SEG2REG: STRUCTURE-PRESERVING APPEARANCE RANDOMIZATION FOR MULTIMODAL RETINAL IMAGE REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal retinal image registration (MRIR) aims to precisely align retinal vascular structures across modalities, supporting clinical diagnosis and treatment planning. However, the development of MRIR is hindered by the scarcity of large-scale annotated multimodal retinal image pairs. To address this issue, we propose a Seg2Reg learning framework that employs Retinal Structure-Preserving Appearance Randomization (RSPAR) to construct a synthetic registration dataset from retinal vessel segmentation data. RSPAR generates image pairs with randomized appearances while preserving vascular structure and exact geometric correspondences. The preserved vascular structures and exact geometric correspondences provide supervision for vessel-aware keypoint detection and feature matching, respectively, enabling registration learning without manually registered pairs. Beyond the limited training data, strong cross-modal appearance variations can cause anatomically corresponding locations to exhibit substantially different local features, leading to ambiguous feature matching. Motivated by this observation, we introduce Retinal Vascular Geometric Consistency (RVGC) to characterize whether candidate correspondences agree with the global retinal vascular geometry, and develop an RVGC Module for learned reliability-guided correspondence filtering and geometric estimation. Extensive experiments on the public CFFA benchmark and the private MICRO dataset demonstrate that Seg2Reg achieves state-of-the-art multimodal retinal image registration performance among the evaluated methods.

1 INTRODUCTION

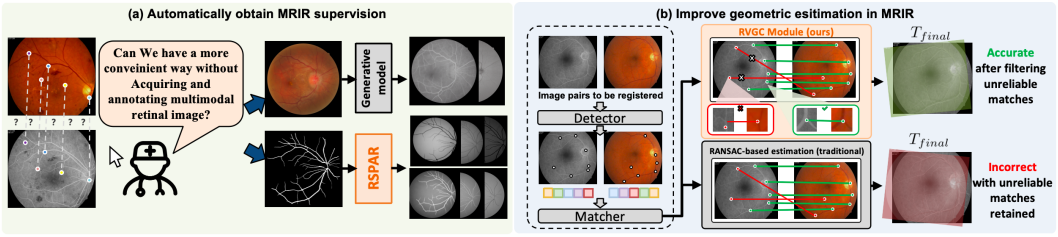

Figure 1: (a) Accurately annotated multimodal pairs in MRIR are scarce. Previous generative models synthesize modality-like appearances, whereas RSPAR randomizes appearances from a shared vessel mask and provides exact vascular correspondences without manually registered pairs. (b) During geometric estimation in MRIR, RANSAC-based estimation can retain unreliable matches and yield an incorrect transformation. The RVGC Module filters these matches and estimates an accurate T_{final}.

Retinal imaging modalities, such as color fundus (CF), fundus fluorescein angiography (FFA), indocyanine green angiography (ICGA), and fundus autofluorescence (FAF), provide complementary observations of the same retinal anatomy (Wang et al., 2024; Liu et al., 2022). Precise multimodal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

retinal image registration (MRIR) aligns retinal vascular structures across these modalities, supporting clinical diagnosis, longitudinal assessment, and cross-modal analysis (Nie et al., 2024).

With the development of MRIR, several retinal image registration datasets have been released for model training and evaluation, including CFFA, which contains 59 color fundus–fluorescein angiography image pairs (Addison Lee et al., 2015). To reduce the dependence on annotated retinal image pairs, previous studies have further explored image generation for retinal registration, as illustrated by the upper path in Fig. 1(a). Sindel et al. (2022) generated synthetic multimodal retinal pairs with preserved vessel structures for feature matching, while VesselCycleGAN (Sindel et al., 2023) translated retinal images across modalities under vessel-structure constraints. These studies show a common trend of preserving retinal vascular structures while generating modality-specific appearances. However, acquiring large-scale multimodal registration annotations remains difficult, and existing generation methods mainly aim to approximate realistic or modality-like retinal images. For MRIR, we observe that effective registration supervision primarily depends on exact vascular geometric correspondence rather than realistic image appearance.

Based on this observation, we propose a Seg2Reg learning framework centered on Retinal Structure-Preserving Appearance Randomization (RSPAR), as illustrated in Fig. 1(a). RSPAR converts retinal vessel segmentation data into synthetic registration pairs with independently randomized appearances while preserving exact vascular correspondences. The retained vascular structure and exact geometric correspondences supervise vessel-aware keypoint detection and feature matching, respectively. Using RSPAR, we construct **Seg2Reg-Syn**, comprising 73,300 fixed–moving pairs generated from 733 vessel segmentation masks. This design extends widely available vessel segmentation data to registration learning, instead of pursuing realistic modality translation.

Beyond training supervision, reliable correspondence selection and geometric estimation are also important for accurate MRIR. Local feature methods such as SuperPoint detect and describe sparse keypoints (DeTone et al., 2018), while SuperGlue and LightGlue further establish cross-image correspondences using contextual information (Sarlin et al., 2020; Lindenberger et al., 2023). Retinal-specific methods such as SuperRetina further adapt sparse feature matching to retinal vascular structures (Liu et al., 2022). As illustrated in Fig. 1(b), these methods typically estimate retinal transformations from sparse matched keypoints, followed by geometric estimation methods such as RANSAC (Fischler & Bolles, 1981). Overall, existing methods focus on estimating transformations from sparse matches, but their matching confidence does not explicitly measure whether each correspondence is geometrically consistent with the global retinal vascular structure. Repeated bifurcations, crossings, and similar vessel patterns can therefore produce high-confidence matches that still violate the global geometry required for stable transformation estimation.

We therefore propose a Retinal Vascular Geometric Consistency (RVGC) Module to replace conventional RANSAC-based geometric estimation after feature matching, as illustrated in Fig. 1(b). Specifically, RVGC employs Vascular Geometry-Aware Attention to jointly model matched keypoint features and pairwise geometric relationships for correspondence reliability assessment. It further introduces Reliability-Guided Geometric Estimation (RGE) to remove unreliable matches, select spatially distributed correspondences, and estimate the final transformation.

The main contributions are summarized as follows:

- We propose a **Seg2Reg** learning framework that extends retinal vessel segmentation data to registration learning. Its RSPAR strategy generates diverse non-realistic appearances while preserving exact vascular correspondences, providing registration supervision without manually registered multimodal image pairs.
- We propose an RVGC Module for reliable geometric estimation after feature matching. Vascular Geometry-Aware Attention models matched keypoint features and pairwise geometric relationships to assess correspondence reliability, while RGE selects reliable and spatially distributed correspondences for robust transformation estimation.
- Based on RSPAR, we construct **Seg2Reg-Syn**, a large-scale synthetic retinal registration dataset containing 73,300 fixed–moving pairs generated from 733 vessel segmentation masks across five retinal datasets. Seg2Reg-Syn provides diverse appearances with exact vascular correspondences, extending existing vessel segmentation resources to registration learning.

- Extensive experiments on the CFFA and MICRO benchmarks demonstrate the effectiveness of Seg2Reg, which achieves state-of-the-art multimodal retinal image registration performance among the evaluated methods.

2 Related Work

Synthetic Supervision for Retinal Image Registration Accurately annotated multimodal retinal image pairs remain scarce, motivating registration methods to reduce their dependence on manually registered training data. As an early synthetic-supervision paradigm, SynthMorph (Hoffmann et al., 2021) learns registration from synthetically generated anatomy with known spatial transformations. Following this idea in retinal imaging, Sindel et al. (2022) construct synthetic multimodal retinal pairs with known homographies to supervise vessel-based feature matching. Meanwhile, retinal image generation has increasingly incorporated structural constraints, as VesselCycleGAN (Sindel et al., 2023) preserves vessel locations during cross-modal appearance translation. Overall, existing studies either introduce known transformations for synthetic registration supervision or preserve vascular structures during modality translation. Seg2Reg builds on this line of work by using RSPAR to convert retinal vessel masks into appearance-diverse registration pairs while preserving exact vascular correspondences, thereby extending existing vessel segmentation data to registration learning.

Sparse Feature Matching and Geometric Estimation Sparse feature matching has become a widely used strategy for establishing correspondences between retinal image pairs. SuperPoint (DeTone et al., 2018) and R2D2 (Revaud et al., 2019) improve local feature detection and description, while SuperGlue (Sarlin et al., 2020) and LightGlue (Lindenberger et al., 2023) further exploit cross-image context for sparse correspondence estimation. With the increasing use of learned features in retinal registration, GLAMpoints (Truong et al., 2019) and SuperRetina (Liu et al., 2022) adapt keypoint learning to retinal vascular structures. After matching, RANSAC (Fischler & Bolles, 1981) remains a common choice for rejecting outliers and estimating geometric transformations. Beyond conventional robust estimation, learned correspondence pruning has been introduced to model relationships within the candidate set. PointCN (Yi et al., 2018) learns correspondence-wise inlier probabilities by combining point-wise features with global context. OANet (Zhang et al., 2019) further captures spatial relationships among correspondences through order-aware filtering. More recent methods model richer consistency among candidate matches: CLNet (Zhao et al., 2021) progressively prunes correspondences through local-to-global consistency learning, while LMC-Net (Liu et al., 2021) exploits local and global motion coherence for reliable correspondence selection. Overall, these methods extend correspondence filtering from individual inlier classification toward relational and set-level consistency modeling. Related retinal registration methods, such as Wang et al. (2021) and An et al. (2022), further combine vascular representations with learned matching and outlier rejection. Overall, existing methods progressively improve feature matching and correspondence filtering, while RVGC combines pairwise geometric relationships among matched keypoints with vessel-detector evidence at their endpoints. Its reliability estimates guide spatially distributed correspondence selection and transformation estimation.

3 Methodology

Given a fixed retinal image I_f and a moving retinal image I_m, MRIR seeks a transformation T_{final} that maps a point $\mathbf{x}_f$ in I_f to its corresponding location $\mathbf{x}_m$ in I_m, i.e., $\mathbf{x}_m = T_{\text{final}}(\mathbf{x}_f)$.

Seg2Reg learns this mapping without manually registered training pairs, as illustrated in Fig. 2. RSPAR constructs synthetic fixed–moving pairs from retinal vessel segmentation data while preserving exact vascular correspondences, thereby providing geometric supervision for registration learning. The generated pairs are then processed by SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023), which serve as the keypoint detector–descriptor and sparse matcher, respectively, and are incorporated into the synthetic training pipeline. The resulting matched keypoints are further processed by the Retinal Vascular Geometric Consistency (RVGC) Module, where Vascular Geometry-Aware Attention estimates correspondence reliability and Reliability-Guided Geometric Estimation (RGE) selects reliable spatially distributed matches to estimate T_{final}.

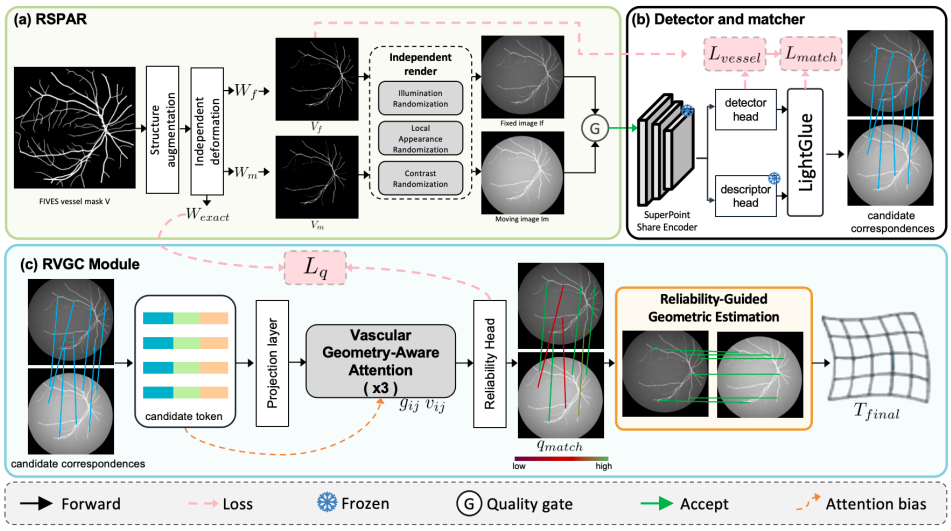

Figure 2: RSPAR-based synthetic training and registration pipeline. (a) A source vessel mask V is structurally augmented and independently deformed by W_f and W_m to generate fixed and moving vessel masks V_f and V_m. Independent rendering randomizes illumination, local appearance, and contrast while preserving the exact warp W_{exact} between the two images. (b) The synthetic pair (I_f, I_m) is processed by the SuperPoint–LightGlue pipeline, with vessel, matching, and reliability supervision $\mathcal{L}_{\text{vessel}}$, $\mathcal{L}_{\text{match}}$, and $\mathcal{L}_q$ used to train the detector, matcher, and RVGC module. (c) LightGlue candidate correspondences are represented as candidate tokens and passed to Vascular Geometry-Aware Attention for reliability prediction. Reliability-Guided Geometric Estimation then selects correspondences and estimates the final transformation.

3.1 SEG2REG FRAMEWORK

Seg2Reg converts retinal vessel segmentation data into registration training pairs with exact geometric supervision. As illustrated in Fig. 2, the framework first uses Retinal Structure-Preserving Appearance Randomization (RSPAR) to generate appearance-diverse fixed–moving image pairs from shared vessel structures. The known vascular geometry and transformations are then used to supervise keypoint detection, feature matching, and correspondence reliability learning.

Retinal Structure-Preserving Appearance Randomization RSPAR generates synthetic registration pairs from retinal vessel segmentation masks while preserving exact vascular correspondences, as shown in Fig. 2(a). A vessel mask V is sampled from the Seg2Reg-Syn source mask pool and augmented by affine, smooth, and topological perturbations to obtain a shared vessel template $\bar{V} = A(V)$, where $A(\cdot)$ denotes the structural augmentation.

The shared template is independently transformed into the fixed and moving image spaces by two deformation models, W_f and W_m:

$$V_f = W_f(\bar{V}), \qquad V_m = W_m(\bar{V}), \tag{1}$$

where V_f and V_m are the corresponding vessel masks. Each deformation contains an affine transformation and a smooth control-point displacement. Samples with invalid deformations or insufficient vessel regions are discarded.

Since V_f and V_m originate from the same vessel template, their geometric correspondence is known exactly. The fixed-to-moving warp is defined as

$$W_{\text{exact}}(\mathbf{x}) = W_m\left(W_f^{-1}(\mathbf{x})\right). \tag{2}$$

RSPAR then renders V_f and V_m independently to generate the synthetic images I_f and I_m. Illumination, contrast, and local appearance are strongly randomized to produce diverse non-realistic

appearances, while the underlying vascular structures and W_{exact} remain unchanged. Therefore, RSPAR increases appearance diversity without breaking the exact correspondence between the two images.

Finally, a quality gate checks whether the generated pair contains sufficient valid keypoint correspondences under W_{exact}. Pairs that do not satisfy this requirement are resampled. In this way, RSPAR converts retinal vessel segmentation data into registration pairs with known vascular geometry, providing exact supervision without manually registered multimodal retinal images.

Seg2Reg Supervision Loss RSPAR provides the vessel masks and exact warp for each synthetic image pair, which are used to construct supervision for vessel-aware keypoint detection, feature matching, and correspondence reliability learning. Accordingly, Seg2Reg contains three complementary loss terms.

For vessel-aware keypoint detection, a binary vessel label $v_{si} \in \{0, 1\}$ is assigned to keypoint i on side $s \in \{f, m\}$ according to the corresponding synthetic vessel mask. Let z_{si} denote the detector logit and $p_{si} = \sigma(z_{si})$ its vessel probability. The vessel supervision loss is defined as

$$\begin{aligned} \mathcal{L}_{\text{vessel}} &= \frac{1}{2}\left(\mathcal{L}^{(f)}_{\text{vessel}} + \mathcal{L}^{(m)}_{\text{vessel}}\right), \\ \mathcal{L}^{(s)}_{\text{vessel}} &= -\frac{1}{N_s}\sum_{i=1}^{N_s}\left[\omega_s v_{si}\log p_{si} + (1 - v_{si})\log(1 - p_{si})\right], \end{aligned} \tag{3}$$

where N_s is the number of detected keypoints on side s, and ω_s balances vessel and background keypoints.

For feature matching, W_{exact} provides the exact target location of each fixed-image keypoint. For a fixed keypoint $\mathbf{x}_{0i}$, its warped location $W_{\text{exact}}(\mathbf{x}_{0i})$ is assigned to the nearest moving-image keypoint when their distance is below r_{match}; otherwise, it is assigned to the dustbin. The matching loss is

$$\mathcal{L}_{\text{match}} = -\frac{1}{N_0}\sum_{i=1}^{N_0}\log P(j_i^* \mid i), \tag{4}$$

where j_i^* denotes the target assignment of keypoint $\mathbf{x}_{0i}$.

The exact warp further provides supervision for correspondence reliability. For a matched keypoint pair $c_i = (\mathbf{x}_{0i}, \mathbf{x}_{1i})$, its geometric error is defined as $e_i = \|W_{\text{exact}}(\mathbf{x}_{0i}) - \mathbf{x}_{1i}\|_2$. Candidates with $e_i \le r_{\text{pos}}$ are labeled as reliable, while those with $e_i \ge r_{\text{neg}}$ are labeled as unreliable. Candidates within the intermediate range are excluded from supervision. Given the predicted reliability logit a_i, the reliability loss is

$$\mathcal{L}_q = -\frac{1}{|\Omega|}\sum_{i\in\Omega}\left[y_i\log\sigma(a_i) + (1 - y_i)\log(1 - \sigma(a_i))\right], \tag{5}$$

where Ω denotes the set of labeled candidates.

The overall Seg2Reg supervision loss is

$$\mathcal{L} = \mathcal{L}_{\text{match}} + \lambda_{\text{vessel}}\mathcal{L}_{\text{vessel}} + \lambda_q\mathcal{L}_q. \tag{6}$$

Here, $\mathcal{L}_{\text{vessel}}$ provides vessel-aware detection supervision, $\mathcal{L}_{\text{match}}$ provides exact keypoint assignment supervision, and $\mathcal{L}_q$ provides correspondence reliability supervision. Together, these losses allow retinal vessel segmentation data to provide the structural and geometric supervision required for Seg2Reg training.

3.2 Sparse Feature Detection and Matching

Seg2Reg adopts SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023) for keypoint detection, description, and sparse feature matching. For SuperPoint, the pretrained encoder and descriptor head are kept frozen, while the detector head is adapted to retinal vessel structures using the RSPAR-generated supervision. LightGlue retains its original architecture and is trained from random initialization on the synthetic image pairs to establish matched keypoints between the fixed and moving images. The resulting matched keypoints and their associated features are then passed to the RVGC Module for correspondence reliability assessment and transformation estimation.

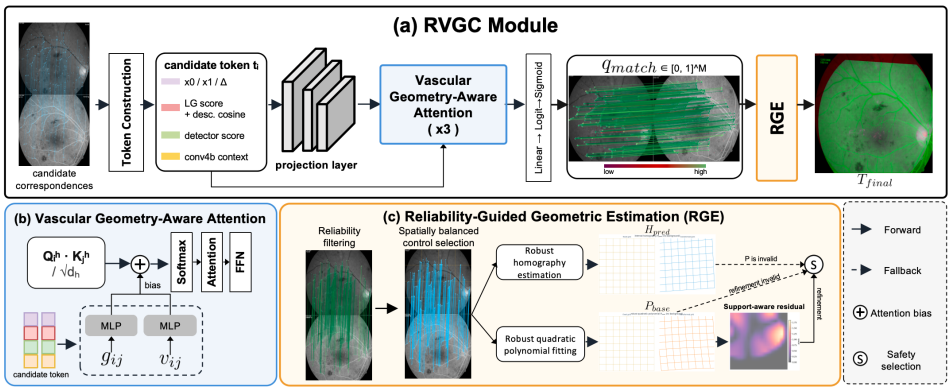

Figure 3: RVGC Module and its two stages. (a) Candidate correspondences are represented as tokens, processed by Vascular Geometry-Aware Attention, and assigned reliability scores q_{match} before entering RGE, which produces T_{final}. (b) Vascular Geometry-Aware Attention combines candidate features, pairwise geometric relationships $\mathbf{g}_{ij}$, and endpoint evidence $\mathbf{v}_{ij}$ through attention biases. (c) Reliability-Guided Geometric Estimation filters unreliable correspondences, selects spatially distributed controls, and estimates the final transformation using robust homography and support-aware polynomial refinement.

3.3 Retinal Vascular Geometric Consistency Module

The matching confidence produced by LightGlue reflects assignment quality, but it does not explicitly indicate whether a matched keypoint pair is consistent with the global retinal vascular geometry. To address this issue, we introduce the Retinal Vascular Geometric Consistency (RVGC) Module, as illustrated in Fig. 3. RVGC contains two stages. Vascular Geometry-Aware Attention estimates the geometric reliability of each matched keypoint pair, while Reliability-Guided Geometric Estimation (RGE) selects reliable and spatially distributed correspondences to estimate the final transformation.

Vascular Geometry-Aware Attention A reliable retinal correspondence should be supported not only by its local matching evidence, but also by its geometric relationship with other matched keypoints. For each candidate $c_i = (\mathbf{x}_{0i}, \mathbf{x}_{1i})$, we construct a feature token using the matched keypoint coordinates, displacement, matching score, descriptor similarity, detector confidence, and local feature context. These features preserve the local information available from the detector and matcher.

To explicitly model retinal vascular geometry, we further describe the relationship between candidates i and j as

$$\mathbf{g}_{ij} = [\Delta_{ij}^{\text{disp}}, \Delta_{ij}^{\text{angle}}, \Delta_{ij}^{\text{dist}}, \Delta_{ij}^{\text{scale}}], \tag{7}$$

where the four terms measure differences in displacement, direction, relative distance, and scale between the two matched keypoint pairs. A lightweight multilayer perceptron maps $\mathbf{g}_{ij}$ to a geometry bias. The detector confidences at the matched endpoints are also used to form an endpoint bias, providing additional vessel-related evidence.

The two biases are incorporated into multi-head self-attention:

$$A_{ij}^h = \frac{(\mathbf{Q}_i^h)^\top \mathbf{K}_j^h}{\sqrt{d_h}} + b_{\text{geom},ij}^h + b_{\text{endpoint},ij}^h. \tag{8}$$

This design allows each candidate to be evaluated jointly with the complete matched-keypoint set under retinal vascular geometry. The output feature of each candidate is finally mapped to a reliability score $q_i = \sigma(a_i)$, where a larger q_i indicates that candidate c_i is more consistent with the retinal vascular geometry. The exact correspondence supervision generated by RSPAR is used to train this reliability prediction.

Reliability-Guided Geometric Estimation High matching reliability alone does not guarantee stable transformation estimation, as reliable correspondences may still concentrate in a small retinal region. RGE therefore first performs reliability filtering and spatially balanced correspondence

selection. Candidates with $q_i \geq \tau_q$ are retained and ranked by reliability. The fixed image is then divided into a $G \times G$ grid, and correspondences are selected across nonempty cells in a round-robin manner. This produces a reliable and spatially distributed control set for transformation estimation.

Based on the selected correspondences, RGE consists of two complementary components: Robust Homography estimation (RH) and Support-Aware Polynomial refinement (SAP). RH denotes the homography-estimation component inside RGE, whereas RANSAC-H denotes the standalone homography baseline used in controlled comparisons. RH estimates a robust homography H_{pred} as a global reference and fallback transformation. The fitting process jointly considers correspondence reliability, spatial distribution, and reprojection errors to reduce the influence of unreliable or locally concentrated matches.

SAP estimates a quadratic base transformation from the selected correspondences to capture smooth retinal deformation beyond a global homography. For a normalized fixed-image coordinate $\mathbf{x} = (u, v)$, the quadratic base transformation is defined as

$$\begin{aligned} \boldsymbol{\phi}(\mathbf{x}) &= [1, u, v, u^2, uv, v^2]^\top, \\ P_{\text{base}}(\mathbf{x}) &= [\mathbf{a}^\top \boldsymbol{\phi}(\mathbf{x}), \mathbf{b}^\top \boldsymbol{\phi}(\mathbf{x})]^\top, \end{aligned} \tag{9}$$

where $\mathbf{a}$ and $\mathbf{b}$ are estimated from the selected correspondences. Correspondences with large fitting errors are iteratively removed to improve the robustness of the base transformation.

To further correct local residual errors, SAP computes a reliability-weighted residual from nearby control points. The correction is reduced in regions with insufficient reliable support and bounded to avoid unstable extrapolation. The final transformation is defined as

$$T_{\text{final}}(\mathbf{x}) = P_{\text{base}}(\mathbf{x}) + \boldsymbol{\Delta}(\mathbf{x}), \tag{10}$$

where $\boldsymbol{\Delta}(\mathbf{x})$ denotes the support-aware local correction. If SAP cannot obtain a valid quadratic transformation, RGE falls back to H_{pred}; if the local correction is unreliable, only P_{base} is retained.

4 Experiments

4.1 Experimental Setup

Datasets We perform zero-shot evaluation on two real-world multimodal retinal datasets, without using any evaluation images for training or fine-tuning. **CFFA** contains 59 color fundus–fluorescein angiography pairs (Addison Lee et al., 2015). **MICRO (private)** is an unpublished internal dataset containing 1,108 preoperative retinal image pairs from 301 eyes across multiple modalities, diseases, and acquisition systems; it is not publicly available at the time of submission. A separate paper providing a detailed description of this dataset is forthcoming. The dataset will be released after publication, subject to ethical and institutional approval.

Seg2Reg-Syn is the synthetic training dataset constructed from 733 vessel segmentation masks, including 600 FIVES masks (Jin et al., 2022), 40 DRIVE masks (Staal et al., 2004), 20 STARE masks (Hoover et al., 2000), 28 CHASE_DB1 masks (Fraz et al., 2012), and 45 HRF masks (Budai et al., 2013). RSPAR expands these masks into 73,300 fixed–moving registration pairs with randomized appearances and exact vascular correspondences.

Experimental Settings Training uses Seg2Reg-Syn without any real benchmark images or landmarks. LightGlue is randomly initialized and jointly trained with the vessel-adapted SuperPoint detector and RVGC Module for 40,000 steps using Adam with a batch size of 2. The learning rates are 10^{-4} and 5×10^{-5} for the main network and detector head, respectively. We set $\lambda_{\text{vessel}} = 0.2$, $\lambda_q = 0.5$, and the positive and negative reliability thresholds to 2.67 and 6.67 pixels, respectively. At inference, $\tau_q = 0.7$, a 4×4 grid, and at most 256 correspondences are used for geometric estimation. Seg2Reg results are reported as mean $\pm$ s.d. over three training seeds (66, 69, and 99), while competing methods are reported from a single evaluation using fixed official implementations or pretrained weights. For each sample, the three Seg2Reg mAUC contributions are averaged before two-sided paired permutation testing with 100,000 permutations against each baseline; image pairs are used as units on CFFA and eyes as cluster units on MICRO, followed by Holm correction within each dataset.

Table 1: Registration results on the multimodal CFFA and MICRO benchmarks. Seg2Reg results are reported as mean ± s.d. over three training seeds; competing methods are reported from a single evaluation. Best results on each dataset are highlighted in bold.

Method	CFFA						MICRO (private)					
	RMSE↓	mAUC↑	AUC@5↑	AUC@10↑	AUC@20↑	SR↑	RMSE↓	mAUC↑	AUC@5↑	AUC@10↑	AUC@20↑	SR↑
SIFT (Lowe, 2004)	18.546	0.016	0.000	0.000	0.010	0.034	5.955	0.096	0.034	0.061	0.087	0.123
SURF-PIIFD-RPM	6.290	0.302	0.078	0.184	0.275	0.390	8.152	0.298	0.049	0.143	0.256	0.427
LoFTR (Sun et al., 2021)	4.058	0.425	0.145	0.304	0.398	0.492	8.151	0.508	0.079	0.237	0.435	0.729
SuperGlue (Sarlin et al., 2020)	5.407	0.193	0.053	0.121	0.179	0.237	7.292	0.426	0.081	0.218	0.371	0.581
LightGlue (Lindenberger et al., 2023)	6.173	0.588	0.169	0.369	0.536	0.763	7.616	0.470	0.090	0.236	0.407	0.654
SuperRetina (Liu et al., 2022)	8.624	0.115	0.018	0.065	0.105	0.170	6.967	0.318	0.060	0.167	0.279	0.426
GLAMpoints (Truong et al., 2019)	–	–	–	–	–	–	8.653	0.283	0.038	0.119	0.235	0.412
XFeat (Potje et al., 2024)	7.509	0.632	0.066	0.296	0.558	0.881	9.017	0.416	0.044	0.171	0.349	0.628
Seg2Reg (ours)	**2.669±0.03**	**0.902±0.01**	**0.477±0.01**	**0.731±0.01**	**0.860±0.01**	**0.989±0.01**	**4.433±0.14**	**0.665±0.02**	**0.248±0.01**	**0.463±0.01**	**0.614±0.02**	**0.787±0.02**

Evaluation Metrics Following Liu et al. (2022), predictions are mapped back to the original image coordinates for evaluation. A registration is successful when the median and maximum landmark errors are below 20 and 50 pixels, respectively. We report RMSE, mAUC over 1–25 pixels, AUC@t ($t \in \{5, 10, 20\}$), and success rate (SR). RMSE is averaged over acceptable samples only, whereas SR is computed over all samples; failed and non-acceptable samples are counted as failures in SR and AUC metrics.

Compared Methods We compare with the classical SIFT (Lowe, 2004) and SURF-PIIFD-RPM methods; generic local features and matchers, including LoFTR (Sun et al., 2021), SuperGlue (Sarlin et al., 2020), LightGlue (Lindenberger et al., 2023), and XFeat (Potje et al., 2024); and retinal registration methods, including SuperRetina (Liu et al., 2022) and GLAMpoints (Truong et al., 2019). Methods with available implementations are re-evaluated using the same 1024×1024 input and original-coordinate evaluation protocol. All MICRO results are reported under the same evaluation protocol. RANSAC-H therefore serves as the common geometric-estimation baseline for controlled comparisons with the proposed RVGC Module.

4.2 Results

Table 1 reports performance on the public CFFA benchmark and private MICRO dataset. In Table 1, the integrated Seg2Reg framework achieves the best results among the evaluated methods across all reported metrics on both datasets. For baselines with complete per-sample results, paired permutation tests confirm significant mAUC improvements on both datasets after Holm correction; the reported adjusted p-values were 0.0001. These results show the performance of Seg2Reg on the evaluated datasets and imaging modalities. Additional analyses in Appendices A.2 and A.3 further evaluate the reliability and generalization of the RVGC Module, respectively, showing more accurate correspondence reliability estimation and effective transfer across different upstream matchers.

Table 2: Ablation study of RSPAR and RVGC components on CFFA and MICRO. VGA and RGE denote Vascular Geometry-Aware Attention and Reliability-Guided Geometric Estimation, respectively. Best results within each dataset are highlighted in bold.

Dataset	RSPAR	RVGC		RMSE↓	mAUC↑	AUC@5↑	AUC@10↑	AUC@20↑	SR↑
		VGA	RGE						
CFFA				6.173	0.5878	0.1689	0.3688	0.5355	0.7627
	✓			3.335	0.8719	0.3680	0.6635	0.8233	0.9831
		✓	✓	6.174	0.4461	0.1419	0.2871	0.4056	0.5763
	✓	✓		3.140	0.8786	0.3827	0.6808	0.8319	0.9831
	✓	✓	✓	**2.644**	**0.9112**	**0.4881**	**0.7420**	**0.8710**	**1.0000**
MICRO				7.616	0.4696	0.0901	0.2357	0.4074	0.6543
	✓			6.201	0.6298	0.1433	0.3586	0.5413	0.7650
		✓	✓	5.093	0.4873	0.1604	0.3194	0.4456	0.5966
	✓	✓		6.054	0.6149	0.1577	0.3683	0.5516	**0.7888**
	✓	✓	✓	**4.312**	**0.6568**	**0.2521**	**0.4642**	**0.6085**	0.7744

4.3 Ablations

In this section, we conduct a series of ablation experiments on the CFFA and MICRO datasets to investigate the effectiveness of RSPAR and RVGC in Seg2Reg. Official LightGlue (Lindenberger et al., 2023) with RANSAC-based homography estimation serves as the baseline.

Effects of RSPAR and RVGC in Seg2Reg Table 2 lists the MRIR results of the baseline, RSPAR, RVGC, and the complete Seg2Reg. RSPAR provides a clear improvement in correspondence quality, reducing RMSE from 6.173 to 3.335 pixels on CFFA and from 7.616 to 6.201 pixels on MICRO. RVGC alone shows limited gains, but when combined with RSPAR, the complete Seg2Reg further reduces RMSE to 2.644 and 4.312 pixels on CFFA and MICRO, respectively, while achieving the best overall mAUC and low-error AUC. These results demonstrate the complementary roles of RSPAR and RVGC in improving multimodal retinal registration.

Effects of RVGC Components We further analyze the components of RVGC while keeping the RSPAR-trained detector and matcher fixed. To separate the effect of correspondence reliability modeling from that of a more flexible transformation model, we first compare RANSAC-H with and without VGA-based correspondence selection under the same homography estimator, and then analyze the transformation components in RGE separately. As shown in Table 2, VGA provides moderate improvements in correspondence selection, whereas RGE delivers the main gains in robust transformation estimation on both datasets. A detailed analysis of the RH and SAP components in RGE is provided in Appendix A.4.

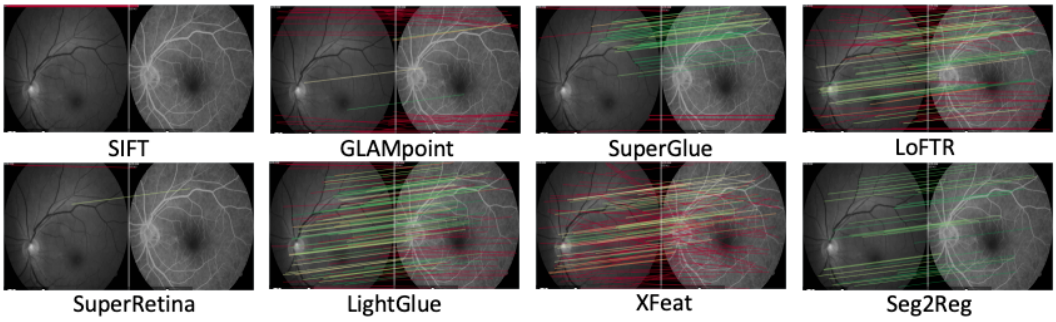

Figure 4: Qualitative comparison of feature matching between Seg2Reg and competing methods on a representative CFFA image pair.

4.4 Qualitative Visualization

Figure 4 shows that Seg2Reg achieves more reliable and vascular structure-consistent matching than the compared methods. It can be observed that Seg2Reg substantially reduces incorrect matches, while most matched keypoints are located on corresponding retinal vessel regions. In contrast, the compared methods exhibit more mismatched or spatially inconsistent correspondences, especially under large cross-modal appearance differences.

5 Conclusion

We present Seg2Reg, a multimodal retinal image registration framework that uses RSPAR to construct Seg2Reg-Syn, a synthetic training dataset of appearance-diverse retinal image pairs with exact vascular correspondences. Seg2Reg uses the preserved vascular structures and known geometric correspondences in Seg2Reg-Syn to supervise registration learning, and further employs the RVGC Module to address cross-modal matching ambiguity by modeling retinal vascular geometric consistency for reliable transformation estimation. Experiments on CFFA and MICRO demonstrate that Seg2Reg achieves state-of-the-art multimodal retinal image registration performance among the evaluated methods.

Reproducibility Statement

The appendix describes RSPAR synthesis and the evaluation protocol. An anonymized source-code ZIP will be submitted as supplementary material, with a README covering training and checkpoint-based evaluation on public CFFA data. The vessel segmentation datasets and CFFA must be obtained from their original providers; trained checkpoints and the private MICRO dataset are not included. The random-deformation implementation in the archive differs from the historical training configuration underlying the reported checkpoints, so retraining may yield different numerical results.

AI use statement

Generative AI tools were used during manuscript preparation for literature retrieval and discovery, research ideation and experimental planning, drafting portions of the manuscript, and improving the wording, structure, and readability of the paper. AI-based generative models were also used in the research workflow to generate synthetic training pairs for the GAN- and diffusion-based comparison experiments, as described in the paper. All AI-assisted outputs, AI-generated data, analyses, and drafted text were reviewed and verified by the authors. The authors take full responsibility for the accuracy, integrity, originality, and interpretation of the work, including all AI-generated content, data, analyses, and conclusions.

References

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*, pp. 1046–1053, 2015.

Cheolhong An, Yiqian Wang, Junkang Zhang, and Truong Q Nguyen. Self-supervised rigid registration for multimodal retinal images. *IEEE Transactions on Image Processing*, 31:5733–5747, 2022.

Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013(1): 154860, 2013.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 337–33712. IEEE, 2018.

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.

Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE transactions on biomedical engineering*, 59 (9):2538–2548, 2012.

Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021.

AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000.

Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data*, 9(1):475, 2022.

Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, and Stewart Lee Zuckerbrod. Attention2angiogan: Synthesizing fluorescein angiography from retinal fundus images using generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9122–9129. IEEE, 2021.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 17581–17592. IEEE, 2023.

Jiazhen Liu, Xirong Li, Qijie Wei, Jie Xu, and Dayong Ding. Semi-supervised keypoint detector and descriptor for retinal image matching. In *European Conference on Computer Vision*, pp. 593–609. Springer, 2022.

Yuan Liu, Lingjie Liu, Cheng Lin, Zhen Dong, and Wenping Wang. Learnable motion coherence for correspondence pruning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3236–3245. IEEE, 2021.

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

Qiushi Nie, Xiaoqing Zhang, Yan Hu, Mingdao Gong, and Jiang Liu. Medical image registration and its application in retinal images: a review. *Visual computing for industry, biomedicine, and art*, 7(1):21, 2024.

Rémi Pautrat, Iago Suárez, Yifan Yu, Marc Pollefeys, and Viktor Larsson. Gluestick: Robust image matching by sticking points and lines together. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9672–9682. IEEE, 2023.

Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2682–2691. IEEE, 2024.

Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019.

Kihwan Ryoo, Hyungtae Lim, and Hyun Myung. Mambaglue: Fast and robust local feature matching with mamba. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5758–5765. IEEE, 2025.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946. IEEE, 2020.

Aline Sindel, Bettina Hohberger, Andreas Maier, and Vincent Christlein. Multi-modal retinal image registration using a keypoint-based vessel structure aligning network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 108–118. Springer, 2022.

Aline Sindel, Andreas Maier, and Vincent Christlein. A vesselsegmentation-based cyclegan for unpaired multi-modal retinal image synthesis. In *BVM Workshop*, pp. 32–37. Springer, 2023.

Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021.

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*, pp. 10732–10741, 2019.

Chiao-Yi Wang, Faranguisse Kakhi Sadrieh, Yi-Ting Shen, Shih-En Chen, Sarah Kim, Victoria Chen, Achyut Raghavendra, Dongyi Wang, Osamah Saeedi, and Yang Tao. Memo: dataset and methods for robust multimodal retinal image registration with large or small vessel density differences. *Biomedical Optics Express*, 15(5):3457–3479, 2024.

Yiqian Wang, Junkang Zhang, Melina Cavichini, Dirk-Uwe G Bartsch, William R Freeman, Truong Q Nguyen, and Cheolhong An. Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework. *IEEE Transactions on Image Processing*, 30:3167–3178, 2021.

Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2666–2674. IEEE, 2018.

Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5845–5854, 2019.

Chen Zhao, Yixiao Ge, Feng Zhu, Rui Zhao, Hongsheng Li, and Mathieu Salzmann. Progressive correspondence pruning by consensus learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6444–6453. IEEE, 2021.

APPENDIX

Here we provide supplementary experimental analyses supporting the main paper:

A.1. RSPAR synthesis and training comparison.

A.2. Correspondence reliability and filtering.

A.3. Cross-matcher generalization and visualization.

A.4. RGE component analysis.

A.5. Inference efficiency.

A.1 RSPAR SYNTHESIS AND TRAINING COMPARISON

Implementation details. RSPAR uses vessel masks resized to 1024×1024. Initial affine augmentation samples rotation $[-25°, 25°]$, scale $[0.85, 1.15]$, and translations of $\pm 10\%$. Smooth displacement is applied with probability 0.8, strength $[1, 3]$ pixels, and Gaussian $\sigma = 0.03$ times the image size. Erosion and dilation each have probability 0.5, followed by 1–3 circular erasures; FOV cropping has probability 0.6.

The moving warp uses random deformations with rotation $\pm 3°$, scale $[0.97, 1.03]$, and translation $\pm 2\%$. The fixed warp uses an anchor similarity with rotation $\pm 1°$, scale $[0.995, 1.005]$, and translation $\pm 0.5\%$. Rendering samples base intensity $[35, 220]$, gamma $[0.75, 1.35]$, vessel contrast, blur, Gaussian noise, and FOV masking. The SuperPoint quality gate uses at most 1024 keypoints, detection threshold 0.0005, at least 100 keypoints per side, a 12-pixel matching radius, at least 80 valid matches, and at most 32 trials. The RVGC training gate requires at least 8 selected matches, precision ≥ 0.85, and a positive–negative gap ≥ 0.25.

Figure 5 shows representative RSPAR-generated fixed–moving image pairs. Each pair is constructed from the same underlying vessel structure, while independent random deformations and appearance rendering create diverse observations with exact geometric correspondence preserved.

Algorithm 1 RSPAR pair generation with exact geometric supervision.

Require: Vessel mask V
1: Resize V and sample affine, smooth, and topology perturbations to obtain $\bar{V}$
2: Sample independent fixed and moving warps W_f, W_m
3: $V_f \leftarrow W_f(\bar{V}); V_m \leftarrow W_m(\bar{V})$
4: Render V_f and V_m independently to obtain I_f and I_m
5: $W_{\text{exact}}(x) \leftarrow W_m(W_f^{-1}(x))$
6: Apply the keypoint/matching quality gate; reject and resample if it fails
7: **return** $(I_f, I_m, V_f, V_m, W_{\text{exact}})$

Comparison with cross-modal generation–based training. We further compare our synthetic training pipeline with LightGlue models trained on image pairs produced by cross-modal generative models. All methods in this comparison use the same SuperPoint (DeTone et al., 2018) frontend, LightGlue (Lindenberger et al., 2023) matcher, and RANSAC-H (Fischler & Bolles, 1981) geometric estimation. **LightGlue-GAN** is trained on 600 pairs formed by FIVES (Jin et al., 2022) color fundus images and their angiography counterparts synthesized by Attention2AngioGAN (Kamran et al., 2021), with random spatial deformations applied to each pair. **LightGlue-Diffusion** is trained on diffusion-generated multimodal pairs with the same deformation protocol. For reference, **LightGlue-RSPAR** denotes our Seg2Reg-Syn training pipeline, which uses Seg2Reg-Syn to train the same LightGlue matcher and RANSAC-H estimation, without the RVGC Module. All models are evaluated under the same protocol described in the Experiments section.

Table 3 reports the results on MICRO. LightGlue-RSPAR uses the same RSPAR-only configuration as the corresponding ablation in Table 2 and follows the same evaluation protocol. LightGlue-RSPAR outperforms both generation-based variants on the reported registration metrics, reducing the RMSE from 6.4203 to 6.201 pixels and improving the mAUC from 0.5483 to 0.6298. The GAN-based and Diffusion-based generation methods perform substantially worse, indicating that appearance-only cross-modal translation provides weaker geometric supervision than the structure-preserving RSPAR synthesis.

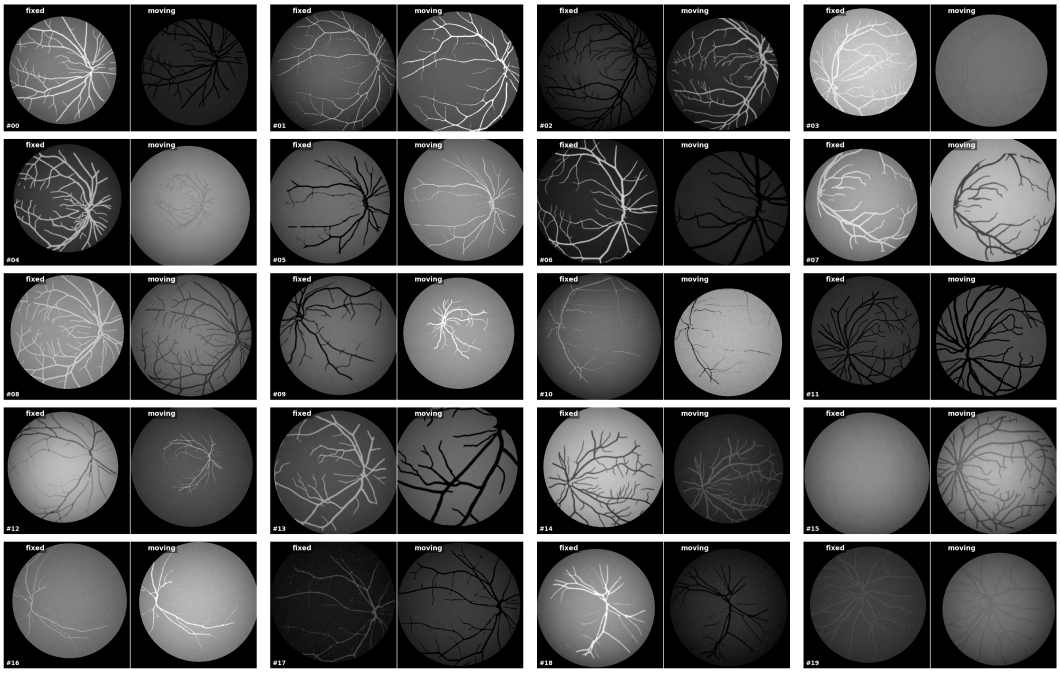

Figure 5: Examples of RSPAR-generated fixed–moving image pairs. Each pair shares the same underlying vessel structure while random deformations and independent appearance rendering produce different observations with exact geometric correspondence.

Table 3: Comparison of LightGlue trained on pairs from different cross-modal generation strategies on MICRO. All models share the same SuperPoint frontend, LightGlue matcher, and RANSAC-H estimation.

Method	RMSE↓	mAUC↑	AUC@5↑	AUC@10↑	AUC@20↑	SR↑
LightGlue-GAN	7.4521	0.3094	0.0682	0.1659	0.2688	0.4251
LightGlue-Diffusion	6.4203	0.5483	0.1324	0.3139	0.4874	0.7157
LightGlue-RSPAR (ours)	**6.201**	**0.6298**	**0.1433**	**0.3586**	**0.5413**	**0.7650**

A.2 Correspondence Reliability and Filtering

We first evaluate whether the reliability score predicted by the RVGC Module can distinguish geometrically correct and incorrect correspondences. We generate 1,000 independent RSPAR image pairs and use the exact warp to calculate the geometric error of each candidate match. Candidates with errors below 2.67 pixels are regarded as reliable, while those above 6.67 pixels are regarded as unreliable.

Table 4 compares q_{match} with the LightGlue matching score, descriptor similarity, and detector confidence. The proposed q_{match} achieves the highest ROC-AUC of 0.9777 and the lowest ECE of 0.0068, substantially outperforming the original LightGlue score. It also obtains the highest Spearman correlation with the negative geometric error. These results show that RVGC provides a more accurate estimate of correspondence reliability than the confidence signals directly produced by the detector and matcher.

Table 4: Candidate reliability on 1,000 independent RSPAR-generated synthetic pairs.

Signal	ROC-AUC↑	ECE↓	Spearman↑
LG match	0.8704	0.3783	0.3651
Desc. sim.	0.6591	0.1205	0.0866
Det. conf.	0.6369	0.0849	0.0834
RVGC Module q_{match}	**0.9777**	**0.0068**	**0.3838**

On this independent synthetic holdout, the predicted q_{match} increases ROC-AUC from 0.8704 for the LightGlue score to 0.9777 and reduces ECE from 0.3783 to 0.0068. It therefore provides a substantially better estimate of whether a candidate is geometrically correct and gives the threshold τ_q a direct reliability interpretation. Its Spearman correlation improves only slightly over the LightGlue score, indicating that q_{match} is more effective at separating correct and incorrect candidates than at ordering the precise residual errors of candidates that have already passed reliability screening.

Comparison with learned correspondence filters. We compare RVGC with generic learned correspondence filters while keeping the RSPAR-trained SuperPoint–LightGlue frontend, cached candidates, input size, and evaluation protocol fixed. PointCN (Yi et al., 2018), OANet (Zhang et al., 2019), CLNet (Zhao et al., 2021), and LMCNet (Liu et al., 2021) use their official pretrained checkpoints without tuning; their selected correspondences are passed to the same RANSAC-H estimator.

As shown in Table 5, RVGC achieves the lowest RMSE and highest mAUC and thresholded AUC on both datasets. On MICRO, raw RANSAC-H achieves a higher SR (0.8375 versus 0.7744) because RVGC removes low-reliability matches before estimation, leaving some pairs with too few or geometrically inadequate correspondences to yield a valid transformation. This conservative filtering lowers SR but improves RMSE and AUCs in some cases.

Table 5: Comparison of correspondence filtering and geometric estimation within the Seg2Reg framework. All methods use the same RSPAR-trained SPDet–LightGlue frontend and candidate correspondences. Best results on each dataset are highlighted in bold.

Dataset	Filter / estimator	RMSE↓	mAUC↑	AUC@5↑	AUC@10↑	AUC@20↑	SR↑
CFFA	Raw candidates + RANSAC-H	3.3380	0.8847	0.3597	0.6727	0.8363	**1.0000**
	PointCN + RANSAC-H	5.0235	0.8176	0.1940	0.5245	0.7555	**1.0000**
	OANet + RANSAC-H	6.1310	0.7756	0.1734	0.4816	0.7111	**1.0000**
	CLNet + RANSAC-H	4.2608	0.8502	0.2443	0.5843	0.7921	**1.0000**
	LMCNet + RANSAC-H	5.6152	0.5817	0.1158	0.3500	0.5311	0.7288
	RVGC (ours)	**2.6436**	**0.9112**	**0.4881**	**0.7420**	**0.8710**	**1.0000**
MICRO (private)	Raw candidates + RANSAC-H	6.3415	0.6434	0.1591	0.3763	0.5740	**0.8375**
	PointCN + RANSAC-H	7.7318	0.4760	0.0777	0.2320	0.4101	0.6652
	OANet + RANSAC-H	8.3318	0.4774	0.0604	0.2128	0.4052	0.6904
	CLNet + RANSAC-H	7.5951	0.5291	0.0882	0.2650	0.4579	0.7329
	LMCNet + RANSAC-H	9.2691	0.2976	0.0276	0.1188	0.2474	0.4558
	RVGC (ours)	**4.3125**	**0.6568**	**0.2521**	**0.4642**	**0.6085**	0.7744

A.3 Cross-Matcher Generalization

We further evaluate whether the RVGC Module trained with LightGlue candidates can generalize to other sparse matchers. Specifically, the trained RVGC Module is frozen and directly applied to candidates generated by SuperGlue (Sarlin et al., 2020), MambaGlue (Ryoo et al., 2025), and GlueStick (Pautrat et al., 2023), without retraining or fine-tuning. For each matcher, RANSAC-H and RVGC use the same candidate correspondences for transformation estimation.

Table 6: Cross-matcher generalization of the frozen RVGC Module on MICRO. Best results for each matcher are highlighted in bold.

Matcher	Estimator	MICRO					
		RMSE↓	mAUC↑	AUC@5↑	AUC@10↑	AUC@20↑	SR↑
SuperGlue (Sarlin et al., 2020)	RANSAC-H	6.985	0.473	0.083	0.247	0.415	**0.635**
	RVGC	**4.768**	**0.526**	**0.174**	**0.352**	**0.483**	0.633
MambaGlue (Ryoo et al., 2025)	RANSAC-H	6.815	0.589	0.110	0.316	0.520	**0.785**
	RVGC	**4.663**	**0.648**	**0.211**	**0.433**	**0.596**	0.776
GlueStick (Pautrat et al., 2023)	RANSAC-H	7.336	0.364	0.059	0.182	0.317	**0.498**
	RVGC	**5.005**	**0.396**	**0.127**	**0.260**	**0.362**	0.481

Table 6 reports the cross-matcher generalization results on MICRO. The frozen RVGC Module consistently improves RMSE, mAUC, and low-error AUC for SuperGlue, MambaGlue, and GlueStick. For example, RVGC reduces the RMSE from 6.985 to 4.768 pixels for SuperGlue and from 6.815 to 4.663 pixels for MambaGlue. These results demonstrate that RVGC can generalize to candidate correspondences from different matchers without retraining or fine-tuning.

Visualization of RVGC across matchers. Fig. 6 visualizes the effect of RVGC on different upstream matchers. The upper row shows the raw candidate correspondences, while the lower row shows the correspondences retained by RVGC. For Seg2Reg, SuperGlue (Sarlin et al., 2020), and MambaGlue (Ryoo et al., 2025), RVGC removes geometrically inconsistent candidates while preserving reliable correspondences distributed across the retinal vascular structure. This similar filtering behavior across different matchers further illustrates the cross-matcher generalization of RVGC.

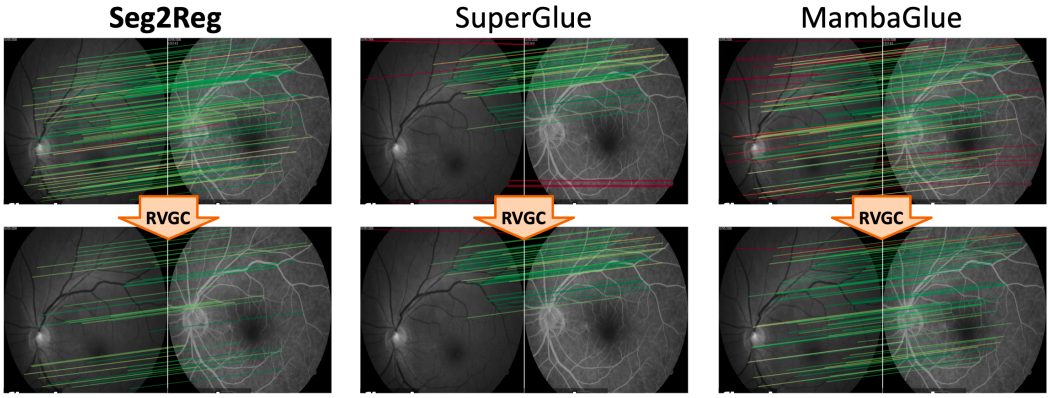

Figure 6: RVGC filtering across different upstream matchers. The upper row shows raw candidate correspondences, and the lower row shows correspondences retained by RVGC.

A.4 RGE Component Analysis

To further investigate the effectiveness of RGE, we evaluate the contributions of Robust Homography estimation (RH) and Support-Aware Polynomial refinement (SAP). RH estimates the global transformation from the reliability-guided correspondences, while SAP further models smooth retinal deformation using quadratic fitting and support-aware local correction. Both settings use the same correspondence reliability and spatial selection results.

Table 7: Ablation study of the RGE components on CFFA and MICRO. RH and SAP denote Robust Homography estimation and Support-Aware Polynomial refinement, respectively. Best results within each dataset are highlighted in bold.

Dataset	RH	SAP	RMSE↓	mAUC↑	AUC@5↑	AUC@10↑	AUC@20↑	SR↑
CFFA	✓		2.777	0.9078	0.4622	0.7287	0.8643	1.0000
	✓	✓	**2.644**	**0.9112**	**0.4881**	**0.7420**	**0.8710**	1.0000
MICRO	✓		4.901	0.6330	0.1976	0.4181	0.5799	0.7671
	✓	✓	**4.312**	**0.6568**	**0.2521**	**0.4642**	**0.6085**	**0.7744**

As shown in Table 7, SAP consistently improves the performance of RH on both datasets. On CFFA, it reduces the RMSE from 2.777 to 2.644 pixels and improves AUC@5 from 0.4622 to 0.4881. On MICRO, the RMSE decreases from 4.901 to 4.312 pixels, while AUC@5 increases from 0.1976 to 0.2521. These results show that support-aware polynomial refinement further improves transformation estimation beyond robust homography.

A.5 INFERENCE EFFICIENCY

Table 8 reports end-to-end inference on an RTX 2080 Ti using 1024×1024 inputs, batch size 1, 50 warm-up samples, and three timing rounds. Both methods share the same SPDet–LightGlue frontend checkpoint. Timing includes device transfer, feature extraction, matching, and geometric estimation, while data loading and I/O are excluded.

Compared with LightGlue + RANSAC-H, Seg2Reg adds 61.19–64.57 ms median latency and reduces throughput from 8.73/8.60 to 5.55/5.32 FPS on CFFA/MICRO. The overhead arises from RVGC reliability inference and support-aware polynomial estimation. Peak memory reaches 1,063.7 MiB on CFFA and 6,740 MiB on MICRO, making memory the main constraint on MICRO. Since $K \in \{64, 128, 256\}$ yields similar latency, we retain $K = 256$. These results quantify the computational cost of the accuracy gains provided by RVGC.

Table 8: End-to-end inference efficiency on a single RTX 2080 Ti. Latency is reported in milliseconds, FPS in frames per second, and peak process GPU memory in MiB.

Dataset	Method	End-to-end median	FPS	Peak memory
CFFA	LightGlue + RANSAC-H	119.32	8.73	591.8
CFFA	Seg2Reg ($K = 256$)	180.51	5.55	1063.7
MICRO	LightGlue + RANSAC-H	124.20	8.60	1036
MICRO	Seg2Reg ($K = 256$)	188.77	5.32	6740