Extremely Dense Point Correspondences using a Learned Feature Descriptor

Xingtong Liu[1], Yiping Zheng[1], Benjamin Killeen[1], Masaru Ishii[2], Gregory D. Hager[1], Russell H. Taylor[1], and Mathias Unberath[1]

[1]Johns Hopkins University
[2]Johns Hopkins Medical Institutions
{xingtongliu, unberath}@jhu.edu

Abstract

High-quality 3D reconstructions from endoscopy video play an important role in many clinical applications, including surgical navigation where they enable direct video-CT registration. While many methods exist for general multi-view 3D reconstruction, these methods often fail to deliver satisfactory performance on endoscopic video. Part of the reason is that local descriptors that establish pairwise point correspondences, and thus drive reconstruction, struggle when confronted with the texture-scarce surface of anatomy. Learning-based dense descriptors usually have larger receptive fields enabling the encoding of global information, which can be used to disambiguate matches. In this work, we present an effective self-supervised training scheme and novel loss design for dense descriptor learning. In direct comparison to recent local and dense descriptors on an in-house sinus endoscopy dataset, we demonstrate that our proposed dense descriptor can generalize to unseen patients and scopes, thereby largely improving the performance of Structure from Motion (SfM) in terms of model density and completeness. We also evaluate our method on a public dense optical flow dataset and a small-scale SfM public dataset to further demonstrate the effectiveness and generality of our method. The source code is available at https://github.com/lppllpp1920/DenseDescriptorLearning-Pytorch.

1. Introduction

Background. In computer vision, correspondence estimation aims to find a match between 2D points in image space and corresponding 3D locations. Many potential applications rely on this fundamental task, such as Structure from Motion (SfM), Simultaneous Localization and Mapping (SLAM), image retrieval, and image-based localization. In particular, SfM and SLAM have been shown to be effective for endoscopy-based surgical navigation [20], video-CT registration [15], and lesion localization [40]. These successes rely on the fact that SfM and SLAM estimate a sparse 3D structure of the observed scene as well as the camera's trajectory from unlabeled video, simultaneously.

The advantages of SLAM and SfM are complementary. In applications that require real-time estimation, *e.g.* surgical navigation, SLAM provides a computationally efficient framework for correspondence estimation. Robust camera tracking requires a dense 3D reconstruction estimated from previous frames, but computational constraints usually limit SLAM to local optimization. This often leads to drifting errors, especially when the trajectory loop is not evident. On the other hand, SfM prioritizes high density and accuracy for the sparse 3D structure. This is due to the time-consuming global optimization used in the bundle adjustment, which limits SfM to applications where offline estimation is acceptable.

In video-CT registration, a markerless approach relies on correspondence estimation to provide a sparse reconstruction and the camera trajectory from the video. The reconstruction is then registered to the CT surface model with a registration algorithm [32]. This requires SfM since it relies on a dense and accurate 3D reconstruction. The accuracy of the estimated camera trajectory is also crucial so that the camera pose of each video frame aligns with the CT surface model. However, when estimating camera trajectory from endoscopic video, typical SfM or SLAM pipeline fails to produce a high-quality reconstruction or accurate camera trajectory. Recent work aims to mitigate this shortcoming through procedural changes in the video capture, which we discuss below. In this work, we focus on developing a more effective feature descriptor, which is used in the feature extraction and matching module of the pipeline, to substantially increase the density of extracted correspondences

(cf. Fig. 1).

Related Work. A local descriptor consists of a feature vector computed from an image patch, whose size and orientation are usually determined by a keypoint detector, such as Harris [10], FAST [29], and DoG [18]. The hand-crafted local descriptor SIFT [18] has been arguably the most popular feature descriptor for correspondence estimation and related tasks. In recent years, advanced variants of SIFT have been proposed, such as RootSIFT [1], RootSIFT-PCA [3], and DSP-SIFT [7]. Some of these outperform the SIFT descriptor in tasks such as fundamental matrix estimation [2], pair-wise feature matching, and multi-view reconstruction [31]. Additionally, learning-based local descriptors have grown in popularity with the advent of deep learning, with recent examples being L2-Net [36], GeoDesc [19], and HardNet [22]. Though learning-based methods have outperformed hand-crafted ones in many areas of computer vision, advanced variants of SIFT continue to perform on par with or better than their learning-based local descriptors [2, 31].

Several dense descriptors have been proposed, such as DAISY [37], UCN [5], and POINT^2 [16]. Compared with local descriptors, which follow a *detect-and-describe* approach [8], dense descriptors extract image information without using a keypoint detector to find specific locations for feature extraction. As a result, dense descriptors have higher computation efficiency than local descriptors in applications that require dense matching. They also avoid the possibility of repeated keypoint detection [8]. On the other hand, learning-based dense descriptors typically show better performance compared with hand-crafted ones. This is because Convolutional Neural Networks (CNN) can encode and fuse high-level context and low-level texture information more effectively than manual rules given enough training data. Our method belongs to the category of learning-based dense descriptors. There are also works that jointly learn a dense descriptor and a keypoint detector, such as SuperPoint [6] and D2-Net [8], or learn a keypoint detector that improves the performance of a local descriptor, such as GLAMpoints [38].

In the field of endoscopy, researchers have applied SfM and SLAM to video from various anatomy, including sinus [15], stomach [40], abdomen [9, 20], and oral cavity [27]. Popular SfM pipelines such as COLMAP [30] and SLAM systems such as ORB-SLAM [24] usually do not achieve satisfactory results in endoscopy without further improvement. Several challenges stand in the way of successful correspondence estimation in endoscopic video. First, tissue deformation, as in video from a colonoscopy, violates the static scene assumption in these pipelines. To mitigate this issue, researchers have proposed SLAM-based methods that tolerate scene deformation [14, 34]. Second, the textures in endoscopy are often smooth and repetitive,

which makes the sparse matching with local descriptors error-prone. Widya *et al*. [40] proposed spreading IC dye in the stomach to manually add texture to the surface, increasing the matching performance of local descriptors. This leads to denser and more complete reconstructions. Qiu *et al*. [27] use a laser projector to project patterns on the surface of the oral cavity to add more textures to improve the performance of a SLAM system. However, introducing additional procedures as above is usually not desired by surgeons because it will interrupt the original workflow. Therefore, instead of adding textures, we develop a dense descriptor that works well on the texture-scarce surface to replace the original local descriptors in these systems.

Contributions. First, to the best of our knowledge, this is the first work that applies learning-based dense descriptors to the task of multi-view reconstruction in endoscopy. Second, we present an effective self-supervised training scheme which includes a novel loss called Relative Response Loss that can train a high-precision dense descriptor with the learning style of keypoint localization. The proposed training scheme outperforms the popular hard negative mining strategy used in various learning-based descriptors [5, 4, 22]. For evaluation, we have conducted extensive comparative studies on the task of pair-wise feature matching and SfM on a sinus endoscopy dataset, pair-wise feature matching on the KITTI Flow 2015 dataset [21], and SfM on a small-scale natural scene dataset [35].

2. Methods

In this section, we describe our self-supervised training scheme for dense descriptor learning, which includes overall network architecture, training scheme, custom layers, loss design, and a dense feature matching method.

Overall Network Architecture. As shown in Fig. 2, the training network is a two-branch Siamese network. The input is a pair of color images, which are used as source and target. The training goal is, given a keypoint location in the source image, finding the correct corresponding keypoint location in the target image. A SfM method [15] with SIFT is applied to video sequences to estimate the sparse 3D reconstructions and camera poses. The groundtruth point correspondences are then generated by projecting the sparse 3D reconstructions onto the image planes using the estimated camera poses. The dense feature extraction module is a fully convolutional DenseNet [13] which takes in a color image and output a dense descriptor map that has the same resolution as the input image and the length of the feature descriptor as the channel dimension. The descriptor map is L2-normalized along the channel dimension to increase the generalizability [39]. For each source keypoint location, the corresponding descriptor is sampled from the source descriptor map. Using the descriptor of the source keypoint as a 1×1 convolution kernel, a 2D convolution is performed on
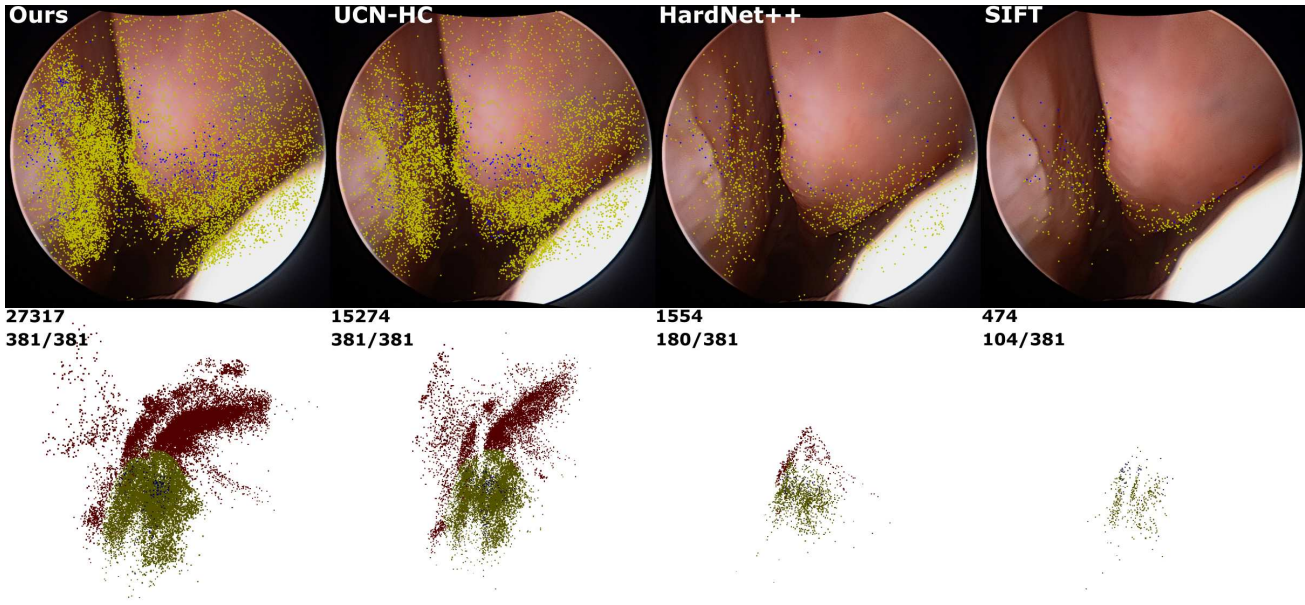
Figure 1. **Qualitative comparison of SfM performance in endoscopy.** The figure shows the performance of different descriptors on the task of SfM on the same sinus endoscopy video sequence. The comparison descriptors are ours, UCN [5] trained with recently proposed Hardest Contrastive Loss [4] on endoscopy data, HardNet++ [22] fined-tuned with endoscopy data, and SIFT [18]. The first row shows the same video frame and the reprojection of the corresponding sparse 3D reconstruction from SfM. The second row displays the sparse reconstructions and relevant statistics; the number in the first row of each image is the number of points in the reconstruction; the two numbers in the second row of each image are the number of registered views and the total number of views in the sequence. The red points are those not visible in the showed frame. The yellow points are in the field of view of the displayed frame but reconstructed by other frames. The triangulation of blue points in the figure involves the displayed frame.
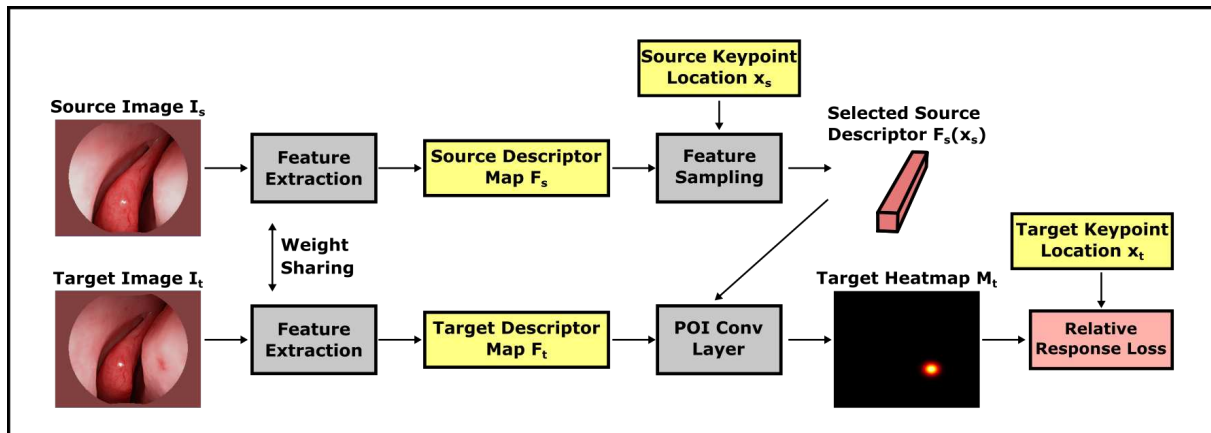

Figure 2. **Overall network architecture.** The training data consists of a pair of source and target images and groundtruth source-target 2D point correspondences. The source and target images are randomly selected from the frames which share observations of the same 3D points. For each pair of images, a certain number of point correspondences are randomly selected from the available ones in each training iteration. For the simplicity of illustration, only one target-source point pair and the corresponding target heatmap are shown in the figure. All concepts in the figure are defined in the *Methods* section.

the target descriptor map in the Point-of-interest (POI) Conv Layer [16]. The computed heatmap represents the similarity between the source keypoint location and every location on the target image. The network is trained with proposed Relative Response Loss (RR) to force the heatmap to present high response only at the groundtruth target location. The

idea of converting the problem of descriptor learning to keypoint localization is proposed by Liao *et al*. [16], which was originally used to solve the problem of X-ray-CT 2D-3D registration.

Point-of-Interest (POI) Conv Layer. This layer is used to convert the problem of descriptor learning to key-

point localization [16]. For a pair of source and target input images, a pair of dense descriptor maps, $\mathbf{F}_s$ and $\mathbf{F}_t$, are generated from the feature extraction module. The size of an input image and a descriptor map are $3 \times H \times W$ and $C \times H \times W$, respectively. For a descriptor at the source keypoint location $\mathbf{x}_s$, the corresponding feature descriptor, $\mathbf{F}_s(\mathbf{x}_s)$, is extracted with the nearest neighbor sampling, which could be changed to other sampling methods if needed. The size of the descriptor is $C \times 1 \times 1$. By treating the sampled feature descriptor as a 1×1 convolution kernel, the 2D convolution operation is performed on $\mathbf{F}_t$ to generate a target heatmap, $\mathbf{M}_t$, storing the similarity between the source descriptor and every target descriptor in $\mathbf{F}_t$.

Relative Response Loss (RR). The loss is proposed with the intuition that a target heatmap should present a high response at the groundtruth target keypoint location and the responses at other locations should be suppressed as much as possible. Besides, we do not want to assume any prior knowledge on the response distribution of the heatmap to preserve the potential of multimodal distribution to respect the matching ambiguity of challenging cases. To this end, we propose to maximize the ratio between the response at the groundtruth location and the summation of all responses of the heatmap. Mathematically, it is defined as,

$$\mathcal{L}_{\mathrm{rr}} = -\log\left(\frac{e^{\sigma\mathbf{M}_t(\mathbf{x}_t)}}{\sum_{\mathbf{x}} e^{\sigma\mathbf{M}_t(\mathbf{x})}}\right) \quad , \text{where} \qquad (1)$$

a scale factor σ is applied to the heatmap $\mathbf{M}_t$ to enlarge the value range which was $[-1, 1]$. A spatial softmax is then calculated at the groundtruth location $\mathbf{x}_t$ of the scaled heatmap, where the denominator is the summation of all elements of the scaled heatmap. The logarithm operation is used to speed up the convergence. We observe that, by only penalizing the value at the groundtruth location after spatial softmax operation, the network learns to reduce the response at the other locations and increase the response at the groundtruth location effectively. We compare the feature matching and SfM performance of dense descriptors trained with different common loss designs that are originally for the task of keypoint localization in the *Experiments* section. A qualitative comparison of target heatmaps generated by different dense descriptors is shown in Fig. 3.

Dense Feature Matching. For each source keypoint location in the source image, a corresponding target heatmap is generated with the method above. The location with the largest response value in the heatmap is selected as the estimated target keypoint location. The descriptor at the estimated target keypoint location then performs the same operation on the source descriptor map to estimate the source keypoint location. Because of the characteristics of dense matching, the traditional mutual nearest neighbor criterion used in the pair-wise feature matching of a local descriptor is too strict. We relax the criterion by accepting the match as

	UCN-C	UCN-HC	Softarg.	Softarg.+BCE	Softmax+BCE	RR+Softarg.	RR
PCK@5px	25.5	58.8	36.5	44.6	35.4	57.9	63.0
PCK@10px	35.0	67.2	54.6	63.1	51.1	68.6	71.9
PCK@20px	47.0	74.0	73.6	77.4	66.0	78.6	80.0

Table 1. **Evaluation of feature matching performance in endoscopy.** This table shows the average percentage of correct keypoints (PCK) with threshold 5px, 10px, and 20px over all 9 sequences from the 3 testing patients. The PCK is calculated on all image pairs whose interval is within 20 frames. For each pair, PCK is computed by comparing the dense matching results with the groundtruth point correspondences from SfM results. The feature matching results in each column are generated by the descriptor whose name is on the first row. From left to right, the evaluated descriptors are UCN trained with Contrastive Loss (UCN-C) [5], UCN trained with Hardest Contrastive Loss (UCN-HC) [4], replacing the proposed Relative Response Loss (RR) with Softargmax [12], replacing RR with Softargmax and Binary Cross Entropy (BCE), replacing RR with spatial softmax and BCE [11], RR and Softargmax, and the proposed training scheme with RR. The model trained with the proposed RR achieves the best average matching accuracy.

long as the estimated source keypoint location is within the vicinity of the original source keypoint location, which we call cycle consistency criterion. The computation of dense matching can be parallelized on modern GPU by treating all sampled source descriptors as a kernel with the size of $N \times L \times 1 \times 1$; N is the number of query source keypoint locations used as the output channel dimension; L is the length of the feature descriptor used as the input channel dimension of a standard 2D convolution operation.

3. Experiments

We evaluate our proposed method on three datasets. Sinus endoscopy dataset is used to evaluate the performance of local and dense descriptors on the task of pair-wise feature matching and SfM in endoscopy. KITTI Flow 2015 dataset [21] is used to evaluate the performance of dense descriptors on the task of pair-wise feature matching in natural scenes. A small-scale dataset with a collection of building photos [35] is used to evaluate the performance of local and dense descriptors on the task of SfM in natural scenes. All experiments are conducted on a workstation with 4 NVIDIA Tesla M60 GPUs, each with 8 GB memory, and the method is implemented using PyTorch [25].

Evaluation on Sinus Endoscopy. The dataset consists of video data collected from 8 patients and 2 cadavers. The overall duration is around 30 minutes. For the ease of experiments, all images are downsampled to 256×320 pixels during both training and testing. For our method, we use a light-weight version of Fully Convolutional DenseNet (FC-DenseNet) [13] with 32 layers and filter growth rate of 10. The length of the output descriptor is 256; the overall number of parameters is 0.53 million. The model is trained with Stochastic Gradient Descent with the cyclic learning
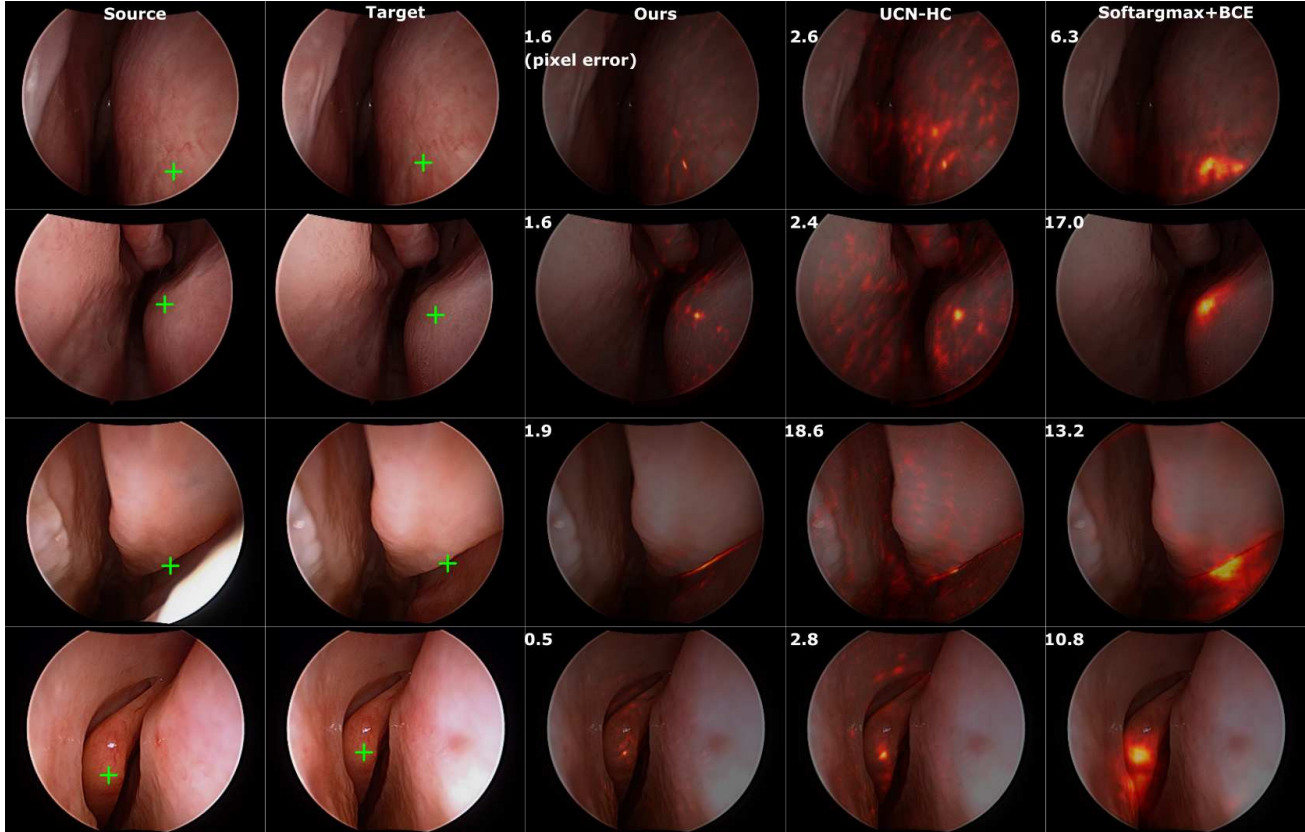
Figure 3. **Qualitative comparison of feature matching performance in endoscopy.** The figure qualitatively shows the performance of three dense descriptors trained with different loss designs on the task of pair-wise feature matching. The first two rows are training images and the rest are testing ones. The first and second columns show the source-target image pairs, where the green crossmarks indicate the groundtruth source-target point correspondences. For each dense descriptor, a target heatmap, as shown in the last three columns, is generated from the POI Conv Layer. To visualize the contrast better, the displayed heatmap is normalized with spatial softmax operation and then with the maximum value of the processed heatmap. The numbers shown in the last three columns are the pixel errors between the estimated target keypoint locations and the groundtruth ones. The fourth column shows the results of UCN [5] trained with recent Hardest Contrastive Loss on the endoscopy dataset. The model in the fifth column is trained with the same method as ours except that the training loss is Softargmax [12] and BCE instead of the proposed Relative Response Loss. The results show that our method produces fewer high responses, which leads to better matching accuracy.

	Seq. 1-1 (381)			Seq. 1-2 (314)			Seq. 1-3 (370)			Seq. 2-1 (455)			Seq. 2-2 (630)			Seq. 2-3 (251)			Seq. 3-1 (90)			Seq. 3-2 (1309)			Seq. 3-3 (336)		
SIFT	104	474	5.62	219	1317	5.58	113	938	5.16	119	751	5.81	295	10384	6.43	122	1896	5.38	48	435	5.09	55	953	5.51	169	2169	5.57
DSP-SIFT	149	783	5.09	235	1918	5.06	132	1228	4.78	404	6557	5.32	296	7322	5.64	167	3450	5.00	42	293	4.81	150	745	5.17	180	1180	5.18
RootSIFT-PCA	104	384	5.89	219	1004	5.67	115	661	5.11	227	821	5.82	295	10147	6.43	128	2025	5.46	50	255	5.18	217	3188	5.35	176	2450	5.62
HardNet++	180	1554	4.63	233	2162	4.81	244	3003	4.65	424	4755	4.65	534	9828	4.85	225	5727	4.56	79	610	4.66	416	4658	4.62	228	3196	4.66
UCN-C	349	13402	4.26	311	13198	4.50	248	8336	4.43	405	11935	4.13	293	8258	4.46	196	9273	3.98	77	2445	4.10	503	16166	4.29	206	3736	4.17
UCN-HC	381	15274	4.84	314	13519	4.84	352	16900	4.89	455	33299	4.67	630	45375	4.81	251	26322	4.37	86	2988	4.39	484	13394	4.39	283	11555	4.39
Softarg.	348	5966	4.74	312	7774	4.74	252	7426	4.63	293	4861	4.50	547	12590	4.24	205	2847	4.22	59	534	4.17	451	7247	4.76	302	6039	4.26
Softarg.+BCE	357	11502	4.47	314	10373	4.57	244	10339	4.55	426	19848	4.34	560	22482	4.19	125	1150	4.04	46	774	4.04	500	12187	4.51	303	6268	4.06
Softmax+BCE	165	2246	4.26	306	8885	4.26	228	8628	4.19	378	8559	4.10	296	12081	4.19	77	1124	3.96	34	353	4.02	261	5024	4.19	181	2973	4.07
RR+Softarg.	381	19921	4.99	314	20375	4.98	256	20550	4.94	455	44388	4.75	630	39752	4.64	244	10055	4.35	87	5071	4.33	507	20906	4.61	312	12856	4.36
RR	381	27317	5.07	314	22898	5.23	367	29734	5.06	455	41380	4.78	630	45654	4.80	251	19645	4.43	89	6763	4.62	507	35645	4.68	313	21703	4.53

Table 2. **Evaluation of SfM performance in endoscopy.** We compare the SfM results of 9 sequences from the 3 testing patients. The SfM results are generated by the descriptors whose names are on the first column. We compare the SfM performance of local and dense descriptors. Starting from the first descriptor, these are SIFT [18], DSP-SIFT [7], RootSIFT-PCA [3], HardNet++ [22] fine-tuned with the endoscopy dataset, UCN trained with Contrastive Loss (UCN-C) [5], UCN trained with Hardest Contrastive Loss (UCN-HC) [4], replacing the proposed RR with Softargmax [12], replacing RR with Softargmax and BCE, replacing RR with spatial softmax and BCE [11], RR and Softargmax, and the proposed training scheme with RR. Each number in the first row represents the number of frames in each sequence. In the following rows, for each sequence and each method, three numbers from left to right are the number of registered views, the number of sparse points, and the average track length of sparse points. It shows that the proposed method (RR) obtains the most number of registered views in all sequences and the densest reconstructions for most of the sequences. SIFT or RootSIFT-PCA achieve highest average track length in all sequences.

rate [33] ranging from 1.0e-4 to 1.0e-3. The scale factor σ used in the Relative Response Loss is empirically set to 20.0. Data from 5 patients and 1 cadaver are used for training; the other cadaver is used for validation; the remaining 3 patients are for testing. Because our evaluation focuses on the loss design, for fairness, we use the same network architecture described above for all dense descriptors to extract features. All models are trained until the performance on the validation data stops improving. The evaluation results of pair-wise feature matching are shown in Table 1. To measure the accuracy of feature matching, we use the percentage of correct keypoints (PCK) with three thresholds, which are 5, 10 and 20 pixels. The matching is determined as correct if the detected target keypoint location is within a specified number of pixels. The results show that our proposed training scheme for the dense descriptor outperforms competing methods for dense descriptor learning, which are Contrastive Loss in [5] and Hardest Contrastive Loss in [4]. Besides, since we convert the problem of descriptor learning to keypoint localization, we also evaluate the performance of several loss functions used in keypoint localization by training the proposed network with these instead of Relative Response Loss. For the proposed method, generating and matching a pair of dense descriptor maps under the current setting takes about 37ms. To evaluate the performance of local and dense descriptors on the task of SfM in endoscopy, we use a simple SfM pipeline [15] which takes in pair-wise feature matches, uses Hierarchical Multi-Affine [26] for geometric verification, and global bundle adjustment [23] for optimization. Pair-wise feature matches are estimated in all image pairs whose interval is within 30 frames. For all local descriptors, DoG [18] is used to extract keypoint locations in both source and target images for sparse feature matching with mutual nearest neighbor (MNN) as the matching criterion. For dense descriptors, DoG is used to extract keypoint locations in only source images and dense matching is performed on the target images for these detected candidate keypoint locations in the source images. The false matches are ruled out using the cycle consistency criterion described in the *Dense Feature Matching* subsection. Because of the texture smoothness of endoscopy, we change the hyperparameters of DoG so that more candidate keypoint locations can be detected. The number of layers in each octave is 8; the contrast threshold is $5.0e^{-5}$; the edge threshold is 100.0; the standard deviation of the Gaussian applied to the image at the first octave is 1.1. All hand-crafted descriptors use the parameter setting recommended by the original authors. The SfM results are shown in Table 2. Note that we build an image patch dataset from SfM results in endoscopy using the same method as [19] to fine-tune the HardNet++ [22] for a fair comparison, which does have better performance compared with the pre-trained model released by the authors.
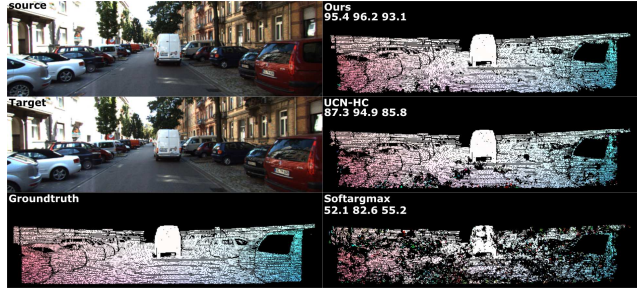

Figure 4. **Qualitative comparison of feature matching performance in KITTI Flow 2015 [21].** The optical flow estimates from three dense descriptors for the same source-target image pair are shown in this figure. The descriptors are our method, UCN-HC [4], and the proposed method trained with Softargmax [12] instead of RR. The numbers shown in each image in the second column represent the putative match ratio, precision, and matching score of the optical flow estimates. We use the cycle consistency criterion with 6px threshold to rule out potential false matches. The images in the first column are source, target, and groundtruth dense optical flow map, where black values mean there are no valid measurements. The second column shows the dense optical flow estimates, where the black pixels include those with no groundtruth measurements or ruled out with cycle consistency criterion.

Evaluation on KITTI Flow 2015 [21]. In this evaluation, we evaluate the performance of dense descriptors on the task of optical flow estimation. First, we estimate the Matching Score $= \frac{\#\text{Inlier Matches}}{\#\text{Features}}$. The #Inlier Matches is the number of matches where the distance between the estimated target keypoint location and the groundtruth target location is within 10 pixels. The #Features is equal to the number of pixels in an image. We also evaluate the Putative Match Ratio $= \frac{\#\text{Putative Matches}}{\#\text{Features}}$ and Precision $= \frac{\#\text{Inlier Matches}}{\#\text{Putative Matches}}$ [31]. The match is determined as putative if it passes the cycle consistency criterion. We follow the same training protocol as [5], where 1000 point correspondences are randomly selected for each image pair in the KITTI dataset and fixed during training. The models trained with the proposed Relative Response Loss, Softargmax loss [12], Contrastive Loss [5], and Hardest Contrastive Loss [4], respectively, are evaluated. To evaluate the performance of different loss designs, we train all models with the same network architecture for feature extraction. We use FC-DenseNet with 38 layers and filter growth rate of 16. The overall number of parameters is 1.68 million. Other parameter settings are the same as the evaluation in endoscopy. The images are downsampled by a factor of 2 during training. Two top-performing results of the comparison methods presented in [5] are cited here. An example of dense optical flow estimates from different trained models is shown in Fig. 4. The quantitative evaluation results are shown in Table 3.

Evaluation on Multi-view Stereo 2008 [35]. The

	DaisyFF	DM	UCN-C	UCN-HC	Softargmax	RR
Putative Match Ratio (%)	100.0	100.0	73.6	80.6	46.0	88.0
Precision (%)	79.6	85.6	73.2	90.9	70.7	89.8
Matching Score (%)	79.6	85.6	61.9	76.9	41.5	81.7

Table 3. **Evaluation of optical flow estimation on KITTI Flow 2015 [21].** DaisyFF [41] and DM [28] use global optimization to estimate a dense optical flow map between a pair of images. The last four methods are introduced in Table 2. It shows that, by removing unconfident matches using cycle consistency criterion with 6 pixels as the threshold, our method achieves marginally lower precision than UCN-HC, whereas UCN-HC has lower putative match ratio than ours. Our method obtains the highest matching score among the last four methods. Note that we assume the first two methods do not discard any matches, which is why the putative match ratios are shown as 100.

dataset consists of several small-scale sequences where the same building is captured from different viewpoints within each sequence. We evaluate the performance of the proposed method on the task of SfM in natural scenes and compare it with the hand-crafted local descriptors. Our model is trained with the SfM results of *gerrard-hall*, *personal-hall*, and *south-building*, which were released by the author of COLMAP [30]. We use FC-DenseNet with 32 layers and filter growth rate as 16. The overall number of parameters is 1.26 million. Other parameter settings are the same as the evaluation in endoscopy. All images for training and testing are downsampled to 256×320. All descriptors use the DoG keypoint detector with the same parameter setting as the evaluation in endoscopy. The evaluation results are shown in Table 4. Most experiments are conducted on the same SfM pipeline [15] as in endoscopy. SIFT and DSP-SIFT are also evaluated with COLMAP.

4. Discussion

Intuition on the Performance Difference of Various Training Schemes for Dense Descriptor Learning. We attribute the performance difference between our method and UCN-HC to the different strategies of training data sampling. For UCN-HC, given a positive point pair, one hardest negative point is obtained in a minibatch for each of the points in the pair to calculate the negative loss. A diameter threshold is also set to avoid mining points that are too close to the positive point. A positive margin threshold and negative margin threshold is also set to avoid penalizing positive pairs that are close enough or negative pairs that are far enough. There are several potential problems with this setting. First, the strategy of hardest sample selection, which was also used similarly in the local descriptor training [22], could potentially lead to training instability, which was also mentioned by the original authors in their Github repository. Because for each iteration of training, only the hardest negative samples in a minibatch provide gradients to the network training with other samples being ignored, the gradient direction may not be helpful to these ignored

	Entry (10)			Fountain (11)			HerzJesu (25)			Castle (30)		
SIFT-COLMAP	10	1557	3.86	11	1566	4.33	25	3637	5.91	30	3718	4.57
DSP-SIFT-COLMAP	10	1849	3.93	11	1769	4.42	25	3650	5.90	30	4203	4.78
SIFT	10	1444	3.54	10	2775	3.63	24	6706	4.59	30	4589	3.76
DSP-SIFT	10	3041	3.79	11	4244	3.85	25	7334	4.35	22	1804	3.96
RootSIFT-PCA	7	1109	3.42	10	2467	3.54	25	6584	4.47	21	3991	3.84
RR	10	2980	3.62	7	6293	3.72	25	12807	3.96	29	3684	3.67
RR-SG	10	5310	3.60	8	7676	3.90	24	15799	4.28	27	3431	3.88

Table 4. **Evaluation of SfM performance on Multi-view Stereo 2008 [35].** Though the scene variation between the training and testing dataset is large, our method (RR) still performs comparably against the hand-crafted local descriptors. However, compared with endoscopy, we do observe a larger amount of false matches in the pair-wise feature matching phase. This potentially means a dense descriptor needs a larger amount of training data or a limited receptive field to avoid overfitting when the scene variation is large. To test the hypothesis, we train another model, which is RR-SG, with 4 times smaller receptive field and gray image as input and similar number of parameters to RR. It shows that RR-SG produces denser reconstructions in three out of four sequences. This might mean, compared with RR-SG, RR overfits to the high-level context information to a larger degree. Compared with the SfM pipeline in [15], COLMAP has more stable performance in terms of the completeness of the camera trajectory but usually smaller number of sparse points. This observation is similar to the results in [30], where they compared COLMAP with other SfM pipelines. It is probably due to the stricter inlier criteria of COLMAP.

samples. This could potentially lead to training oscillation where hardest samples jump among different samples but the network never converges to the optimal solution. The results of instability can be found in Fig. 3, where many high responses are scattered in the heatmap. Second, the manually specified diameter and margin thresholds could also lead to a suboptimal solution. Because samples that are within the diameter of a selected sample are not considered as negative ones, the network will never try to push nearby samples away from the selected one. Therefore, this limits the matching precision of the descriptor. This again can be observed in Fig. 3, where the high-response clusters around the groundtruth target locations appear to be wider than our proposed method. The margin thresholds in the loss design also remove the possibility of further pushing away negative samples from the positive ones and pulling positive pairs closer, which could be another reason for obtaining such heatmaps. As a comparison, in our method, for each sampled point in the source image, all points in the target image are observed in one training iteration. Only the groundtruth target point is considered a positive point and all other points are considered negative ones. This avoids the oscillation related to the descriptor distance between the selected source point and all points in the target image. The reason why this training scheme will not suffer from the problem of data imbalance is due to the proposed Relative Response Loss (RR). The goal of RR is to make the ratio between the response at the groundtruth target location and the summation of all responses in the target image as high as possible. By doing this, the network will try to suppress all

responses except the one at the target groundtruth location. It does not assume any prior distribution of the response heatmap and conveys the goal of precise feature matching clearly, which we believe improves the expressivity of the network.

We have also evaluated some common losses used in the task of keypoint localization, such as spatial softmax + BCE and Softargmax [12]. Spatial softmax + BCE is used for heatmap regression so that the network produces a similar heatmap as the groundtruth one. However, because the groundtruth distribution is usually assumed to be Gaussian with a manually specified standard deviation, this limits the expressivity of the network in cases where Gaussian distribution is not optimal. This can be observed in the third row in Fig. 3, where the model trained with Softargmax + BCE tries to infer a Gaussian-like distribution around the groundtruth location. As a comparison, the learned descriptor in our proposed method naturally produces high response along the edge of the surface, which is where the most ambiguities come from. Besides, BCE also suffers from the data imbalance problem for the case where positive and negative samples are highly unbalanced, which is also observed in [17]. Softargmax converts the task of keypoint localization to a position regression task where the network tries to produce a heatmap so that the centroid of the heatmap is close to the groundtruth target location. However, this suffers from the fact that any distribution where the centroid is equal to the target location will not be further penalized. Therefore, Softargmax makes the network easily trapped in suboptimal solutions of learning a discriminative descriptor, whereas there are no such training ambiguities in RR. Though this ambiguity can be reduced by combining Softargmax with BCE, the performance is still worse than RR, as observed in Table 1 and 2, because of the unimodal distribution assumption.

Local Descriptor vs. Learning-based Dense Descriptor. We observe that learning-based dense descriptors usually perform better than local descriptors in the experiments related to SfM in sinus endoscopy. We attribute this to two reasons. First, local descriptors usually need a keypoint detector, such as DoG [18], to detect candidate keypoints before sparse feature matching. The lack of repeatability in the keypoint detector makes many true matches unable to be found because either source or target locations for these matches are not detected as candidate keypoints in the keypoint detection phase. As observed in [8], the unstable detection is because the detector usually uses low-level information, which is often significantly affected by changes such as viewpoint and illumination. Second, the smooth and repetitive textures in endoscopy make it challenging for the local descriptors that have a limited receptive field to find correct matches even if all points in the true matches are detected by the keypoint detector. On the other hand, learning-based dense descriptors do not rely on the keypoint detector to produce repeatable keypoint locations and have a larger receptive field.

Compared with local descriptors, dense descriptors also have disadvantages. First, a dense descriptor is more memory-demanding. This is because, to parallelize the dense matching procedure with many keypoint locations, the descriptors need to be organized in the form described in the *Dense Feature Matching* subsection. This requires memory to store a response target heatmap for each source keypoint location before the target location is estimated from the heatmap. Though a sparse matching can also be performed with a dense descriptor, the performance will degrade because of the reliance on a repeatable keypoint detector. Therefore, the practical usage of a dense descriptor on a low-cost embedded system is limited. Second, learning-based dense descriptors seem to be more overfitting-prone compared with learning-based local descriptors. This is because the dense descriptor network relies on both high-level and low-level image information to generate a descriptor map. Because high-level information, presumably, has more variation compared with low-level texture information that learning-based local descriptors only need, more training data is probably needed for a dense descriptor. The reason why dense descriptors seem to generalize well in endoscopy could be due to the lower anatomical variation compared with the variation in natural scenes.

5. Conclusion

In this work, we propose an effective self-supervised training scheme with a novel loss design for the learning-based dense descriptor. To the best of our knowledge, this is the first work that applies a learning-based dense descriptor to endoscopy for multi-view reconstruction. We evaluate our method on both endoscopy and natural scene dataset on the task of pair-wise feature matching and SfM, where our proposed method outperforms other local and dense descriptors on a sinus endoscopy dataset and outperforms recent dense descriptors in a dense optical flow public dataset. The extensive comparison study helps to gain more insights on the difference between local and dense descriptors, and the effects of different loss designs on the overall performance of a dense descriptor. Because SfM is an offline method, it is not able to support real-time localization and mapping. We plan to extend this work to incorporate a learning-based dense descriptor into an existing SLAM system in the future to make it more accurate and robust for surgical navigation in endoscopy. We also plan to adopt a bootstrapping training method to train the dense descriptor because of the observation that a descriptor model trained with sparse SfM results helps the SfM estimate denser reconstructions from both testing and training sequences.

References

[1] R. Arandjelovic. Three things everyone should know to improve object retrieval. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2911–2918, Washington, DC, USA, 2012. IEEE Computer Society. 2

[2] J.-W. Bian, Y.-H. Wu, J. Zhao, Y. Liu, L. Zhang, M.-M. Cheng, and I. Reid. An evaluation of feature matchers for fundamental matrix estimation. In *British Machine Vision Conference (BMVC)*, 2019. 2

[3] A. Bursuc, G. Tolias, and H. Jégou. Kernel local descriptors with implicit rotation matching. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 595–598, New York, NY, USA, 2015. ACM. 2, 5

[4] C. Choy, J. Park, and V. Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8958–8966, 2019. 2, 3, 4, 5, 6

[5] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems*, pages 2414–2422, 2016. 2, 3, 4, 5, 6

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 2

[7] J. Dong and S. Soatto. Domain-size pooling in local descriptors: Dsp-sift. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5097–5106, 2014. 2, 5

[8] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-net: A trainable cnn for joint detection and description of local features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 8

[9] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. Montiel. Visual slam for handheld monocular endoscope. *IEEE transactions on medical imaging*, 33(1):135–146, 2013. 2

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988. 2

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4, 5

[12] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018. 4, 5, 6, 8

[13] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017. 2, 4

[14] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel. Defslam: Tracking and mapping of deforming scenes from monocular sequences. *arXiv preprint arXiv:1908.08918*, 2019. 2

[15] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G. L. Gallia, R. H. Taylor, et al. Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data. 37(10):2185–2195, Oct. 2018. 1, 2, 6, 7

[16] H. Liao, W. Lin, J. Zhang, J. Zhang, J. Luo, and S. K. Zhou. Multiview 2d/3d rigid registration via a point-of-interest network for tracking and triangulation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12638–12647. Computer Vision Foundation / IEEE, 2019. 2, 3, 4

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 8

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 2, 3, 5, 6, 8

[19] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 168–183, 2018. 2, 6

[20] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. M. M. Montiel. Orbslam-based endoscope tracking and 3d reconstruction. In *CARE@MICCAI*, 2016. 1, 2

[21] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 4, 6, 7

[22] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, pages 4826–4837, 2017. 2, 3, 5, 6, 7

[23] P. Moulon, P. Monasse, R. Perrot, and R. Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016. 6

[24] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 4

[26] G. A. Puerto-Souza and G. L. Mariottini. Hierarchical multi-affine (hma) algorithm for fast and accurate feature matching in minimally-invasive surgical images. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2007–2012. IEEE, 2012. 6

[27] L. Qiu and H. Ren. Endoscope navigation and 3d reconstruction of oral cavity by visual slam with mitigated data scarcity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2197–2204, 2018. 2

[28] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision*, 120(3):300–323, 2016. 7

[29] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, pages 430–443, Berlin, Heidelberg, 2006. Springer-Verlag. 2

[30] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 7

[31] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2017. 2, 6

[32] A. Sinha, S. D. Billings, A. Reiter, X. Liu, M. Ishii, G. D. Hager, and R. H. Taylor. The deformable most-likely-point paradigm. *Medical Image Analysis*, 55:148 – 164, 2019. 1

[33] L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017. 6

[34] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake. Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. *IEEE Robotics and Automation Letters*, 3(4):4068–4075, 2018. 2

[35] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Ieee, 2008. 2, 4, 6, 7

[36] Y. Tian, B. Fan, and F. Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 661–669, 2017. 2

[37] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009. 2

[38] P. Truong, S. Apostolopoulos, A. Mosinska, S. Stucky, C. Ciller, and S. D. Zanet. Glampoints: Greedily learned accurate match points. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10732–10741, 2019. 2

[39] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille. Normface: l 2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049. ACM, 2017. 2

[40] A. R. Widya, Y. Monno, K. Imahori, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki. 3d reconstruction of whole stomach from endoscope video using structure-from-motion. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3900–3904, 2019. 1, 2

[41] H. Yang, W.-Y. Lin, and J. Lu. Daisy filter flow: A generalized discrete approach to dense correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3406–3413, 2014. 7