BEDOI: BENCHMARKS FOR DETERMINING OVERLAPPING IMAGES WITH PHOTOGRAMMETRIC INFORMATION

H. Zhan [a], Y.F. Yu [a], Y.W. Xu [a], Q.B. Hou [a], R., Xia [a], X. Wang [a] *, Y. Feng [b], Z.Q. Zhan [a], M.L. Li [c], M. Gruber [d], R. Hänsch [e], C. Heipke [f]

[a] School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430072, People's Republic of China
(zhanhao2020, yfyu2020,xywjohn_SGG2020,qianbao-hou)@whu.edu.cn, (xwang,zqzhan)@sgg.whu.edu.cn
[b] Chair of Cartography and Visual Analytics, Technical University of Munich, Germany – y.feng@tum.de
[c] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, People's Republic of China – minglei_li@nuaa.edu.cn
[d] Vexcel Imaging GmbH, Austria – michael.gruber@vexcel-imaging.com
[e] Microwaves and Radar Institute, German Aerospace Center (DLR), Germany – rww.haensch@gmail.com
[f] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany – Heipke@ipi.uni-hannover.de

KEY WORDS: Structure from Motion (SfM), Image Matching, Overlapping image pairs, Image Retrieval, Photogrammetric Information

ABSTRACT: For conventional SfM pipeline, image matching is enduring limitation when considering the time efficiency. In the last few years, to speed up image matching procedure, many image retrieval works were proposed to fast find overlapping image pairs, e.g., bag-of-word that clusters hand-crafted local features in a hierarchical way for efficient similar image retrieval, or learning-based global features (such as, VGG or ResNet) are used to represent image in a global compact manner. However, there are rarely benchmarks with referenced overlapping information to: first, evaluate the retrieval performance; second, fine tune deep-learning models along the direction that is more capable to deal with overlapping image pairs. In this work, based on traditional photogrammetric procedures, relevant photogrammetric information is obtained including image orientation parameters, 3D mesh model and etc., we then generate a benchmark for determining Overlapping Images - BeDOI, in which referenced pairwise overlapping relationships are estimated via rigorous photogrammetric geometry. To extend the generality, in total, BeDOI contains 13667 images which are basically UAV and close-range images of various scene categories, e.g., urban cities, campus, village, historical relics, green land, buildings and etc. Lastly, to demonstrate the efficacy of the proposed BeDOI, several image retrieval methods are tested and the experimental results are reported as a competition challenge[1].

1. INTRODUCTION

Thanks to the development of sensor and other relevant technologies, image acquisition becomes easier and lower-cost. This results in increasing demands on Structure from Motion (SfM) pipelines regarding the processing of large-scale image datasets. In general, SfM mainly includes two steps with high computational costs (Wang et al., 2019): Image matching and image orientation. Image matching typically requires more time, especially for datasets with a large number of input images.

To accelerate and lighten image matching, visually similar images are first time-efficiently identified as overlapping image pairs via image retrieval techniques. The extracted local features are then matched to generate correspondences. Currently, there are two main categories of methods: First, building an efficient indexing structure. The most popular solution is the so-called BoW method (Bag-of-Word, Nistér and Stewenius, 2006) and its variants (Havlena and Schindler, 2014; Schönberger et al., 2016; Zhan et al., 2017). The key idea is to train a hierarchical tree structure with hand-crafted local features (SIFT (Lowe, 2004), ORB (Rublee et al., 2011)) using unsupervised clustering algorithms. Wang et al. (2017 and 2019) build a random kd-forest with SIFT descriptors where the pairwise similarity is based on the Euclidean distance of neighbouring features. Approaches of the second category extract learning-based compact global features. Convolutional Neural Networks (CNN) have shown excellent performance in many computer vision tasks, such as image classification, object detection, semantic segmentation, etc. Tolias et al. (2016) and Radenovic et al. (2016) employed the feature maps of several pre-trained CNN architectures to yield a compact global feature. Similar image pairs are identified by investigating the distance of two images in the global latent feature space. However, several limitations exist that are yet to receive the necessary attention. There are barely any benchmark dataset that provide geometrically correct overlap information. Most of the retrieval methods are evaluated based on manually annotated (Philbin et al., 2008) or geospatial referenced benchmarks (Arandjelović et al., 2016) which potentially contain incorrect similar images due to the subjective judgement or wrong geospatial labels. SfM results are often analysed to indirectly demonstrate the effectiveness of overlapping image pairs methods. In addition, many of the leveraged pre-trained CNN models were trained using ImageNet (Deng et al., 2009) and predict the semantic category of an object present in the image. This significantly differs from the task of detecting overlapping image pairs in the context of SfM or image orientation where the goal is to identify two images which partially cover the same area of the 3D object space and should share similar geometric characteristics (Hou et al., 2023). Therefore, a domain gap between overlapping image pairs and image classification exists.

To cope with scarce benchmarks and the mentioned domain gap, this paper makes two main contributions: First, in line with classical photogrammetric procedures, we provide a benchmark with geometrically correct references of overlapping image relationships - BeDOI, including 13,667 images of several different content (such as urban buildings, countryside, forest, etc). It cannot only be applied for evaluating performance of relevant overlapping image pairs retrieval algorithms, but also cast as training data for learning-based global feature extractors to boost the sensitivity for pairwise overlapping information. Second, several popular image retrieval methods are explored, including SIFT-based KD-forest (Wang et al., 2019), learning-based global features (VGG, ResNet), as well as learning-based

* Corresponding author

[1] The competition challenge and datasets can be online accessed via https://github.com/WHUHaoZhan/BeDOI

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1685-2023 | © Author(s) 2023. CC BY 4.0 License.

image matching mechanism (SuperGlue, Sarlin et al., 2020), whose retrieval results are extensively studied based on the generated benchmark.

2. RELATED WORK

The identification of overlapping image pairs is analogous to an image retrieval problem as images with overlapping regions typically look similar and mostly aims at accelerating image matching for SfM or image orientation. This section reviews three related aspects, i.e., conventional methods with local features (including both hand-crafted and learning-based solutions), CNN-based methods with global features, and relevant benchmarks.

Conventional methods with local features. Since the emergence of canonical handcrafted local features (such as SIFT, SURF, ORB, etc.), they are widely used in image localization (Li et al., 2012), SfM (Schonberger and Frahm, 2016; Zhu et al., 2018), SLAM (Engel et al., 2014; Mur-Artal et al., 2015), etc. However, most of these handcrafted features fail to deal with large baseline stereo pairs, weak texture, etc. DeTone et al. (2018) proposed an end-to-end and self-supervised feature detection descriptor estimation method using simulated training data, in which the encoding layer is followed by two decoding layers for feature detection and description. Recently, Chen and Heipke (2022) presented an architecture to predict the affine transform between local feature patch pairs enabling to learn a more discriminative descriptor for increased matching performance. A more comprehensive review of local feature methods is presented in Chen et al. (2021).

Given extracted local features, many works were proposed to efficiently handle the matching problem. One standard approach is to obtain matched features from two images using approximate nearest neighbours (ANN) based on a well-designed indexing structure, such as k-d tree or random k-d forest (Arya et al., 1998; Muja and Lowe, 2014).

To further speed up the retrieval, leveraging the fact that local features that are similar could be simplified by a compact representation, various BoW models cluster local features of the entire image and form one single compact descriptor, which are commonly used in many SfM (Schonberger and Frahm, 2016; Zhu et al., 2018) and SLAM (Mur-Artal et al., 2015). Then, Fisher Vectors (Perronnin et al., 2010) and VLAD (Jégou et al., 2012) works focus on decreasing quantization errors, reducing memory requirements, and increasing retrieval efficiency. Havlena and Schindler (2014) proposed VocMatch, a 2-layer vocabulary tree codebook is built with the first and second layer consisting of 4096 and 4096×4096 sub-clusters, respectively. In principle, the features which are quantized into the same vocabulary are matchable points. This also means if two cluster centres are very close to each other, the results may be ambiguous. In contrast to the hand-designed matching heuristics, image matching can also be implemented in a supervised fashion via a matching network, e.g., proposed by Sarlin et al. (2019), the so-called SuperGlue. It takes the position and visual descriptor of the extracted local features (both handcrafted and learning-based features can be used) as input. The contextual and positional information are considered via an attentional graph neural network and matching points are generated by using a differentiable partial assignment solution, i.e., the Sinkhorn algorithm (Luise et al., 2018). Real-time matching performance can be achieved using GPUs, but the designed architecture can only deal with local features from two images, which means SuperGlue must be run $N(N-1)/2$ times for N unordered images.

CNN-based methods with global features. A variety of studies applied CNN activations to the task of image retrieval. The gained superior achievement demonstrates their corresponding capability. There are mainly two categories of methods depending on whether pre-trained CNN models are used directly as they are or their parameters are fine-tuned for the given task. Based on off-the-shelf models, various works aim at aggregating CNN activations for improving the discrimination of global compact features. Two common strategies – sum/average pooling and max pooling are widely used to aggregate CNN feature maps. For example, Razavian et al. (2016) attempts to perform spatial max pooling on the feature maps of an off-the-shelf CNN model. Babenko et al. (2014) proposes sum-pooling convolutional features (SPoC) utilising a Gaussian centre prior to obtain compact descriptors. An alternative idea is to pool some local regions in an activation feature map (Babenko et al., 2015), which is identical to R-MAC (Tolias et al., 2016). In addition to the convolutional layers, Gong et al. (2014) investigated the use of Fully Connected (FC) layer activations, whereas subsequent studies (Tolias et al., 2016) showed that FC layers are typically inferior to using CNN layers alone for image retrieval. To generate better retrieval results, the original CNN model is fine-tuned and updated according to specific retrieval tasks. Radenovic et al. (2016) fine-tunes the convolution layer of AlexNet and VGG according to the sparse SfM reconstruction results generated via BoW retrieval results. Siamese networks and contrastive loss are adopted by considering both matched image pairs and non-matched image pairs. Motivated by the discrete feature embedding VLAD (Jégou et al., 2012), Arandjelović et al. (2016) train a differentiable pooling layer (together with several convolutional layers), namely NetVLAD that approximates the inherent discreteness of VLAD by a soft assignment. Triplet loss is used to take care of the influence of positive and negative training image pairs, which are obtained by geo-tagged information.

Relevant benchmarks. In the computer vision field, many famous benchmarks have been established for evaluating the performance of image retrieval (Please note that while there are ample image retrieval benchmarks worth reviewing, this review section only lists a few popular and relevant works). Oxford5K (Philbin et al., 2007) and Pairs6K (Philbin et al., 2008) are the two of the most well-known examples. However, they contain false positives and false negatives due to incorrect annotations. Radenović et al. (2018) purify these two datasets using more manually corrections and publish two refined benchmarks of $\mathcal{R}$Oxford5K and $\mathcal{R}$Pairs6K, in which images of three different retrieval difficulties are suggested. Zheng et al. (2020) propose University-1625 which provides ground-truth relationships among images from various sources, i.e., satellite, drone-based and ground images. Their goal is to geospatially localise drones via cross-view image retrieval. To the best of our knowledge, the proposed BeDOI is similar to GL3D (Shen et al., 2018) and LOIP (Hou et al., 2023), but we make two extensions: First, more images with extra regions are included; Second, not only the overlapping relationships are provided, but also the estimated similarities of every image pair and the relevant photogrammetric information (including orientation parameters, 3D mesh models and etc.) are available. More detailed comparison can be found in Section 3.3.

3. BEDOI GENERATION WITH REFERENCED OVERLAPPING RELATIONSHIPS

In this section, we first give an overview introduction of our BeDOI dataset. Then, the automatic procedure for generating

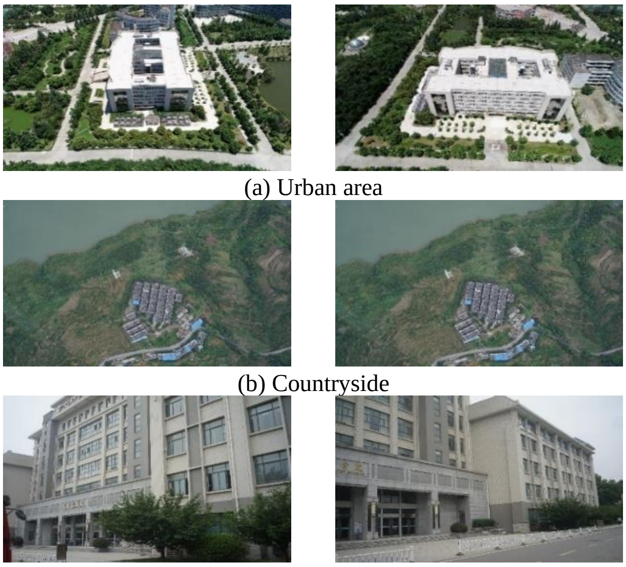
BeDOI is explained. Finally, we compare several relevant benchmarks.

3.1 Introduction of BeDOI

In general, BeDOI is composed of 11 high-resolution image datasets, including UAV images captured via a nadir camera and oblique photogrammetric images with multiple cameras, as well as manually self-collected close-range images with different overlap degrees, which is tailored for overlapping image pair identification on photogrammetric image datasets. More specifically, as Tab.1 lists, in total, 13,667 images covering various categories of areas are collected, such as urban buildings, woodland, countryside, scenic spots, etc. Fig. 1 shows several examples.

Name	Image Num.	Source	Category
SKFX	60	Close range	Historic Relics
GB	68	UAV	Scenic Spot
GRAZ	250	Oblique	Urban City
YD	374	UAV	Scenic Spot
NH	606	UAV	Building
TZH	1060	UAV	Countryside
SXKQ	1185	UAV	Forest
JYYL	1429	Close range	Building
XHSD	2133	Oblique	Urban City
WHU	2652	UAV	University
SHHY	3850	Oblique	Village
BeDOI	13667	Multi-sources	Multi-categories

Table 1. Information of each dataset in BeDOI.


(a) Urban area

(b) Countryside

(c) High-resolution images of ground building

Figure 1. Example images of BeDOI.

3.2 Automatic annotation for generating BeDOI

The overall pipeline to automatically generate BeDOI is illustrated in Fig. 2, in which pre-processing is for obtaining 3D mesh model and image orientation parameters, and automatic annotation is for estimating referenced overlapping relationships:

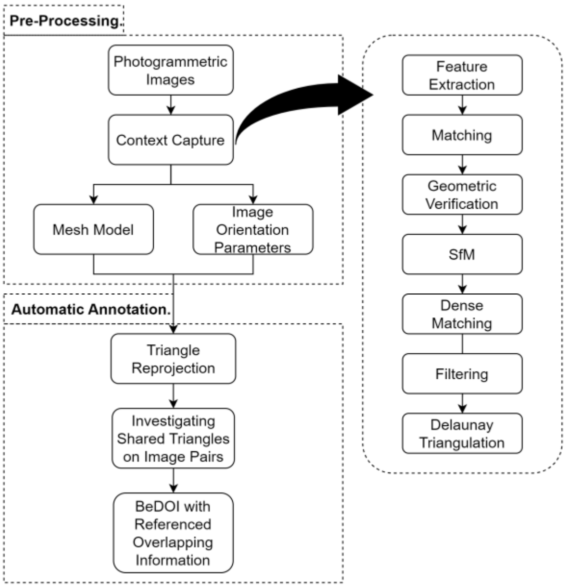


Figure 2. Flowchart of BeDOI generation.

Pre-Processing. Given a set of collected images, this step is to generate corresponding photogrammetric information, i.e., 3D mesh models and image orientation parameters. Following the canonical photogrammetric processing, several consecutive procedures are required: feature extraction and matching, SfM, stereo dense matching and multi-view fusion, filtering and 3D mesh construction (including Delaunay triangulation, texture re-organization etc.). Note that orientation information is computed after SfM. This BeDOI processing chain might seem counterintuitive since image matching is usually completed before the 3D mesh model is built. However, leveraging a 3D mesh model for identifying real overlapping image pairs is not only a viable but also highly advantageous solution, as most local features are typically not invariant to large view angle change, e.g., oblique images. Such a procedure is beneficial even for state-of-the-art learning-based local feature extractors can only slight improve the matching performance (Yi et al., 2016). This motivates us to explore 3D mesh models for estimating correct overlapping information in a geometrically rigorous manner[2]. One sample mesh model of JYYL is shown in Fig. 3.

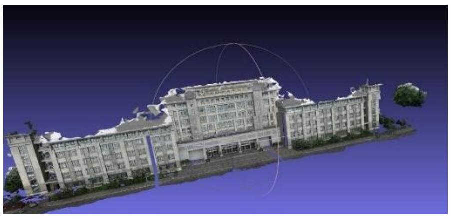


Figure 3. 3D mesh model of JYYL.

Automatic Annotation. Based on the collinearity equation, we present an automatic annotation method for geometrically correct referenced overlapping image pairs using the generated 3D mesh model and image orientation parameters. The basic idea is to reproject every triangle on every image. Shared triangles between two images are explored for determining the corresponding overlapping degree. The more common reprojected triangles, the larger the corresponding overlapping area will be. To estimate accurate triangle reprojection, it is necessary to deal with

[2] Mind that we select the professional commercial software for our pre-processing, i.e., ContextCapture.

occlusions. Fig. 4 shows that there are many incorrectly identified overlapping areas without occlusion detection which can lead to incorrect results in BeDOI. In this work, occlusion is detected by the number of triangles that the corresponding ray (from the camera center to the center of the target triangle) passes through. No occlusion happens if and only if the number is zero. Furthermore, in order to enhance occlusion detection speed, we construct an AABB tree for the mesh model. After the occlusion detection, the correct triangle information of the image can be obtained.

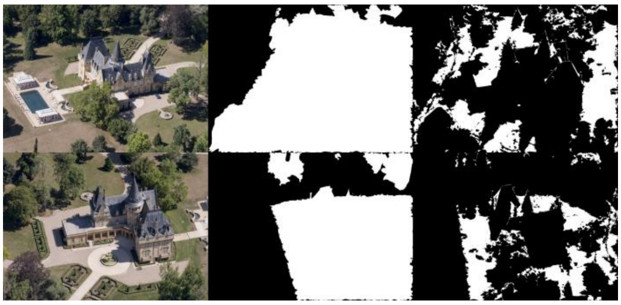
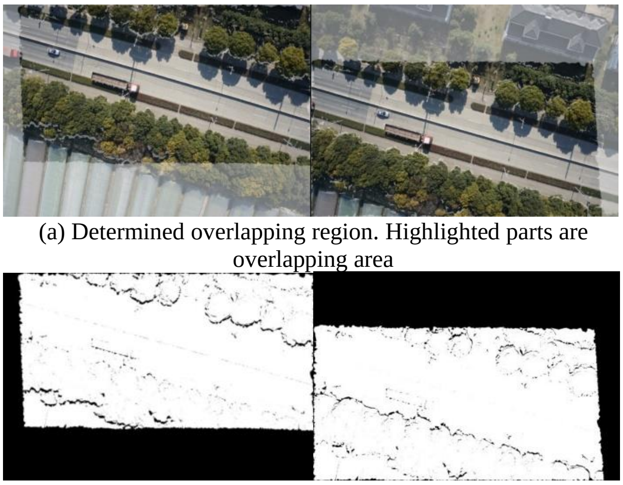


Figure 4. With occlusion (WO) vs. Without occlusion (WoO). White pixels indicate the overlapping area via the proposed triangle reprojections.

After triangle reprojection and occlusion detection, the similar or overlapping degree of image pair (i, j) can be computed as follows:

$$OI_{ij} = \sqrt{\frac{|TR(i) \cap TR(j)|_n}{|TR(i)|_n} \cdot \frac{|TR(i) \cap TR(j)|_n}{|TR(j)|_n}} \qquad (1)$$

where $|.|_n$ returns the number of triangles, $TR(\mathtt{i}) \cap TR(\mathrm{j})$ represents the set of triangles that can be observed in both image i and j. Straightforwardly, the larger the value of OI_{ij} is, the more similar the image pair (i, j) is. Based on the conventional photogrammetric regularity, image pairs i and j can be identified as overlapping if OI_{ij} values exceed 0.3. Fig. 5 qualitatively shows the determined overlapping region, where the highlighted part in Fig. 5(a) is the overlapping area of the two images, Fig. 5(b) is a binary image with the white region corresponding to the highlighted area Fig. 5(a).



(a) Determined overlapping region. Highlighted parts are overlapping area

(b) Binary results of overlapping region. White regions indicate overlapping area.

Figure 5. Qualitative results of determined overlapping region.

Ultimately, the overlap or similarity degree among all image pairs can be calculated by equation (1). In this paper, we sorted the values of OI_{ij} in descending order based on the number of overlapping patches. For a binary classification, image pairs with OI_{ij} values exceeding 0.3 are the referenced overlapping ones.

3.3 Comparison with relevant benchmarks

This section reviews several popular image retrieval benchmarks, i.e., Oxford5k, Paris6k, GL3D, LOIP. The first two datasets are widely used to evaluate the result of image retrieval algorithms in the computer vision field, whereas, the other two are typically used for finding overlapping image pairs, which are partially identical with our work. A detailed comparison of these 4 public datasets and BeDOI is listed in Tab. 2, mainly including the source of images (SoI), Strong semantics (SS), mesh models (MM), High resolution (HR), dense reconstruction (DR), overlapping degree rank (ODR), the number of images (NoI).

	BeDOI	Oxford5k	Paris6k	LOIP	GL3D
SoI	Pg.	Cs.	Cs.	Cs.	Pg.
SS	×	√	√	√	×
DR	√	×	×	×	√
MM	√	×	×	×	√
HR	√	×	×	√	√
ODR	√	×	×	×	×
NoI	13.6K	5k	6.3k	1.8k	90.5k

Cs. = Crowdsourced Images, Pg. = Photogrammetric Images

Table 2. Comparison of several public datasets and BeDOI.

In the proposed BeDOI, 80% of the images are UAV images, and the rest are close-range images taken by digital cameras. Both groups contain various degrees of overlap. Different from Oxford5k and Paris6K, BeDOI does not address a semantic task The collected images are required to be geometrically oriented and only contain weak semantic content corresponding to various categories. Similar to LOIP and GL3D, we provide references of overlapping image pairs. However, BeDOI offers several extensions: First, in contrast to GL3D, all the orientation parameters and 3D mesh models are provided which can be used for other tasks (such as image orientation, 3D mesh model generation, etc.). Second, based on LIOP, another six datasets of various regions are included. Furthermore, besides the overlapping relationship, the overlapping or similar degree values are also included, which are supposed to be beneficial for training better learning-based global feature models. Fig. 6 gives a qualitative overview of each dataset in the generated BeDOI, including sample images, overlapping or similarity degree values and 3D mesh model.

4. EXPERIEMNTS AND EVALUATIONS

4.1 Experimental Settings

This section verifies the efficacy of the generated BeDOI. We conduct overlapping image pair identification experiments and the corresponding retrieval performance is reported. In particular, 100 sample images were randomly selected as queries, whose referenced overlapping relationships are inherently available in BeDOI. Then, Top-N similar images for these selected 100 query images are found by four popular image retrieval methods, which are as follows:

VGG-16. VGG series have been widely used as backbone in many various computer vision tasks. In this paper, VGG-16

(Simonyan and Zisserman, 2014) with 13 convolutional layers is employed, specifically, we utilize the output of the last max pooling layer as the learning-based global image feature, whose feature descriptor dimension is 512×7×7 = 25,088. The similarity degree of two images is then calculated by Euclidean distance of VGG-16 feature descriptors.

ResNet-18. Another popular CNN-based global feature, ResNet-18 (He et al., 2016), is investigated as well. Similar to VGG-16, we extract global image features via the last average pooling layer of ResNet-18, whose dimension is 512. The similarity degree of two images is also estimated by Euclidean distance of ResNet-18 global feature descriptors.

SuperGlue is one of the state-of-the-art image matching methods, which was demonstrated to be able to provide accurate correspondences in real time. SuperGlue applies a graph neural network to predict a matching score for each feature. The best matched features are selected based on the matching scores.

In this experiment, 1024 superpoints are extracted for each image, SuperGlue is used to match all potential pairs and the similarity degrees are computed by summing the scores of all matched features.

Random k-d Forest is an efficient indexing structure with several independent k-d trees. In this paper, the input number of SIFT features for building k-d forest is 1000 per image, and four k-d trees are built for efficient retrieval. According to Wang et al. (2019), the similarity measuring any two images is estimated by equation (2)

$$S_{ij} = log_{10}P_{ij} \times \left(\frac{1}{e}\right)^{D_{ij}} \tag{2}$$

where, S_{ij} is the similarity measure between two images, P_{ij} is the number of the neighboring features from an image pair (i, j). D_{ij} is the average Euclidean distance between all the corresponding neighboring features.

Note that we use a pre-trained model provided by the corresponding authors for all the employed learning-based architectures. All the experiments were tested on a machine with ten 3.7Hz Intel Core i9-10900X processors and dual GPUs of GTX1080ti.

Evaluation Metrics. The retrieval performance is quantitatively evaluated by precision and recall. Based on the estimated similarity degrees, for the selected 100 sample query images, Top-N (Top-5, 10, 20, 30 40, 50, 100) similar images are determined from BeDOI, the averaging precision and recall are investigated. Additionally, the time efficiency is also compared for these four methods.

4.2 Results

	Random k-d Forest	SuperGlue	ResNet-18	VGG-16
Top-5	0.861	**0.907**	0.780	0.334
Top-10	0.766	**0.828**	0.658	0.265
Top-20	0.652	**0.736**	0.552	0.190
Top-30	0.577	**0.691**	0.497	0.154
Top-40	0.540	**0.661**	0.469	0.140
Top-50	0.513	**0.648**	0.457	0.130
Top-100	0.457	**0.630**	0.430	0.101

Table 3. Average Precision.

Tab. 3 lists the average precision values of various Top-N results from four investigated methods. It can be easily found that as more top similar images are considered, the precision tends to decrease which means more false positives are found. This can be explained by the fact that due to the limited representation capability of the corresponding feature descriptors; incorrect similar images are often more likely to be found when considering more candidate similar images. In addition, methods using CNN-based global features are typically inferior to that using local features. This can be expected, as the applied CNN models are pre-trained on ImageNet benchmarks which results in the extracted global features not sensitive to overlapping information.

	Random k-d Forest	SuperGlue	ResNet-18	VGG-16
Top-5	0.146	**0.163**	0.126	0.026
Top-10	0.221	**0.255**	0.175	0.037
Top-20	0.294	**0.356**	0.232	0.047
Top-30	0.321	**0.405**	0.261	0.051
Top-40	0.338	**0.433**	0.279	0.057
Top-50	0.347	**0.451**	0.292	0.060
Top-100	0.375	**0.511**	0.334	0.067

Table 4. Average Recall.

Tab. 4 provides the average recall results. In contrast to Tab. 3, recall values have an increasing tendency as more similar candidate images are retrieved. This is due to the fact that more true positives can be established if Top-N becomes larger. When comparing these four methods, a similar conclusion can be drawn as Tab. 3, Random k-d forest and SuperGlue are always superior to ResNet-18 and VGG-16. Looking into the magnitude of recalls, even the best top-100 obtain just 0.511. This is because the reported recall value is averaged on the selected 100 query images and some of them have nearly 300 referenced overlapping images whose recall values are just around 0.333 even if the corresponding Top-100 precision value is 1.0.

	Random k-d Forest	SuperGlue	ResNet-18	VGG-16
Time(s)	0.076	0.05	**0.007**	0.013

Table 5. Cost time (in second) for one image pair.

The cost time of estimating similarity degree for one image pair is shown in Tab. 5, ResNet-18 and VGG-16 are typically faster than Random k-d forest and SuperGlue. The former two methods extract a global feature for one image for calculating similarity degree and the other two need more computations for dealing with numbers of high-dimensional local features.

Synthesis. In general, learning-based global features can provide a faster image retrieval solution than local features can, but due to the domain gap between overlapping image pair identification and semantic object image retrieval, the performance of precision and recall is typically worse. This naturally motivates the possibility of using BeDOI as an extra training dataset to improve learning-based global features.

5. CONCLUSIONS

In this paper, we introduce a benchmark with referenced overlapping relationships of potential image pairs - BeDOI. In general, BeDOI consists of 11 datasets with 13,667 images in total. Various categories of scenes are included, i.e., forest, urban cities, countryside, buildings, and etc. The canonical photogrammetric processing is employed to obtain 3D mesh model and image orientation parameters, which are then used for estimating geometrically correct overlapping degrees for every image pair.

Our BeDOI is supposed to be a very beneficial extension benchmark for evaluating image retrieval or overlapping image determination methods. Furthermore, it can also be utilized as a training dataset for fine-tuning learning-based global feature extractors. In the next step, we would like to carry out corresponding investigations on several backbone architectures, e.g., ViT (Dosovitskiy et al. 2021) or SwinT (Liu et al. 2021).

ACKNOWLEDGEMENTS

This work was supported by ISPRS Scientific Initiatives 2023, Natural Science Foundation of Hubei Province, China (No. 2022CFB727) and National Natural Science Foundation of China (No. 61871295).

REFERENCES

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5297-5307.

Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y., 1998. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. Journal of the ACM, 45(6), pp. 891-923.

Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V., 2014. Neural codes for image retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 584-599.

Babenko, A., Lempitsky, V., 2015. Aggregating local deep features for image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1269-1277.

Chen, L., Heipke, C., 2022. Deep learning feature representation for image matching under large viewpoint and viewing direction change. ISPRS Journal of Photogrammetry & Remote Sensing, 190, pp. 94-112.

Chen, L., Rottensteiner, R., Heipke, C., 2021. Feature detection and description for image matching: from hand-crafted design to deep learning. Geo-spatial Information Science, 24(1), pp. 58-74.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.F., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 248-255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., etc., 2021. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In arXiv preprint, arXiv:2010.11929.

Gong, Y., Wang, L., Guo, R., Lazebnik, S., 2014. Multi-scale order less pooling of deep convolutional activation features. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 392-407.

Havlena, M., Schindler, K., 2014. VocMatch: Efficient Multiview Correspondence for Structure from Motion. In: Proceedings of the European Conference on Computer Vision, 46-60.

He, K.M., Zhang, X. Y., Ren, S.Q., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

Hou, Q.B., Xia R., Zhang, J.H., Feng, Y., Zhan, Z.Q., Wang, X., 2023. Learning visual overlapping image pairs for SfM via CNN fine-tuning with photogrammetry geometry information. International Journal of Applied Earth Observation and Geoinformation, 116:103162.

Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C., 2012. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9), pp. 1704-1716.

Krizhevsky, A., Sutskever, I. Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the Neural Information Processing Systems (NeurIPS), pp. 1097-1105.

Li, Y.P., Snavely, N., Huttenlocher, D., Fua, P., 2012. Worldwide Pose Estimation Using 3D Point Clouds. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 15-29.

Liu, Z., Lin, Y., Cao, Y., etc., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 10012–10022.

Lowe, D.G., 2004. Distinctive Image Features from Scale invariant Keypoints. International Journal of Computer Vision, 60(2), 91–110.

Luise, G., Rudi, A., Pontil, M., Ciliberto, C., 2018. Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance. In: Proceedings of the Neural Information Processing Systems (NeurIPS), pp.5859-5870.

Muja, M., Lowe, D. G., 2014. Scalable nearest neighbor algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, pp. 2227-2240.

Mur-Artal, R., Montiel, J.M.M., Tardos, J.D., 2015. Orb-slam: a versatile and accurate monocular slam system. IEEE Transactions on Robotics, 31(5), pp. 1147-1163.

Nistér, D., Stewenius, H., 2006. Scalable Recognition with a Vocabulary Tree. In: Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 2161-2168.

Perronnin, F., Liu, Y., Sanchez, J., Poirier, H., 2010. Large-scale image retrieval with compressed Fisher vectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 245-256.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2007. Object retrieval withlarge vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Radenović, F., Tolias, G., Chum, O., 2016. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard example. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-20.

Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O., 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), pp. 5706-5715.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. In: Proceedings of International Conference on Computer Vision, pp. 2564-257.

Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4937-4946.

Schönberger, J.L., Price, T., Sattler, T., Frahm, J.M., Pollefeys, M., 2016. A Vote-and-Verify Strategy for Fast Spatial Verification in Image Retrieval. In: Proceedings of the Asian Conference on Computer Vision (ACCV), pp. 321-337.

Schönberger, J.L., Frahm, J.M., 2016. Structure-from-Motion Revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104-4113.

Shen, T.W., Luo, Z.X., Zhou, L., Zhang, R.Z., Zhu, S.Y., Fang, T., Quan, L.,2018. Matchable Image Retrieval by Learning from Surface Reconstruction. In: Proceedings of the Asian Conference on Computer Vision (ACCV), pp. 415-431.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: arXiv preprint arXiv:1409.1556

Tolias, G., Sicre, R., Jégou, H., 2016. Particular object retrieval with integral max-pooling of CNN activations. In: Proceedings of the International Conference on Learning Representations (ICLR).

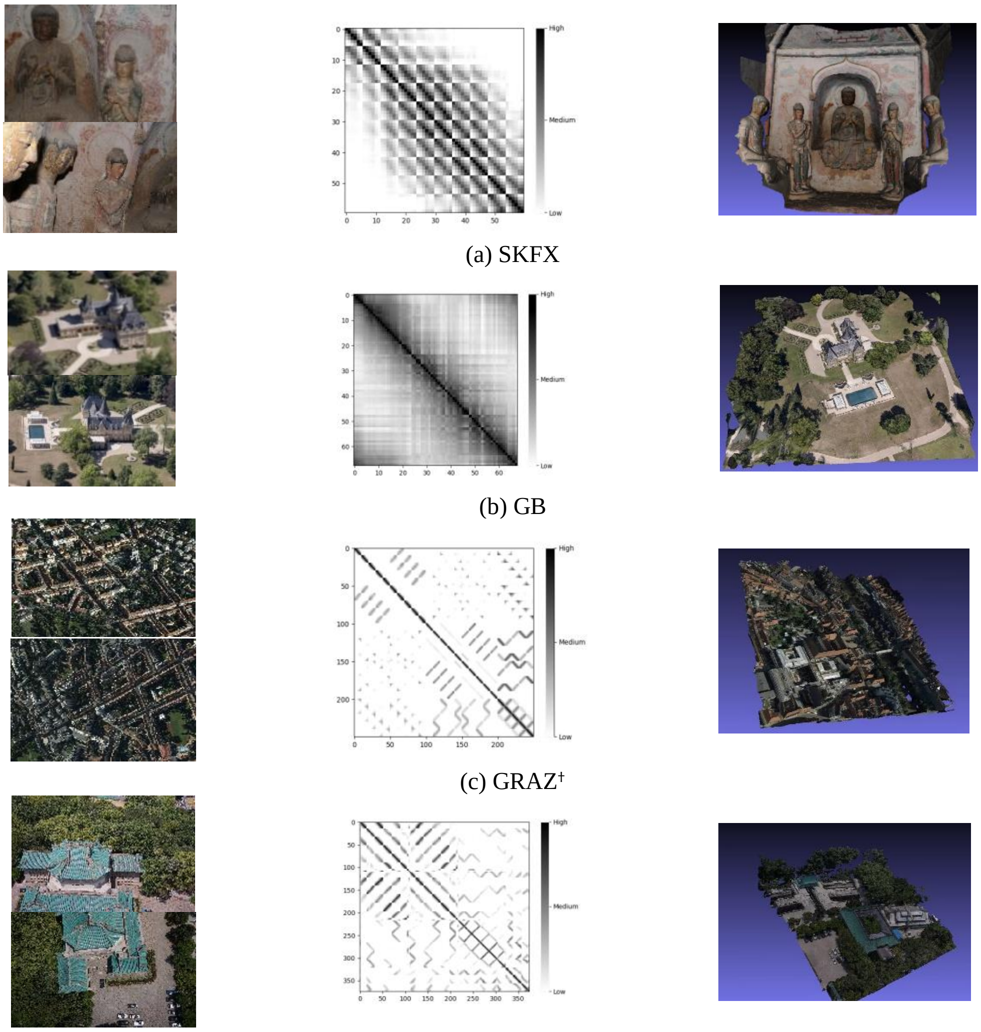
Wang, X., Zhan, Z.Q., Heipke, C., 2017. An efficient method to detect mutual overlap of a large set of unordered images for Structure-from-Motion. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-1-W1, pp.191-198.

Wang, X., Rottensteiner, F., Heipke, C., 2019. Structure from Motion for ordered and unordered image sets based on random k-d forests and global pose estimation. ISPRS Journal of Photogrammetry & Remote Sensing, 147, pp. 19–41.

Yi, K.M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned invariant feature transform. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 467-483.

Zheng, Z., Wei, Y., Yang, Y., 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1395–1403.

Zhu, S., Zhang, R.Z., Zhou, L., Shen, T.W., Fang, T., Tan, P., Quan, L., 2018. Very Large-Scale Global SfM by Distributed Motion Averaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4568-4577.

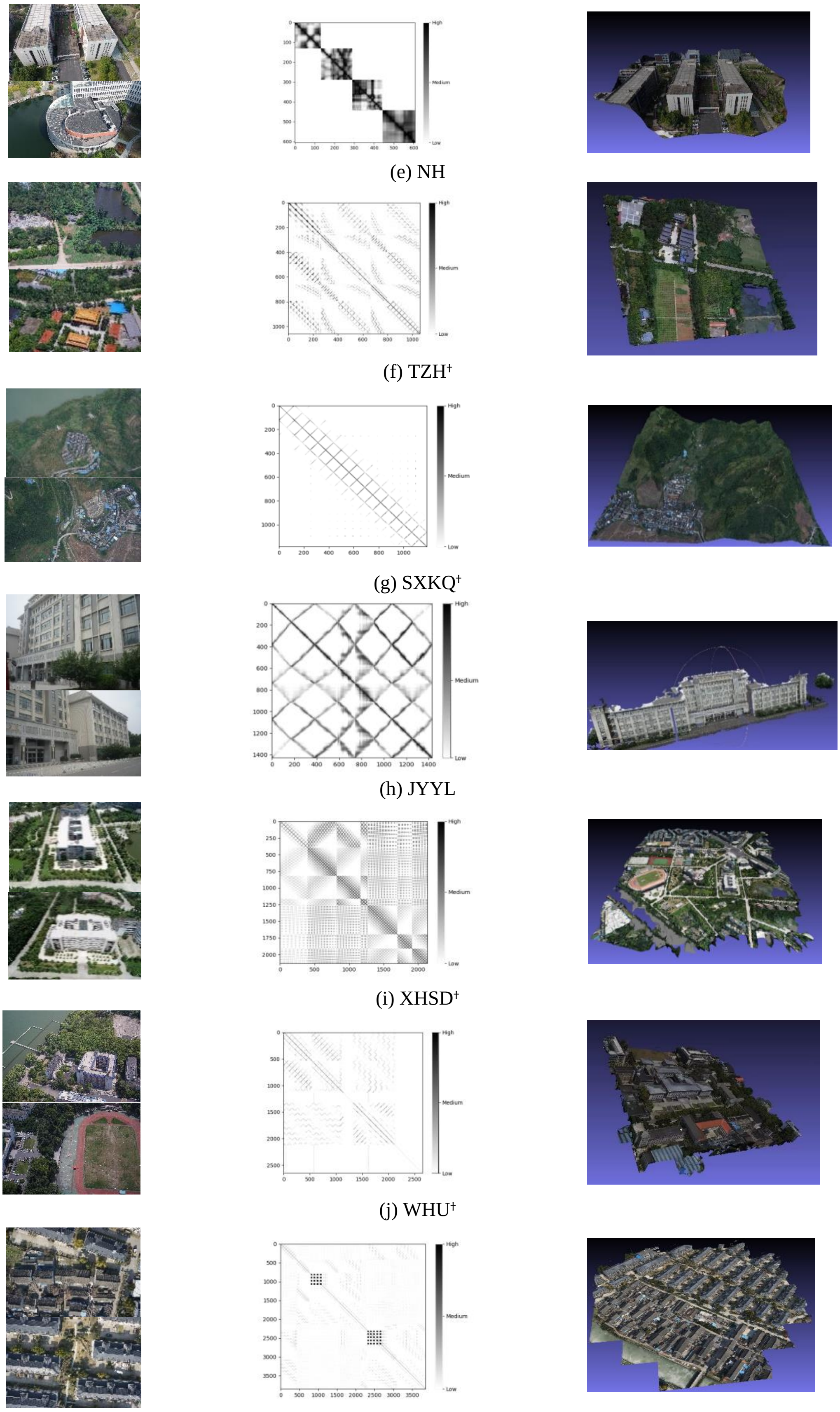


(a) SKFX

(b) GB

(c) GRAZ†

(d) YD



(e) NH

(f) TZH†

(g) SXKQ†

(h) JYYL

(i) XHSD†

(j) WHU†

(k) SHHY†

Figure 6. Qualitative overview of BeDOI. First column shows two sample images of each dataset, second column denotes the overlapping relationship graph in which higher similarity degree is indicated by darker colour, the third column is the generated 3D mesh model († means that only parts of the 3D mesh model are shown in the corresponding dataset).