Auto-Regressive Transformation for Image Alignment

Kanggeon Lee[1], Soochahn Lee[2]*, Kyoung Mu Lee[1]*
[1]ASRI, Dept. of ECE, Seoul National University, Korea
[2]Dept. of Electronics Engineering, Kookmin University, Korea
dlrkdrjs97@snu.ac.kr, sclee@kookmin.ac.kr, kyoungmu@snu.ac.kr

Abstract

*Existing methods for image alignment struggle in cases involving feature-sparse regions, extreme scale and field-of-view differences, and large deformations, often resulting in suboptimal accuracy. Robustness to these challenges can be improved through iterative refinement of the transform field while focusing on critical regions in multi-scale image representations. We thus propose Auto-Regressive Transformation (*ART*), a novel method that iteratively estimates the coarse-to-fine transformations through an auto-regressive pipeline. Leveraging hierarchical multi-scale features, our network refines the transform field parameters using randomly sampled points at each scale. By incorporating guidance from the cross-attention layer, the model focuses on critical regions, ensuring accurate alignment even in challenging, feature-limited conditions. Extensive experiments demonstrate that* ART *significantly outperforms state-of-the-art methods on planar images and achieves comparable performance on 3D scene images, establishing it as a powerful and versatile solution for precise image alignment.*

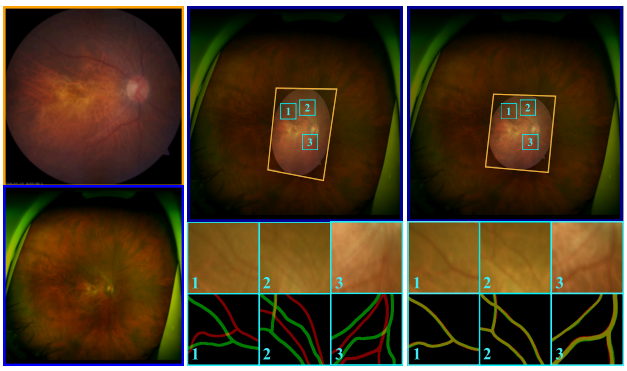


Figure 1. **Alignment Results in Challenging Scenarios.** For image pairs with sparse features, scale differences, deformations, degradations, and domain shifts, our method performs coarse-to-fine auto-regressive transformation refinement, achieving accurate alignment even in challenging scenarios where state-of-the-art methods struggle. The zoomed-in boxes show the local alignment results, and the highlighted vessel image below illustrates the intersection (yellow) between the two images (red and green).

1. Introduction

Image alignment is a fundamental problem in computer vision that involves registering images captured from different perspectives, times, or modalities. The process is essential for achieving seamless integration and analysis of images. However, scale variations, structural deformations, and indistinct features complicate accurate alignment, requiring robust and adaptive methods.

Existing methods for image alignment often fail when (1) feature-based methods [24, 34, 35, 52, 59] struggle to detect keypoints due to homogeneous textures, low contrast, or weak features; (2) intensity-based methods [2, 11, 21, 28, 29, 42] cannot handle large scale differences or deformations beyond their effective range; or (3) iterative refinement-based methods [5, 12, 14, 30, 39, 43, 64–66] suffer from poor initialization, leading to slow convergence or suboptimal alignment results.

To reduce reliance on local feature matching, which primarily depends on tentative one-to-one correspondences, matching can be performed over larger appearance regions. This can be addressed by jointly estimating correspondences for sets of points. To handle large scale differences, it is crucial to search across a wide range of scales. This can be achieved by learning to infer transform field parameters within a coarse-to-fine framework, enabling the network to iteratively refine its estimates. To improve robustness against poor initialization in iterative pipelines, non-parametric conditions should guide the refinement process. This can be done by incorporating global appearance cues from the input image pair into the sampling process.

Auto-Regressive Transformation (ART) is a novel image alignment framework robust to image pairs with large scale and field-of-view differences, deformations, and limited distinctive features, as shown in Fig. 1. ART employs

*Corresponding authors

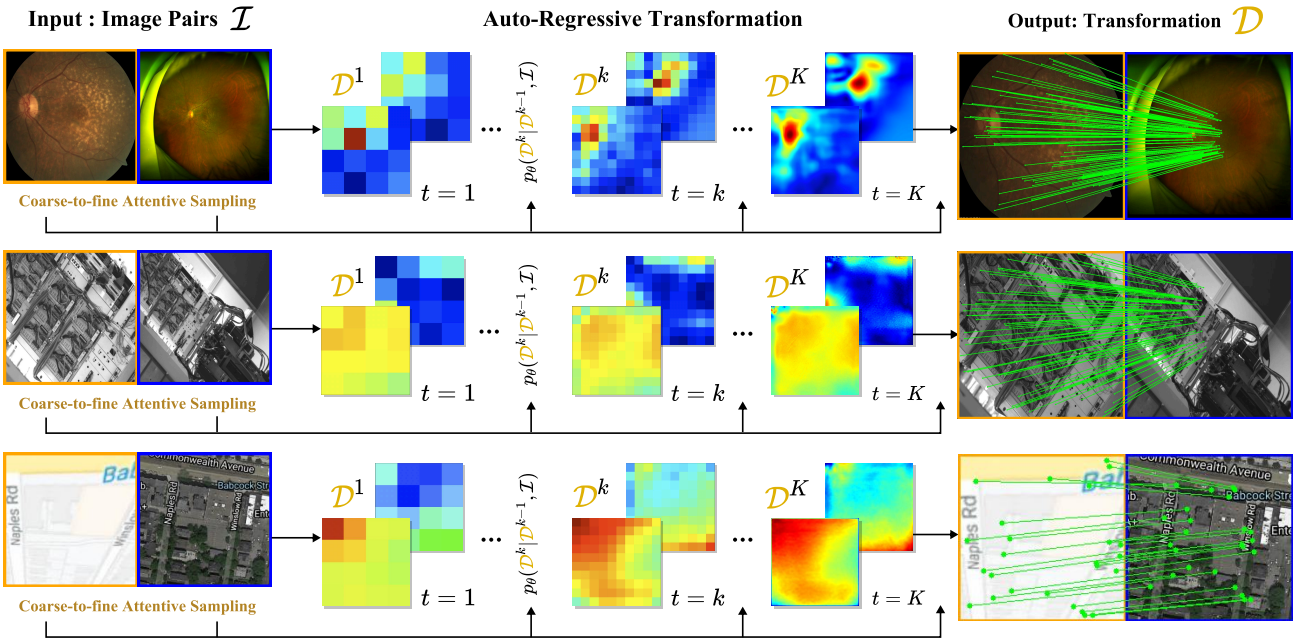


Figure 2. **Method Overview.** Auto-Regressive Transformation (ART) iteratively refines the transformation $\mathcal{D}$ for image pairs $\mathcal{I}$ in a coarse-to-fine manner. Its sampling strategy enables effective operation across diverse domains and datasets.

an auto-regressive approach, iteratively sampling and refining local transform parameters by joint estimation for a set of points in a coarse-to-fine manner guided by multi-scale representations from a pyramid feature extraction network, as depicted in Fig. 2. Moreover, by leveraging global appearance cues from the entire image pair as conditioning signals, ART achieves robustness to initialization. Extensive evaluations demonstrate that ART significantly outperforms existing feature-based [10, 16, 24, 34, 35, 47, 49, 50, 52, 55], intensity-based [7, 14], and iterative refinement-based methods [7, 35, 50, 65, 66] across various datasets.

Our contributions are as follows:

- Coarse-to-fine auto-regressive modeling enables ART to handle substantial transformations between images.
- ART demonstrates state-of-the-art performance for a wide range of datasets with limited features, scale difference, large deformation, and considerable domain shift.
- ART can adapt to different complexity requirements by controlling the number of inference iterations.

2. Related Works

Feature-based methods align images by detecting and matching keypoints to estimate transformations. Traditional approaches [4, 6, 37, 48] are widely used for their robustness, while deep learning-based methods [16, 35, 46, 55] improve keypoint detection and description.

Further advancements enhance alignment performance. SuperGlue [49] introduces graph neural networks for robust correspondence, while LightGlue [33] improves efficiency with a lightweight design. However, these methods struggle in feature-sparse regions and under extreme distortions.

Detector-free methods estimate transformations directly from image pairs. Deep homography estimation [15] pioneered this approach, followed by NCNet [47], which optimizes efficiency with sparse convolutions. Optical flow-based models [26, 60] estimate dense correspondences. Transformer-based methods such as LoFTR [52], GeoFormer [34], and RoMa [18], as well as diffusion-based approaches like RetinaRegNet [50] built on DIFT [53], further enhance spatial reasoning. However, these models often demand substantial computational resources and large-scale datasets for effective generalization.

Intensity-based methods align images by optimizing a transformation that minimizes pixel intensity differences using similarity metrics [56, 67]. Traditional methods refine transform field parameters iteratively [40, 54].

Deep learning improves these approaches by directly predicting transformations, as seen in Deep Image Homography Estimation [15] and Spatial Transformer Networks [27], while ISTN [30] and REMPE [24] enhance flexibility and robustness. These methods are widely applied in optical flow estimation [26, 60, 62] and medical image registration [2, 9, 25, 28, 41, 61]. However, they struggle with brightness variations, contrast differences, and modality changes, and can be computationally expensive for high-resolution images or complex transformations.

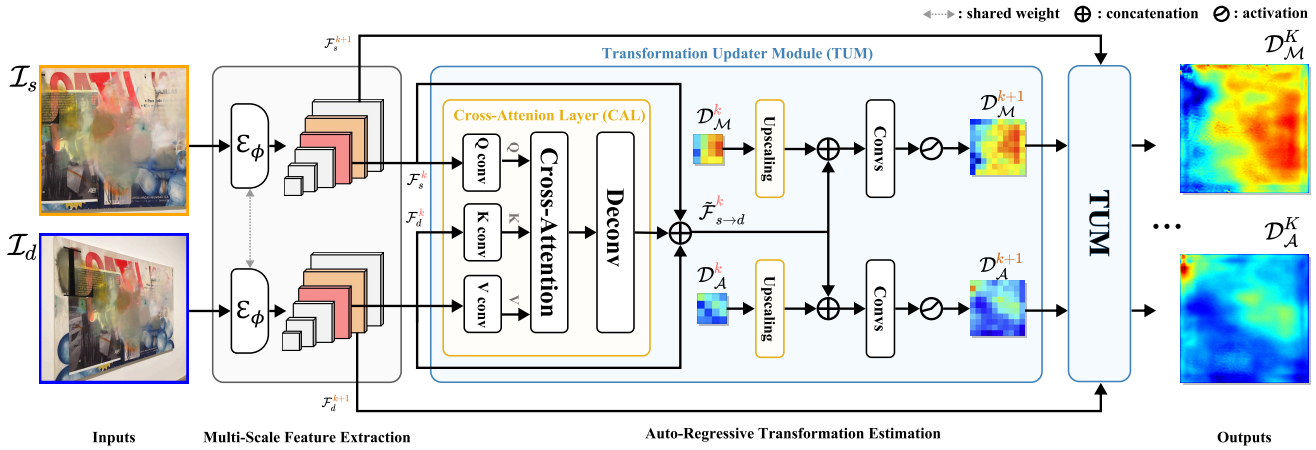


Figure 3. **Overall Framework.** ART first extracts multi-scale features $\mathcal{F}_s$ and $\mathcal{F}_d$ from the input image pair $\mathcal{I}_s$ and $\mathcal{I}_d$. At each sampling step k, the corresponding features, $\mathcal{F}_s^k$ and $\mathcal{F}_d^k$, are passed through the Cross-Attention Layer (CAL) to identify the correlated features that guide the network's focus on regions requiring refinement. The attentive feature map $\tilde{\mathcal{F}}_{s\to d}^k$ is then used to refine the transform field parameters $\mathcal{D}_\mathcal{M}^k$ and $\mathcal{D}_\mathcal{A}^k$ to $\mathcal{D}_\mathcal{M}^{k+1}$ and $\mathcal{D}_\mathcal{A}^{k+1}$ through multiple convolutional neural networks. This auto-regressive process continues until the initialized transform field parameters $\mathcal{D}_\mathcal{M}^0$ and $\mathcal{D}_\mathcal{A}^0$ reach the full resolution of the input image pair $\mathcal{I}_s$ and $\mathcal{I}_d$.

Iterative refinement-based methods progressively adjust transformations to improve alignment, inspired by traditional frameworks like RANSAC [20] and ICP [5]. Early methods, such as Lucas-Kanade [38], employed gradient-based optimization but struggle with large deformations. Deep learning models [7, 65, 66] refine alignment by employing multi-stage or recurrent processes. Diffusion-based approaches [57, 63] further improve accuracy.

3. Proposed Method

3.1. Problem Formulation

Given a source image $\mathcal{I}_s$ and a destination image $\mathcal{I}_d$, both with spatial resolution (H, W), we denote their respective set of point coordinates as $\mathcal{P}_s = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and $\mathcal{P}_d = \{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}$, respectively.

We can then define the locally linear point-wise transformations between $\mathcal{P}_s$ and $\mathcal{P}_d$ as:

$$\mathcal{P}_d = \mathcal{D}_\mathcal{M} \cdot \mathcal{P}_s + \mathcal{D}_\mathcal{A}, \tag{1}$$

where $\mathcal{D}_\mathcal{M}$ and $\mathcal{D}_\mathcal{A}$ represent the multiplicative and additive transform field parameters for point-wise scaling and translation, and the operations $\cdot$ and $+$ denote element-wise multiplication and addition, respectively. Both $\mathcal{D}_\mathcal{M}$ and $\mathcal{D}_\mathcal{A}$ have shape $(H, W, 2)$, where the last dimension corresponds to the x and y-axis components of the transformation. That is, a point (x_i, y_j) in $\mathcal{P}_s$ is mapped to its corresponding point (x'_i, y'_j) in $\mathcal{P}_d$ as follows:

$$\begin{aligned} x'_i &= \mathcal{D}_\mathcal{M}[x_i, y_j, 0] \times x_i + \mathcal{D}_\mathcal{A}[x_i, y_j, 0], \\ y'_j &= \mathcal{D}_\mathcal{M}[x_i, y_j, 1] \times y_j + \mathcal{D}_\mathcal{A}[x_i, y_j, 1]. \end{aligned} \tag{2}$$

While each point transform is individually simple, the field as a whole can represent complex and flexible free-form deformations.

3.2. Auto-Regressive Transformation

To accurately estimate the transform field parameters $\mathcal{D}_\mathcal{M}$ and $\mathcal{D}_\mathcal{A}$ between $\mathcal{I}_s$ and $\mathcal{I}_d$, ART employs an auto-regressive coarse-to-fine refinement strategy, where transform field parameters are progressively updated through multiple steps, as was depicted in Fig. 2.

In the most coarse level, the $(H_0, W_0, 2)$ shaped transform field parameters $\mathcal{D}_\mathcal{M}^0$ and $\mathcal{D}_\mathcal{A}^0$ are initialized to $\mathbf{1}_{H_0 \times W_0 \times 2}$ and $\mathbf{0}_{H_0 \times W_0 \times 2}$, respectively. Every iteration doubles the spatial resolution, so after k steps, $\mathcal{D}_\mathcal{M}^k$ and $\mathcal{D}_\mathcal{A}^k$ reach a spatial size 2^k times larger than $\mathcal{D}_\mathcal{M}^0$ and $\mathcal{D}_\mathcal{A}^0$, enabling the estimation of finer details. This iterative refinement process enables the model to incrementally improve precise estimation at each sampling step k until reaching the final step K, as follows:

$$(\mathcal{D}_\mathcal{M}^{k+1}, \mathcal{D}_\mathcal{A}^{k+1}) = \text{ART}(\mathcal{D}_\mathcal{M}^k, \mathcal{D}_\mathcal{A}^k | \mathcal{I}_s, \mathcal{I}_d). \tag{3}$$

3.3. Architectural Details

ART consists of two main components: (1) A multi-scale feature extractor that captures details from coarse to fine levels; (2) A transformation updater module that autoregressively refines transform field parameters. The entire network structure is depicted in Fig. 3. Further details will be discussed in the following sections.

Multi-Scale Feature Extractor The network $\mathcal{E}_\phi$ as in [1, 66] extracts multi-scale features with progressively in-

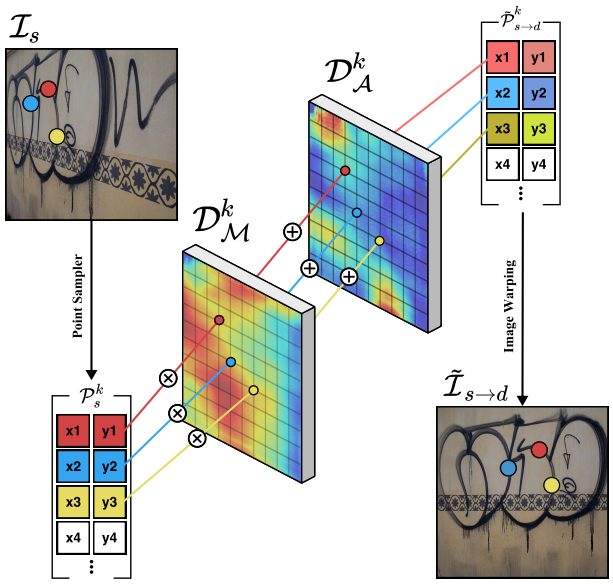


Figure 4. **Point-based Image Warping.** At sampling step k, the extracted source points set $\mathcal{P}^k_s$ is warped to $\tilde{\mathcal{P}}^k_{s\to d}$ by sequentially multiplying with the corresponding values of the transform field parameter $\mathcal{D}^k_{\mathcal{M}}$ and adding $\mathcal{D}^k_{\mathcal{A}}$ for each point. These point pairs are then used to compute the warped image $\tilde{\mathcal{I}}_{s\to d}$.

creasing spatial resolutions from $\mathcal{I}_s$ and $\mathcal{I}_d$, as follows:

$$\begin{aligned}(\mathcal{F}^0_s, ..., \mathcal{F}^k_s, ..., \mathcal{F}^K_s) &= \mathcal{E}_\phi(\mathcal{I}_s),\\ (\mathcal{F}^0_d, ..., \mathcal{F}^k_d, ..., \mathcal{F}^K_d) &= \mathcal{E}_\phi(\mathcal{I}_d),\end{aligned} \tag{4}$$

where k is the current and K is the maximum transform field parameters sampling step. Each output's spatial resolution doubles from the previous one, while the number of channels remains fixed. At each scale, feature maps from the previous resolution are progressively integrated, ensuring effective multi-scale feature fusion. This process enables the network to capture both fine-grained details and broader contextual information across resolutions.

Transformation Updater Module At sampling step k, the Cross-Attention Layer (CAL) first extracts the corresponding attentive feature map $\tilde{\mathcal{F}}^k_{s\to d}$ by using the source feature map $\mathcal{F}^k_s$ as the query and the destination feature map $\mathcal{F}^k_d$ as the key and value as follows:

$$\tilde{\mathcal{F}}^k_{s\to d} = \texttt{Concat}[\mathcal{F}^k_s, \mathcal{F}^k_d, \text{CAL}(\mathcal{F}^k_s, \mathcal{F}^k_d)], \tag{5}$$

where `Concat` represents concatenation of tensors. To ensure computational efficiency with respect to the spatial size of the feature map, CAL applies a downsampling convolution to the query, key, and value features, followed by an upsampling deconvolution at the final stage.

The extracted attentive feature map $\tilde{\mathcal{F}}^k_{s\to d}$ guides the subsequent network in determining where to focus, enabling the iterative update of $\mathcal{D}^{k+1}_{\mathcal{M}}$ and $\mathcal{D}^{k+1}_{\mathcal{A}}$ based on $\mathcal{D}^k_{\mathcal{M}}$ and $\mathcal{D}^k_{\mathcal{A}}$, which were derived from the previous step $k-1$. This process can be expressed as follows:

$$\begin{aligned}\mathcal{D}^{k+1}_{\mathcal{M}} &= \texttt{Conv}(\texttt{Concat}[^{\times 2}\mathcal{D}^k_{\mathcal{M}}, \tilde{\mathcal{F}}^k_{s\to d}]),\\ \mathcal{D}^{k+1}_{\mathcal{A}} &= \texttt{Conv}(\texttt{Concat}[^{\times 2}\mathcal{D}^k_{\mathcal{A}}, \tilde{\mathcal{F}}^k_{s\to d}]),\end{aligned} \tag{6}$$

Here, `Conv` denotes multiple convolutional layers with Leaky ReLU, `Concat` denotes tensor concatenation, and $^{\times 2}$ indicates bilinear upscaling by a factor of 2.

This iterative process is repeated in a coarse-to-fine manner to obtain the final transform field parameters $\mathcal{D}^K_{\mathcal{M}}$ and $\mathcal{D}^K_{\mathcal{A}}$, having the same spatial resolution as the input image.

3.4. Image Warping

At any k-th sampling step during autoregressive estimation, a set of source points $\mathcal{P}^k_s$ selected by a point sampler can be warped as $\tilde{\mathcal{P}}^k_{s\to d} = \mathcal{D}^k_{\mathcal{M}} \cdot \mathcal{P}^k_s + \mathcal{D}^k_{\mathcal{A}}$, as depicted in Fig. 4.

We can utilize these sets of source points $\mathcal{P}^k_s$ and the corresponding points $\tilde{\mathcal{P}}^k_{s\to d}$ to model the transform function from the source image $\mathcal{I}_s$ to the destination image $\mathcal{I}_d$ to get warped image $\tilde{\mathcal{I}}_{s\to d}$. This can be represented either as a linear warp [22] for global changes or a quadratic warp [29] for both global and local deformations.

3.5. Training ART

The end-to-end training loss $\mathcal{L}$ of ART is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{P}} + \lambda_{\text{R}}\mathcal{L}_{\text{R}}, \tag{7}$$

where $\mathcal{L}_{\text{P}}$ and $\mathcal{L}_{\text{R}}$ are the pixel matching loss, and regularization loss, respectively. λ_{R} controls the relative importance of the regularization loss.

Stochastic Pixel Matching Loss $\mathcal{L}_{\text{P}}$ computes the difference of warped source points set $\tilde{\mathcal{P}}^k_{s\to d}$ with ground-truth destination points set $\mathcal{P}^k_d$ for all $0 < k \leq K$ as follows:

$$\mathcal{L}_{\text{P}} = \mathbb{E}_k \left\| \tilde{\mathcal{P}}^k_{s\to d} - \mathcal{P}^k_d \right\|^2_2. \tag{8}$$

Note that we can utilize the point sampler used for image warping to stochastically select the source points set $\mathcal{P}^k_s$. This stochastic sampling, instead of using regular grid points or conventional keypoint detection techniques [37], plays a key role in enabling the network to learn to robustly estimate the transforms at any particular scale.

Regularization Loss To complement the pixel matching loss, we define a regularization term to ensure that all warped points converge to their appropriate positions using homography matrix $\mathcal{H}$, shaping the distribution of estimated correspondence points rather than directly predicting ground-truth coordinates, for all $0 < k \leq K$ as follows:

$$\mathcal{L}_{\text{R}} = \mathbb{E}_k \left\| \mathcal{H}^k - \mathcal{H}^k_{GT} \right\|^1_1, \tag{9}$$

where $\mathcal{H}^k$ is computed from differentiable RANSAC [19] with inlier threshold 2 and $\mathcal{H}^k_{GT}$ is the ground-truth with sampling step k.

Table 1. **Datasets for Evaluation.**

Category	Type	Dataset	Image Content	Training Type
Retinal	HR	KBSMC	Image pairs of SFI and UWFI	FS
		FIRE [23]	Image pairs of SFI and SFI	SS
		FLORI21 [17]	Image pairs of UWFI and UWFI	SS
Scene	HR	HPatches [3]	Planar images under varying illumination and viewpoint	SS
		MegaDepth-1500 [31]	Outdoor scenery images under different lighting and perspective conditions	SS
		ScanNet-1500 [13]	Indoor scenery images with real-world viewpoint and lighting variations	SS
	LR	GoogleEarth [65]	Satellite images of the earth's surface	FS
		GoogleMap [65]	Navigation map with satellite images	FS
		MSCOCO [32]	Common images in natural context	FS

FS and SS denote fully-supervised and self-supervised, respectively.
HR and LR denote high-resolution and low-resolution, respectively.

4. Experiments

4.1. Datasets

Evaluation of ART is performed across retinal and scene categories, as described in Tab. 1.

For retinal images, we evaluate ART on three datasets, comprising standard fundus images (SFI) and ultra-wide fundus images (UWFI). For cross-domain alignment, we use a private dataset from the Kangbuk Samsung Medical Center (KBSMC) Ophthalmology Department, collected between 2017 and 2019[1], consisting of 3,744 SFI-UWFI pairs with scale differences of approximately $\times 1 \sim \times 4$, where ground truth transformation was manually annotated. Additionally, we utilize the public datasets FIRE [23] and FLORI21 [17] for in-domain alignment of SFI-SFI and UWFI-UWFI pairs.

For scene categories, we evaluate ART on HPatches [3] (planar images), MegaDepth-1500 [31] (outdoor images), ScanNet-1500 [13] (indoor images), GoogleEarth [65] (satellite images), GoogleMap [65] (navigation maps), and MSCOCO [32] (common images).

4.2. Implementation Details

Here, we define high-resolution (HR) and low-resolution (LR) images as 768×768 and 192×192, respectively, based on their spatial dimensions. KBSMC, FIRE [23], FLORI21 [17], HPatches [3], MegaDepth-1500 [31], and ScanNet-1500 [13] are assigned as HR type, while GoogleEarth [65], GoogleMap [65], and MSCOCO [32] are LR, respectively. The original images may be resized to meet these definitions. The number of inference steps K is set to 6 for HR and 4 for LR images, respectively. The point sampler selected 100 points, randomly for training and via a feature detector [37] for consistent testing.

Common Setup We used the AdamW [36] optimizer with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ to train ART, applying weight decay every 100K iterations with a decay rate of 0.01. The model was trained for more than 1K epochs using an NVIDIA A100 GPU. We set λ_R to 0.5. For the fully-supervised (FS) training strategy in Tab. 1, we apply data augmentation by introducing random rotations and random photometric distortions, including variations in illumination, contrast, blur, and noise. For the self-supervised (SS) training strategy in Tab. 1, we additionally apply random transformations to the training image, along with the aforementioned augmentation. We normalize the points in $\mathcal{P}_s$ to have values in the range $[-1, 1]$. We set the spatial dimensions of the initialized transform field parameters $\mathcal{D}^0_{\mathcal{M}}$ and $\mathcal{D}^0_{\mathcal{A}}$ to $H_0 = 12$ and $W_0 = 12$.

[1]This study adhered to the tenets of the Declaration of Helsinki and was approved by the Institutional Review Boards (IRB) of Kangbuk Samsung Hospital (No. KBSMC 2019-08-031). The study is a retrospective review of medical records, and the data were fully anonymized prior to processing. The IRB waived the requirement for informed consent.

4.3. Evaluation on Retinal Categories

Baselines for Comparison We compare ART with SuperPoint [16], GLAMpoints [55], ISTN [30], NCNet [47], SuperGlue [49], REMPE [24], DLKFM [65], LoFTR [52], IHN [7], SuperRetina [35], ASPanFormer [10], MCNet [66], GeoFormer [34], and RetinaRegNet [50].

Evaluation Metrics To evaluate alignment performance, we use the CEM approach [51] to calculate the median error (MEE) and maximum error (MAE). The results are categorized as follows: i) *Acceptable* (MAE < 50 and MEE < 20), ii) *Inaccurate* (others). We also calculated the Area Under Curve (AUC) score [23], with mean AUC (mAUC).

Discussion We randomly split the KBSMC dataset into 3,370 training and 374 test pairs and trained the model in a FS manner. For the FIRE [23] and FLORI21 [17] datasets, ART was trained in a SS manner using SFIs from KBSMC and FIRE, as well as UWFIs from KBSMC and FLORI21, with warped pairs synthesized via random transformations.

The results in Tab. 2 clearly demonstrate the superiority of the proposed ART method across multiple retinal datasets. Compared to state-of-the-art methods, ART consistently achieves the highest *Acceptable* rate and mAUC, confirming its effectiveness in retinal image alignment.

On the challenging KBSMC dataset, ART achieves an *Acceptable* rate of 64.71% and an mAUC of 40.1, outperforming GeoFormer [34]. These results highlight the capability of ART to handle complex retinal image transformations. On the FIRE and FLORI21 datasets, ART achieves near-perfect *Acceptable* rates of 99.25% and 100%, and state-of-the-art mAUCs of 78.5 and 92.5, respectively.

Although methods such as DLKFM [65] and MCNet [66] adopt iterative point refinement, their reliance on only four points for homography estimation limits their performance, especially on complex, high-resolution pairs like SFI-UWFI. In contrast, ART uses a more expressive transformation model, achieving accurate and reliable alignment in challenging scenarios.

Fig. 5 presents qualitative results, further demonstrating the effectiveness of ART compared to state-of-the-art methods such as GeoFormer [34] and RetinaRegNet [50].

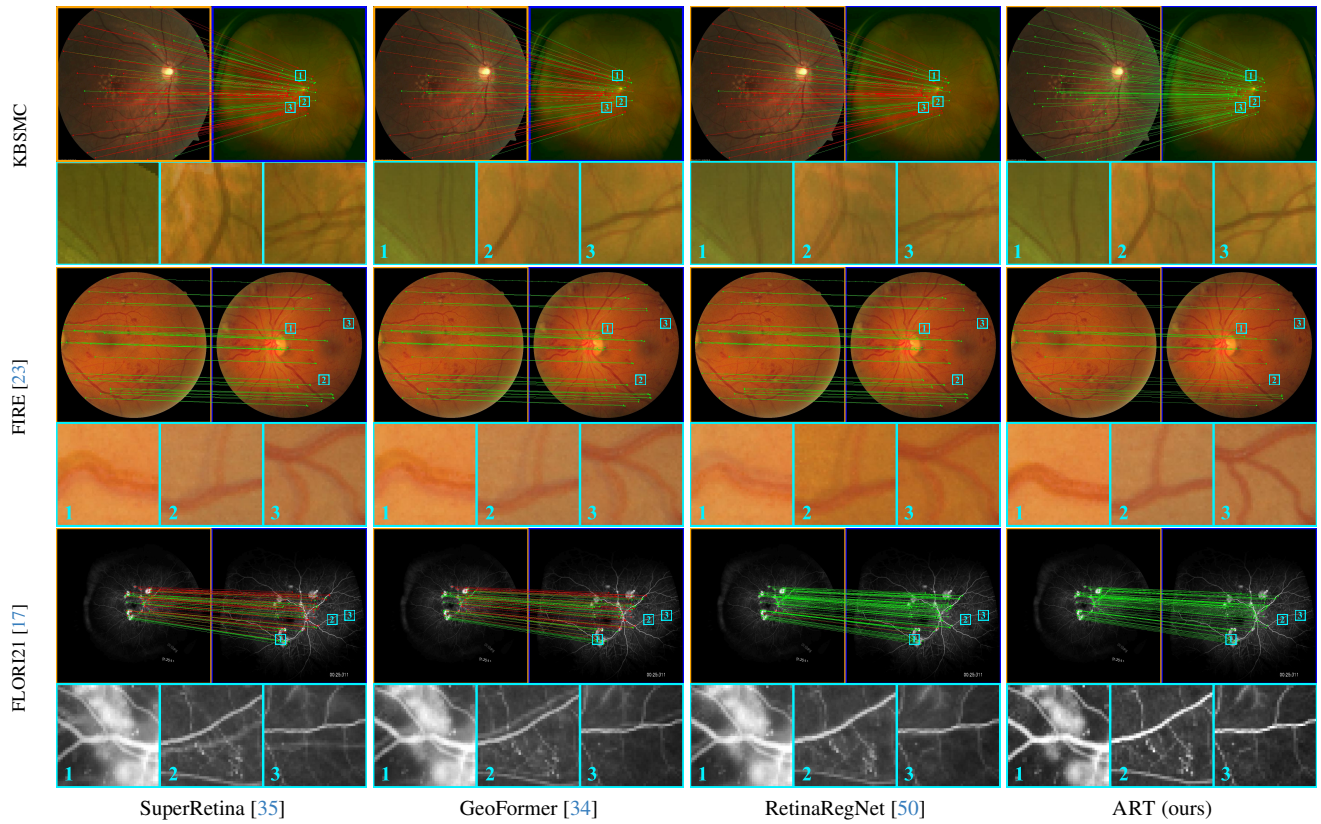


Figure 5. **Qualitative Evaluation on Retinal Datasets.** Across various domains, ART robustly identifies sufficient matches compared to SuperRetina [35], GeoFormer [34], and RetinaRegNet [50]. Correct and incorrect matches are shown in green and red, respectively. The zoomed-in boxes highlight overlaid local regions.

Table 2. **Quantitative Evaluation on Retinal Datasets.**

Methods	KBSMC		FIRE		FLORI21	
	Acceptable$_\uparrow$(%)	**mAUC**$_\uparrow$	*Acceptable*$_\uparrow$(%)	**mAUC**$_\uparrow$	*Acceptable*$_\uparrow$(%)	**mAUC**$_\uparrow$
SuperPoint [16]	9.09	8.7	94.78	67.3	40	39.1
GLAMpoints [55]	9.89	8.4	93.28	61.9	33.33	34.4
ISTN [30]	20.86	12.1	86.57	60.9	53.33	52.5
NCNet [47]	12.30	9.6	86.57	61.4	53.33	50.8
SuperGlue [49]	24.06	15.3	95.52	68.7	80	59.8
REMPE [24]	22.46	15.0	97.01	72.1	73.33	60.0
DLKFM [65]	22.73	13.5	86.57	61.4	40	40.1
LoFTR [52]	26.20	16.9	97.01	71.5	66.67	51.5
IHN [7]	23.80	14.5	88.81	63.5	60	50.0
SuperRetina [35]	34.76	22.3	<u>98.51</u>	75.5	80	65.0
ASPanFormer [10]	24.87	16.2	92.54	70.4	73.33	62.8
MCNet [66]	32.89	20.9	92.54	69.3	60	48.6
GeoFormer [34]	36.10	24.1	<u>98.51</u>	75.6	<u>93.33</u>	71.4
RetinaRegNet [50]	31.28	20.3	**99.25**	77.9	**100**	86.8
ART w/o CAL (ours)	<u>51.87</u>	<u>37.2</u>	**99.25**	<u>78.2</u>	**100**	<u>92.3</u>
ART w/ CAL (ours)	**64.71**	**40.1**	**99.25**	**78.5**	**100**	**92.5**

The bold and underline values denote the best and second best results, respectively.

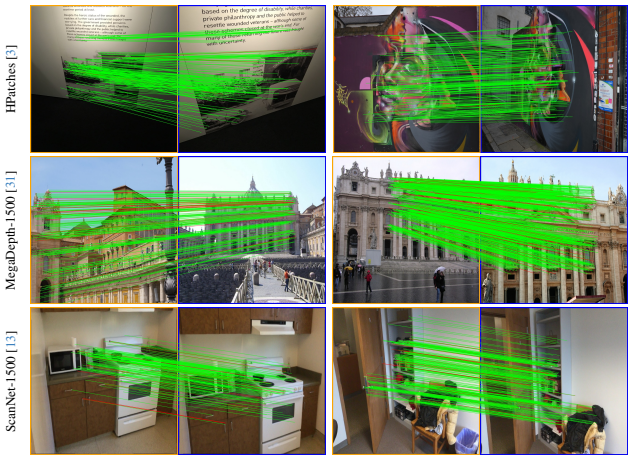


Figure 6. **Qualitative Evaluation on Scene-HR Datasets.** To evaluate our method under diverse conditions, we visualize correspondences on the Scene-HR datasets, including HPatches [3], MegaDepth-1500 [31], and ScanNet-1500 [13]. Correct and incorrect matches are shown in green and red, respectively.

Table 3. **Quantitative Evaluation on Scene-HR Datasets.**

Methods	mAUC↑		
	HPatches	MegaDepth-1500	ScanNet-1500
LoFTR [52]	75.4	67.7	40.7
LightGlue [33]	77.5	72.3	48.2
MatchFormer [58]	78.1	68.2	43.2
RoMa [18]	<u>78.4</u>	**75.2**	**52.0**
ART w/o CAL (ours)	75.3	68.8	45.8
ART w/ CAL (ours)	**78.6**	<u>74.9</u>	<u>51.1</u>

The bold and underline values denote the best and second best results, respectively.

4.4. Evaluation on Scene Categories

Baselines for Comparison We compare ART with LoFTR [52], LightGlue [33], MatchFormer [58], and RoMa [18] for HR datasets and DLKFM [65], IHN [7], and MCNet [66] for LR datasets.

Evaluation Metrics For 2D geometric transformation datasets, we follow prior works [52, 65] and report the average corner error (ACE) on GoogleEarth [65], GoogleMap [65], and MSCOCO [32], and the mAUC of ACE on HPatches [3] at thresholds of 3, 5, and 10 pixels. For two-view transformations (MegaDepth-1500 [31], ScanNet-1500 [13]), we follow prior works [18, 52] and report the mAUC of pose error at 5°, 10°, and 20°, defined as the maximum angular deviation in rotation and translation.

Discussion For the HR datasets, we pretrained the ART using images from the Oxford-Paris datasets [44, 45] and finetuned it in a SS manner on the HR datasets. For the LR datasets, we trained the model in a FS manner.

Tab. 3 and 4 present comparative quantitative evaluations, demonstrating the effectiveness of the proposed ART across HR and LR datasets with varying characteristics.

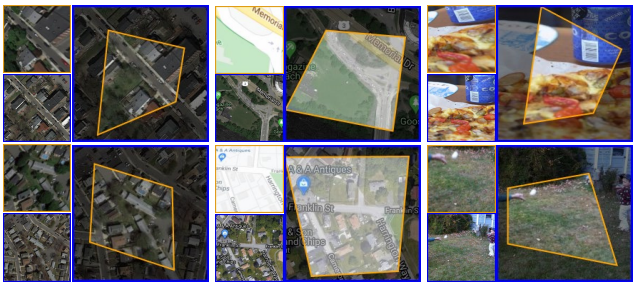


Figure 7. **Qualitative Evaluation on Scene-LR Datasets.** On the GoogleEarth [65], GoogleMap [65], and MSCOCO [32] datasets, ART successfully finds the correct transformation between input image pairs, even with sparse features from low resolution, domain gaps, and scale differences.

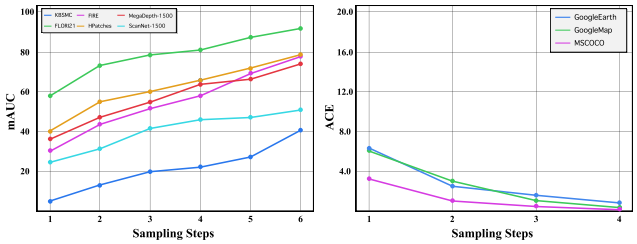
Table 4. **Quantitative Evaluation on Scene-LR Datasets.**

Methods	ACE↓		
	GoogleEarth	GoogleMap	MSCOCO
DLKFM [65]	3.88	4.41	0.55
IHN [7]	1.60	0.92	0.19
MCNet [66]	<u>0.60</u>	<u>0.23</u>	**0.03**
ART w/o CAL (ours)	0.65	0.96	<u>0.05</u>
ART w/ CAL (ours)	**0.17**	**0.19**	**0.03**

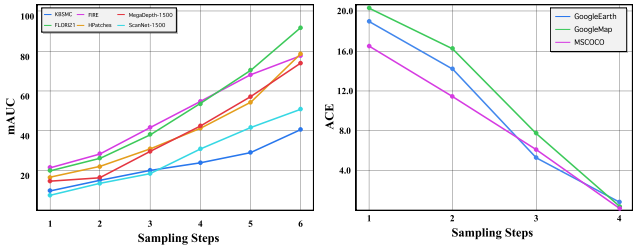
The bold and underline values denote the best and second best results, respectively.

For the HR datasets, ART achieves state-of-the-art performance on HPatches [3], which features mostly planar surfaces with rich structural details and minimal domain shifts. ART also achieves performance comparable to state-of-the-art on estimating two-view 3D geometric transformations from the MegaDepth-1500 [31] and ScanNet-1500 [13] datasets. We believe the lack of improvement compared to RoMa [18] is due to challenges like limited overlap, repetitive structures, and severe degradations that are hard to fully simulate despite extensive training augmentations (Sec. 4.2). Fig. 6 shows challenging cases where ART accurately estimates correspondences across datasets.

In contrast to HR datasets, LR datasets introduce additional challenges including significant feature loss and resolution discrepancies. GoogleMap [65] images also exhibit domain shift, adding to the alignment difficulty. As shown in Tab. 4, traditional iterative deep homography estimation methods [7, 65] struggle under these conditions. ART achieves the lowest ACE across all datasets, demonstrating superior adaptability to low-resolution and cross-domain alignment tasks. This suggests that our method's contextual feature refinement contributes to robust alignment, even in scenarios with large-scale variations and domain mismatches. Fig. 7 further highlights these capabilities, depicting cases where ART successfully estimates transformations for LR datasets.



(a) Performance change with standard initialization



(b) Performance change with varying initialization

Figure 8. **Ablation Study on Sampling.** ART performance varies with (a) the number of sampling steps and (b) different initialization strategies, across HR (left) and LR (right) datasets.

4.5. Understanding ART

Here, we present ablation studies to gain a deeper understanding of the key components that constitute ART.

Sampling Efficiency The aforementioned number of iteration steps, 6 for HR images and 4 for LR images, can be reduced to improve efficiency during inference.

For instance, starting with the initialized transform field parameters $\mathcal{D}^0_{\mathcal{M}} = \mathbf{1}_{12\times12\times2}$ and $\mathcal{D}^0_{\mathcal{A}} = \mathbf{0}_{12\times12\times2}$, only a few model inference steps are needed, followed by upsampling to achieve the full resolution of the transform field parameters $\mathcal{D}^K_{\mathcal{M}}$ and $\mathcal{D}^K_{\mathcal{A}}$. The results in Fig. 8 (a) show that the network estimates the transform field parameters early in the autoregressive process for both HR and LR datasets without requiring the full number of refinement steps. This suggests that, compared to other iterative refinement-based approaches [7, 8, 66], our autoregressive process is not only more effective but also adaptable to accelerated sampling.

Alternatively, increasing the spatial size of the initialized transform field parameters $\mathcal{D}^0_{\mathcal{M}}$ and $\mathcal{D}^0_{\mathcal{A}}$ can also allow ART to estimate the final parameters $\mathcal{D}^K_{\mathcal{M}}$ and $\mathcal{D}^K_{\mathcal{A}}$ at full resolution with fewer sampling steps. In Fig. 8 (b), the transform field parameters are initialized at resolutions of $\mathcal{D}^0_{\mathcal{M}} = \mathbf{1}_{H/(2^k)\times W/(2^k)\times2}$ and $\mathcal{D}^0_{\mathcal{A}} = \mathbf{0}_{H/(2^k)\times W/(2^k)\times2}$, where H and W denote the spatial resolution of the input image. It is evident that this limits the model's capacity for wide-range coarse estimation as well as fine-grained local refinements, resulting in poor performance.

Importance of CAL Comparative results with and without the use of CAL are reported in Tab. 2, 3, and 4 across all datasets. Without CAL, $\tilde{\mathcal{F}}^k_{s\to d}$ is computed as $\texttt{Concat}[\mathcal{F}^k_s, \mathcal{F}^k_d]$, lacking the attention necessary for refining the transform field parameters.

For image pairs in datasets such as FIRE [23], FLORI21 [17], or HPatches [3], which contain rich details and abundant features, CAL offers only marginal improvement. In contrast, for LR datasets with sparse features, such as GoogleEarth [65], MSCOCO [32], and GoogleMap [65], CAL significantly improves performance. In the challenging KBSMC dataset, characterized by ambiguous features, large scale variation, and a significant domain gap, CAL notably improves the *Acceptable* rate.

Computational Cost ART offers markedly better runtime efficiency compared to coarse-to-fine baselines such as LoFTR [52] (1.101s) and GeoFormer [34] (1.150s), due to its lightweight multi-scale design. On an NVIDIA A100 GPU, ART runs in 0.16s using 261MB of memory.

Transform Field Representation We empirically observed that predicting only additive parameters for translation degrades performance due to their wide dynamic range. Introducing multiplicative parameters for scaling stabilizes model optimization by normalizing spatial displacements, reducing translation variance, and ultimately enhancing performance. In contrast, directly predicting full affine or projective transform parameters often resulted in non-convergence due to the model's excessive complexity.

Limitations In this paper, we have considered only two types of datasets: HR and LR, with spatial resolutions of 768×768 and 192×192, respectively. This setting inherently constrains the input image size, and the required resizing can directly degrade the network's performance. For example, when estimating correspondences between larger images, the predefined 6 sampling steps may be insufficient to determine the transformation parameters accurately. Future research should explore adaptive sampling strategies that adjust dynamically to the input resolution.

5. Conclusion

ART tackles the challenging problem of image alignment, where existing methods struggle due to homogeneous textures, large scale differences, and weak feature regions. ART effectively mitigates issues related to poor initialization and scale dependency, achieving precise alignment even in difficult scenarios. Through extensive evaluations across diverse datasets, we demonstrated that ART significantly outperforms existing feature-based, intensity-based, and iterative refinement-based approaches. We believe that ART provides a strong foundation for future research in image alignment, particularly for applications such as medical imaging, remote sensing, and scene analysis.

Acknowledgements This work was supported in part by the IITP grants [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University), No. 2021-0-02068, No.2023-0-00156, and No.RS-2025-02219317], the NRF grant (RS-2025-00521972, AI Star Fellowship (Kookmin University)), and the Industrial Technology Alchemist Project [No. RS-2024-00432410] funded by MOTIE, Korea.

References

[1] Yechao Bai, Ziyuan Huang, Lyuyu Shen, Hongliang Guo, Marcelo H. Ang Jr, and Daniela Rus. Multi-scale feature aggregation by cross-scale pixel-to-region relation operation for semantic segmentation. *IEEE Robotics and Automation Letters*, 6(3):5889–5896, 2021. 3

[2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38, 2019. 1, 2

[3] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors, 2017. 5, 7, 8

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 2

[5] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 1992. 1, 3

[6] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *ECCV*, 2010. 2

[7] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *CVPR*, 2022. 2, 3, 5, 6, 7, 8

[8] Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, and Hui-Liang Shen. Recurrent homography estimation using homography-guided image warping and focus transformer. In *CVPR*, 2023. 8

[9] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *MICCAI*, 2017. 2

[10] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *ECCV*, 2022. 2, 5, 6

[11] Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022. 1

[12] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61, 1995. 1

[13] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5, 7

[14] Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52, 2019. 1, 2

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint*, 2016. 2

[16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 2, 5, 6

[17] Li Ding, Tony Kang, Ajay Kuriyan, Rajeev Ramchandran, Charles Wykoff, and Gaurav Sharma. Flori21: Fluorescein angiography longitudinal retinal image registration dataset, 2021. 5, 6, 8

[18] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 7

[19] Brachmann et al. Dsac* - differentiable ransac for camera localization. In *CVPR*, 2019. 4

[20] 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. 3

[21] Morteza Ghahremani, Mohammad Khateri, Bailiang Jian, Benedikt Wiestler, Ehsan Adeli, and Christian Wachinger. H-ViT: A Hierarchical Vision Transformer for Deformable Image Registration . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11513–11523, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1

[22] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge, 2003. 4

[23] Carlos Hernandez-Matas, Xenophon Zabulis, Areti Triantafyllou, Panagiota Anyfanti, Stella Douma, and Antonis A Argyros. Fire: Fundus image registration dataset. *Journal for Modeling in Ophthalmology*, 1, 2017. 5, 6, 8

[24] 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, 2020. 1, 2, 5, 6

[25] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical Image Analysis*, 49, 2018. 2

[26] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *ECCV*, 2022. 2

[27] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint*, 2016. 2

[28] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: cycle con-

sistent unsupervised deformable image registration. *Medical Image Analysis*, 71, 2021. 1, 2

[29] Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: unsupervised deformable image registration using diffusion model. In *ECCV*, 2022. 1, 4

[30] Matthew C.H. Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. Image-and-spatial transformer networks for structure-guided image registration. In *MICCAI*, 2019. 1, 2, 5, 6

[31] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 7

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5, 7, 8

[33] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 2, 7

[34] Jiazhen Liu and Xirong Li. Geometrized transformer for self-supervised homography estimation. In *ICCV*, 2023. 1, 2, 5, 6, 8

[35] Jiazhen Liu, Xirong Li, Qijie Wei, Jie Xu, and Dayong Ding. Semi-supervised keypoint detector and descriptor for retinal image matching. In *ECCV*, 2022. 1, 2, 5, 6

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 5

[37] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2004. 2, 4, 5

[38] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981. 3

[39] Tai Ma, Suwei Zhang, Jiafeng Li, and Ying Wen. Iirp-net: Iterative inference residual pyramid network for enhanced image registration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11546–11555, 2024. 1

[40] Frederik Maes, André Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997. 2

[41] Mingyuan Meng, Lei Bi, Michael Fulham, Dagan Feng, and Jinman Kim. *Non-iterative Coarse-to-Fine Transformer Networks for Joint Affine and Deformable Image Registration*, page 750–760. Springer Nature Switzerland, 2023. 2

[42] Mingyuan Meng, Lei Bi, Michael Fulham, Dagan Feng, and Jinman Kim. *Non-iterative Coarse-to-Fine Transformer Networks for Joint Affine and Deformable Image Registration*, page 750–760. Springer Nature Switzerland, 2023. 1

[43] Mingyuan Meng, Dagan Feng, Lei Bi, and Jinman Kim. Correlation-aware coarse-to-fine mlps for deformable medical image registration, 2024. 1

[44] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 7

[45] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 7

[46] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Repeatable and reliable detector and descriptor. *arXiv preprint*, 2019. 2

[47] 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, 2020. 2, 5, 6

[48] Edward Rosten, Reid Porter, and Tom Drummond. Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2008. 2

[49] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 5, 6

[50] Vishal Balaji Sivaraman, Muhammad Imran, Qingyue Wei, Preethika Muralidharan, Michelle R. Tamplin, Isabella M . Grumbach, Randy H. Kardon, Jui-Kai Wang, Yuyin Zhou, and Wei Shao. Retinaregnet: A zero-shot approach for retinal image registration, 2024. 2, 5, 6

[51] Charles Stewart, Chia-Ling Tsai, and Badrinath Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE Transactions on Medical Imaging*, 22, 2003. 5

[52] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 1, 2, 5, 6, 7, 8

[53] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[54] Jean-Philippe Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998. 2

[55] Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, and Sandro De Zanet. Glampoints: Greedily learned accurate match points. In *ICCV*, 2019. 2, 5, 6

[56] Paul Viola and William M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997. 2

[57] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 3

[58] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Asian Conference on Computer Vision*, 2022. 7

[59] Yu Wang, Xiaoye Wang, Zaiwang Gu, Weide Liu, Wee Siong Ng, Weimin Huang, and Jun Cheng. Superjunction:

Learning-based junction detection for retinal image registration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):292–300, 2024. 1

[60] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, 2022. 2

[61] Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *MICCAI*, 2019. 2

[62] Feihu Zhang, Oliver J Woodford, Victor Adrian Prisacariu, and Philip HS Torr. Separable flow: Learning motion cost volumes for optical flow estimation. In *ICCV*, 2021. 2

[63] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *ICLR*, 2024. 3

[64] Xiaoran Zhang, John C. Stendahl, Lawrence Staib, Albert J. Sinusas, Alex Wong, and James S. Duncan. Adaptive correspondence scoring for unsupervised medical image registration, 2024. 1

[65] Yiming Zhao, Xinming Huang, and Ziming Zhang. Deep lucas-kanade homography for multimodal image alignment. In *CVPR*, 2021. 2, 3, 5, 6, 7, 8

[66] Haokai Zhu, Si-Yuan Cao, Jianxin Hu, Sitong Zuo, Beinan Yu, Jiacheng Ying, Junwei Li, and Hui-Liang Shen. Mcnet: Rethinking the core ingredients for accurate and efficient homography estimation. In *CVPR*, 2024. 1, 2, 3, 5, 6, 7, 8

[67] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003. 2