

Scale and Rotation Estimation of Similarity-Transformed Images via Cross-Correlation Maximization Based on Auxiliary Function Method

Shinji Yamashita*, Yuma Kinoshita*[†], and Hitoshi Kiya[†]

* Tokai University, Japan

E-mail: {5CEIM050, ykinoshita}@tokai.ac.jp

[†] Tokyo Metropolitan University, Japan

E-mail: kiya@tmu.ac.jp

Abstract—This paper introduces a highly efficient algorithm capable of jointly estimating scale and rotation between two images with sub-pixel precision. Image alignment serves as a critical process for spatially registering images captured from different viewpoints, and finds extensive use in domains such as medical imaging and computer vision. Traditional phase-correlation techniques are effective in determining translational shifts; however, they are inadequate when addressing scale and rotation changes, which often arise due to camera zooming or rotational movements. In this paper, we propose a novel algorithm that integrates scale and rotation estimation based on the Fourier transform in log-polar coordinates with a cross-correlation maximization strategy, leveraging the auxiliary function method. By incorporating sub-pixel-level cross-correlation our method enables precise estimation of both scale and rotation. Experimental results demonstrate that the proposed method achieves lower mean estimation errors for scale and rotation than conventional Fourier transform-based techniques that rely on discrete cross-correlation.

I. INTRODUCTION

Image alignment is a technique that spatially registers images by correcting geometric transformations between them. This technique is essential in a wide range of applications, including medical image processing [1], panorama stitching [2], 3D reconstruction [3], and high dynamic range (HDR) image generation [4], [5]. In recent years, there has been an increasing demand for fast and accurate alignment even on resource-constrained devices such as smartphones, highlighting the importance of low-cost algorithms.

There are two principal approaches to image alignment: homography-based methods and intensity-based methods. Homography-based approaches estimate a homography matrix between two images by leveraging feature descriptors such as SIFT [6], SURF [7], and A-KAZE [8], often in conjunction with the RANSAC algorithm [9]. Recently, deep learning-based techniques for homography estimation have also been investigated [10], [11]. Intensity-based methods generally focus on estimating optical flow [12]–[15]. However, both homography and optical flow estimation methods typically incur considerable computational costs.

To address these limitations, two-dimensional (2D) cross-correlation-based methods [16], [17] are frequently employed for real-time image alignment [5]. These methods aim to maximize the 2D cross-correlation function between two images to estimate the translational displacement. The cross-correlation for discrete pixel shifts can be computed efficiently using the fast Fourier transform (FFT), and its maximum can be readily identified. Furthermore, Kinoshita et al. proposed an efficient algorithm for maximizing the 2D cross-correlation function with subpixel accuracy, based on the auxiliary function method, also known as majorization-minimization (MM) [18].

Nevertheless, methods based on maximizing the 2D cross-correlation function typically assume that the displacement between images is purely translational and estimate only the translation parameters. In practice, however, images often exhibit differences in scale and rotation due to camera zoom or rotation, resulting in misalignments beyond simple translation.

In response to this issue, we aim to estimate scale and rotation changes with high accuracy from a pair of images including scale, rotation, and translation differences. To achieve this goal, we combine a Fourier transform-based scale and rotation estimation [19] with the cross-correlation maximization algorithm proposed by Kinoshita et al [18]. In this method, scale and rotation are estimated by maximizing the 2D cross-correlation between the amplitude spectra represented in log-polar coordinate system. For maximizing the 2D cross-correlation, the use of the algorithm developed by Kinoshita et al. enables us to estimate scale and rotation with high accuracy.

We evaluate the estimation accuracy of the proposed method and the baseline estimation with the standard discrete cross-correlation maximization. Experimental results demonstrate that the proposed method outperforms the baseline method in terms of absolute errors between the ground truth and the estimated values.

II. PRELIMINARIES

In this section, we briefly formalize the image alignment problem and present the mathematical framework that underpins the proposed methods.

A. Problem Statement

This paper focuses on the task of aligning two-dimensional discrete signals, particularly images, represented by x and y . The pixel value at position $\mathbf{p} = (p_1, p_2)^\top$ is denoted by $x[\mathbf{p}]$ and $y[\mathbf{p}]$, where the superscript \top indicates the transpose of a vector or matrix. If the spatial misalignment between the images x and y can be described by a combination of translation $\Delta\mathbf{p} \in \mathbb{R}^2$, rotation $\mathbf{R}(\theta)$, and scaling $s \in \mathbb{R}^+$, then the relationship can be formulated as

$$y[\mathbf{p}] = x[s\mathbf{R}(\theta)\mathbf{p} + \Delta\mathbf{p}], \quad (1)$$

where

$$\mathbf{R}(\theta) = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}. \quad (2)$$

The aim of this paper is to estimate the parameters θ , and s from the given pair (x, y) because once scale and rotation are aligned the parameter $\Delta\mathbf{p}$ can be easily estimated by cross-correlation methods.

B. Maximization of 2D cross-correlation

When the displacement between the images is due solely to translation (i.e., $s = 1$ and $\theta = 0$), the parameter $\Delta\mathbf{p}$ can be determined by maximizing the generalized two-dimensional cross-correlation function. Let \hat{x} and \hat{y} denote the two-dimensional discrete Fourier transforms of x and y , respectively, both of size $N \times M$. We assume \hat{x} and \hat{y} are strictly band-limited, meaning $\hat{x}(\boldsymbol{\omega}) = \hat{y}(\boldsymbol{\omega}) = 0$ for angular frequency $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$ where at least one component equals π . Utilizing the 2D cross-spectrum is defined as $\hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl}) = \hat{x}^*(\boldsymbol{\omega}_{kl})\hat{y}(\boldsymbol{\omega}_{kl})$, the generalized two-dimensional cross-correlation function between x and y is given by

$$\check{\Phi}_2^{(xy)}[\mathbf{p}] = \frac{1}{NM} \sum_{k \in K} \sum_{l \in L} w_{kl} \hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl}) \exp(j\boldsymbol{\omega}_{kl}^\top \mathbf{p}), \quad (3)$$

where j donates the imaginary unit, $w_{kl} \in \mathbb{R}^+$ denotes an arbitrary positive weighting coefficient, and $\boldsymbol{\omega}_{kl} = (\omega^{(k)}, \omega^{(l)})^\top = (\frac{2\pi k}{N}, \frac{2\pi l}{M})^\top$. The sets K and L are defined as $K = \{-N/2 + 1, -N/2 + 2, \dots, N/2\}$ and $L = \{-M/2 + 1, -M/2 + 2, \dots, M/2\}$, respectively. When $w_{kl} = 1$, the function reduces to the standard cross-correlation. In the case where $w_{kl} = |\hat{\Phi}_2^{(xy)}(\boldsymbol{\omega}_{kl})|^{-1}$, the method becomes equivalent to phase-only correlation (POC) [20], which is also referred to as generalized cross-correlation with phase transform (GCC-PHAT) [16] in the context of acoustic signal processing.

Kinoshita et al. [18] proposed an algorithm that regards \mathbf{p} in Eq. (3) as a real-valued vector, considers the continuous function $\Phi_2^{(xy)}$, and maximizes

$$\tilde{\Delta\mathbf{p}} = \arg \max_{\mathbf{p} \in \mathbb{R}^2} \Phi_2^{(xy)}(\mathbf{p}) \quad (4)$$

using the auxiliary function method, thereby achieving fast convergence to a local optimum and subpixel accuracy in displacement estimation. Instead of directly maximizing the

objective function in Eq. (4), the algorithm iteratively maximizes an auxiliary function $Q(\mathbf{p}, \theta)$ that serves as its lower bound.

$$\boldsymbol{\theta}^{(i)} = f(\mathbf{p}^{(i)}), \quad \mathbf{p}^{(i+1)} = \arg \max_{\mathbf{p} \in \mathbb{R}^2} Q(\mathbf{p}, \boldsymbol{\theta}^{(i)}). \quad (5)$$

Here, i denotes the iteration number.

C. Fourier Transform-Based Scale and Rotation Estimation

The spectra \hat{x} and \hat{y} of the signals $x(\mathbf{p})$ and $y(\mathbf{p}) = x(s\mathbf{R}(\theta)\mathbf{p} + \Delta\mathbf{p})$ satisfy the following relationship:

$$\hat{y}(\boldsymbol{\omega}) = \exp(j2\pi\boldsymbol{\omega}^\top \mathbf{R}(-\theta)\Delta\mathbf{p}) \frac{1}{s^2} \hat{x}\left(\frac{1}{s}\mathbf{R}(\theta)\boldsymbol{\omega}\right). \quad (6)$$

Consequently, the amplitude spectrum is invariant to translation and, after appropriate scaling and rotation, coincides with the amplitude spectrum of the other signal. When the amplitude spectrum is represented in log-polar coordinates (ρ, ϕ) , where $\rho = \log \|\boldsymbol{\omega}\|$ and $\phi = \tan^{-1}(\omega_2/\omega_1)$, scaling corresponds to a translation along the ρ axis, whereas rotation corresponds to a translation along the ϕ axis. By exploiting this property, the scale factor s and rotation angle θ can be estimated by maximizing the cross-correlation between the amplitude spectra in the log-polar domain [19]. The derivation of Eq. (6) is presented below.

When $y(\mathbf{p})$ is an affine transformation of $x(\mathbf{p})$, that is, $y(\mathbf{p}) = x(\mathbf{A}\mathbf{p} + \mathbf{b})$ with $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ and $\mathbf{b} \in \mathbb{R}^2$, the Fourier transform is given by

$$\hat{y}(\boldsymbol{\omega}) = \iint_{\mathbb{R}^2} y(\mathbf{p}) \exp(-j2\pi\boldsymbol{\omega}^\top \mathbf{p}) d\mathbf{p} \quad (7)$$

$$= \iint_{\mathbb{R}^2} x(\mathbf{A}\mathbf{p} + \mathbf{b}) \exp(-j2\pi\boldsymbol{\omega}^\top \mathbf{p}) d\mathbf{p} \quad (8)$$

By applying the change of variables $\mathbf{q} = \mathbf{A}\mathbf{p} + \mathbf{b}$ yields $\mathbf{p} = \mathbf{A}^{-1}(\mathbf{q} - \mathbf{b})$, and therefore,

$$d\mathbf{p} = \left| \frac{\partial \mathbf{p}}{\partial \mathbf{q}} \right| d\mathbf{q} = \frac{1}{|\mathbf{A}|} d\mathbf{q}. \quad (9)$$

Thus,

$$\hat{y}(\boldsymbol{\omega}) = \frac{1}{|\mathbf{A}|} \iint_{\mathbb{R}^2} x(\mathbf{q}) \exp(-j2\pi\boldsymbol{\omega}^\top \mathbf{A}^{-1}(\mathbf{q} - \mathbf{b})) d\mathbf{q} \quad (10)$$

$$= \exp(j2\pi\boldsymbol{\omega}^\top \mathbf{A}^{-1}\mathbf{b}) \frac{1}{|\mathbf{A}|} \hat{x}((\mathbf{A}^{-1})^\top \boldsymbol{\omega}), \quad (11)$$

where

$$\hat{x}((\mathbf{A}^{-1})^\top \boldsymbol{\omega}) = \iint_{\mathbb{R}^2} x(\mathbf{q}) \exp(-j2\pi\boldsymbol{\omega}^\top \mathbf{A}^{-1}\mathbf{q}) d\mathbf{q}. \quad (12)$$

By substituting $\mathbf{A} = s\mathbf{R}(\theta)$ and $\mathbf{b} = \Delta\mathbf{p}$ into Eq. (11), we recover the expression in Eq. (6).

In this paper, we combine the Fourier transform-based scale and rotation estimation with the cross-correlation maximization algorithm proposed by Kinoshita et al., enabling highly accurate estimation of the rotation angle θ and the scale factor s .

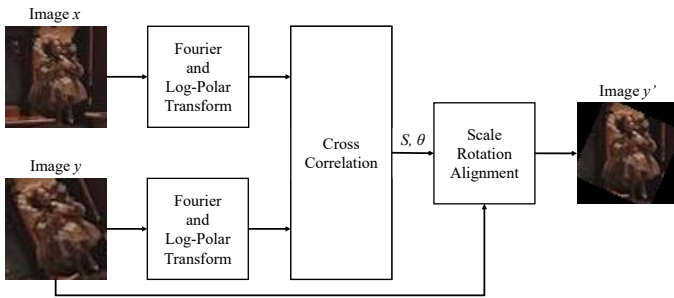


Fig. 1. Procedure of proposed method

III. PROPOSED METHOD

In this section, we present our algorithm for estimating the displacement in scale and rotation between two images, x and y .

A. Overview

The procedure for the proposed method is illustrated in Fig. 1. To estimate the displacement in scale and rotation, we maximize the cross-correlation between the amplitude spectra of the images represented in log-polar coordinates, since scaling and rotation correspond to translation in this transformed domain, as described in section II-C. Here, we adopt the continuous cross-correlation and iteratively optimize its auxiliary function, as discussed in section II-B.

The proposed procedure consists of the following steps:

- 1) Compute the amplitude spectra $|\hat{x}|$ and $|\hat{y}|$ of the images x and y , respectively.
- 2) Transform the amplitude spectra into log-polar coordinates.
- 3) Estimate the scale and rotation parameters by maximizing the continuous cross-correlation between the spectra in the log-polar domain.

Each step is explained in detail below.

B. Calculation of Amplitude Spectrum

To obtain the amplitude spectrum of image x , we compute the discrete Fourier transform using a window function ϖ ,

$$\hat{x}[\boldsymbol{\omega}] = \sum_{p_1=0}^{N-1} \sum_{p_2=0}^{M-1} \varpi[\mathbf{p}]x[\mathbf{p}] \exp(-j\boldsymbol{\omega}^\top \mathbf{p}), \quad (13)$$

where we use a 2D Gaussian window with parameters $\sigma_1 = N/5$ and $\sigma_2 = M/5$ as ϖ . The amplitude $|\hat{x}|$ is then computed. The amplitude spectrum of image y is calculated in the same manner.

C. Log-Polar Transform

The amplitude spectra $|\hat{x}|$ and $|\hat{y}|$ in Cartesian coordinates are transformed into log-polar coordinates. This transformation is referred to as the log-polar transform. The log-polar transformation is defined for $\mathbf{r} = (\rho, \phi)^\top$ as follows:

$$|\hat{X}[\mathbf{r}]| = |\hat{x}[(e^\rho \cos \phi, e^\rho \sin \phi)^\top]|, \quad (14)$$



Fig. 2. Original high-resolution image “MusicBox” used in the experiment

where $\rho = \log \|\boldsymbol{\omega}\|$ and $\phi = \tan^{-1}(\omega_2/\omega_1)$. We evaluate Eq. (14) at discrete points $\rho \in \{\frac{\log(\min(N,M))}{N}i : i = 1, \dots, N\}$ and $\phi \in \{\frac{2\pi}{M}i : i = 0, \dots, M-1\}$. Since $|\hat{x}|$ is a discrete signal, bi-cubic interpolation is applied to $|\hat{x}|$ for resampling it. The amplitude spectrum $|\hat{Y}|$ of image $|\hat{y}|$ is computed in the same manner.

D. Estimation of Scale and Rotation

The cross-correlation function between $|\hat{X}|$ and $|\hat{Y}|$ is maximized using the algorithm of Kinoshita et al., as described in section II-B. When calculating the cross-spectrum $\hat{\Phi}^{(XY)}$ of $|\hat{X}|$ and $|\hat{Y}|$, both $|\hat{X}|$ and $|\hat{Y}|$ are zero-mean normalized and processed with the Gaussian window ϖ . As a result, the estimated translational displacement $\tilde{\Delta \mathbf{r}} = (\tilde{\rho}, \tilde{\phi})^\top$ in the log-polar domain yields the estimates for scale and rotation, \tilde{s} and $\tilde{\theta}$, as follows:

$$\tilde{s} = \exp\left(\frac{\log(\min(N, M))}{N} \tilde{\rho}\right), \quad \tilde{\theta} = \frac{2\pi}{M} \tilde{\phi} \quad (15)$$

IV. EXPERIMENTS

To evaluate the performance of the proposed method in terms of estimation accuracy, we conducted a simulation experiment.

A. Experimental Setup

Five pairs of images were prepared by simulating similarity transformations. In this experiment, random similarity transformations with parameters s , θ , and $\Delta \mathbf{p}$ were applied to a high-resolution image, “MusicBox” (shown in Fig. 2), which was obtained from the “Ultra-high Definition/Wide-Color-Gamut Standard Images” dataset¹. Image pairs were subsequently generated by randomly cropping 64×64 pixel regions from the both original and the transformed images at corresponding locations.

The parameters, namely the scaling factor s , rotation angle θ , and translation vector $\Delta \mathbf{p}$, were randomly sampled from uniform distributions over the intervals listed in Table I. The absolute errors between the ground-truth and the estimated values were computed to evaluate the estimation accuracy.

¹<https://www.ite.or.jp/content/test-materials/uhdvtv/>

TABLE I
RANGES OF TRANSFORMATION PARAMETERS USED IN THE EXPERIMENT.

Parameter	Range
Rotation angle θ ($^\circ$)	-30 – 30
Scaling factor s	0.8 – 1.2
Horizontal translation p_1 (px.)	-5 – 5
Vertical translation p_2 (px.)	-5 – 5

TABLE II
ABSOLUTE ESTIMATION ERRORS FOR SCALE AND ROTATION ANGLE:
COMPARISON BETWEEN THE BASELINE AND PROPOSED METHODS

	Baseline		Ours	
	Scale	Angle	Scale	Angle
Image 1	0.038	0.676	0.031	0.924
Image 2	0.002	1.987	0.001	0.702
Image 3	0.193	2.085	0.192	0.472
Image 4	0.074	1.430	0.040	3.095
Image 5	0.072	2.281	0.012	0.783
Average	0.076	1.692	0.055	1.195
Variance	0.003	0.281	0.004	0.769

The proposed method was compared against the conventional Fourier transform-based scale and rotation estimation approach, which maximizes the standard discrete cross-correlation (Baseline).

B. Experimental Results

The experimental results are summarized in Table II. For each method, the table reports the estimation errors in scale and rotation angle for five image pairs, along with their corresponding averages and variances. As shown in Table II, the proposed method consistently achieves lower average errors in both scale and rotation angle estimation than the baseline method, indicating improved estimation accuracy.

For Images 2 and 3, the input images x and y , as well as the alignment results based on scale and rotation estimation by both the baseline and proposed methods, are illustrated in Figs. 3 and 4, respectively. As seen in Fig. 3, both the baseline and the proposed methods achieve accurate scale alignment. Furthermore, the proposed method demonstrated superior rotation estimation compared to the baseline, as reflected in Table II. As seen in Fig. 4, although the proposed method outperforms the baseline in terms of rotation estimation, both methods yield suboptimal estimations of the scale factor, as also indicated in Table II.

These results indicate that while the proposed method generally provides higher estimation accuracy than the baseline, there remain certain images for which the estimation is not sufficiently reliable. Achieving robust and highly accurate estimation for arbitrary input images remains an important area for future work.

V. CONCLUSIONS

In this paper, we have proposed a novel approach that combines Fourier transform-based scale and rotation estimation with the cross-correlation maximization algorithm developed by Kinoshita et al. The integration of Kinoshita's algorithm

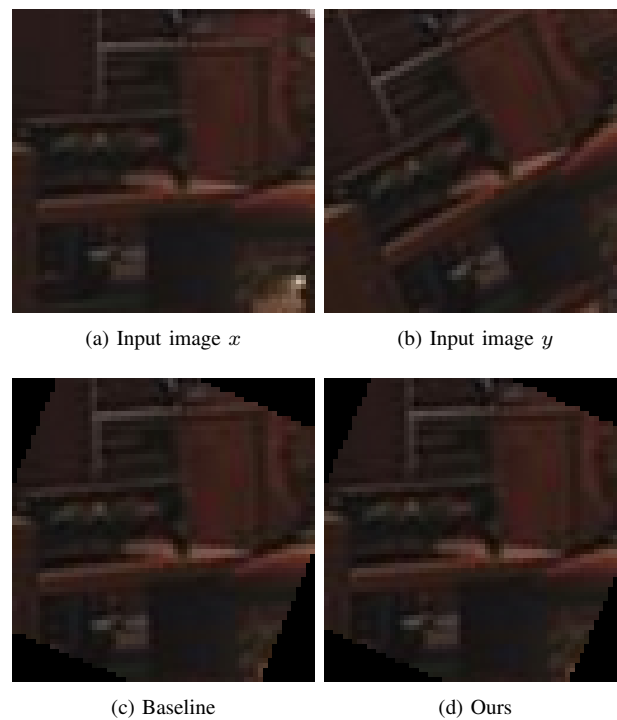


Fig. 3. Alignment results for Image 2

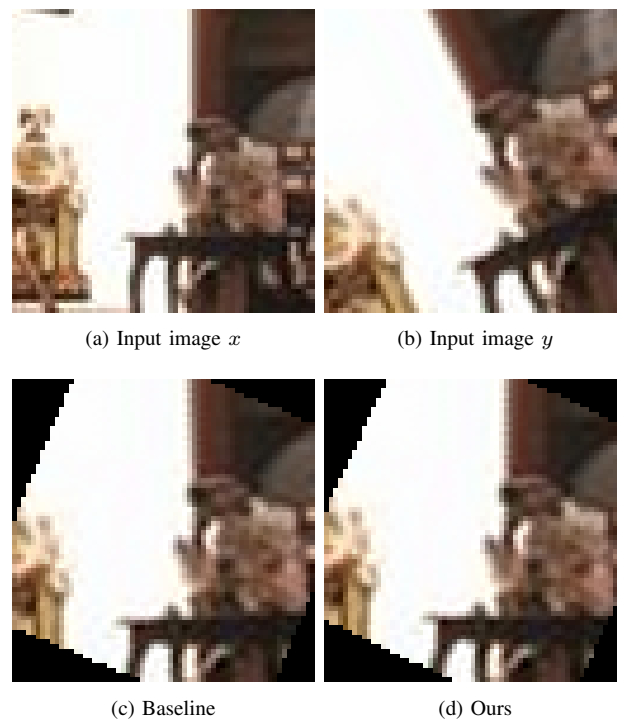


Fig. 4. Alignment results for Image 3

enables highly accurate estimation of both scale and rotation parameters. Experimental results demonstrated that the proposed method generally achieves superior estimation accuracy compared to the baseline method; however, suboptimal estimations were observed for certain image pairs.

For future work, we aim to further enhance estimation

accuracy across a wider variety of input images, and to extend the alignment framework to simultaneously estimate translation parameters in addition to scale and rotation.

REFERENCES

- [1] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: A survey," *Machine Vision and Applications*, vol. 31, no. 1, p. 8, Jan. 2020, ISSN: 1432-1769.
- [2] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, Aug. 2007, ISSN: 1573-1405.
- [3] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *International Journal of Computer Vision*, vol. 80, no. 2, pp. 189–210, Nov. 2008, ISSN: 1573-1405.
- [4] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005, ISBN: 0125852630.
- [5] S. W. Hasinoff, D. Sharlet, R. Geiss, *et al.*, "Burst Photography for High Dynamic Range and Low-Light Imaging on Mobile Cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Nov. 2016.
- [6] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [7] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Proc. ECCV*, vol. 3951, May 2006, pp. 404–417.
- [8] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE Features," in *Proc. ECCV*, vol. 7577, Oct. 2012, pp. 214–227.
- [9] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
- [10] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative Deep Homography Estimation," in *Proc. IEEE/CVF CVPR*, Jun. 2022, pp. 1869–1878.
- [11] M. Hong, Y. Lu, N. Ye, C. Lin, Q. Zhao, and S. Liu, "Unsupervised Homography Estimation with Coplanarity-Aware GAN," in *Proc. IEEE/CVF CVPR*, Jun. 2022, pp. 17 642–17 651.
- [12] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proc. DARPA Image Underst. Workshop*, Apr. 1981, pp. 121–130.
- [13] A. Dosovitskiy, P. Fischer, E. Ilg, *et al.*, "FlowNet: Learning Optical Flow with Convolutional Networks," in *Proc. IEEE ICCV*, Dec. 2015, pp. 2758–2766.
- [14] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks," in *Proc. IEEE/CVF CVPR*, Jul. 2017, pp. 2462–2470.
- [15] H. Jung, Z. Hui, L. Luo, *et al.* "Anyflow: Arbitrary Scale Optical Flow with Implicit Neural Representation." arXiv: 2303.16493. (Mar. 29, 2023), [Online]. Available: <http://arxiv.org/abs/2303.16493>.
- [16] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
- [17] K. Takita, T. Higuchi, and K. Kobayashi, "High-Accuracy Subpixel Image Registration Based on Phase-Only Correlation," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. 86, no. 8, pp. 1925–1934, Aug. 1, 2003.
- [18] K. Yuma, Y. Kouei, and H. Kiya, "2d cross-correlation maximization based on auxiliary function method," *Proc. APSIPA ASC*, Nov. 2023.
- [19] B. Reddy and B. Chatterji, "An fft-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. Image Process.*, vol. 5, no. 8, pp. 1266–1271, 1996. DOI: 10.1109/83.506761.
- [20] C. D. Kuglin, "The Phase Correlation Image Alignment Method," in *Proc. Int. Conf. Cybern. Soc.*, Sep. 1975, pp. 163–165.