be integrated into the proposed SDAIM.

3) Extensive experiments confirm that SDAIM can remarkably enhance the performance of local features and local feature matching methods under large image scale changes and that Scale-Net has good generalization ability and high scale ratio estimation accuracy.

II. RELATED WORK

Since the purpose of this paper is to mitigate the scale problem of local features, below we briefly review local feature extraction and local feature matching in the literature. Because image scale ratio estimation is an essential step in our method, we also review some scale ratio estimation methods.

A. Local Feature Extraction

Local feature extraction consists of keypoint detection and feature description. Various handcrafted keypoint detectors have been proposed over the past few decades. FAST [26] and Harris [27] are pioneering corner detectors. But neither of them is invariant to large scale changes. A simple way to cope with scale changes is to extract keypoints at several scales. Mikolajczyk and Schmid used the extrema over scale of the Laplacian-of-Gaussian to select the scale of interest points picked by multi-scale Harris detector [28]. SIFT uses Difference-of-Gaussian operator to select blobs over multiple scale levels [5]. Gao et al. devised two corner detection methods based on a multi-scale Log-Gabor wavelet transform [21]. KAZE applies the Hessian detector to a nonlinear diffusion scale space [29]. In recent years, several learning-based detectors have been proposed [30], [31], [32]. TILDE focuses on devising a keypoint detector which is robust to imaging changes of weather and lighting conditions [30]. Savinov et al. proposed a transformation-invariant keypoint detector that can be trained in an unsupervised manner [31]. Key.Net uses a multi-scale neural network to detect robust features [32]. Given keypoints, high level information can be captured from patches around keypoints by descriptors. Handcrafted descriptors encode image patches based on grayscale and gradient information [5], [33], [34]. Compared with handcrafted descriptors, learning-based descriptors achieve better discriminative ability by using convolutional neural networks to dig out higher level information from image patches [35], [36], [37], [38], [39], [40], [41], [42]. But few of descriptors focus on scale invariance. GIFT uses group convolutions to fuse features extracted from the transformed versions of an image to obtain a descriptor which is robust to scale changes [38]. There are also several detect-and-describe methods [8], [9], [10], [11], [12], [43], [44], [45]. ASLFeat [12], LF-Net [8] and RF-Net [43] adopt different strategies to deal with scale changes. ASLFeat utilizes the inherent pyramidal features of a CNN [12], LF-Net uses a ResNet [46] to generate a feature map from an image and resizes the feature map several times to simulate several scale levels [8]. RF-Net treats different layers of a CNN as different scale levels [43]. But all the above methods only take several scale levels near original image scale levels into account. Given an image pair with large scale difference, most scale levels sampled by the above methods of one image are unrelated with those of another image. In that case, all the above keypoint detectors have difficulty in detecting consistent keypoints [17], [18]. And the scale invariance of existing descriptors is not enough under large scale changes [38]. Thus, few inlier correspondences can be established between the image pair. To solve this problem, the proposed SDAIM reduces image scale difference of the image pair before sampling scale levels so that most sampled scale levels of one image are related with those of another image.

B. Local Feature Matching

The most widely used matching procedure contains the following steps: mutual nearest neighbour matching (MNN), ratio test [5] and RANSAC verification [47], [48], [49], [50]. In recent years, several new matching methods have been proposed. Zhao et al. utilized motion prior to obtain an efficient learning-based local feature matching method [51]. But motion prior is provided by inertial measurement unit integration, which is not always available in practical applications. Given putative correspondences of feature points between two views, BB-Homography refines the set of correspondences and estimates a homography matrix by solving a second-order bipartite graph problem [52]. BB-Homography only applies to the homography estimation problem. PointCN uses a deep neural network to tell inlier correspondences from outlier ones [14]. OANet introduces both local and global context of sparse correspondences into learned outlier rejection methods [15]. AdaLAM is a handcrafted method, which utilizes spatial consistency in a hierarchical pipeline for effective outlier rejection [13]. But PointCN, OANet and AdaLAM are all outlier rejection methods. When facing large scale changes, there are very few inlier correspondences among putative correspondences computed by MNN. In that case, under the help of PointCN, OANet and AdaLAM, the number of inlier correspondences is still not sufficient for downstream tasks. SuperGlue is a ground-breaking matching method. Instead of taking the putative correspondences established by MNN as input, it supplements the discriminative power of local descriptors with learned spatial verification inside its network [16]. However, under large scale changes, due to the low repeatability of keypoint detectors, SuperGlue is not able to establish many inlier correspondences, either. In summary, under large scale changes, because local features are extracted from mismatched scale levels, existing matching methods can not help local features out of trouble. In order to make sure that most local features are extracted from related scale levels, the proposed SDAIM narrows scale differences before local feature extraction. Therefore, SDAIM can significantly boost
the performance of local feature matching methods when facing large scale changes.

C. Image Scale Ratio Estimation

There are very few methods of estimating the scale ratio of an image pair. Rau et al. introduced an Image Box Embedding (IBE) to estimate the image scale ratio [24]. But the model of Rau et al. is scene-specific. An image box embedding model can only make sense in the training scene. Zhou et al. used a Scale Level Matching method based on Bag-of-Features model (SLM-BoF) to determine the scale ratio of an image pair [17]. Given a pair of images, SLM-BoF samples several scale levels from every image. Then the Bag-of-Features framework is used to generate representation vectors for all scale levels of the two images. At last, SLM-BoF exhaustively matches scale levels by comparing those representation vectors and estimates the scale ratio of the image pair. By contrast, we propose a learning-based method to estimate the scale ratio, which has much better generalization ability than IBE [24] and much higher accuracy than SLM-BoF [17].

III. DEEP SCALE-DIFFERENCE-AWARE IMAGE MATCHING

A. Definitions

Our goal is narrowing the scale difference between two images before feature extraction. How to determine that there is no scale difference? We adopt the following definition [24]:

There is no scale difference between two images when visual overlaps approximately occupy the same number of pixels in each of the two images. Visual overlaps are image areas picturing the same 3D object surfaces, which are marked by red rectangles in Figure 2(a).

We also make a definition for the image scale ratio: If the scale ratio between image $I_1$ and image $I_2$ is $s$, $I_1$ should be resized to $s^{-1}$ times its original size so that there is nearly no scale difference between the resized $I_1$ and $I_2$. The scale ratio definition is denoted by $\phi(\cdot, \cdot)$, as shown in:

$$\phi (I_1, I_2) = s, \quad \phi (I_2, I_1) = \frac{1}{s}. \quad (1)$$

Note that the scale ratio definition is asymmetric.

The above scale ratio is a global scale ratio. Due to viewpoint changes, given an image pair, the local scale ratios of different image areas are not always identical with the global scale ratio. However, compared with our proposed Scale-Difference-Aware Image Matching (SDAIM) method based on the global scale ratio (Section III-B), using local scale ratios to reduce scale differences is not straightforward for those local feature extractors which take a whole image as input [9], [12], [38], [45], [53]. Moreover, experimental results in Section IV-D and IV-E demonstrate that our SDAIM suffices to greatly boost the performance of local feature extractors and local feature matching methods under large scale changes.

B. Scale-Difference-Aware Image Matching

The Multi-Scale Feature Extraction method only extracts local features from the scale levels near original image scale levels [5], [10], [12], [20]. By contrast, we propose a Scale-Difference-Aware Image Matching (SDAIM) method that reduces the scale difference before local feature extraction, as illustrated in Figure 2(a). Given two images, $I_1$ and $I_2$, assume that the estimated scale ratio between $I_1$ and $I_2$ is $s$. $I_1$ is resized to $s^{-0.5}$ times its original size. $I_2$ is resized to $s^{0.5}$ times its original size. Lanczos interpolation is used to resize images. Then local features are extracted from the resized image pair.

The advantages of the proposed SDAIM over the Multi-Scale Feature Extraction method are depicted in Figure 3. There are four subfigures. In each subfigure, a white triangle represents the Gaussian Pyramid of an image. If there is no scale difference between image $I_1$ and image $I_2$, their scale levels, $s_1$ and $s_2$, are at the same height. In that case, $s_1$ and $s_2$ are related scale levels. Without loss of generality, we assume that images are only down-sampled in the Multi-Scale Feature Extraction method. It only samples neighbouring scale levels of original image scale levels, as shown in Figure 3(a) and (c). Local features are only extracted at those sampled scale levels. If the scale difference between $I_1$ and $I_2$ is not very large, as shown in Figure 3(a), some sampled scale levels of $I_1$ are related with sampled scale levels of $I_2$. The related scale levels are marked in blue. However, when the scale difference is large, all the sampled scale levels are unrelated, as shown in Figure 3(c). In that case, very few matches are right, as vividly depicted in Figure 1(a). Our method is capable of solving the above problem. As shown in Figure 3(b) and (d), no matter how large the scale difference between $I_1$ and $I_2$ is, the scale difference between resized $I_1$ and resized $I_2$ is small, so that almost all sampled scale levels of $I_1$ are related with those of $I_2$, which results in sufficient inlier matches for downstream tasks, as shown in Figure 1(b).

C. Scale Ratio Estimation Network

The proposed SDAIM needs a scale ratio estimation method. A neural network, termed as Scale-Net, is proposed to estimate the scale ratio of a given image pair. Its overall architecture is shown in Figure 2(b). It consists of three components: Multi-Scale Feature Extraction Module, Covisibility-Attention-Reinforced Matching module...
(CVARM) and Scale Ratio Regressor. Given two images, \( I_1 \) and \( I_2 \), firstly, the Multi-Scale Feature Extraction Module is used to extract dense feature maps from \( I_1 \) and \( I_2 \). The L2-normalized dense feature maps from \( I_1 \) and \( I_2 \) are denoted by \( F_1 \) and \( F_2 \). Secondly, CVARM takes \( F_1 \) and \( F_2 \) as input and outputs a correlation map. At last, the Scale Ratio Regressor takes the correlation map as input and calculates the scale ratio of the image pair.

1) Multi-Scale Feature Extraction Module: This module is devised to extract multi-scale dense feature maps from an image. In this module, the input image is downsampled once and upsampled once. And a three-level Gaussian Pyramid is obtained. A CNN is used to extract dense feature maps from each level of the pyramid. Then, the dense feature maps from three levels are resized. The sizes of these resized dense feature maps are same. The final output of this module is the L2-normalized weighted sum of these resized dense feature maps.

2) Covisibility-Attention-Reinforced Matching Module (CVARM): According to the scale ratio definition (Section III-A), the scale ratio depends on the visual overlaps between \( I_1 \) and \( I_2 \). Therefore, Scale-Net should pay much attention to the information of covisible areas for accurate scale ratio estimation. To this end, Covisibility-Attention-Reinforced Matching Module (CVARM) is proposed. The diagram of this module is shown in Figure 2(b). Given the dense feature maps of \( I_1 \) and \( I_2 \), \( F_1, F_2 \in \mathbb{R}^{h \times w \times c} \), there are \((h \times w) \times (h \times w)\) pairs of local feature descriptors between \( F_1 \) and \( F_2 \) in total. The inner products of all these descriptor pairs are computed and form a correlation map \( C_{12} \in \mathbb{R}^{h \times w \times (h \times w)} \) [54]. The calculation process of \( C_{12} \) is shown in:

\[
C_{12}(i, j, k) = F_1(i, j)^T F_2(ik, jk),
\]

where \( k = h(i_1 - 1) + j_1, i \in [1, h], i_1 \in [1, h], j \in [1, w], j_1 \in [1, w] \). Each \( c \)-dimensional vector in \( F_1 \) or \( F_2 \) is a descriptor of a grid patch in image \( I_1 \) or image \( I_2 \). Therefore, the raw correlation map, \( C_{12} \), is constructed by exhaustively comparing patches of \( I_1 \) and \( I_2 \). \( C_{12}(i, j, k) \) is just the inner product of the covisible similarity between patch \((i, j)\) in \( I_1 \) and patch \((ik, jk)\) in \( I_2 \). Thus, \( C_{12} \) contains much information of non-covisible areas. The information of non-covisible areas may have a bad influence on scale ratio estimation. Therefore, we aim to lay more stress on the covisible areas and suppress the disturbance resulting from non-covisible areas.

In what follows, each element of \( C_{12} \) is called as the similarity value between two patches. Assume a certain patch \( A \) is in \( I_1 \) and a certain patch \( B \) is in \( I_2 \). After \( C_{12} \) is obtained, patch \( A \) has been compared with all the patches in \( I_2 \). \( C_{12} \) contains \((h \times w)\) similarity values related to \( A \), each of which is the similarity value between patch \( A \) and a certain patch in \( I_2 \). Generally, if patch \( A \) and patch \( B \) picture the same 3D surface, their similarity value is high. Furthermore, if patch \( A \) belongs to the covisible areas, the highest similarity value among all the \((h \times w)\) similarity values related to \( A \) is large. Thus, the Highest Similarity Value of a certain patch can indicate whether this patch belongs to covisible areas.

Based on the above analysis, we propose two covisibility-attention branches to dig out covisible areas in \( I_1 \) and \( I_2 \) from the raw correlation map \( C_{12} \). The two Covisibility-Attention Branches are called as \( CAB_1 \) and \( CAB_2 \) in Figure 2(b). In \( I_1 \), the corresponding patch of descriptor \( F_1(i, j) \) is denoted by \( P_{1ij} \). In \( I_{12} \), the \((h \times w)\)-dimensional vector \( C_{12}(i, j) \) stores all the similarity values related to \( P_{1ij} \). In \( CAB_1 \), \( C_{12} \) is max-pooled along the channel axis. Thereafter, the Highest Similarity Values of all patches in \( I_1 \) are obtained and form a similarity map \( S_1 \in \mathbb{R}^{h \times w \times 1} \). Then, a convolution operation with the filter size of \( 5 \times 5 \), denoted by \( f^{5 \times 5} \), and a sigmoid function, denoted by \( \sigma \), are used to refine \( S_1 \). Finally, a covisibility score map \( M_1 \in \mathbb{R}^{h \times w \times 1} \) is obtained. \( M_1(i, j) \) represents the probability of belonging to covisible areas of \( P_{1ij} \). In another word, \( M_1 \) is a soft mask that indicates covisible areas in image \( I_1 \). The calculation process of \( M_1 \) is summarized as below:

\[
M_1 = \sigma \left(f^{5 \times 5}(\text{MaxPool}_{\text{Channel}}(C_{12}))\right).
\]

In \( I_2 \), the corresponding patch of descriptor \( F_2(ik, jk) \) is denoted by \( P_{2ikj} \). The \( k \)-th channel of \( C_{12} \) is a matrix denoted by \( T_k \). \( T_k \) stores all the similarity values related to \( P_{2ikj} \). In \( CAB_2 \), \( C_{12} \) is max-pooled globally. Thereafter, the Highest Similarity Values of all patches in \( I_2 \) are obtained and form a \((h \times w)\)-dimensional vector \( S_2_{\text{ec}} \). According to Equation (2), \( S_2_{\text{ec}} \) can be reshaped to a similarity map \( S_2 \in \mathbb{R}^{h \times w \times 1} \). This reshaping operation is denoted by \( \text{Rsp}_1 \), as shown in:

\[
\text{Rsp}_1(k) : S_2(ik, jk) \leftarrow S_2_{\text{ec}}(k), k = h(i_1 - 1) + j_1.
\]

The following step is using a convolution operation with the filter size of \( 5 \times 5 \), denoted by \( f^{5 \times 5} \), and a sigmoid function, denoted by \( \sigma \), to refine \( S_2 \). Note that the filter \( f^{5 \times 5} \) here and the filter \( f^{5 \times 5} \) in \( CAB_1 \) share the same set of weights. The refined \( S_2 \) is called as \( M_2 \), which is a soft mask that indicates covisible areas in image \( I_2 \). At last, \( M_2 \) is reshaped to \( M_{2_{\text{ec}}} \in \mathbb{R}^{1 \times 1 \times (h \times w)} \). This reshaping operation is denoted by \( \text{Rsp}_2 \), which is shown in:

\[
\text{Rsp}_2(k) : M_{2_{\text{ec}}}(k) \leftarrow M_2(ik, jk), k = h(i_1 - 1) + j_1.
\]

The calculation process of \( M_{2_{\text{ec}}} \) is summarized as below:

\[
M_{2_{\text{ec}}} = \text{Rsp}_2\left(\sigma\left(f^{5 \times 5}(\text{Rsp}_1(\text{MaxPool}(C_{12})))\right)\right).
\]

\( M_1 \) and \( M_{2_{\text{ec}}} \) are used to reduce the information of non-covisible areas in \( I_1 \) and \( I_2 \) respectively. They are combined to emphasize covisibility information in \( C_{12} \), as shown in:

\[
C_{12, \text{CR}} = M_{2_{\text{ec}}} \odot (M_1 \odot C_{12}),
\]

where \( \odot \) denotes element-wise multiplication. During element-wise multiplication, \( M_1 \) is broadcasted along the channel axis, \( M_{2_{\text{ec}}} \) is broadcasted along the spatial dimension. Then, a covisibility-reinforced correlation map \( C_{12, \text{CR}} \) is obtained. In the end, a Scale Ratio Regressor takes \( C_{12, \text{CR}} \) as input and outputs a scale ratio between \( I_1 \) and \( I_2 \). The network architecture of the Scale Ratio Regressor is shown in Figure 4. Note that the output of the fully connected layer in the Scale
be close to the known ground truth scale ratios. The reason scale ratio of (ld, l2, s) i is more rational in a self-supervised manner. Scale-COCO dataset. We use Scale-COCO dataset to train image pairs with scale differences. We call this dataset as Scale-Net and SDAIM. The above datasets are detailed as below.

1) Scale-COCO: We randomly select a real number, m ∈ [0, 7], and get a scale ratio s = 2m. Two different images, Ib1 and Ib2 are selected as background images from the training set of ImageNet ILSVRC dataset [56]. We call this dataset as Scale-PT dataset. ES-HP dataset, proposed by Liu et al. [38], contains considerable image scale differences. We resort to ES-HP dataset and Scale-PT dataset to demonstrate the effect of Scale-Net and SDAIM. The above datasets are detailed as below.

2) Scale-PT: 15 outdoor scenes with image poses, camera intrinsic parameters and semi-dense depth maps, are chosen from PhotoTourism dataset. The depth maps are generated by COLMAP. The scale ratio of an image pair depends heavily on the visual overlaps. And visual overlaps can be found accurately by means of the method proposed by Rau et al. [24] when image poses, camera intrinsic parameters and semi-dense depth maps are known. Thus, we make use of the method proposed by Rau et al. to generate image pairs with ground truth scale ratios. The procedure of ground truth scale ratio annotation is detailed as below.

Assume that there are visual overlaps between image I1 and image I2. Firstly, visible point clouds of I1 and I2 in the our Scale-Net. But Scale-COCO dataset does not contain self-occlusion, which is an important challenge in correspondence estimation. Therefore, we create another dataset based on PhotoTourism dataset [25]. We call this dataset as Scale-PT dataset. ES-HP dataset, proposed by Liu et al. [38], contains considerable image scale differences. We resort to ES-HP dataset and Scale-PT dataset to demonstrate the effect of Scale-Net and SDAIM. The above datasets are detailed as below.

3) Dual Consistent Loss: D ≡ ((I1, I2, s) i)i=N 1 represents a training set, whose number of samples is N. In a training sample (I1, I2, s) i, I1 and I2 form an image pair, whose ground truth scale ratio is s i. Assume that the scale ratio of the image pair (I1, I2) estimated by Scale-Net is ˆs i. Given the training sample (I1, I2, s) i, there is a dual training sample (I1, I2, 1/s i) according to Equation (1). Assume that the scale ratio of (I2, I1) estimated by Scale-Net is ˆs i'. Then, we propose a dual loss l d, as shown in Equation (8).

\[
l_d = \frac{1}{2N} \sum_{i=1}^{N} \left[ \left( \log_2 \hat{s}_i - \log_2 s_i \right)^2 + \left( \log_2 \hat{s}_i' - \log_2 \frac{1}{s_i} \right)^2 \right].
\]  

(8)

In l d, the scale ratios predicted by Scale-Net are forced to be close to the known ground truth scale ratios. The reason why l d is defined in log scale is that l d should be symmetric on both sides of a ground truth scale ratio in the scale space [18]. Besides, ˆs i and ˆs i' are expected to be consistent with each other according to Equation (1). In another word, ˆs i ˆs i' should be close to 1. Thus, we devise a consistent loss l c, as shown in Equation (9).

\[
l_c = \frac{1}{N} \sum_{i=1}^{N} \left( \log_2 \hat{s}_i + \log_2 \hat{s}_i' \right)^2.
\]  

(9)

Although solely using l c cannot enable Scale-Net to predict scale ratios accurately, l c is able to serve as a regularization term. l c explicitly utilizes the consistency constraint of scale ratio definition to make the estimation results of Scale-Net more rational in a self-supervised manner. l d and l c are combined to train our Scale-Net, as shown in:

\[
L = \lambda_d l_d + \lambda_c l_c,
\]  

(10)

where \( \lambda_d \) and \( \lambda_c \) are scalar weights.

IV. EXPERIMENTS

A. Datasets Containing Large Scale Differences

To our knowledge, there does not exist a dataset containing large scale differences, whose samples are sufficient to train a neural network. Thus, we create a large dataset based on MS-COCO 2014 dataset [55] by synthetically generating image pairs with scale differences. We call this dataset as Scale-COCO dataset. We use Scale-COCO dataset to train the Scale-Ratio Regressor, m, is a scalar. \( 2^m \) is the scale ratio estimated by the Scale Ratio Regressor.

Fig. 5. Training image pairs. (a) is a sample of Scale-COCO dataset. (b) is a sample of Scale-PT dataset. Visual overlaps are roughly marked by yellow rectangles. The number below each image pair is the ground truth scale ratio of the image pair. Best viewed in color.
world coordinate system can be easily obtained using their poses, their camera intrinsic parameters and their semi-dense depth maps. Visible point clouds of $I_1$ and $I_2$ are denoted by $P_1$ and $P_2$, respectively. Secondly, given a 3D point $p$ in $P_1$, we compute the Euclidean distances between $p$ and all 3D points in $P_2$ and call the shortest one as the distance between $p$ and $P_2$. If the distance between $p$ and $P_2$ is smaller than a threshold $r$, $p$ is visible in $I_2$. Then, in $P_1$, we compute the number of 3D points that are visible in $I_2$. This number is called as $V_1$. In $P_2$, we also compute the number of 3D points that are visible in $I_1$. This number is called as $V_2$. At last, the scale ratio between $I_1$ and $I_2$ is $V_1/V_2$, i.e., $\phi(I_1, I_2) = V_1/V_2$, where $\phi(\cdot, \cdot)$ has the same meaning as the one in Equation (1). An annotation result is shown in Figure 5(b). The scale ratio range of Scale-PT dataset is $[0.002, 1) \cup (1, 512]$. The image scale ratio distribution histogram of Scale-PT dataset is shown in Figure 6.

Our Scale-PT dataset is divided into a training set and a test set. There are 446685 image pairs in the training set. There are 39778 image pairs in the test set. Training image pairs are selected from the following scenes: Brandenburg Gate, Buckingham Palace, Colosseum Exterior, Grand Place Brussels, Pantheon Exterior, Prague Old Town Square, Reichstag, Taj Mahal, Temple Nara Japan, Westminster Abbey, Trevi Fountain. Test image pairs are selected from the following scenes: Notre Dame Fre Face (NFDF), Palace of Westminster (PW), Sacre Coeur (SC), St. Peter’s Square (SPS).

3) MegaDepth-Rau: MegaDepth is a large-scale outdoor dataset [57]. Rau et al. arranged four outdoor scenes based on MegaDepth dataset [24]. The four scenes are Notre-Dame (ND), Big Ben (BB), Venice (Ve) and Florence (Fl) respectively. The depth maps in this dataset are generated by a modified MVS algorithm based on COLMAP and are filtered by semantic segmentation. We use the above scale ratio annotation method to compute the ground truth scale ratios of image pairs in the four scenes. We call this dataset including four outdoor scenes as MegaDepth-Rau dataset. Its scale ratio range is $[0.002, 1) \cup (1, 512]$. The average $L_1$ discrepancy between the predicted and ground truth scale ratios as the evaluation metric here. Given ground truth scale ratios $[s_i]_{i=1}^N$ and predicted scale ratios $[\hat{s}_i]_{i=1}^N$, the average $L_1$ discrepancy, $E_s$, is shown as below:

$$E_s = \frac{1}{N} \sum_{i=1}^{N} |\log_2 s_i - \log_2 \hat{s}_i|. \quad (11)$$

The average $L_1$ discrepancies of IBE, SLM-BoF and Scale-Net are shown in Table I. Scale-Net_COCO is trained with Scale-COCO dataset. We further finetune Scale-Net on the training set of Scale-PT dataset for 1 epoch.

C. Image Scale Ratio Estimation

We resort to MegaDepth-Rau dataset [24] and Scale-PT dataset to compare the scale ratio estimation accuracy of IBE [24], SLM-BoF [17] and our Scale-Net. The area pictured in PW of Scale-PT dataset is a part of that in BB of MegaDepth-Rau dataset. The area pictured in NDF of Scale-PT dataset is a part of that in ND of MegaDepth-Rau dataset. But PW and NDF contain more image pairs with large scale differences than BB and ND. We recompute the enclosure and concentration values in MegaDepth-Rau dataset without weighting points by normals. And we use the recomputed MegaDepth-Rau dataset to retrain IBE models. Because IBE is not able to generalize across scenes, it is not evaluated in SC and SPS. To reimplement SLM-BoF [17], the training images of Scale-PT dataset are used to build a vocabulary tree. 8K SIFT feature points are extracted from every image when building the vocabulary tree. When using SLM-BoF to estimate image scale ratios, we still extract 8K SIFT feature points from every image. We adopt the average $L_1$ discrepancy between the predicted and ground truth scale ratios as the evaluation metric here. Given ground truth scale ratios $[s_i]_{i=1}^N$ and predicted scale ratios $[\hat{s}_i]_{i=1}^N$, the average $L_1$ discrepancy, $E_s$, is shown as below:

$$E_s = \frac{1}{N} \sum_{i=1}^{N} |\log_2 s_i - \log_2 \hat{s}_i|. \quad (11)$$

The average $L_1$ discrepancies of IBE, SLM-BoF and Scale-Net are shown in Table I. Scale-Net_COCO is trained with Scale-COCO dataset. We further finetune Scale-Net_COCO on the training set of Scale-PT dataset for only 1 epoch and obtain Scale-Net_ft_PT. The curves of the average $L_1$ discrepancies over scale differences of image pairs are shown in Figure 7. The meaning of a point $(x \sim (x + 1), y)$ on a curve is illustrated as follows: when only those image pairs whose ground truth scale ratios belong to $(2^{-x}2^{-x+1}) \cup (2^{x}2^{x+1})$ are taken into account, the average $L_1$ discrepancy is $y$. As shown in Table I, our Scale-Net_ft_PT is comparable with IBE on MegaDepth-Rau dataset. Note that an IBE model can only apply to the single scene used for training it. Therefore, in this experiment, there are four IBE models.
TABLE I

| Methods          | MegaDepth-Rau | Scale-PT |
|------------------|---------------|----------|
| BB (17)          | 1.53 1.56 1.79 1.36 | 3.04 2.34 2.71 2.69 |
| IBE [24]         | 1.51 1.23 1.43 1.28 | 4.64 5.25 / / |
| Scale-Net_COCO  | 1.52 1.42 1.72 1.39 | 2.24 1.71 1.58 1.83 |
| Scale-Net_ft_PT | 1.32 1.26 1.51 1.06 | 1.08 1.12 0.74 0.86 |

Fig. 7. The curves of scale ratio estimation error over scale differences of image pairs. (a) The evaluation results of IBE, SLM-BoF, Scale-Net_COCO and Scale-Net_ft_PT in NDFF and PW. (b) The evaluation results of SLM-BoF, Scale-Net_COCO and Scale-Net_ft_PT in SC and SPS.

trained with the four scenes of MegaDepth-Rau dataset respectively. By contrast, our Scale-Net_ft_PT is not trained with MegaDepth-Rau dataset. Thus, compared with IBE, Scale-Net has much better generalization ability. Scale-Net_ft_PT achieves better results than Scale-Net_COCO, which confirms that perspective transformations can not totally replace the real viewing angle changes. As shown in Figure 7, compared with IBE and SLM-BoF, Scale-Net_ft_PT has much lower estimation error when facing large scale differences. Two qualitative results of our Scale-Net_ft_PT are shown in Figure 8.

D. Image Matching

In this section, we utilize ES-HP dataset [38] to confirm that our SDAIM is able to boost the image matching performance of local features under considerable scale changes. The ground truth homography matrices of image pairs are contained in ES-HP dataset. We select SIFT [5], ASLFeat [12], GIFT-SP (SuperPoint + GIFT) [38], KAZE [29], ALIKE [45], and CD-UNet [53] as baselines.

Evaluation Protocols: Because IBE is not able to generalize across scenes, only SLM-BoF [17] and the proposed Scale-Net_COCO are used to estimate scale ratios of image pairs for the proposed SDAIM. We assess the performance of the following two combinations: (SLM-BoF + SDAIM) and (Scale-Net_COCO + SDAIM). For fair comparison, we build a new vocabulary tree for SLM-BoF using the training images of MS-COCO 2014 dataset. 8K SIFT feature points are extracted from every image during building the vocabulary tree. Given an image, 2K feature points are extracted by each local feature method. Following GIFT [38], we use Percentage of Correctly Matched Keypoints (PCK) to quantify the performance for correspondence estimation. PCK is defined as the ratio between the number of correct matches and the total number of interest points [38], [60]. If a match conforms to the ground truth homography matrix within a Euclidean distance error tolerance $\beta$, this match is regarded as a correct match. $\beta$ is set to 5 pixels. We also use keypoint repeatability to compare the performance of all methods. Keypoint repeatability is the ratio between possible matches and the minimum number of keypoints in the shared view. The threshold used to determine a possible match is 3 pixels. Evaluation results are shown in Table II. “Illum.” means illumination changes. “View.” means viewpoint changes.

As shown in Table II, the results of the above local features are all improved by SDAIM. And SDAIM with Scale-Net_COCO is able to achieve much better results than SDAIM with SLM-BoF. Table II also confirms that our Scale-Net is
robust to viewpoint and illumination changes. As shown in Figure 9(a) and (b), in ES-HP dataset, the proposed SDAIM and Scale-Net are able to help ASLFeat establish much more inlier correspondences.

E. Relative Pose Estimation

1) Enhancement to Local Features: In this section, we demonstrate that SDAIM is able to improve the relative pose estimation precision of local features under large scale changes. We evaluate the performance of SDAIM on the test set of Scale-PT dataset, which consists of 39778 image pairs whose ground truth relative camera poses are known. IBE [24], SLM-BoF [17] and our Scale-Net are used to estimate scale ratios for SDAIM. We assess the performance of all the following combinations: (IBE + SDAIM), (SLM-BoF + SDAIM), (Scale-Net_COCO + SDAIM), (Scale-Net_ft_PT + SDAIM). In this section, SLM-BoF uses the vocabulary tree built in Section IV-C. Because IBE is not able to generalize across scenes, it is not evaluated in SC and SPS.

Evaluation Protocols: We select SIFT [5], ASLFeat [12], GIFT-SP [38], ALIKE [45], and CD-UNet [53] as baselines. 2K feature points are extracted from every image. The feature matching procedure contains the following steps: mutual nearest neighbour matching (MNN), ratio test and RANSAC verification. The RANSAC algorithm proposed by Fischler and Bolles [47] is used. The reprojection threshold is 1.5 pixels. The confidence threshold is 0.999. The maximum number of iterations for random sample selection is 10^5. Because the stereo relative pose estimation problem is defined up to a scale factor, we compute the angular difference between the estimated and ground truth translation vectors, \( E_t \), and the angular difference between the estimated and ground truth rotation matrices, \( E_{rot} \). Following Yi et al. [14], we adopt the following definitions of \( E_t \) and \( E_{rot} \): Assume there is an estimated translation vector \( t_e \) and a ground truth translation vector \( t_g \). \( E_t \) is the closest arc distance between the normalized \( t_e \) and the normalized \( t_g \). Assume there is an estimated rotation matrix \( R_e \) and a ground truth rotation matrix \( R_g \). \( q_e \) and \( q_g \) are four-dimensional unit vectors. \( q_e \) stores the four coefficients of the quaternion representation of \( R_e \). \( q_g \) stores the four coefficients of the quaternion representation of \( R_g \).

\[
E_{rot} = \arccos \left( 2(q_e^T q_g)^2 - 1 \right) \quad \text{[14]},
\]

\( E_t \) and \( E_{rot} \) are defined in degrees. Following Yi et al. [14], we report mean Average Accuracy (mAA) here. \( mAA(\theta^\circ) \) is obtained by integrating the translation and rotation errors to \( \theta^\circ \). The larger the mAA is, the better the performance is. Repeatability scores of local features are also reported here.

Table III shows that our (Scale-Net_COCO + SDAIM) and our (Scale-Net_ft_PT + SDAIM) are able to raise the relative pose estimation precision of all the above local features. Both our (Scale-Net_COCO + SDAIM) and our (Scale-Net_ft_PT + SDAIM) perform better than (SLM-BoF + SDAIM). (IBE + SDAIM) slightly worsen the relative pose estimation performance of local features. Combining Table I and Table III, we can find that the performance of our SDAIM relies heavily on the image scale ratio estimation precision. The higher the scale ratio estimation precision is, the better the performance of our SDAIM is. And the low scale ratio estimation precision of IBE results in the unsatisfactory performance of (IBE + SDAIM).

The curves of relative pose estimation accuracy rate over scale differences of image pairs are drawn in Figure 10. If the angular error is smaller than 10^\circ, the estimated rotation matrix and translation vector are regarded as accurate. Because existing IBE models are not able to generalize to SC and SPS, these curves are based on the results in NDFF and PW. In Figure 10, we take SIFT and ASLFeat for example. The accuracy rate curves of other local features are similar to those of SIFT and ASLFeat. According to Figure 10, when scale ratios of image pairs are larger than 4 or smaller than 0.25, our (Scale-Net_COCO + SDAIM) and (Scale-Net_ft_PT + SDAIM) are able to raise the relative pose estimation accuracy of SIFT and ASLFeat. Compared with SLM-BoF, the proposed Scale-Net_ft_PT achieves much better results.

2) Enhancement to Local Feature Matching Methods: In previous experiments, only mutual nearest neighbor matching (MNN) and ratio test are used to match local features. In this experiment, we resort to the test set of Scale-PT dataset to confirm that our Scale-Net and SDAIM are able to greatly enhance relative pose estimation performance of state-of-the-art local feature matching methods under large scale changes.

Evaluation Protocols: We select Scale-Invariant Image Matching (SIIM) [17], AdaLAM [13], OANet [15] and...
**TABLE III**
EVALUATION RESULTS OF SEVERAL LOCAL FEATURES ON THE TEST SET OF SCALE-PT DATASET

| Methods                  | Rotation mAA(5°) | Rotation mAA(10°) | Translation mAA(5°) | Translation mAA(10°) | Repeatability($) | Repeatability(%) |
|--------------------------|------------------|-------------------|--------------------|----------------------|------------------|------------------|
|                          | NDFP  | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS |
| SIFT (5)                 | 0.103 | 0.197 | 0.219 | 0.094 | 0.158 | 0.138 | 0.287 | 0.154 | 0.060 | 0.094 | 0.174 | 0.054 | 0.110 | 0.105 | 0.249 | 0.098 | 0.276 | 0.278 | 0.298 | 0.341 |
| SIFT + (IBE [24] + SDAIM) | 0.099 | 0.078 | / | 0.157 | 0.117 | / | 0.061 | 0.055 | / | 0.114 | 0.099 | / | 0.235 | 0.174 | / | 0.257 | 0.174 | / | 0.257 | 0.174 | / | 0.257 | 0.174 |
| SIFT + (SLM-BoF [17] + SDAIM) | 0.126 | 0.102 | 0.234 | 0.002 | 0.033 | 0.155 | 0.131 | 0.193 | 0.073 | 0.063 | 0.198 | 0.055 | 0.141 | 0.119 | 0.283 | 0.118 | 0.354 | 0.354 | 0.354 | 0.354 |
| SIFT + (Scale-Net COCO + SDAIM) | 0.141 | 0.109 | 0.241 | 0.111 | 0.228 | 0.146 | 0.348 | 0.212 | 0.083 | 0.087 | 0.214 | 0.058 | 0.159 | 0.127 | 0.313 | 0.129 | 0.376 | 0.368 | 0.376 | 0.376 |
| SIFT + (Scale-Net COCO + PT) | 0.154 | 0.138 | 0.264 | 0.132 | 0.254 | 0.204 | 0.383 | 0.249 | 0.094 | 0.093 | 0.242 | 0.072 | 0.178 | 0.161 | 0.340 | 0.152 | 0.395 | 0.371 | 0.405 | 0.404 |

**TABLE IV**
EVALUATION RESULTS OF SEVERAL REPRESENTATIVE LOCAL FEATURE MATCHING METHODS ON THE TEST SET OF SCALE-PT DATASET

| Local Features          | SDAIM                  | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS | NDFP | PW | SC | SPS |
|-------------------------|------------------------|------|----|----|----|------|----|----|----|------|----|----|----|------|----|----|----|------|----|----|----|------|----|----|----|
| MNN + Ratio test        | /                      | 0.103 | 0.107 | 0.219 | 0.094 | 0.156 | 0.138 | 0.267 | 0.154 | 0.060 | 0.094 | 0.174 | 0.054 | 0.110 | 0.105 | 0.249 | 0.098 | 0.276 | 0.278 | 0.298 | 0.341 |
| SIFT                    | /                      | 0.115 | 0.119 | 0.225 | 0.105 | 0.184 | 0.171 | 0.296 | 0.190 | 0.074 | 0.076 | 0.192 | 0.060 | 0.133 | 0.134 | 0.268 | 0.123 |
| SIIM [17]               | /                      | 0.156 | 0.138 | 0.288 | 0.103 | 0.234 | 0.204 | 0.382 | 0.249 | 0.094 | 0.093 | 0.242 | 0.072 | 0.178 | 0.161 | 0.340 | 0.152 |
| SIFT                    | /                      | 0.163 | 0.144 | 0.279 | 0.143 | 0.257 | 0.210 | 0.388 | 0.257 | 0.100 | 0.097 | 0.234 | 0.078 | 0.183 | 0.166 | 0.328 | 0.164 |
| MNN + AdaLAM [13]       | /                      | 0.123 | 0.119 | 0.221 | 0.109 | 0.178 | 0.165 | 0.284 | 0.185 | 0.081 | 0.084 | 0.190 | 0.063 | 0.133 | 0.122 | 0.252 | 0.117 |
| MNN + AdaLAM [13]       | /                      | 0.182 | 0.159 | 0.297 | 0.162 | 0.281 | 0.222 | 0.390 | 0.266 | 0.114 | 0.100 | 0.246 | 0.099 | 0.199 | 0.164 | 0.336 | 0.174 |
| MNN + OANet [15]        | /                      | 0.147 | 0.146 | 0.233 | 0.137 | 0.216 | 0.211 | 0.301 | 0.229 | 0.101 | 0.111 | 0.201 | 0.092 | 0.187 | 0.177 | 0.273 | 0.156 |
| MNN + OANet [15]        | /                      | 0.216 | 0.195 | 0.322 | 0.191 | 0.332 | 0.282 | 0.422 | 0.335 | 0.143 | 0.148 | 0.268 | 0.114 | 0.230 | 0.236 | 0.368 | 0.221 |
| SuperPoint              | /                      | 0.066 | 0.065 | 0.091 | 0.049 | 0.096 | 0.095 | 0.125 | 0.083 | 0.042 | 0.047 | 0.073 | 0.028 | 0.073 | 0.078 | 0.113 | 0.057 |
| SuperPoint              | /                      | 0.113 | 0.087 | 0.155 | 0.069 | 0.178 | 0.151 | 0.222 | 0.164 | 0.065 | 0.062 | 0.120 | 0.047 | 0.133 | 0.103 | 0.184 | 0.101 |
| SuperGlue [16]          | /                      | 0.150 | 0.210 | 0.303 | 0.170 | 0.287 | 0.287 | 0.369 | 0.381 | 0.153 | 0.189 | 0.284 | 0.117 | 0.249 | 0.268 | 0.374 | 0.312 |

Fig. 10. The accuracy rate curves of the estimated relative camera poses of SIFT and ASLFeat in NDFP and PW. If the angular error is smaller than 10°, the estimated rotation matrices and translation vectors are regarded as accurate. Best viewed in color with 300% zoom in.

SuperGlue [16] as baselines. To the best of our knowledge, SIIM proposed by Zhou et al. [17] is the only existing local feature matching method focusing on image scale changes. AdaLAM [13] is the existing best handcrafted outlier rejection method. OANet [15] is the existing best learned outlier rejection method. SuperGlue [16] is the existing best learned local feature matching method. The adopted local features are SIFT [5] and SuperPoint [9]. 2K feature points are extracted from every image. We report mAA(5°) and mAA(10°) in Table IV. In Table IV, / means that SDAIM is not used. In another word, / means that image pairs are not resized before local feature extraction. For convenience, we use (m + Scale-Net) to denote the matching method, m, assisted by SDAIM and Scale-Net_ft_PT. (m + /) means that m is the adopted matching method and that image pairs are not resized before local feature extraction.

As shown in Table IV, compared with (SIIM + /), (MNN + Ratio test + Scale-Net) achieves better results. By comparing the results of (SIIM + /) and (SIIM + Scale-Net), we can find that SDAIM and Scale-Net are able to boost the performance of SIIM. The proposed SDAIM and Scale-Net can also greatly improve the performance of AdaLAM and OANet. As shown in Table IV, our SDAIM and Scale-Net can still drastically enhance the performance of SuperGlue. Qualitative results of...
the remarkable boost from (Scale-Net + SDAIM) are shown in Figure 11. Our Scale-Net and SDAIM greatly raise the number of inlier correspondences established by SuperGlue.

Following Section IV-E.1, we also draw the curves of relative pose estimation accuracy over scale differences of image pairs, as shown in Figure 12. When the scale ratios of image pairs are larger than 4 or smaller than 0.25, our proposed SDAIM and Scale-Net are able to drastically raise the relative pose estimation accuracy of all the above local feature matching methods.

**F. Ablation Study**

In this section, firstly, we confirm the effectiveness of our Covisibility-Attention-Reinforced Matching module, abbreviated as CVARM. Secondly, we demonstrate that resizing both images of an image pair is able to achieve better performance than resizing only a single image in the proposed SDAIM. At last, we compare the proposed SDAIM with a Quasi Inlier based Scale Estimation method, which is detailed as follows: Given two images, \( I_1 \) and \( I_2 \), we build a Gaussian pyramid with 6 scale levels for each image. In each Gaussian pyramid, the first level is the original image. The adopted scale factor is 2. In another word, if the size of the first level is \( H \times W \), the size of the second level is \( H/2 \times W/2 \). Thus, the scale ratio between the 6th level and the 1st level is \( 2^{10} \), which can cover the scale changes in Scale-PT dataset. We use \( l_j (i = 1, 2; j = 1, 2, 3, 4, 5, 6) \) to denote the \( j \)-th level of \( I_j \). Between the Gaussian pyramids of \( I_1 \) and \( I_2 \), these 11 level pairs are considered: \((l_{11}, l_{23})\), \((l_{11}, l_{24})\), \((l_{11}, l_{25})\), \((l_{11}, l_{26})\), \((l_{12}, l_{21})\), \((l_{13}, l_{21})\), \((l_{14}, l_{21})\), \((l_{15}, l_{21})\), \((l_{16}, l_{21})\). Within a level pair, local feature correspondences are estimated by MNN and ratio test. Then RANSAC [47] is used to select out quasi inlier correspondences. We compute the quasi inlier numbers of all the above 11 level pairs and find the level pair with the largest inlier number. If \((l_{1m}, l_{2n})\) has the largest quasi inlier number, the scale ratio between \( I_1 \) and \( I_2 \) is \( 2^{(m-n)} \). If \((l_{11}, l_{2n})\) has the largest quasi inlier number, the scale ratio between \( I_1 \) and \( I_2 \) is \( 2^{(1-n)} \).

1) The Effectiveness of CVARM: We remove \( CAB_1 \) and \( CAB_2 \) in Figure 2(b) and obtain Scale-Net without CVARM, denoted by Scale-Net_plain. And Scale-Net_plain is trained in the same way as Scale-Net_COCO. Then, Scale-Net_plain is evaluated on the test set of Scale-PT dataset and compared with Scale-Net_COCO. Scale-Net_plain and Scale-Net_COCO are used to estimate image scale ratios for SDAIM respectively. (Scale-Net_plain + SDAIM) and (Scale-Net_COCO + SDAIM) are used to boost the relative pose estimation performance of SuperGlue [16]. The evaluation results of Scale-Net_plain and Scale-Net_COCO are shown in Table V. \( E_s \) denotes the scale ratio estimation error, as shown in Equation (11). The mAA\(^{10}\) here is identical with that proposed by Yi et al. [14], [25]. As shown in Table V, Scale-Net_COCO has higher scale ratio estimation accuracy than Scale-Net_plain in the four scenes. And Scale-Net_COCO achieves higher relative pose estimation accuracy than Scale-Net_plain. These results demonstrate that the proposed CVARM is able to improve the scale ratio estimation accuracy of Scale-Net. In Figure 13, four image pairs overlaid with covisibility attention maps, i.e., \( M_1 \) and \( M_2 \) in Figure 2(b), are displayed. Figure 13 shows that the proposed CVARM pays more attention to the visual overlaps. Figure 13 also confirms that the dense feature maps put into CVARM are robust to viewpoint changes.

**TABLE V**

| Networks         | R\(_s\) | mAA\(^{10}\) | R\(_s\) | mAA\(^{10}\) | R\(_s\) | mAA\(^{10}\) | R\(_s\) | mAA\(^{10}\) |
|------------------|---------|-------------|---------|-------------|---------|-------------|---------|-------------|
| Scale-Net_plain  | 0.06    | 0.277       | 1.98    | 0.394       | 2.15    | 0.453       | 2.38    | 0.442       |
| Scale-Net_COCO   | 2.24    | 0.291       | 1.71    | 0.321       | 1.58    | 0.403       | 1.63    | 0.253       |

![Fig. 11. Visualization of the correspondences established by SuperGlue with or without the help of (Scale-Net + SDAIM). Only those correspondences conforming to the ground truth epipolar geometry are drawn. SuperGlue is used to match SuperPoint feature points.](image1)

![Fig. 12. The accuracy rate curves of the estimated relative camera poses of various local feature matching methods. If the angular error is smaller than \(10^\circ\), the estimated rotation matrices and translation vectors are regarded as accurate. Best viewed in color with 400% zoom in.](image2)

![Fig. 13. Visualization of Covisibility Attention maps. There are four image pairs overlaid with covisibility attention maps. (a) and (b) are two image pairs in the test set of Scale-PT dataset. (c) and (d) are two image pairs in ES-HP dataset [38]. Best viewed in color.](image3)
TABLE VI
COMPARATIVE EXPERIMENTAL RESULTS OF UPSAMPLING OR DOWNSAMPLING A SINGLE IMAGE, RESIZING BOTH IMAGES AND USING QUASI INLIER BASED SCALE ESTIMATION METHOD IN SDAIM ON SCALE-PT DATASET

| Features          | SDAIM | nMA (10⁻¹) | Repeatability (%) | Scale Estimation Runtime (ms) |
|-------------------|-------|------------|-------------------|-----------------------------|
|                   |       |            |                   |                             |
|                    |        | 0.979      | 0.84            | 20.71                       |
| Scale-Net + Downsample | 0.250  | 0.893      | 0.58            | 24.71                       |
| SIPT (5)          |       | 0.237      | 0.171           | 0.32            | 24.71                       |
| Scale-Net + Upsample | 0.272  | 0.288      | 0.59            | 24.71                       |
| Inlier + Both     |       | 0.239      | 0.169           | 0.307           | 2020.40                     |
|                   |       | 0.182      | 0.181           | 0                     | 24.71                       |
| Scale-Net + Downsample | 0.344  | 0.282      | 0.352           | 24.71                       |
| SIPT (5)          |       | 0.263      | 0.223           | 0.34            | 24.71                       |
| Scale-Net + Both (Ours) | 0.351  | 0.291      | 0.393           | 24.71                       |
| Inlier + Both     |       | 0.352      | 0.294           | 0.378           | 820.18                      |

2) The Effectiveness of Resizing Both Images: The comparative results on Scale-PT dataset between resizing both images (Ours) and resizing only one image are shown in Table VI. Given two images I₁ and I₂, assume that the scale ratio between I₁ and I₂ is s, s > 1. In Table VI, “Downsample” means only downsampling I₁ to s⁻¹ times its original size. “Upsample” means only upsampling I₂ to s times its original size. “Both” means our method illustrated in Figure 2(a). Our Scale-Net is used to estimate the scale ratios. As shown in Table VI, our method achieves better results than resizing only a single image.

3) Comparison Between Scale-Net and Quasi Inlier Based Scale Estimation Method: In Table VI, “Inlier” means that Quasi Inlier based Scale Estimation method is used to estimate scale ratios in SDAIM. The running time of scale estimation is shown in the last column in Table VI. As shown in Table VI, the running time of Quasi Inlier based Scale Estimation method is much larger than that of Scale-Net. The reason is that feature extraction from all levels of the two Gaussian pyramids and RANSAC verifications between level pairs are time-consuming, as confirmed in Table VII. Moreover, the performance of Quasi Inlier based Scale Estimation method using SIFT features is lower than that of Scale-Net. Quasi Inlier based Scale Estimation method using ALIKE features performs only slightly better than Scale-Net. Thus, our Scale-Net is generally better considering both performance and efficiency.

G. Running Time Evaluation

In this section, we report the running time of the proposed Scale-Net, SDAIM and all steps in several competitive image matching methods mentioned in Section IV-E. The running time is measured on a PC equipped with an 8-core Intel Core i7-9700K CPU (3.6GHz), 64GB of RAM and one NVIDIA GeForce RTX 2080 Ti GPU. The experimental results are listed in Table VII. In Table VII, “Feature” means extracting local features from an image pair. “Matching” means feature matching. “Pose” means pose recovering. / means that the proposed SDAIM is not used. In the last column of Table VII, the ratios between running time change and the original running time without the proposed SDAIM and Scale-Net are marked with brackets. There are 14 rows in Table VII.

TABLE VII
RUNNING TIME [MS] OF SEVERAL REPRESENTATIVE IMAGE MATCHING METHODS ON SCALE-PT DATASET

| Methods          | Scale-Net + Ratio Test | SIFT + Sift-Net + SDAIM | CD-Net + Sift-Net | CD-Net + OANet + SDAIM | CD-Net + SuperGlue + SDAIM | CD-Net + OANet + SIFT | CD-Net + OANet + SIFT-Net |
|------------------|------------------------|-------------------------|------------------|------------------------|-----------------------------|------------------------|--------------------------|
| SIFT (5)         | 24.71                  | 47.58                   | 2.24             | 358.72                 | 23.99                       | 764.52                 | 47.58                     |
| SIFT + Sift-Net + SDAIM | 24.71                  | 11.17                   | 2.24             | 358.72                 | 23.99                       | 764.52                 | 47.58 | 11.17 | 2.24 | 358.72 | 23.99 | 764.52 |
| CD-Net + Sift-Net | 24.71                  | 11.17                   | 2.24             | 358.72                 | 23.99                       | 764.52                 | 47.58 | 11.17 | 2.24 | 358.72 | 23.99 | 764.52 |
| CD-Net + OANet + SDAIM | 24.71                  | 11.17                   | 2.24             | 358.72                 | 23.99                       | 764.52                 | 47.58 | 11.17 | 2.24 | 358.72 | 23.99 | 764.52 |
| CD-Net + SuperGlue + SDAIM | 24.71                  | 11.17                   | 2.24             | 358.72                 | 23.99                       | 764.52                 | 47.58 | 11.17 | 2.24 | 358.72 | 23.99 | 764.52 |
| CD-Net + OANet + SIFT | 24.71                  | 11.17                   | 2.24             | 358.72                 | 23.99                       | 764.52                 | 47.58 | 11.17 | 2.24 | 358.72 | 23.99 | 764.52 |
| CD-Net + OANet + SIFT-Net | 24.71                  | 11.17                   | 2.24             | 358.72                 | 23.99                       | 764.52                 | 47.58 | 11.17 | 2.24 | 358.72 | 23.99 | 764.52 |

As shown in Table VII, SDAIM does not always increase the running time. It reduces the running time of SIFT + MNN + Ratio test, CD-UNet and SIFT + MNN + OANet. The reasons are as follows: The proposed SDAIM can increase the inlier ratios of matches output by the above three methods. Thus, much less sampling loops are needed to achieve termination conditions in RANSAC under the help of SDAIM.

H. Evaluation on Original Hpatches Dataset

In this section, we evaluate the proposed Scale-Net and SDAIM on original Hpatches Dataset [58]. Scale-Net_COCO is used to estimate the scale ratios of image pairs. 5K feature points are extracted from every image. Evaluation results are shown in Table VIII. The thresholds of PCK and repeatability here are identical with those in Section IV-D. Most image pairs in original Hpatches Dataset only contain small or even no scale changes. Table VIII shows that Scale-Net and SDAIM only result in slight fluctuations when there are small or even...
no scale changes. Scale-Net and SDAIM are useful under the circumstances of considerable scale changes, which is confirmed in Figure 10 and Figure 12.

V. CONCLUSION

In this paper, we propose a Scale-Difference-Aware Image Matching (SDAIM) method that is able to greatly improve the performance of local features under large scale changes in images. Given an image pair, firstly, the proposed SDAIM estimates the scale ratio of this image pair. Secondly, it resizes both images to drastically reduce the scale difference within the image pair before local feature extraction so that it can still establish sufficient inlier correspondences for downstream tasks when facing large scale changes. We also propose a novel neural network, termed as Scale-Net, based on our proposed Covisibility-Attention-Reinforced Matching Module (CVARM), to accurately estimate the scale ratios of image pairs for the proposed SDAIM. Rigorous evaluations on image matching and relative pose estimation tasks have demonstrated the remarkable performance of SDAIM, the good generalization ability and the high scale ratio estimation accuracy of Scale-Net. In the future, we plan to enable the proposed Scale-Net to estimate local scale ratios.

REFERENCES

[1] C. Toft et al., “Long-term visual localization revisited,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 4, pp. 2074–2088, Apr. 2022.
[2] P. Zhang, C. Zhang, B. Liu, and Y. Wu, “Leveraging local and global descriptors in parallel to search correspondences for visual localization,” Pattern Recognit., vol. 122, Feb. 2022, Art. no. 108344.
[3] X. Guo, S. Shen, L. Zhu, T. Shi, Z. Wang, and Z. Hu, “Complete scene reconstruction by merging images and laser scans,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 10, pp. 3688–3701, Oct. 2020.
[4] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multitrip slam,” IEEE Trans. Robot., vol. 37, no. 6, pp. 1874–1890, Dec. 2021.
[5] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Feb. 2004.
[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 2564–2571.
[7] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “LIFT: Learned invariant feature transform,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 467–483.
[8] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, “LF-Net: Learning local features from images,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2018, pp. 1–11.
[9] 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. 224–236.
[10] M. Dusmanu et al., “D2-Net: A trainable CNN for joint detection and detection of local features,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8092–8101.
[11] J. Revuǎ, C. De Souza, M. Humenberger, and P. Weinzaepfel, “R2D2: Reliable and repeatable detector and descriptor,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2019, pp. 12405–12415.
[12] Z. Luo et al., “ASLFeat: Learning local features of accurate shape and localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6589–6598.
[13] L. Cavalli, V. Larsson, M. R. Osvald, T. Sattler, and M. Pollefeys, “Handcrafted outlier detection revisited,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 770–787.
[14] 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.
[41] J. Ye, S. Zhang, T. Huang, and Y. Rui, “CDbin: Compact discriminative binary descriptor learned with efficient neural network,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 3, pp. 862–874, Mar. 2020.

[42] H. Pan, Y. Chen, Z. He, F. Meng, and N. Fan, “TCDesc: Learning topology consistent descriptors for image matching,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 5, pp. 2845–2855, May 2022.

[43] X. Shen et al., “RF-net: An end-to-end image matching network based on receptive field,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8132–8140.

[44] M. Tyszkiewicz, P. Fua, and E. Trulls, “DLKS: Learning local features with policy gradient,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 33, 2020, pp. 14254–14265.

[45] X. Zhao, X. Wu, J. Miao, W. Chen, P. C. Y. Chen, and Z. Li, “ALIKE: Accurate and lightweight keypoint detection and descriptor extraction,” IEEE Trans. Multimedia, early access, Mar. 3, 2022, doi: 10.1109/TMM.2022.315927.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[47] M. A. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” Commun. ACM, vol. 24, no. 6, pp. 381–395, 1981.

[48] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, “USAC: A universal framework for random sample consensus,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 2022–2038, Aug. 2013.

[49] D. Barath, J. Matas, and J. Noskova, “MAGSAC: Marginalizing sample consensus,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10197–10205.

[50] D. Barath and J. Matas, “Graph-cut RANSAC,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6733–6741.

[51] X. Zhao, J. Liu, X. Wu, W. Chen, F. Guo, and Z. Li, “Probabilistic spatial distribution prior based attentional keypoints matching network,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1313–1327, Mar. 2022.

[52] S. Liu, H. Wang, Y. Wei, and C. Pan, “BB-homography: Joint binary features and bipartite graph matching for homography estimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 2, pp. 239–250, Feb. 2015.

[53] O. Wiles, S. Ehrhardt, and A. Zisserman, “Co-attention for conditioned image matching,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 15920–15929.

[54] I. Rocco, R. Arandjelovic, and J. Sivic, “Convolutional neural network architecture for geometric matching,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 11, pp. 2553–2567, Nov. 2019.

[55] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 740–755.

[56] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[57] Z. Li and N. Snavely, “MegaDepth: Learning single-view depth pre-