Learning-based Correspondences for Ophthalmic Image Registration

Chen Zhao
chen.zhao@epfl.ch

Ziyi Guan
ziyi.guan@epfl.ch

***Abstract*—We present our work on ophthalmic image registration which aims to register healthy eye images, diseased eye images with the same image type, and diseased eye images with different image types, in this report. The appearance variation caused by contrast and illumination changes, disease, and multiple modalities makes the registration challenging. We address this issue by using a pixel-level correspondence-based registration strategy. We employ deep learning networks to locate pixel-level keypoints and match the repeatable keypoints over the pairwise images. The registration is achieved by estimating a geometric transformation from the generated correspondences and then deforming the source image based on this transformation. Experiments show that the learning-based method is more robust to appearance variation than the hand-crafted one, and able to generalize well to the three different tasks.**

***Keywords*-Ophthalmic image, image registration, feature correspondence**

I. Introduction

Ophthalmic image registration, which aims to align consistent visual patterns over two images, is getting more and more attention in recent years [6], [20], [11], [1]. The registered images can provide additional information to ophthalmologists to deliver efficient patient care. In practice, ophthalmic images are often taken at different time and even with different devices, which could result in deformation and appearance variation. As shown in Fig. 1, the locations of consistent visual patterns, e.g., OHN and vessels, and appearance, e.g., color and texture, of the source image (a) and target image (b) are significantly different[1]. Moreover,

(a) Source image (b) Target image (c) Warped image

Figure 1. **Illustration of ophthalmic image registration.** Given two ophthalmic images, source image (a) and target image (b), the registration aims to align the consistent visual patterns, e.g., OHN and vessels. After warping the source image to the coordinate system of the target image by using an estimated geometric transformation, the consistent visual patterns are well registered (c).

as the two images tend to be partially overlapped, the non-overlapping regions are actually noises and should be ignored during the registration. Therefore, it is non-trivial to robustly detect the consistent visual patterns in different scenarios.

To address this issue, we propose to register images by using pixel-level correspondences. Specifically, we formalize the underlying deformation between two images as an affine transformation. The affine transformation has 6 degrees of freedom, which is capable of handling rotation, translation, and scale changes. In order to estimation this transformation, we first locate some keypoints on the images, and extract features from the local pattern around each keypoint. Given the location of each keypoint on the source image as well as its local feature description, we then find out the corresponding point on the target image, according to the local feature similarities. Therefore, the affine transformation can be computed over the generated correspondences.

Due to the existence of noises, the correspondences may contain some outliers, i.e., false matches, which could affect the accuracy of the estimated affine transformation. Consequently, how to get reliable correspondences is critical in our method. Most existing works propose to use hand-craft methods such as SIFT [13] to detect keypoints and describe their local features. However, we found that hand-crafted approaches cannot effectively establish correspondences in some scenarios, such as textureless images and images with varied appearance. Therefore, we propose to detect keypoints and describe local features by using a deep learning method called SuperPoint [4]. The acquired keypoints are then matched by employing another deep learning network SuperGlue [18]. The details can be found in Sec.III. We use these two methods to build up correspondences for all three tasks, i.e., healthy eye images, diseased eye images with the same image type, and diseased eye images with different image types. Experimental results demonstrate that the combination of SuperPoint and SuperGlue is superior to the hand-crafted method SIFT. The learning-based registration is also generalized well to different tasks even without retraining.

II. Related Works

Registration by correspondences. Correspondences have been widely used for image registration in the past twenty years. The generation of correspondences requires high-accuracy keypoint localization and distinctive local feature description. Lowe et al. [13], presented SIFT which achieves sub-pixel keypoint localization by detecting local maximum of the difference-of-Gaussian function. Local features around a keypoint are described as a 128d vector by computing the gradient orientation histogram of a local patch. The

[1]The images are from the published work [8]

local maximum detection in multi-scale spaces and gradient orientation histogram result in scale and rotation invariance, respectively. Morel et al. [14], modified SIFT to an affine invariant version by detecting keypoints and describing features in an affine augmented space. Rublee et al. [17], sped up SIFT by proposing a binarized feature descriptor. Such hand-crafted methods are able to effectively locate keypoints and distinctively extract local features when the images are well textured. However, for ophthalmic images, the problem of being textureless is inevitable. Moreover, the appearance variation caused by contrast and illumination changes, disease, and multiple modalities could also confuse the hand-crafted descriptors. In such context, many efforts have been put into employing deep learning networks to perform keypoint detection and feature description. Detone et al. [4] presented a self-supervised interest point detection and description network called SuperPoint. The network architecture includes an encoder and two separate decoders. The keypoint locations are marked after a channel-wise softmax function in the detector decoder, and the corresponding local features are described in the descriptor decoder. Since SuperGlue can be trained on indoor datasets where most images are textureless, this method shows superiority compared with hand-crafted methods facing the texturelessness challenge.

Registration by deep learning networks. Deep learning has been dominating most fields of computer vision and image processing such as object classification [9], [12], detection [16], [15], and segmentation [7], [10]. In the field of image registration, some efforts also have been done to predict a geometric transformation from an image pair by employing deep learning networks. Detone et al. [3], proposed to use a deep convolutional neural network to predict a homography matrix from two images. The pairwise training data are acquired by deforming the source image based on a randomly sampled homography matrix. Since the domain gap between the simulated data and real data is inevitable, such method has difficulty in generalizing to real-world applications.

Ophthalmic image registration. In the field of ophthalmic image registration, existing works align two images by two steps [11]. First, two images are registered based on the estimated geometric transformation. Second, a displacement is assigned to each pixel of the registered image to handle the local deformation. In this project, we found that the underlying transformation between two image can be properly represented by an affine matrix, and the local displacement is prone to resulting in over deformation. Therefore, we focus on affine matrix estimation in the following sections.

III. Method

A. Problem Formulation

As shown in Fig. 2, our method includes four steps: keypoint detection and feature description, feature matching, correspondence pruning, and image registration. Specifically, given a pair of images $(\mathbf{I}, \mathbf{I}^{'})$, two sets of keypoints $(\mathcal{P}, \mathcal{P}^{'})$ are detected on $(\mathbf{I}, \mathbf{I}^{'})$ separately by the decoder of SuperPoint. The corresponding local features $(\mathcal{D}, \mathcal{D}^{'})$ are extracted by another decoder of SuperPoint. Subsequently, a set of initial correspondences $\mathcal{C}$ is builded with $(\mathcal{P}, \mathcal{D}; \mathcal{P}^{'}, \mathcal{D}^{'})$ by using the matching network of SuperGlue. One may observe from Fig. 2 that some mismatches occur in $\mathcal{C}$. Therefore, a pruning strategy is utilized to remove false matches (outliers) and preserve correct matches (inliers). Finally, an affine matrix $\mathbf{A} \in \mathbb{R}^{2\times3}$ is estimated over the selected inliers, and image $\mathbf{I}$ is warped to the coordinate system of image $\mathbf{I}^{'}$ using the transformation $\mathbf{A}$. The details of each step will be introduced in the following subsections.

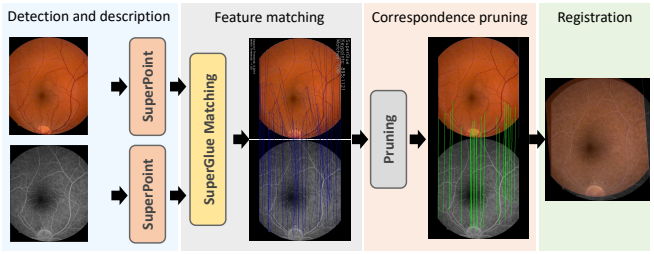

Figure 2. **The framework of our method.**

B. Feature Detection and Description

Fig. 3 illustrates the architecture of SuperPoint. Given a grayscale image $\mathbf{I}$ with the size (W, H), a VGG-style encoder is leveraged to extract features and down sample $\mathbf{I}$ into its counterpart with the size of (W_c, H_c) $(W_c = \frac{W}{8}, H_c = \frac{H}{8})$. Two decoder networks are then used to detect interest points (keypoints) and describe their local features separately. The interest point detector head computes $\mathcal{X} \in \mathbb{R}^{H_c\times W_c\times 65}$ and outputs a tensor sized $\mathbb{R}^{H\times W}$. The 65 channels correspond to non-overlapping 8×8 grid regions of pixels on $\mathbf{I}$ plus an extra "no interest point" dustbin. After a channel-wise softmax function, the local maximum is located in each 8×8 grid regions as the location of keypoint on $\mathbf{I}$. The descriptor head computes $\mathcal{D} \in \mathbb{R}^{H_c\times W_c\times D}$ and outputs a tensor sized $\mathbb{R}^{H\times W\times D}$. The down sampled feature map $\mathcal{D}$ is upsampled by an interpolation process, which brings the representation back to $\mathbb{R}^{H\times W}$. L2 normalization is performed over each D dimension descriptor vector to constraint the length to be unit.

In order to train the encoder and decoder networks, SuperPoint presents a self-supervised training strategy. Specifically, the encoder and interest point head are trained on synthetic data, where ground-truth keypoint locations can be easily acquired. After pre-training, the networks of SuperPoint are refined on real data. Given an unlabeled image, a sequence of randomly sampled homography transformations are performed to augment the raw image. the pre-trained encoder and keypoint decoder are employed on these data to generate pseudo-ground truth keypoints. Taking the raw image and its augmented counterpart as input, the encoder and two decoders can be jointly refined using the pseudo-ground truth labels. The loss function of SuperPoint is

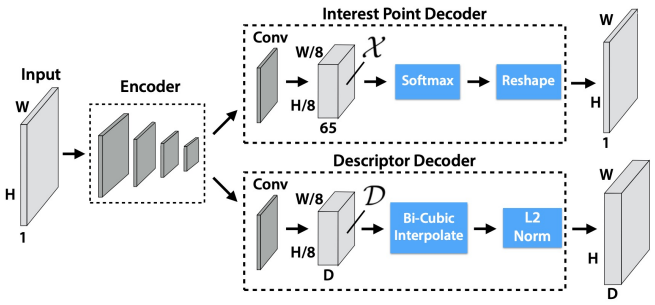

Figure 3. **The architecture of SuperPoint.**

formalized as

$$L(\mathcal{X},\mathcal{X}^{'},\mathcal{D},\mathcal{D}^{'};\mathcal{Y},\mathcal{Y}^{'}) = L_p(\mathcal{X},\mathcal{Y}) + L_p(\mathcal{X}^{'},\mathcal{Y}^{'}) + \lambda L_d(\mathcal{D},\mathcal{D}^{'}), \tag{1}$$

where $(\mathcal{Y},\mathcal{Y}^{'})$ represent the pseudo-ground truth labels on $(\mathbf{I},\mathbf{I}^{'})$, L_p and L_d denote detector loss and descriptor loss, respectively. L_p is computed as

$$L_p(\mathcal{X},\mathcal{Y}) = \frac{1}{H_cW_c}\sum_{h,w=1}^{H_c,W_c} -log(\frac{\exp(\mathbf{x}_{hwy})}{\sum_{k=1}^{65}\exp(\mathbf{x}_{hwk})}). \tag{2}$$

L_d is a contrastive loss which is denoted as

$$L_d(\mathcal{D},\mathcal{D}^{'}) = \frac{1}{(H_cW_c)^2}\sum\sum s*\max(0,m_p-\mathbf{d}^T\mathbf{d}^{'}) + (1-s)*\max(0,\mathbf{d}^T\mathbf{d}^{'}-m_n), \tag{3}$$

where s indicates $(\mathbf{d},\mathbf{d}^{'})$ is a pair of positive samplings, m_p and m_n are predefined positive and negative margins, respectively.

C. Feature Matching

After getting the keypoint sets $(\mathcal{P},\mathcal{P}^{'})$ and the corresponding features $(\mathcal{D},\mathcal{D}^{'})$, the next step is generating correspondences $\mathcal{C}$ over $(\mathcal{P},\mathcal{P}^{'})$. A straightforward solution is performing brute force matching for each $\mathbf{p}\in\mathcal{P}$. In particular, for each $\mathbf{p}\in\mathcal{P}$, we estimate the feature similarities with all keypoints in $\mathcal{P}^{'}$ and search for the most similar one $\mathbf{p}^{'}$. However, such strategy only utilizes the information of pairwise descriptors. Both keypoint locations and information among multiple elements in $(\mathcal{D},\mathcal{D}^{'})$ are ignored. To address this issue, we propose to employ SuperGlue to generate correspondences.

Fig. 4 demonstrates the architecture of SuperGlue. The network takes keypoint locations $\mathbf{p}$ and local features $\mathbf{d}$ as input. The intra information of $(\mathcal{P},\mathcal{D})$ and $(\mathcal{P}^{'},\mathcal{D}^{'})$, as well as inter information of $(\mathcal{P},\mathcal{P}^{'};\mathcal{D},\mathcal{D}^{'})$ are extracted by a self-attention network and cross-attention network, respectively. After the attention process, the output matching descriptors $(\mathbf{f}_i\in\mathcal{F},\mathbf{f}_j^{'}\in\mathcal{F}^{'})$ are capable of perceiving both intra and inter information. The score matrix $\mathbf{S}$ is then computed over the matching descriptors as

$$\mathbf{S}_{i,j} =< \mathbf{f}_i,\mathbf{f}_j^{'} >, \forall\mathbf{f}_i\in\mathcal{F},\mathbf{f}_j^{'}\in\mathcal{F}^{'}. \tag{4}$$

In order to match keypoint sets $(\mathcal{P},\mathcal{P}^{'})$, SuperGlue formalizes the matching step as an optimization problem and employs Sinkhorn Algorithm [19] to estimate assignments for keypoints in $\mathcal{P}$. Since some of the keypoints may locate on non-overlapping regions of the two images, which cannot be matched, a predefined threshold (0.2 in our experiments) is used to remove correspondences with the assignment scores lower than the threshold.

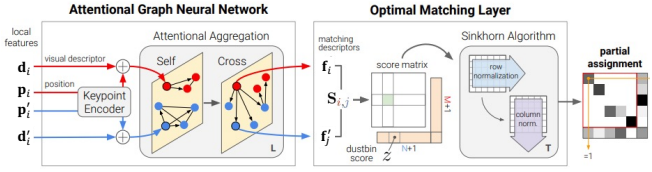

Figure 4. **The architecture of SuperGlue.**

D. Correspondence Pruning

As illustrated in Fig. 2, even with the combination of SuperPoint and SuperGlue, the initial correspondences still include some outliers. Therefore, a correspondence pruning strategy [21], [22] is crucial. In this paper, we propose to employ RANSAC [5] to prune initial correspondences and robustly estimate the affine transformation.

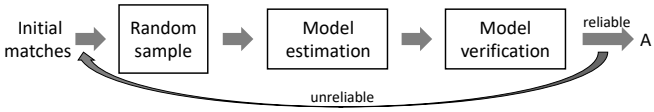

Figure 5. **The pipeline of RANSAC.**

Fig. 5 shows the pipeline of RANSAC which includes three steps. Taking the initial correspondences $\mathcal{C} = \{(\mathbf{p}_i,\mathbf{p}_j^{'}),\mathbf{p}_i\in\mathcal{P},\mathbf{p}_j^{'}\in\mathcal{P}^{'}\}$ as input, RANSAC first randomly samples a subset of correspondences (3 for affine matrix estimation). Second, an affine matrix is estimated by solving a system of linear equations as

$$\begin{cases} a_{11}*x_1+a_{21}*y_1+a_{13}=x_1^{'} \\ a_{12}*x_1+a_{22}*y_1+a_{23}=y_1^{'} \\ a_{11}*x_2+a_{21}*y_2+a_{13}=x_2^{'} \\ a_{12}*x_2+a_{22}*y_2+a_{23}=y_2^{'} \\ a_{11}*x_3+a_{21}*y_3+a_{13}=x_3^{'} \\ a_{12}*x_3+a_{22}*y_3+a_{23}=y_3^{'}, \end{cases} \tag{5}$$

where $\mathbf{p}=[x,y]^T$, $\mathbf{p}^{'}=[x^{'},y^{'}]^T$, and a indicates the element of $\mathbf{A}$ which is defined as

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23}. \end{bmatrix} \tag{6}$$

Since the subsampled correspondences may contain outliers, the estimated $\mathbf{A}$ could be unreliable. Then RANSAC suggests evaluating $\mathbf{A}$ by counting the number of correspondences which are consistent under the constraint of $\mathbf{A}$. These three steps are iteratively performed till the verification step confirms the reliability of $\mathbf{A}$. The correspondences which are consistent with the verified $\mathbf{A}$ are identified as inliers.

E. Image Registration

After getting the inliers from the initial set of correspondences, $\mathbf{A}$ is re-estimated over the acquired inliers by solving the system of linear equations which is similar to Eq. 5. The registration is implemented by warping all pixels of $\mathbf{I}$ to the coordinate system of $\mathbf{I}^{'}$ as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \mathbf{p} + \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} \tag{7}$$

IV. Experiments

A. Experimental Setups

We conducts experiments on three tasks as follows:

- **Task1**: registration for healthy eye images.
- **Task2**: registration for diseased eye images with the same image type on the same date and across different dates.
- **Task3**: registration for diseased eye images with different image types on the same date and across different dates.

Since the dataset contains ophthalmic images of several patients, the images should not be registered in the following cases: images of different patients, images of left and right eyes, and images centered on OHN and macula. Therefore, we classify images into separates sets according to the patient ID, left/right eye, and image center. Pairwise images are randomly sampled from a specific set based on the task requirement.

B. Implementation Details

Due to the privacy concerns, we can only access the data on a specific server without GPUs. Moreover, the provided ophthalmic images are not labelled with ground-truth information about the keypoint locations or affine matrices. Therefore, it is impractical to retrain SuperPoint and SuperGlue networks. Alternatively, in our experiments, we employ the models of SuperPoint and SuperGlue pretrained on ScanNet [2]. To tackle to problem of different resolutions, we resize images to a constant resolution (1024×1024 in our experiments) before registration.

C. Qualitative Results

Fig. 6 demonstrates the qualitative results of SuperPoint+SuperGlue (SP+SG) and SIFT on the three tasks [2]. For the task1 and task2, both SP+SG and SIFT can generate promising warped results. However, the selected inliers are still inconsistent for SIFT on task3, where the appearance between two images is significantly different. By contrast, SP+SG is still capable of acquiring reliable inliers in such context, which leads to decent registration results.

[2]The results are masked because of the privacy concerns, please contact with Tomasoni Mattia(mattia.tomasoni@fa2.ch) to access the results.

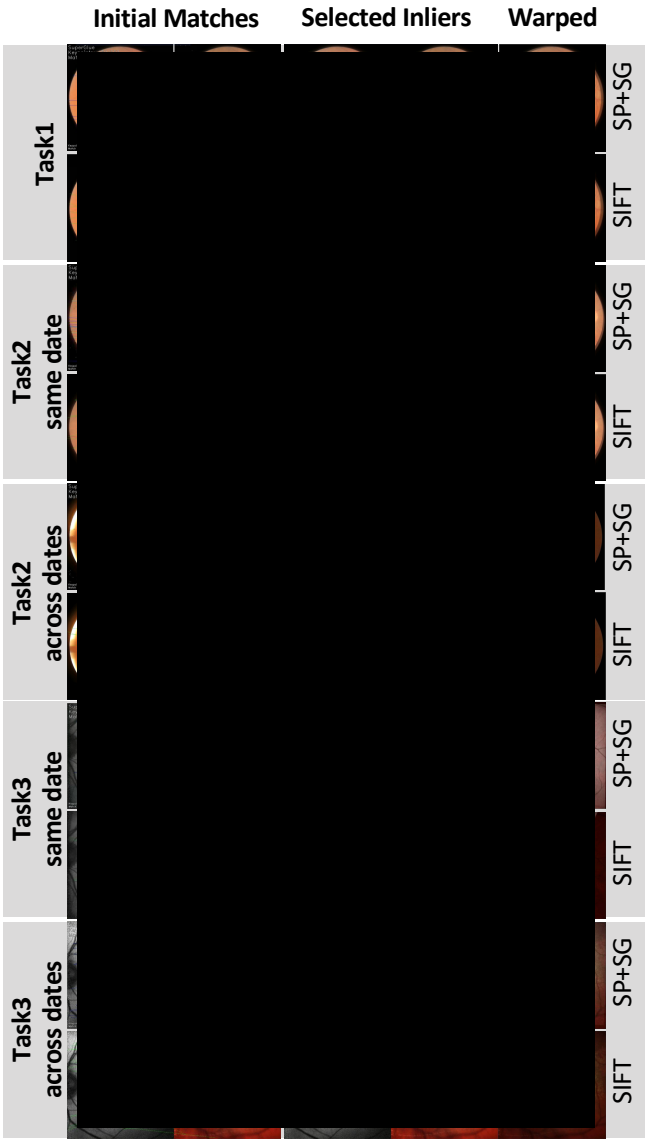

Figure 6. **Qualitative Results of SuperPoint+SuperGlue (SP+SG) and SIFT on three tasks.** For the warped images, we average RGB values of the warped source image and original target image for better visualization.

V. Conclusion

In this report, we tackled the problem of ophthalmic image registration by employing learning-based correspondences. A combination of SuperPoint and SuperGlue is utilized to establish correspondences between pairwise images. We used RANSAC to prune the initial correspondences and robustly estimate the geometric transformation. Experiments have shown that SuperPoint+SuperGlue is more robust to appearance changes compared with SIFT.

REFERENCES

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew 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.

[5] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[6] Yunguan Fu, Nina Montaña Brown, Shaheer U Saeed, Adrià Casamitjana, Zachary Baum, Rémi Delaunay, Qianye Yang, Alexander Grimwood, Zhe Min, Stefano B Blumberg, et al. Deepreg: a deep learning toolkit for medical image registration. *arXiv preprint arXiv:2011.02580*, 2020.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[8] Álvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. *Procedia Computer Science*, 126:97–104, 2018.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[10] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[11] Zhang Li, Fan Huang, Jiong Zhang, Behdad Dashtbozorg, Samaneh Abbasi-Sureshjani, Yue Sun, Xi Long, Qifeng Yu, Bart ter Haar Romeny, and Tao Tan. Multi-modal and multi-vendor retina image registration. *Biomedical optics express*, 9(2):410–422, 2018.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[14] Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM journal on imaging sciences*, 2(2):438–469, 2009.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[17] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[18] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[19] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[20] Shan Suthaharan, Ethan A Rossi, Valerie Snyder, Jay Chhablani, Raphael Lejoyeux, José-Alain Sahel, and Kunal K Dansingani. Laplacian feature detection and feature alignment for multimodal ophthalmic image registration using phase correlation and hessian affine feature space. *Signal processing*, 177:107733, 2020.

[21] Chen Zhao, Zhiguo Cao, Chi Li, Xin Li, and Jiaqi Yang. Nm-net: Mining reliable neighbors for robust feature correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 215–224, 2019.

[22] Chen Zhao, Yixiao Ge, Feng Zhu, Rui Zhao, Hongsheng Li, and Mathieu Salzmann. Progressive correspondence pruning by consensus learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6464–6473, 2021.