Reverse Knowledge Distillation: Training a Large Model using a Small One for Retinal Image Matching on Limited Data

Sahar Almahfouz Nasser* Nihar Gupte*
Amit Sethi
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India
194072001@iitb.ac.in, 213070002@iitb.ac.in, asethi@iitb.ac.in

[1]

Abstract

Retinal image matching (RIM) plays a crucial role in monitoring disease progression and treatment response as retina is the only tissue where blood vessels can be directly observed. However, datasets with matched keypoints between temporally separated pairs of images are not available in abundance to train transformer-based models. Firstly, we release keypoint annotations for retinal images from multiple datasets to aid further research on RIM. Secondly, we propose a novel approach based on reverse knowledge distillation to train large models with limited data while preventing overfitting. We propose architectural modifications to a CNN-based semi-supervised method called SuperRetina [22] that helps improve its results on a publicly available dataset. We train a computationally heavier model based on a vision transformer encoder, utilizing the lighter CNN-based model. This approach, which we call reverse knowledge distillation (RKD), further improves the matching results even though it contrasts with the conventional knowledge distillation where lighter models are trained based on heavier ones is the norm. Further, we show that our technique generalizes to other domains, such as facial landmark matching.

1. Introduction

Keypoint detection and matching, often referred to as feature point extraction or feature detection, can be used as a foundational task in computer vision to aid higher-level tasks, such as object recognition, fine-grained matching, image registration, image stitching, pose estimation, facial recognition, depth estimation from stereo image pairs, and augmented reality. The core objectives in keypoint detection and matching are the identification and localization of salient points or landmarks in images, and matching corresponding keypoints across images. Alongside location information, each point has an associated feature descriptor that aids in recognition or matching with corresponding points in other images. A keypoint, therefore, must exhibit characteristics that are locally distinctive and robust to image augmentations. That is, keypoint detection algorithms must pinpoint image locations with local uniqueness in terms of intensity, color, or texture. Furthermore, these points must be associated with features that are invariant to changes in scale, rotation, and illumination.

Over the years, a multitude of methods have been developed for keypoint detection. These methods span classical techniques like the Harris corner detector [31], scale-invariant feature transform (SIFT) [24], and speeded-up robust features (SURF) [8], to more recent approaches rooted in deep learning, such as oriented fast and rotated BRIEF (ORB) [30] and SuperPoint [12]. Various methods have been proposed for keypoint matching in retinal images. Addison et al. [6] introduced the low-dimensional step pattern analysis (LoSAP) technique for image registration. LoSAP adeptly handles intensity changes and remains invariant to rotation. However, the SPA descriptor used in LoSAP lacks the discriminatory power required for identifying specific eye identities. Truong et al. [37] presented a semi-supervised CNN-based feature point detector known as Greedily-Learned Accurate Match points (GLAMpoints), designed specifically for matching and registering retinal images. GLAMpoints utilizes deep learning techniques to enhance the accuracy and precision of keypoint matching. Another approach, proposed by Hernandez et al. [17], involves a registration framework based on eye modeling. This framework concurrently estimates eye pose and shape and addresses the registration problem as a 3D pose estimation task, utilizing corresponding points in the retinal images. We present a technique to train a large model for keypoints detection with limited data.

[1] * Indicates equal contribution

Our work makes several contributions. Firstly, we release annotations for a meticulously curated dataset exclusively tailored for the detection of keypoints in retinal images. Secondly, we introduce an architecture for keypoint detection and matching that outperforms the state-of-the-art (SOTA) SuperRetina model [22]. Thirdly, we compare a CNN with heavier transformer model on limited data training regime. Lastly, we train a large model to emulate a smaller model on limited data using reverse knowledge distillation (RKD) and show that the former generalizes better even on limited data when trained this way. In this approach, the larger model (student) imbibes knowledge from the smaller model (teacher) to enhance its own performance, ultimately surpassing its teacher's capabilities. Our findings are further corroborated by our experimentation on images from a vastly different domain, namely, keypoints detection in facial images.

2. Related work

In this section, we shall examine notable techniques for detecting keypoints. We will also furnish an outline of vision transformers, underlining their import in the domain of computer vision. Moreover, we will explore diverse strategies utilized for training vision transformers under conditions of constrained data availability.

2.1. Keypoint detection

Traditional keypoint detection algorithms have held prominence in computer vision applications for decades. These algorithms identify keypoints in images that remain invariant to scaling, rotation, and lighting variations. Subsequently, they characterize the local image patch around these keypoints using a set of features, facilitating the matching of keypoints between disparate images and object recognition. However, these techniques possess certain drawbacks, such as high computational complexity, diminished accuracy in the face of extreme lighting and viewpoint changes, and challenges in managing occlusions and cluttered backgrounds.

In recent years, deep learning-based keypoint detection algorithms have emerged as a promising alternative. These algorithms possess the capability to autonomously learn robust and discriminative features directly from data. As a result, they are better equipped to handle intricate and diverse image variations. This has led to their application in various domains, including object detection, semantic segmentation, and image retrieval.

In the domain of deep learning, various types of keypoint detection algorithms exist, encompassing supervised, semi-supervised, self-supervised, and unsupervised techniques. Supervised techniques necessitate annotated data, where keypoints are manually labeled in training images. Such algorithms prove advantageous in scenarios where a substantial volume of labeled data is available, as observed in facial recognition or object detection. Conversely, unsupervised techniques function independently of labeled data. Instead, the network learns to identify keypoints by maximizing specific objectives, such as information preservation during feature extraction. These methods are particularly valuable in contexts where obtaining labeled data is challenging or expensive, as in medical imaging or remote sensing.

Prominent deep learning-based keypoint detection methods include UnsuperPoint, SuperPoint, GLAMpoints, and SuperRetina. UnsuperPoint [10] introduces an innovative unsupervised training approach, employing a blend of differentiable soft nearest neighbor loss and unsupervised clustering loss. SuperPoint [12] represents a self-supervised deep learning-based algorithm for keypoint detection and description. It employs a novel loss function for training on unlabeled images, rendering it more adaptable and scalable across diverse applications. The loss functions, including geometric consistency loss and descriptor matching loss, prompt the network to learn predicting the spatial placement of keypoints and their descriptors devoid of supervision. It relies on a convolutional neural network (CNN) to extract keypoints and descriptors from images.

The primary distinction between UnsuperPoint and SuperPoint lies in their training methodologies. While SuperPoint adopts a self-supervised approach, UnsuperPoint follows an unsupervised path. Additionally, UnsuperPoint achieves the SOTA performance across various benchmarks, even surpassing SuperPoint in challenging scenarios marked by significant viewpoint alterations and illumination shifts. GLAMpoints [37] emerges as a semi-supervised deep learning-based algorithm for interest point detection and description. It employs a unique greedy training strategy for end-to-end learning of keypoint detection and description. This strategy involves learning to select the most precise keypoints and their descriptors, yielding heightened accuracy and efficiency. GLAMpoints outperforms both SuperPoint and UnsuperPoint in accuracy and efficiency, particularly in demanding scenarios encompassing substantial viewpoint changes, scaling, rotation, and benchmarks like HPatches [7]. Furthermore, GLAMpoints is adept at accommodating multiple object instances within the same image, making it suitable for multi-object tracking and matching. SuperRetina [22] signifies a semi-supervised approach for keypoint detection and description in retinal images. The technique leverages both labeled and unlabeled data to enhance the performance of the keypoint detector and descriptor. It comprises three main components: a supervised keypoint detector, an unsupervised keypoint descriptor, and a semi-supervised loss function that amalgamates labeled and unlabeled data.

The proposed method employs an iterative refinement

process to increase the accuracy and robustness of keypoint matches. This process entails the removal of outlier matches and the addition of new matches based on geometric constraints.

2.2. Vision transformers

Inspired by successful transformer models in natural language processing, vision transformers adopt self-attention for visual data processing [23,40]. Treating images as token sequences, they excel at capturing global dependencies. Despite their performance, training vision transformers with limited data is challenging due to their complexity and risk of overfitting.

Strategies to address this include data augmentation, generating examples through transformations [35]. Transfer learning leverages pre-trained models, fine-tuning on smaller datasets [38]. Regularization methods like dropout and weight decay prevent overfitting [35], enhancing generalization.

2.3. Knowledge distillation

Knowledge distillation [15] is a technique wherein a pre-trained model (referred to as the teacher model) is employed to guide the training of another model (the student model). The student model learns to replicate the predictions or internal data representations (features) of the teacher model. Typically, this is done to transfer the knowledge and generalization capabilities of the larger teacher model into a smaller student model.

In our work, we adopt reverse distillation: a small CNN-based model serves as the teacher model, while a large transformer-based model functions as the student model. We hypothesize that larger models can encounter overfitting when fitting a smaller-dimensional output. However, this issue can potentially be circumvented if they are trained to accommodate a larger dimensional representation (feature vector) [21].

3. Datasets

For training a retinal image keypoint detection network, precise keypoint labels are vital. We used the FIRE dataset for testing, as previous methods did. Yet, private dataset access was elusive, so we formed training annotations from available datasets meant for other tasks like retinal disease classification. Given our limited expertise in the domain, our focus centered on normal images. This approach aimed to bolster keypoint detection skills, encompassing both normal and abnormal images, including those within the FIRE dataset. This section details our dataset, annotations, and the FIRE dataset.

3.1. MeDAL-Retina Dataset

Our dataset consists of 261 retinal images curated from multiple public datasets, and it is divided into 208 for training and 61 for validation [14]. These images were meticulously annotated to identify keypoints at intersections, crossovers, and bifurcations, with detected keypoints ranging from 18 to 86 per image, averaging 42.96 ± 14.03. The distribution is visually depicted in the supplementary data distribution figure [32]. To compile the dataset, we merged 201 normal images from the e-ophtha dataset [1] and 60 images from the retinal disease classification dataset [5], as shown in the supplemental material's dataset figure [32]. An annotation team of five engineering students executed the process, taking approximately five minutes per image and eight minutes for a pair. A Python script facilitated the annotation.

In Section 4, we explore the use of Swin UNETR [16] as a network backbone. We undertook self-supervised Swin UNETR training, necessitating a sizable dataset due to significant distribution differences from ImageNet [11]. We sourced $\sim 1.9K$ images from various online resources [2–4, 18, 33]. This dataset also served for descriptor decoder training.

Preprocessing involved z-score normalization, followed by contrast limited adaptive histogram equalization (CLAHE) and gamma correction. The preprocessed images were normalized by dividing by 255. The green channel was consistently used for its high information content in retinal images.

3.2. FIRE Dataset

The FIRE dataset centers on fundus image registration, comprising 129 retinal images [18]. These images were categorized into 134 pairs based on overlap and deformation levels, each assigned to specific categories: S, P, and A. In the S category, 71 image pairs exhibit substantial overlap ($> 75\%$) and minimal anatomical differences, showing brightness changes, slight shifts, and rotations. The P category involves 49 pairs with smaller overlaps, displaying significant shifts and rotations. The A category holds 14 pairs with large overlaps, acquired at different times, resulting in notable anatomical changes like spots, cotton-wool patches, and increased vessel tortuosity.

Images possess a resolution of 2912×2912 pixels and a $45°$ field of view in both dimensions. These images were sourced from 39 patients. Examples from our dataset [14] and the FIRE dataset are shown in the supplemental material's dataset figure [32].

3.3. Wider Facial Landmarks in-the-wild dataset

We additionally tested our method on facial landmark detection using the Wider Facial Landmarks in-the-wild

(WFLW) dataset.It comprises 10,000 faces, with 7,500 designated for training and 2,500 for testing [39], where each face has 98 landmarks. For more details, please see the supplementary materials.

4. Proposed Method

SuperRetina [22] represents a cutting-edge technique for identifying crucial keypoints in retinal images. Derived from the SuperPoint model [12], SuperRetina is a tailored version for robust retinal image analysis. It uses a semi-supervised learning framework that adeptly combines supervised and unsupervised techniques to maximize the utilization of limited labeled retinal image data. The network architecture comprises an encoder for extracting downsampled feature maps, and two decoders – one for detecting keypoints and another for generating descriptors for these keypoints. The keypoint detector is trained with a blend of labeled and unlabeled data, while the descriptor training employs self-supervised learning.

Rigorous experimentation on benchmark retinal image datasets demonstrates its superior performance in keypoint detection and matching accuracy, surpassing existing methods [26].

4.1. UNet-empowered SuperRetina

SuperRetina's architecture follows U-Net [29]. Its shallow encoder includes one convolutional layer followed by three blocks, each with two convolutional layers, a 2×2 max pooling layer, and ReLU activation. For keypoint decoding, three blocks house two convolutional layers each, utilizing encoder skip connections for bilinear upsampling, ReLU activation, and concatenation, yielding feature maps of the input image's size. The detection map (P) is generated through a convolutional block with three convolutional layers and a sigmoid activation.

In descriptor decoding, encoder feature maps downsize to $\frac{w}{16} \times \frac{h}{16} \times d$. A transposed convolutional block upsamples to match input image size, yielding a full-sized descriptor tensor (D) of dimensions $h \times w \times d$, L2-normalized.

We enhanced SuperRetina by refining the encoder. This entails architectural adjustments using CNN and transformer approaches to boost overall outcomes.

4.2. Large kernel-empowered SuperRetina

Inspired by Jia et al.'s work [20], which effectively boosts a basic U-Net to rival the potent transformer architecture, our approach involves embedding kernels of varying sizes into each layer of SuperRetina's encoder. This effectively captures long-range dependencies in retinal image matching. Our modification focuses on SuperRetina's encoder architecture. Instead of using a 3×3 kernel in each layer, we employ three kernels of different sizes: 1×1, 3×3, and 5×5. These changes propel the enhanced SuperRetina beyond the SOTA method for RIM. It outperforms all prior approaches assessed on the FIRE dataset, excelling across all evaluation metrics and establishing its supremacy.

4.3. Swin UNETR-empowered SuperRetina

After observing promising outcomes from experiments involving larger kernels to expand SuperRetina's encoder receptive field, we considered boosting performance further by introducing a transformer-based encoder. This choice aligns with transformers' inherent ability to capture extended dependencies, advantageous for our task. Nonetheless, training a transformer with limited data poses substantial challenges, detailed in the following paragraphs.

To comprehensively convey our modifications to SuperRetina's architecture, we first introduce Swin Transformer [23] and Swin UNETR [16] concepts, which serve as foundational references. We then detail our specific architectural adjustments to SuperRetina and outline the distinctive approach used to train this computationally intensive model on our small dataset.

4.3.1 Swin transformer and Swin UNETR

The core reason behind the Swin Transformer's success lies in its hierarchical structure [23]. Rather than processing the entire image as a single entity, it breaks the image into non-overlapping patches, treating each patch as a token. It introduced the notion of shifted windows, where tokens selectively interact with a restricted nearby set of tokens, avoiding attention to all tokens. Utilizing a multi-stage hierarchical design, the Swin Transformer adeptly captures extensive dependencies while keeping computational complexity manageable.

Swin UNETR was purpose-built for semantic segmentation, fusing Swin Transformer and CNNs in a UNet-style setup for pixel-level segmentation [16]. The UNet's prominent advantage is its deployment of skip connections. In our research, we swapped SuperRetina's encoder with Swin UNETR's encoder.

4.3.2 Reverse knowledge distillation

In our research, we addressed the challenge of training a transformer model when faced with limited data. Our aim was to develop models capable of handling complex dependencies over extended sequences. Despite our diligent efforts to utilize self-supervision and transfer learning techniques, the performance of our transformer model consistently lagged behind that of a CNN model. To tackle this issue, we turned to a strategy known as knowledge distillation.

In machine learning, "knowledge distillation" typically involves the process of transferring knowledge from a larger

and more complex model (referred to as the "teacher") to a smaller and simpler model (the "student") [15]. This usage of the term is the most prevalent and aligns with the concept of distillation in the usual sense, where the distillate is smaller than the original substance. However, when the knowledge transfer goes from a smaller model to a larger one, it deviates from this conventional definition, and we propose using the term reverse knowledge distillation (RKD). That is, RKD uses the simplified learning in a smaller teacher model to enhance the performance of a larger student model having more capacity. The simplified knowledge structure learned by teacher may have a regularization effect on the student. In practice, we found that a larger student can generalize better than the teacher.

The loss function architecture supporting our reverse knowledge distillation strategy consists of a weighted sum of two essential components. Initially, we compute the loss between the predictions of the student network and the actual output, akin to conventional loss functions. Simultaneously, we introduce a distillation loss between the outputs of the student network and the teacher network. This innovative paradigm introduces additional steps in each training iteration. In addition to the standard training procedures, our approach incorporates keypoint heatmap generation using the teacher model. The subsequent computation of the dice loss between the keypoint heatmaps of the student and teacher models is denoted as "l_{clf}^{RKD}." Furthermore, we integrate contrastive matching between descriptors from both the teacher and student models, which we refer to as "l_{des}^{RKD}." The integration of both the detection RKD loss l_{clf}^{RKD} and descriptor RKD loss l_{des}^{RKD} seamlessly fits into the original detector and descriptor loss functions. For a comprehensive understanding of the original SuperRetina loss functions, please refer to [22].

Additionally, the distinction between traditional knowledge distillation and reverse knowledge distillation resonates with the insights presented in the work of Jiang et al. [21]. This work showcases the potential of knowledge transfer and demonstrates the effectiveness of their reverse knowledge distillation technique in various classification tasks. Their experiments, which employ shallower CNNs as teachers and deeper CNNs as students, underscore the significance of loss calibration in achieving superior performance. While traditional knowledge distillation primarily aims to improve the accuracy of smaller models using insights from larger models, their focus lies in enhancing the confidence calibration within larger, complex models by drawing on insights from smaller models. This nuanced distinction aligns harmoniously with the fundamental motivations and methodologies of our proposed technique.

Equations 1, 2, 3, and 4 represent the detector loss of our Swin UNETR-boosted SuperRetina model, with SuperRetina/LK-SuperRetina as a teacher. See Fig. 1.

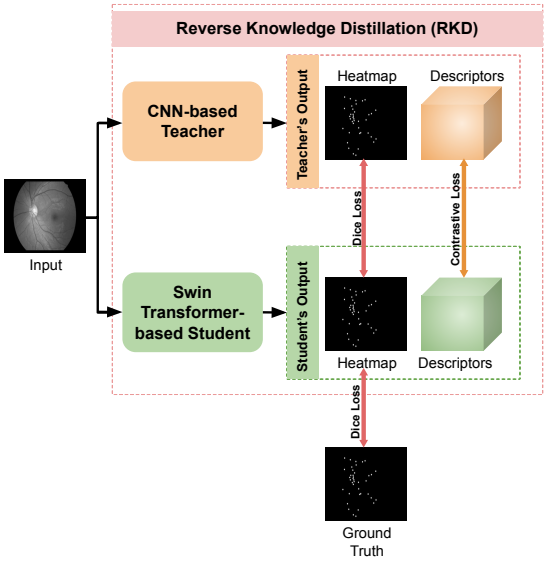

Figure 1. The architecture of the proposed method.

The total loss of the detector 1 is

$$l_{det} = l^{'}_{clf} + l_{geo} \tag{1}$$

$$l^{'}_{clf} = l_{clf} + l^{RKD}_{clf} \tag{2}$$

$$l_{clf}(I;Y) = 1 - \frac{2.\sum_{i,j}(P \circ \tilde{Y})_{i,j}}{\sum_{i,j}(P \circ P)_{i,j} + \sum_{i,j}(\tilde{Y} \circ \tilde{Y})_{i,j}} \tag{3}$$

where $\tilde{Y}$ is the smoothed version of the binary ground truth labels Y of the keypoints after blurring them with a 2D Gaussian.

$$l^{RKD}_{clf}(I_S;I_T) = 1 - \frac{2.\sum_{i,j}(P_S \circ P_T)_{i,j}}{\sum_{i,j}(P_S \circ P_S)_{i,j} + \sum_{i,j}(P_T \circ P_T)_{i,j}}, \tag{4}$$

where P_S stands for the keypoint heatmap of the student, and P_T refers to the keypoint heatmap of the teacher model, l_{geo} is the Dice loss between the output heatmap of the student model when the input is the image I, and the inverse projection of the heatmap produced by the student when the input to it is the augmented version of the image I , $I^{'}$. Similarly, the new descriptor loss is a combination of the original descriptor loss and the reverse knowledge distillation loss as in 5

$$l_{Des} = l_{des} + l^{RKD}_{des} \tag{5}$$

When feeding the image I and its augmented version $I^{'}$ to the student network, we optain two tensors for the descriptors D, and $D^{'}$. For each keypoint (i,j) in the non-maximum supressed keypoint set $\tilde{P}$, two distances are computed $\Phi^{rand}_{i,j}$ between the descriptors of (i,j) in the set $\tilde{P}$

and a random point from registered heatmap $H(\tilde{P})$. And $\Phi_{i,j}^{hard}$ the minimal distance. As 6 depicts

$$l_{des}(I,H) = \sum_{(i,j)\in\tilde{P}} max(0, m + \Phi_{i,j} - \frac{1}{2}(\Phi_{i,j}^{rand} + \Phi_{i,j}^{hard})) \tag{6}$$

Similar to l_{des}, we compute l_{des}^{RKD} between the descriptors generated when passing I to the student model, and the descriptors generated when passing I to the teacher model. For further details on the reverse knowledge distillation method and the loss functions.

5. Experiments

We rigorously evaluated our proposed technique by comparing it to various approaches in the retinal image matching task. Table 1 presents a comparison between our leading technique and alternative methods for retinal image matching. This encompasses both traditional and deep learning-based approaches, and our results clearly indicate the effectiveness of our method, surpassing all others.

The evaluation metrics comprise two aspects: failure rate and acceptance rate. The failure rate is determined by the number of matches between a query image and its reference. A registration is considered unsuccessful if the matches are fewer than 4, the minimum required for estimating a homography, H. Conversely, the acceptance rate is computed for each query point in the image. It involves calculating the L2 distance between a registered point and its corresponding reference point in the reference image. The median distance defines the median error (MEE) for each query image, with the maximum distance denoting the maximum error (MAE). For acceptance, MEE must be under 20, and MAE must be under 50. Otherwise, the registration is deemed inaccurate, please refer to MAE and MEE figures in the supplementary material.

To assess a method's overall performance, we report the area under the receiver operating characteristic curve (AUC). AUC estimates the acceptance rates' expectation concerning the decision threshold, reflecting performance across all methods. Additionally, AUC is separately computed for each category (Easy, Mod, Hard), and their mean (mAUC) is used as an overall measure.

In conclusion, the superior method exhibits a higher acceptance rate or AUC and lower inaccuracies or failures. For analyzing the impact of various encoder modifications, diverse techniques for training the Swin UNETR encoder, and differing kernel sizes of the large kernel-boosted SuperRetina, we conducted ablation studies, detailed below.

5.1. Different kernel sizes

By conducting an ablation study centered around kernel size, we found that a blend of kernels with dimensions of 1×1, 3×3, and 5×5 yielded the most favorable results for the large kernel-enhanced SuperRetina. Refer to the ablation studies in Table 2.

5.2. Transfer learning

To mitigate the challenge of training a transformer model with limited data, we turned to transfer learning. Our approach involved amassing a substantial dataset of retinal images from online sources. This dataset was leveraged to train a Swin UNETR model across diverse tasks, including image inpainting and angle prediction. The pretrained encoder weights from this model were then adopted as initial weights for the SuperRetina's encoder. The outcomes of employing a pretrained Swin UNETR as the backbone of SuperRetina are presented in Tab. 2. Although this model outperforms others in terms of one specific evaluation metric, namely AUC-Mod, the overall performance is not consistent across all metrics.

5.3. Reverse knowledge distillation

As highlighted in the study by Dosovitskiy et al. (2020) [13], transformers have a high demand for extensive training data and tend to perform less effectively than CNNs when dealing with limited data. Reverse knowledge distillation involves using the knowledge acquired by a smaller model, such as a CNN, to train a larger model, like a transformer. Typically, the knowledge of a larger model is employed to train a smaller model in knowledge distillation, as discussed in works like Chen et al. (2022) [9], Touvron et al. (2021) [36], and Hinton et al. (2015) [19].

In our research, the CNN serves as the "teacher" model, previously trained for the keypoint detection task. The goal is to transfer the CNN's knowledge and generalization abilities to a transformer model, referred to as the "student" model. The distillation process entails training the student model to replicate the behavior of the teacher model, often by using the output probabilities or feature representations of the teacher model as soft targets during the student model's training. By emulating the teacher's predictions, the student model effectively captures the teacher's knowledge and decision-making process.

While we initially anticipated that distilling knowledge from a CNN to a transformer could harness both the CNN's local feature extraction abilities and the transformer's long-range dependency modeling, our experimental results indicate that even after knowledge distillation, the transformer model's performance fell short of our expectations. Please refer to Table 2. To address this, as evident in Table 2, we introduced a 50% dropout, resulting in a significant performance boost for the Swin UNETR-empowered SuperRetina. This adjustment led to 100% accuracy on the testing dataset. The improvement can be attributed to the network's enhanced generalization on testing data, achieved by reduc-

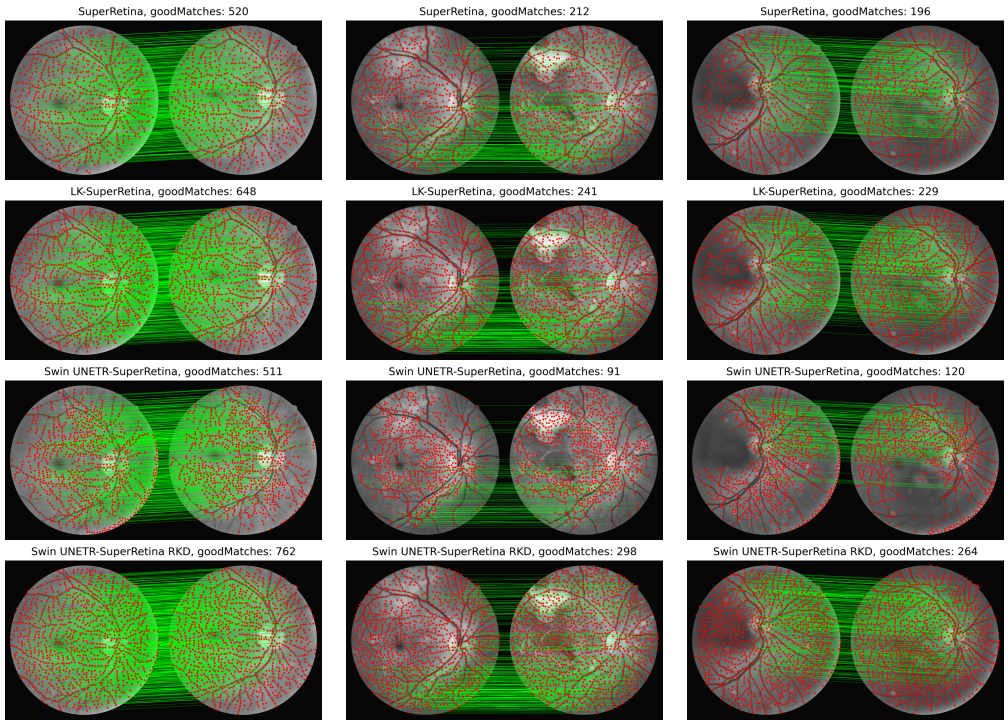

Figure 2. Comparison of our proposed methods on three example scenarios from FIRE dataset [18]: class S (easy), class A (moderate), and class P (hard) from left to right. LK stands for large kernel, RKD refers to Reverse Knowledge Distillation with 50% Dropout.

Method	Failed	Inaccurate	Acceptable	AUC-Easy	AUC-Mod	AUC-Hard	mAUC
SIFT, IJCV04 [24]	**0.00%**	20.15%	79.85%	0.903	0.474	0.341	0.573
PBO, ICIP10 [25]	0.75%	28.36%	70.89%	0.844	0.691	0.122	0.552
REMPE, JBHI20 [17]	**0.00%**	02.99%	97.01%	**0.958**	0.660	0.542	0.720
SuperPoint, CVPRW18 [12]	**0.00%**	05.22%	94.78%	0.882	0.649	0.490	0.674
GLAMpoints, ICCV19 [37]	**0.00%**	07.46%	92.54%	0.850	0.543	0.474	0.622
R2D2, NIPS19 [27]	**0.00%**	12.69%	87.31%	0.900	0.517	0.386	0.601
SuperGlue, CVPR20 [34]	0.75%	03.73%	95.52%	0.885	0.689	0.488	0.687
NCNet, TPAMI22 [28]	**0.00%**	37.31%	62.69%	0.588	0.386	0.077	0.350
SuperRetina [22]	**0.00%**	01.50%	98.50%	0.940	**0.783**	0.542	0.755
Ours-1 (Large kernel-SuperRetina)	**0.00%**	00.75%	99.25%	0.942	**0.783**	**0.558**	**0.761**
Ours-2 (Swin UNETR-SuperRetina)	**0.00%**	**00.00%**	**100.0%**	0.935	0.780	0.550	0.755

Table 1. A comparison among various techniques for retinal image matching, specifically focusing on the results obtained when testing the methods on the FIRE dataset [18]. Our proposed method demonstrates superior performance when compared to both traditional and deep learning approaches. Ours-1 refers to large-kernel-empowered SuperRetina, while Ours-2 refers to Swin UNETR-empowered SuperRetina with SuperRetina as a teacher and drop out 50%. In the table we provide the percentage values [%] of failed, inaccurate, and acceptable.

ing overfitting on the training data, coupled with reverse knowledge distillation. In conclusion, regularization strategies like dropout play a crucial role in reverse knowledge distillation. The dropout technique showcased an improved generalization capability in the student model, enabling it to surpass its teacher model on the testing dataset. For a visual

Method	Failed	Inaccurate	Acceptable	AUC-Easy	AUC-Mod	AUC-Hard	mAUC
SuperRetina [22], KS 3×3	**0.00%**	01.50%	98.50%	0.940	**0.783**	0.542	0.755
LK-SuperRetina, KS $1 \times 1, 3 \times 3, 5 \times 5$	**0.00%**	00.75%	99.25%	0.942	**0.783**	0.558	**0.761**
LK-SuperRetina, KS $1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7$	**0.00%**	02.25%	97.74%	0.925	0.717	0.502	0.714
Swin UNETR-SuperRetina, Trained from scratch	**0.00%**	16.55%	83.45%	0.891	0.649	0.318	0.619
Swin UNETR-SuperRetina, SuperRetina as teacher w/o dropout (DO)	**0.00%**	01.50%	98.50%	**0.947**	0.769	0.549	0.755
Swin UNETR-SuperRetina, SuperRetina as teacher, DO 50%	**0.00%**	**00.00%**	**100.0%**	0.935	0.780	0.550	0.755
Swin UNETR-SuperRetina, LK-SuperRetina as teacher, DO 50%	**0.00%**	00.75%	99.25%	0.914	0.774	0.558	0.749
Pretrained Swin UNETR-SuperRet., LK-SuperRet. as teacher, DO 50%	**0.00%**	00.75%	99.25%	0.928	0.774	**0.559**	0.754

Table 2. Ablation studies on FIRE dataset [18], where KS represents the kernel size, and DO is the drop out percentage

Method	SuperRetina	Swin U-SR	RKD-SR
NME(%)	20.43	11.15	**10.92**

Table 3. Results on facial landmarks, where Swin U-SR is Swin UNETR-SuperRetina, and RKD-SR is RKD-SuperRetina. We got RKD's results for $\lambda = 10$.

comparison between our proposed methods, kindly refer to Fig. 2.

5.4. Facial landmarks detection

To rule out the possibility that our proposed method's success on the RIM task is due to the simplicity of the FIRE benchmark, we conducted experiments on facial landmark detection using the WFLW dataset [39]. For additional information regarding the dataset and task, please refer to our supplementary material. We also assessed other methods using normalized mean error for a facial landmarks detection task, as shown in Table 3. We adopted the Mean Squared Error (MSE) loss as the detector loss for both SuperRetina and the transformer-based SuperRetina. However, for the RKD-based SuperRetina, our approach involves combining the predicted output's MSE loss with the Reverse Knowledge Distillation (RKD) loss. This RKD loss, which calculates the MSE between the coordinates of the student's predicted keypoints and those of the teacher's predicted keypoints, is depicted in Equation 7. Formally, the modified MSE loss, denoted as $l^{'}_{mse}$, is computed as follows:

$$l^{'}_{mse} = l_{mse} + \lambda l^{RKD}_{mse}, \tag{7}$$

where λ represents a balancing factor that guides the influence of the RKD loss in the overall detector loss calculation.

In Figure 3, the top row illustrates the effective performance of SuperRetina in contrast to Swin UNETR SuperRetina, particularly in eyebrow keypoints. Conversely, for nose keypoints, the situation is reversed, with SuperRetina performing well. Remarkably, RKD-SR combines the favorable aspects of both models. Moving to the second row, it's noteworthy that only RKD-SR demonstrates robustness against outliers.

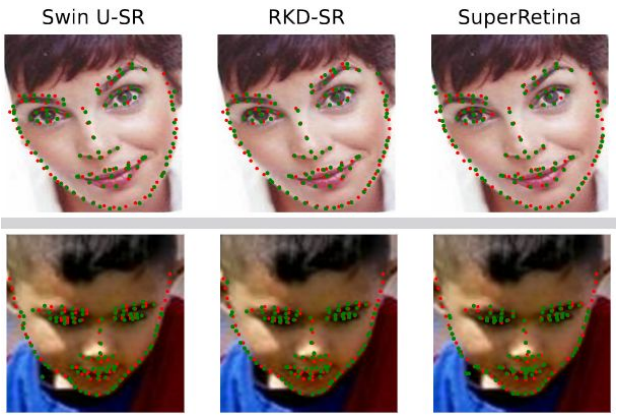

Figure 3. The visual outcomes for facial landmarks by Swin UNETR-SuperRetina (Swin U-SR) RKD-SuperRetina (RKD-SR, using $\lambda = 10$), and SuperRetina. Red and green points denote ground truth and predictions, respectively, with latter on top in case of an overlap.

6. Conclusion

In our study, we aimed to improve SuperRetina method of retinal image matching. Our targeted architectural adjustments in CNN encoders led to improvement of keypoint detection for retinal images over the previous state-of-the-art by effectively capturing keypoints.

We also addressed the challenge of training larger models, such as transformers, using limited data using reverse knowledge distillation, from a smaller CNN teacher model to a larger transformer student model. Implementing reverse knowledge distillation in our SuperRetina model led to a notable 2.5% accuracy boost over the baseline. Our findings in RKD-based keypoint detection was further confirmed through facial landmarks detection, achieving a 9.51% reduction in normalized mean error compared to the baseline SuperRetina. Moreover, we contributed to the research community by providing a public dataset with annotations for retinal image applications to foster algorithm development.

References

[1] e-ophtha dataset. https://www.adcis.net/en/third-party/e-ophtha/. Accessed: 2023-04-20. 3

[2] Kaggle diabetic retinopathy detection competition dataset. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data?select=train.zip.005. Accessed: 01/06/20. 3

[3] Kaggle glaucoma datasets. https://www.kaggle.com/datasets/arnavjain1/glaucoma-datasets. Accessed: 01/06/2023. 3

[4] Kaggle retinal disease classification dataset. https://www.kaggle.com/datasets/andrewmvd/retinal-disease-classification. Accessed: 01/06/2023. 3

[5] Retinal disease classification dataset. https://www.kaggle.com/datasets/andrewmvd/retinal-disease-classification. Accessed: 2023-04-20. 3

[6] Jimmy Addison Lee, Jun Cheng, Beng Hai Lee, Ee Ping Ong, Guozhen Xu, Damon Wing Kee Wong, Jiang Liu, Augustinus Laude, and Tock Han Lim. A low-dimensional step pattern analysis algorithm with application to multimodal retinal image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1046–1053, 2015. 1

[7] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017. 2

[8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Lecture notes in computer science*, 3951:404–417, 2006. 1

[9] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: data-efficient early knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12052–12062, 2022. 6

[10] Peter Hviid Christiansen, Mikkel Fly Kragh, Yury Brodskiy, and Henrik Karstoft. Unsuperpoint: End-to-end unsupervised interest point detector and descriptor. *arXiv preprint arXiv:1907.04011*, 2019. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[12] 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. 1, 2, 4, 7

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[14] Nihar Gupte, Sahar Almahfouz Nasser, Prateek Garg, Keshav Singhal, Tanmay Jain, Aditya, Ravi Kumar, and Amit Sethi. MeDAL-Retina. https://www.dropbox.com/sh/o8q84e2eg54ay3d/AADiAkNr6bFQDoFaKeEjpYtra?dl=0, 2023. Dataset. 3

[15] Gousia Habib, Tausifa Jan Saleem, and Brejesh Lall. Knowledge distillation in vision transformers: A critical review. *arXiv preprint arXiv:2302.02108*, 2023. 3, 5

[16] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021. 3, 4

[17] Carlos Hernandez-Matas, Xenophon Zabulis, and Antonis A Argyros. Rempe: Registration of retinal images through eye modelling and pose estimation. *IEEE journal of biomedical and health informatics*, 24(12):3362–3373, 2020. 1, 7

[18] Carlos Hernandez-Matas, Xenophon Zabulis, Areti Triantafyllou, Panagiota Anyfanti, Stella Douma, and Antonis A Argyros. Fire: fundus image registration dataset. *Modeling and Artificial Intelligence in Ophthalmology*, 1(4):16–28, 2017. 3, 7, 8

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 6

[20] Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? In *Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pages 151–160. Springer, 2022. 4

[21] Xianhui Jiang and Xiaogang Deng. Knowledge reverse distillation based confidence calibration for deep neural networks. *Neural Processing Letters*, 55(1):345–360, 2023. 3, 5

[22] Jiazhen Liu, Xirong Li, Qijie Wei, Jie Xu, and Dayong Ding. Semi-supervised keypoint detector and descriptor for retinal image matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 593–609. Springer, 2022. 1, 2, 4, 5, 7, 8

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3, 4

[24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1, 7

[25] Hannu Oinonen, Heikki Forsvik, Pekka Ruusuvuori, Olli Yli-Harja, Ville Voipio, and Heikki Huttunen. Identity verification based on vessel matching from fundus images. In *2010 IEEE International Conference on Image Processing*, pages 4089–4092. IEEE, 2010. 7

[26] Papers With Code. Fire: Framework for information retrieval evaluation. https://paperswithcode.com/dataset/fire, Accessed: 2023. 4

[27] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019. 7

[28] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Ncnet: Neighbourhood consensus networks for estimating image correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1020–1034, 2020. 7

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[30] 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. 1

[31] JB Ryu, CG Lee, and HH Park. Formula for harris corner detector. *Electronics letters*, 47(3):1, 2011. 1

[32] Nihar Gupte Sahar Almahfouz Nasser and Amit Sethi. Supplementary material of the paper reverse knowledge distillation: Training a large model using a small one for retinal image matching on limited data, 2023. Supplied as supplemental material `supplemental802.pdf`. 3

[33] Abdullah Sarhan, Jon Rokne, Reda Alhajj, and Andrew Crichton. Transfer learning through weighted loss function and group normalization for vessel segmentation from retinal images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9211–9218. IEEE, 2021. 3

[34] 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. 7

[35] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 3

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 6

[37] Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, and Sandro De Zanet. Glampoints: Greedily learned accurate match points. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10732–10741, 2019. 1, 2, 7

[38] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016. 3

[39] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 4, 8

[40] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 3