

Ultra-Precision 6DoF Pose Estimation Using 2-D Interpolated Discrete Fourier Transform

Guowei Shi*, Zian Mao*, Peisen Huang†

UM-SJTU Joint Institute, Shanghai Jiao Tong University

{shiguowei, Asakura_kukii, peisen.huang}@sjtu.edu.cn

Abstract

Ultra-precision estimation of 6DoF pose is essential in applications such as semiconductor manufacturing and nanoscale manipulation. Conventional vision-based techniques are often hampered by sensitivity to defocus and limited estimation accuracy. In this paper, we propose a novel two-dimensional interpolated Discrete Fourier Transform (2D-IpDFT) method for robust 6DoF pose estimation using periodic patterns. We further develop a mathematical framework that links image parameters—phase and frequency—to 6DoF pose, which is applicable to both orthographic and quasi-orthographic imaging systems. Extensive experiments on a low-cost setup, featuring an industrial camera and an etched checkerboard pattern, demonstrate translation estimation accuracy at the nanometer level and rotation estimation accuracy at the microradian level.

1. Introduction

Vision-based 6DoF pose estimation plays a pivotal role in industrial automation and precision engineering, particularly in applications demanding ultra-high precision such as semiconductor manufacturing (e.g., wafer/mask alignment, lithography stage control), microrobotics, and nanoscale manipulation (e.g., integrating with AFMs/SEMs for sample positioning). These applications require cost-effective solutions capable of achieving sub-micrometer or even nanometer-level accuracy within compact workspaces, often utilizing periodic patterns such as 2D optical gratings. Traditional non-contact measurement techniques, such as laser interferometry or capacitive sensing, often face limitations in multi-DoF capability, flexibility, cost, and device footprint. In contrast, vision-based methods, especially those leveraging periodic patterns, offer attractive alternatives that are multi-DoF, ultra-precise, low-cost, and compact.

Existing vision-based pose estimation methods can be divided into spatial-domain and frequency-domain approaches. Spatial-domain methods include template matching [11, 16, 25], feature matching [13, 17], and optical flow analysis [9]. Among these, the Perspective-n-Point (PnP) method [5, 10, 21], a feature matching technique, is widely used for general pose estimation. However, PnP can suffer significant performance degradation as a result of image defocus. It is also incompatible with telecentric or near-telecentric lenses often employed in high-precision micro/nano-scale applications to minimize perspective distortion.

In contrast, frequency-domain methods [1, 2, 4, 6, 7, 18], which typically use periodic patterns, are inherently suited for higher precision in such small-scale, controlled environments. These methods rely heavily on the accuracy of frequency and phase estimation from the captured pattern. A critical challenge arises when the number of pattern periods captured in the image is insufficient: the spectral lines of the image become closely spaced, leading to spectral leakage and significant bias in the estimated frequency and phase. This bias can also change substantially with variations in the pattern's frequency and initial phase, making it difficult to eliminate consistently. Consequently, existing frequency-domain methods are often limited to applications where the measurement range is short or the image contains many cycles of the pattern, restricting their broader use.

To address these challenges, we propose a novel 2-D IpDFT algorithm that significantly improves the accuracy of frequency and phase estimation, particularly in scenarios where the spectrum lines of the image are close to each other. This algorithm considers both positive and negative frequency contributions to eliminate spectrum leakage and uses windowing techniques to reduce bias. As a result, its robustness to frequency and phase changes is significantly enhanced. Furthermore, we derive the mathematical relationship between the frequency and phase of the pattern image and the 6DoF pose of the pattern, applicable to both orthographic and quasi-orthographic projection systems common in micro/nano metrology. Simula-

*Equal contribution.

†Corresponding author.

tion, experimental validation, and comparisons with existing frequency-domain methods demonstrate the effectiveness of our approach, achieving nanometer-level displacement and microradian-level rotation estimation accuracy at low cost with a standard checkerboard pattern. Our contributions are summarized as follows:

- A new 2-D IpDFT algorithm that significantly enhances 6DoF pose estimation accuracy with a periodic pattern, achieving estimation accuracy at the nanometer-level for translation and microradian-level for rotation.
- A novel formulation that connects the frequency and phase of a periodic pattern with its 6DoF pose, applicable to all orthogonal and quasi-orthogonal camera systems.

2. Related Work

Vision-based precision measurement systems typically rely on algorithms to analyze captured images and extract relevant metrics to measure displacement and rotation. They are broadly categorized into spatial- and frequency-domain methods.

Spatial-Domain Methods. Template matching [11, 16, 25] and feature matching [13, 17] dominate spatial-domain methods due to their simplicity. Techniques like SIFT [13] and ORB [17] extract invariant descriptors for pose estimation but struggle with the defocus problem. Active contour models [22] and optical flow [9] methods can handle deformable objects but require high computational resources. Recent advances in deep learning (DL) for pose estimation [3, 8, 12, 14, 19, 23] show promise for general scenes and non-periodic textures. However, they typically require extensive training data and specialized hardware and may not be able to guarantee ultra-high precision needed for metrological applications. In general, DL methods aim for broad generalizability, whereas our work focuses on maximizing precision in the specific context of periodic pattern-based metrology.

Frequency-Domain Methods. These techniques exploit image frequency and phase information from periodic patterns (e.g., sinusoidal gratings, grids [1, 2], or checkerboards [18]) to estimate pose. Such patterns are widely adopted in micro/nano-scale precision measurements due to their well-defined spectral properties. While leveraging these properties can yield high precision, a key challenge is maintaining accuracy when only a few pattern periods are visible in the image, leading to closely spaced spectral lines. In such a case, significant bias in frequency and phase estimation arises due to spectral leakage, primarily caused by image boundary effects, negative frequency components, and pattern harmonics (particularly with non-sinusoidal patterns such as checkerboards). For small relative motions, this bias may be approximately uniform and compensable. However, in absolute positioning or large-scale displacement measurement where frequency and phase vary signif-

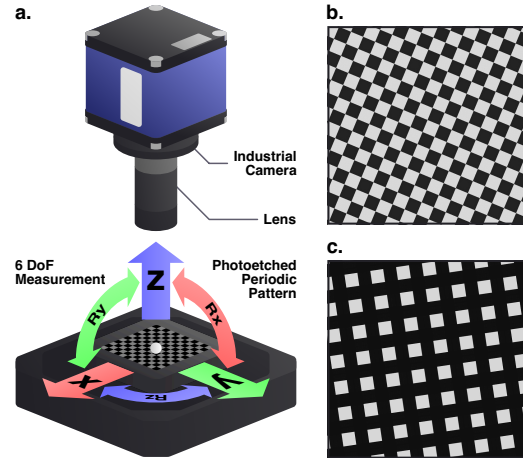


Figure 1. 6DoF pose estimation setup. (a) Overview of the setup, (b) checkerboard pattern, and (c) grid pattern.

icantly, the bias becomes non-negligible and harder to correct. Consequently, existing frequency-domain algorithms achieve optimal performance only under conditions of minimal frequency and phase variation, which either restricts their practical operating range or necessitates a large number of pattern periods within the field of view. Our work addresses these limitations by introducing a more robust frequency and phase estimation algorithm.

3. Camera and Pattern Models

A vision-based 6DoF pose estimation system is shown in Fig. 1. A periodic pattern is mounted on a 6-axis nanopositioning stage, which is positioned beneath a fixed camera equipped with a low-distortion industrial lens. As the stage moves, the frequency and phase of the pattern in the captured images change accordingly. The estimation of 6DoF poses from the images is based on mathematical models of the camera and the pattern, as detailed in this section.

3.1. Camera Model

The standard pinhole camera model is used. Assuming the pattern plane coincides with the $x_w y_w$ plane of the world coordinate frame ($z_w = 0$), the projection model relating world coordinates (x_w, y_w) to image coordinates (x_i, y_i) is

$$s \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \mathbf{t}] \begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} \quad (1)$$

where f is the focal length, s is a scale factor, $\mathbf{r}_1, \mathbf{r}_2$ are the first two columns of the rotation matrix \mathbf{R} (parameterized by angles α, β, γ), and $\mathbf{t} = [t_x, t_y, t_z]^T$ is the translation vector. The full expansion of \mathbf{R} and the initial pinhole equation are standard and are provided in the Supplementary Material (Sec. A.1).

3.2. Pattern Model

Although sinusoidal patterns with no harmonics are ideal, checkerboard or grid patterns (Fig. 1) are often preferred because they are easier to fabricate. These patterns can be mathematically represented as superpositions of 2D sinusoidal components using Fourier series. For instance, a checkerboard pattern is composed of two orthogonal 2D sinusoids at the fundamental frequency and their corresponding harmonics. Detailed Fourier expansions of the checkerboard and grid patterns are described in the Supplementary Material (Sec. A.2). To make the derivation of equations for pose estimation easier, we model the checkerboard pattern as a 2D cosine pattern composed of two non-collinear fundamental components:

$$g(x_w, y_w) = \cos\left(2\pi\frac{x_w}{T_x} + 2\pi\frac{y_w}{T_y}\right) + \cos\left(2\pi\frac{x_w}{T_x} - 2\pi\frac{y_w}{T_y}\right) \quad (2)$$

where T_x, T_y are the periods in the world frame.

4. Pose Estimation Principle

This section outlines the principle for 6DoF pose estimation. First, we establish the analytical relationship between the pattern's frequency-domain parameters (observed frequency ω and phase φ in the image) and its 6DoF pose. Second, we describe our 2D-IpDFT algorithm for accurately extracting these parameters from captured images.

4.1. Pose Estimation from Image Parameters

Under the quasi-orthographic projection assumption (where object distance t_z is much larger than the image dimensions, as is typical in micro/nano metrology), the relationship between the world coordinates (x_w, y_w) and scaled image coordinates $(x_i/f, y_i/f)$ can be expressed as follows (see Sec. B.1 of the Supplementary Material for details of the derivation):

$$\frac{x_i}{f} = \frac{x_w r_{11} + y_w r_{12} + t_x}{t_z} \quad (3)$$

$$\frac{y_i}{f} = \frac{x_w r_{21} + y_w r_{22} + t_y}{t_z}$$

where $r_{jk}(j, k = 1, 2)$ are elements of the rotation matrix \mathbf{R} . Substituting Eq. (3) into Eq. (2), we obtain $g(x_i, y_i)$, the pattern observed in the image frame, as the sum of two 2D sinusoids with angular frequencies $(\omega_{x_1}, \omega_{y_1}), (\omega_{x_2}, -\omega_{y_2})$ and phases φ_1, φ_2 . That is

$$g(x_i, y_i) = \cos(\omega_{x_1}x_i + \omega_{y_1}y_i + \varphi_1) + \cos(\omega_{x_2}x_i - \omega_{y_2}y_i + \varphi_2) \quad (4)$$

The parameters $\omega_{x_1}, \omega_{y_1}, \omega_{x_2}, \omega_{y_2}, \varphi_1, \varphi_2$ are functions of the 6DoF pose parameters $(\alpha, \beta, \gamma, t_x, t_y, t_z)$ and the pattern periods T_x, T_y . Their detailed expressions are provided

in the Supplementary Material (Eqs. (S12a)-(S12f) in Sec. B.2). From these relationships, the 6DoF pose can be determined. The in-plane rotation angle α is

$$\tan \alpha = (\omega_{x_2} - \omega_{x_1})/(\omega_{y_1} + \omega_{y_2}) \quad (5)$$

Let $\eta = 4\pi t_z/f$, which can be found by solving a quadratic equation $\rho_1 \eta^4 - \rho_2 \eta^2 + 1 = 0$ (see Sec. B.3 of the Supplementary Material for expressions of the coefficients ρ_1, ρ_2). Then the out-of-plane rotation angles β and γ can be determined by the following equations:

$$\cos \beta = \frac{\eta}{T_x(\omega_{x_1} + \omega_{x_2}) \cos \alpha - T_x(\omega_{y_2} - \omega_{y_1}) \sin \alpha}$$

$$\cos \gamma = \frac{\eta \cos \alpha}{T_y(\omega_{y_1} + \omega_{y_2})} \quad (6)$$

The sign ambiguities for β, γ can be resolved as in [2]. The translations t_x, t_y, t_z are

$$t_x = -\frac{T_x(\varphi_1 + \varphi_2) \cos \alpha \cos \beta + T_y(\varphi_1 - \varphi_2) r'_{12}}{4\pi}$$

$$t_y = -\frac{T_x(\varphi_1 + \varphi_2) \sin \alpha \cos \beta + T_y(\varphi_1 - \varphi_2) r'_{22}}{4\pi} \quad (7)$$

$$t_z = \frac{\eta f}{4\pi}$$

where $r'_{12} = \cos \alpha \sin \beta \sin \gamma - \sin \alpha \cos \gamma$ and $r'_{22} = \sin \alpha \sin \beta \sin \gamma + \cos \alpha \cos \gamma$.

The above principle can be extended to other periodic patterns by extracting their fundamental frequency components.

4.2. Image Parameter Estimation

To estimate the frequency (λ, μ) and phase (φ) of the pattern from a captured image of $M \times N$ pixels, we apply the proposed 2D-IpDFT method. The sampled and windowed image signal $g(u, v)$ containing two principal components can be expressed as

$$g(u, v) = w(u, v) \left[\cos\left(2\pi\frac{\lambda_1}{M}u + 2\pi\frac{\mu_1}{N}v + \varphi_1\right) + \cos\left(2\pi\frac{\lambda_2}{M}u - 2\pi\frac{\mu_2}{N}v + \varphi_2\right) \right] \quad (8)$$

where $w(u, v)$ is a 2D maximum sidelobe decay window function [15] (see details in Supplementary Material Sec. C.1). The DFT of $g(u, v)$, denoted $G(k, l)$, is a superposition of the DFTs of the window function centered at the component frequencies (see the full expression in Supplementary Material Sec. C.2, Eq. S22). By evaluating $G(k, l)$ at DFT bins $k-1, k, k+1$ around a spectral peak (similarly for l), we formulate a system of linear equations $\mathbf{A}\mathbf{d} = \mathbf{b}$ (see details in Supplementary Material Sec. C.3). Requiring $\det(\mathbf{A}) = 0$ for a non-trivial solution allows solving for,

e.g., λ_1^2 , as follows:

$$\lambda_1^2 = \frac{\begin{vmatrix} (2H-1)H & (2H-1) & G(k-1,l)-G(k,l) \\ -H^2-k^2 & 2k & G(k,l) \\ (2H-1)H & -(2H-1) & G(k+1,l)-G(k,l) \end{vmatrix}}{\begin{vmatrix} (2H-1) & G(k-1,l)-G(k,l) \\ -(2H-1) & G(k+1,l)-G(k,l) \end{vmatrix}} \quad (9)$$

where H is the window order. Parameters μ_1, λ_2, μ_2 are found analogously. Phases φ_1, φ_2 are then obtained as

$$\varphi_1 = \text{angle} \left\{ \frac{G(\lambda_1, \mu_1)}{W(0,0)} \right\}, \varphi_2 = \text{angle} \left\{ \frac{G(\lambda_2, -\mu_2)}{W(0,0)} \right\} \quad (10)$$

Using DFT values around the spectral peaks with the highest amplitudes mitigates noise.

5. Performance Analysis

To evaluate the proposed pose estimation algorithm, we performed a series of simulations. The proposed algorithm was first evaluated against other algorithms in terms of pose estimation accuracy under varying signal-to-noise ratios (SNRs), pattern pitches, and degrees of defocus. Subsequently, its computational efficiency was compared.

The algorithms selected for comparison include QSE (a frequency and phase estimator) [20], linear regression [2], zero-padding with quadrupled signal length, and zero-padding with octupled signal length. For the proposed algorithm, Hanning window was employed. The accuracy of the estimation was evaluated using the root mean square error (RMSE).

5.1. Influence of Signal-to-Noise Ratio

Simulations were conducted assuming the use of a camera with a resolution of 640×480 pixels and a pixel size of $9.9 \mu\text{m}$. The pattern pitch was set to $450 \mu\text{m}$ and the focal length of the lens f to 28 mm. SNR was varied from 10 to 50 dB in increments of 10 dB. The in-plane displacements t_x and t_y were randomly selected within one pitch. The out-of-plane displacement t_z was randomly chosen within the range of $0.99f$ to $1.01f$ (approximately $560 \mu\text{m}$ variation) to simulate a realistic depth of field. This range reflects typical optical hardware setups for high-precision applications and is not a constraint. Our method works in a much wider range of t_z . The in-plane rotation angle α was randomly selected from $[0, 2\pi]$, whereas the out-of-plane rotation angles β and γ were randomly chosen from $[0, \pi/8]$. Under these conditions, the captured images of the pattern contain approximately ten periods.

A total of 1,000 independent simulations were conducted. The RMSEs for $t_x, t_y, t_z, \alpha, \beta,$ and γ are denoted as $e_{t_x}, e_{t_y}, e_{t_z}, e_\alpha, e_\beta,$ and e_γ , respectively. The results are shown in Fig. 2. The proposed algorithm achieves progressively higher accuracy with increasing SNR, consistently outperforming other methods at all SNR levels. We

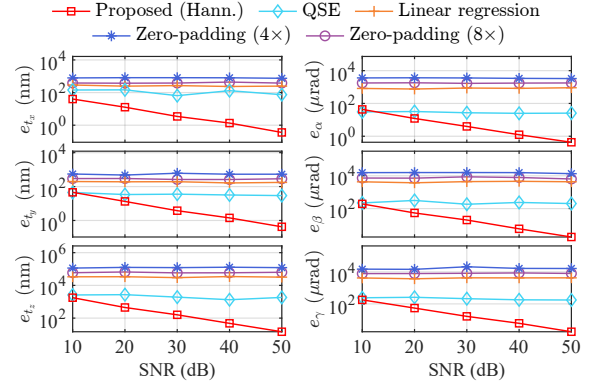


Figure 2. Estimation error versus SNR.

also observe from the results that the errors of QSE, linear regression, and zero-padding methods are almost insensitive to SNR variations. This is because the dominant error source of these methods is spectral leakage, which is non-negligible in this case.

5.2. Influence of Number of Pattern Periods

In this simulation, we evaluated the performance of the algorithms as we varied the number of pattern periods in the image, a parameter that directly affects spectral leakage. We changed the number of periods along the y -axis from 6 to 30 in increments of 6, which corresponds to pattern pitches from approximately $750 \mu\text{m}$ down to $150 \mu\text{m}$. The number of x -axis pattern periods scaled proportionally to the image aspect ratio. The SNR was fixed at 40 dB and the remaining parameters matching those in Sec. 5.1.

The resulting RMSE values are presented in Fig. 3. As the number of pattern periods decreases (i.e., pattern pitch increases), the accuracy of all methods degrades. This is attributed to the reduced number of cycles within the analysis window, leading to more closely spaced spectral lines and consequently, increased spectral leakage. Across all tested numbers of periods, the proposed 2D-IpDFT algorithm consistently demonstrates superior performance compared to the baseline methods, highlighting its robustness even when fewer pattern periods are available.

5.3. Influence of Defocus Blur

To evaluate the impact of image defocus, varying levels of Gaussian blur were added to the images by convolving them with Gaussian kernels of increasing standard deviation σ from 2 to 10 in steps of 1 [24]. SNR was fixed at 40 dB and all other parameters were identical to those in Sec. 5.1. Figure 4 illustrates the estimation errors as functions of the Gaussian blur level. Again, the proposed algorithm shows the best performance. The estimation errors of the proposed algorithm increase with σ because image blur reduces image contrast or the amplitude of the fundamental frequency component. For all other algorithms, the primary

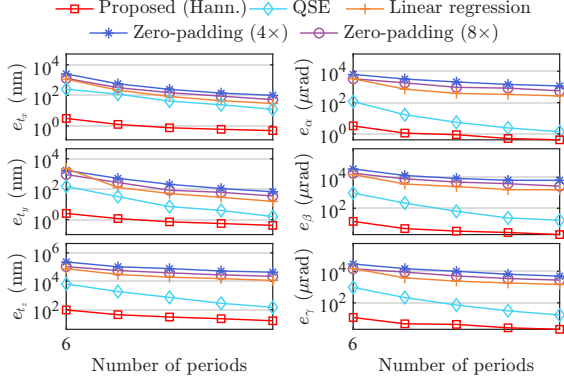


Figure 3. Estimation error versus number of periods.

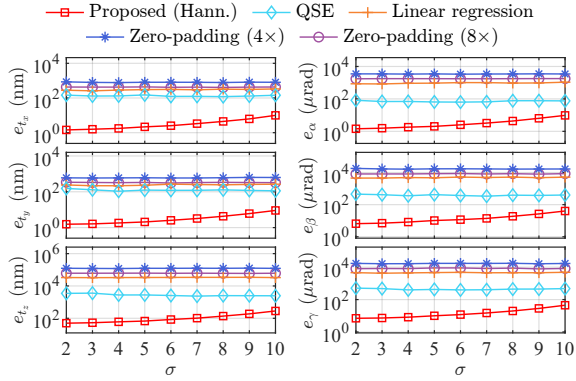


Figure 4. Estimation error versus level of blur.

Table 1. Average computational times. (Unit: ms)

Algorithm	640 × 480	1600 × 1200
Proposed (Hann.)	12.4	77.5
QSE	296	1546
Linear regression	89.5	502
Zero-padding (4×)	90.6	534
Zero-padding (8×)	332	2059

error source is still spectral leakage, rendering their performance less sensitive to image blur.

5.4. Computational Efficiency

To evaluate computational efficiency, we compared the processing times of the proposed algorithm and benchmark methods. All algorithms were implemented in MATLAB R2023a and executed on a desktop computer with a 4.5 GHz CPU and 32 GB RAM. Average computational times for image resolutions of 640 × 480 and 1600 × 1200 are summarized in Tab. 1. The proposed algorithm demonstrates significantly lower computational times compared to QSE, linear regression, and zero-padding methods for both image resolutions, indicating its superior computational efficiency.

6. Experiment

To evaluate the effectiveness of the proposed pose estimation method in real-world scenarios, a series of experiments were carried out using the experimental setup shown in Fig. 5, which includes a black-and-white industrial camera (Balluff mvBlueCOUGAR-X120bG/C) and a checkerboard pattern affixed to a 6-axis nanopositioning stage (Physik Instrumente P-562.6CD). The camera has a pixel resolution of 640 × 480, a pixel size of 9.9 μm, and a low-distortion lens (Opto Engineering MC100X). The checkerboard pattern has a pitch of 150, 450, or 750 μm. The stage has a linear displacement resolution of 1 nm and an angular resolution of 0.1 μrad. The motions provided by the stage were considered the ground truth in the experiments. The performance of the proposed and benchmark methods was evaluated under various motion trajectories, exposure times, pattern pitches, and defocus blur levels.

6.1. Trajectory Tracking Performance

This subsection evaluates the trajectory tracking performance of the proposed algorithm by analyzing the RMSE of the estimated pose across multiple motion profiles. Using a 450 μm-pitch pattern and 30 ms exposure time, we first assessed single-axis motion estimation accuracy for each degree of freedom. The stage was displaced by 100 μm along each translational axis and rotated by 300 μrad around each rotational axis over 200 steps.

Figure 6 shows the estimation results for single-axis motions using the proposed algorithm with a Hanning window. All results align well with the ground truth, especially for in-plane translations and rotations. Table 2 presents the estimation errors of all algorithms. It is evident that our proposed algorithm achieves significantly higher accuracy than other algorithms.

For 2D trajectory tracking, we generated an “ICCV” path with a length of 100 μm. Figure 7 shows the estimated trajectories obtained using the proposed algorithm. The corresponding RMSE values were 5 nm and 7 nm for t_x and t_y respectively. These values are significantly smaller than those of other algorithms, as can be seen in Tab. 3.

Finally, we generated a helical trajectory to evaluate the 3D trajectory tracking performance. Figure 8 displays the estimated helical trajectory using the proposed algorithm,

Table 2. Comparison of RMSE values for single-axis motion estimation. (Units: nm for translation and μrad for rotation)

Algorithm	t_x	t_y	t_z	R_z	R_y	R_x
Proposed (Hann.)	21	27	244	8	15	15
QSE	236	409	2101	86	528	1716
Linear regression	493	680	30906	497	3130	1167
Zero-padding (4×)	152	430	29011	-	-	-
Zero-padding (8×)	484	494	49369	-	-	-

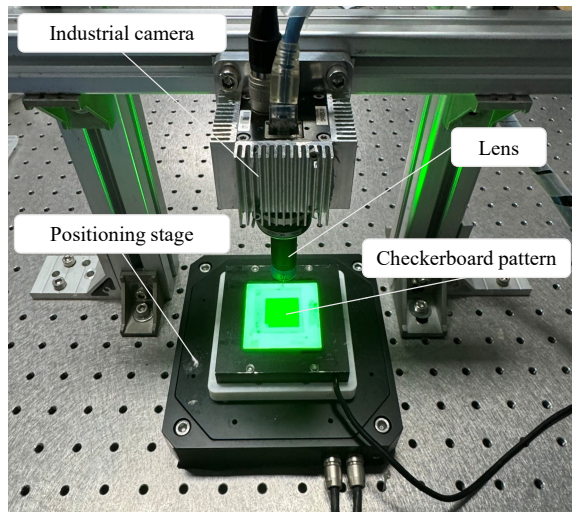


Figure 5. Schematic of the experimental setup.

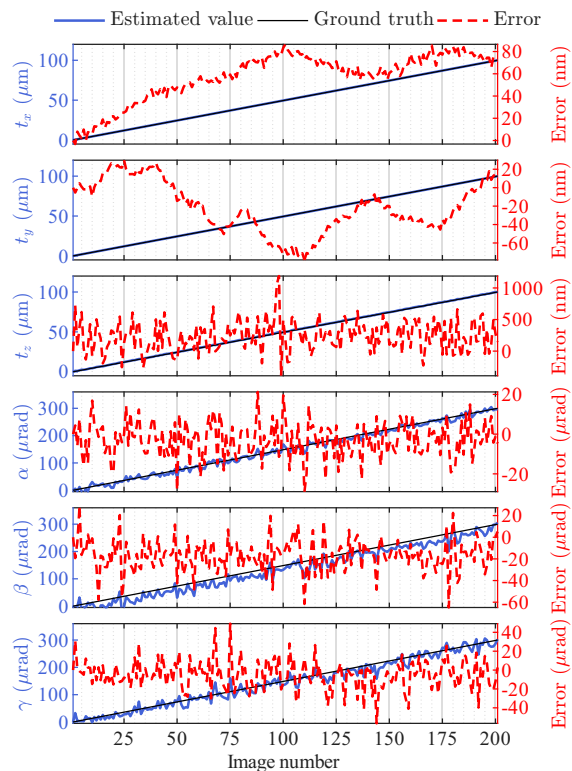


Figure 6. Results of single-axis motion estimation using the proposed algorithm.

which shows close alignment with the ground truth. The RMSE values for all algorithms are detailed in Tab. 4. The proposed algorithm again outperforms all other algorithms by a significant margin.

6.2. Influence of Exposure Time

To assess the impact of exposure time on estimation accuracy, experiments were conducted using three different exposure times: 10, 20, and 30 ms. A circular trajectory

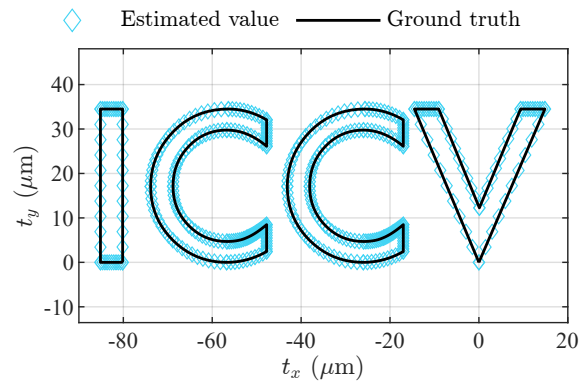


Figure 7. Result of “ICCV” trajectory estimation using the proposed algorithm.

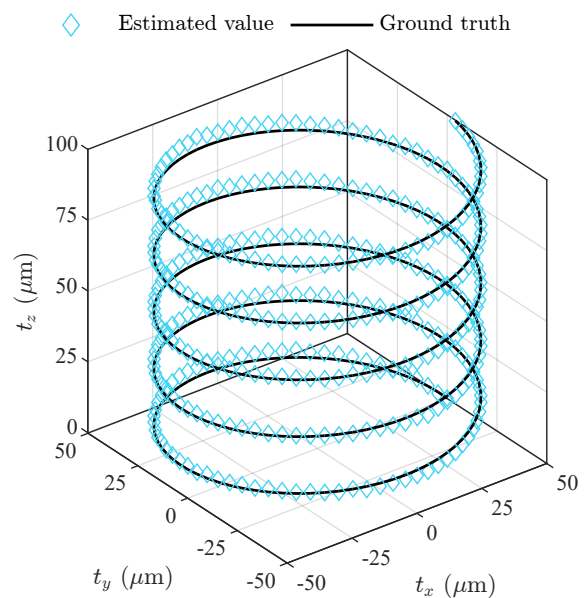


Figure 8. Result of helical trajectory estimation using the proposed algorithm.

Table 3. Comparison of RMSE values for “ICCV” trajectory estimation. (Unit: nm)

Algorithm	t_x	t_y
Proposed (Hann.)	5	7
QSE	201	465
Linear regression	354	577
Zero-padding (4×)	168	538
Zero-padding (8×)	132	430

with a diameter of 100 μm was used. Table 5 summarizes the errors of trajectory estimation. The proposed algorithm achieves the highest accuracy for all exposure times and shows a slight but continuous improvement as the exposure time increases. For all other methods, the errors increase slightly as the exposure time increases.

Table 4. Comparison of RMSE values for helical trajectory estimation. (Unit: nm)

Algorithm	t_x	t_y	t_z
Proposed (Hann.)	34	37	1664
QSE	337	950	16037
Linear regression	539	898	13634
Zero-padding (4×)	293	1150	28948
Zero-padding (8×)	719	1092	41662

Table 5. Comparison of RMSE values for circular trajectory estimation under varying exposure times. (Unit: nm)

Algorithm	10 ms	20 ms	30 ms
Proposed (Hann.)	38	37	35
QSE	182	184	185
Linear regression	582	585	587
Zero-padding (4×)	461	466	468
Zero-padding (8×)	273	276	278

Table 6. Comparison of RMSE values for circular trajectory estimation under varying number of pattern periods.

Algorithm	32	10.7	6.4
Proposed (Hann.)	8	37	57
QSE	16	583	467
Linear regression	17	758	20895
Zero-padding (4×)	84	531	859
Zero-padding (8×)	76	754	618

Table 7. Comparison of RMSE values for circular trajectory estimation under varying levels of defocus blur. (Unit: nm)

Algorithm	Light blur	Heavy blur
Proposed (Hann.)	197	348
QSE	437	1051
Linear regression	342	5668
Zero-padding (4×)	653	1037
Zero-padding (8×)	366	1037

6.3. Influence of Number of Pattern Periods

To experimentally evaluate the influence of the number of pattern periods on performance, we used checkerboard patterns with pitches of 150, 450, and 750 μm . Given our camera’s field of view, these pitches correspond to approximately 32, 10.7, and 6.4 periods along the y -axis, respectively. The exposure time was fixed at 30 ms. Table 6 presents the RMSE values for estimating the same circular trajectory described in Sec. 6.2, with varying numbers of pattern periods in the image. Consistent with simulation results, all methods exhibit improved accuracy as the number of pattern periods increases. The proposed algorithm achieves significantly lower errors than the benchmark algorithms.

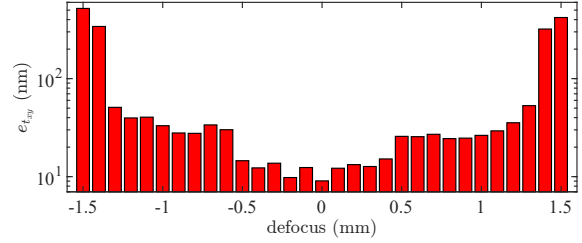


Figure 9. Errors of in-plane trajectory estimation using the proposed method at different defocus distances.

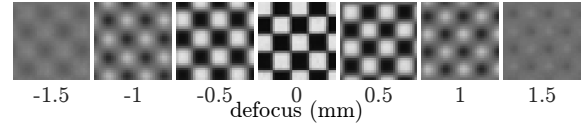


Figure 10. Checkerboard pattern images captured at different defocus distances.

6.4. Influence of Defocus Blur

The impact of defocus blur was investigated by positioning the checkerboard pattern outside the camera’s depth of field. The experiments used a 30 ms exposure time and a 450 μm pattern pitch, with the same circular trajectory as in Sec. 6.2. As shown in Tab. 7, estimation accuracy decreases with increasing defocus blur, consistent with simulation results. This degradation stems from reduced image contrast due to defocus blur, which lowers the SNR. Compared to all benchmark algorithms, the proposed method maintains the highest accuracy under both light and heavy blur conditions.

6.5. Operational Range of Defocus

To determine the operational range of defocus for our proposed method, we translated the checkerboard pattern (150 μm pitch) axially in 0.1 mm increments from the in-focus position to ± 1.5 mm, which is significantly beyond the camera’s depth of field. At each of the 31 axial positions, the pattern was moved along the same circular trajectory as in Sec. 6.2 and the RMSE of in-plane tracking $e_{t_{xy}}$ was evaluated. The camera exposure time was set to 30 ms.

Figure 9 plots $e_{t_{xy}}$ versus defocus distance, while Figure 10 shows example images at selected defocus distances, illustrating different levels of blur. As expected, accuracy degrades with defocus due to decreased contrast and pattern blurring, which reduces the SNR of fundamental frequency components. However, our method still maintains reasonably good accuracy, with $e_{t_{xy}}$ below 500 nm in the full defocus range of ± 1.5 mm and around 10 nm within ± 0.4 mm. These results demonstrate our method’s robustness to axial displacements, validating utility in applications where focal control is challenging.

6.6. Vision-Based Position Control

An additional experiment was conducted to demonstrate vision-based closed-loop feedback control of motion. The

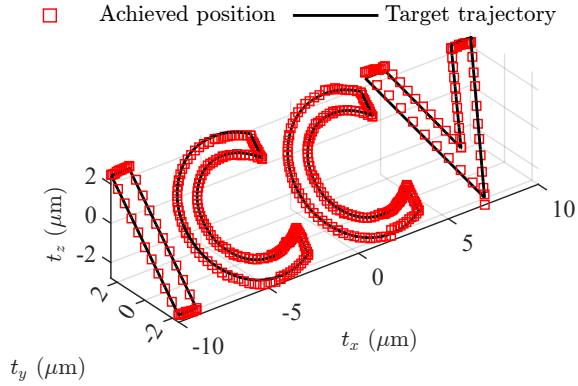


Figure 11. Result of vision-based feedback control for tracking a 3D “ICCV” trajectory.

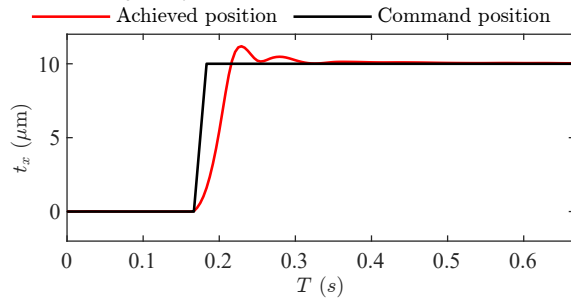


Figure 12. Step response of the vision-based feedback control system.

goal was to utilize the high-accuracy pose estimated by our algorithm to guide the nanopositioning stage to follow a predefined trajectory.

The stage operated in open-loop mode for all axes, while our algorithm provided real-time checkerboard pose feedback. A proportional (P) controller then generated corrective commands at 60 Hz to minimize pose error.

First, we evaluated tracking of a scaled/rotated 3D “ICCV” trajectory (see Sec. 6.1). Figure 11 compares the target trajectory (black line) and achieved positions (red squares). The RMSE values for this trajectory tracking are 6.0 nm (t_x), 8.9 nm (t_y), and 108.6 nm (t_z). Next, we characterized the dynamic response via a 10 μm step input (Figure 12). The system converged with an approximate overshoot of 11% and a settling time of 0.25 s.

These results validate our proposed method as both a high-accuracy pose estimator and a robust feedback source for closed-loop pose control and demonstrate its potential for real-time applications such as robotic microassembly and optical alignment.

7. Conclusion

This paper presented a novel 2D-IpDFT algorithm designed to address key challenges in vision-based 6DoF pose estimation. By incorporating advanced windowing techniques and negative frequency compensation strategies, the proposed method effectively mitigates spectral leakage and

minimizes estimation bias. As a result, it demonstrates robust performance across a wide range of conditions, including variations in SNR, pattern periods, and defocus blur, consistently outperforming existing approaches. This paper also introduced a mathematical framework that establishes a direct mapping between the image spectral parameters and the 6DoF pose, applicable to both orthogonal and quasi-orthogonal imaging configurations. Extensive simulation and experimental results demonstrate pose estimation accuracy at the nanometer level in translations and microradian level in rotations. This work provides a novel solution for ultra-precision 6DoF pose estimation, with significant potential applications in precision engineering, nanoscale robotics, and advanced manufacturing.

References

- [1] Antoine N André, Patrick Sandoz, Benjamin Mauzé, Maxime Jacquot, and Guillaume J Laurent. Sensing one nanometer over ten centimeters: A microencoded target for visual in-plane position measurement. *IEEE/ASME Transactions on Mechatronics*, 25(3):1193–1201, 2020. 1, 2
- [2] Antoine N André, Patrick Sandoz, Maxime Jacquot, and Guillaume J Laurent. Pose measurement at small scale by spectral analysis of periodic patterns. *International Journal of Computer Vision*, 130(6):1566–1582, 2022. 1, 2, 3, 4
- [3] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017. 2
- [4] Zonghao Chen and Peisen S Huang. A vision-based method for planar position measurement. *Measurement Science and Technology*, 27(12):125018, 2016. 1
- [5] Luis Ferraz, Xavier Binefa, and Francesc Moreno-Noguer. Very fast solution to the PnP problem with algebraic outlier rejection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–508, 2014. 1
- [6] Hassan Foroosh, Josiane B Zerubia, and Marc Berthod. Extension of phase correlation to subpixel registration. *IEEE transactions on image processing*, 11(3):188–200, 2002. 1
- [7] Valerian Guelpa, Patrick Sandoz, Miguel Asmad Vergara, Cédric Clévy, Nadine Le Fort-Piat, and Guillaume J Laurent. 2D visual micro-position measurement based on intertwined twin-scale patterns. *Sensors and Actuators A: Physical*, 248: 272–280, 2016. 1
- [8] Yang Hai, Rui Song, Jiaojiao Li, David Ferstl, and Yinlin Hu. Pseudo flow consistency for self-supervised 6D object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14075–14085, 2023. 2
- [9] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 1, 2
- [10] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate o (n) solution to the PnP problem. *In-*

- ternational journal of computer vision*, 81:155–166, 2009. 1
- [11] Hai Li, Xianmin Zhang, Benliang Zhu, and Sergej Fatikow. Online precise motion measurement of 3-DOF nanopositioners based on image correlation. *IEEE Transactions on Instrumentation and Measurement*, 68(3):782–790, 2018. 1, 2
- [12] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 2
- [13] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1, 2
- [14] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Genflow: Generalizable recurrent flow for 6D pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024. 2
- [15] Albert Nuttall. Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(1):84–91, 1981. 3
- [16] Bing Pan, Kemao Qian, Huimin Xie, and Anand Asundi. Two-dimensional digital image correlation for in-plane displacement and strain measurement: a review. *Measurement science and technology*, 20(6):062001, 2009. 1, 2
- [17] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 1, 2
- [18] Guowei Shi, Zian Mao, Jiayue Ding, and Peisen Huang. 2-D interpolated discrete Fourier transform for high-accuracy in-plane displacement and rotation measurement. In *2024 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6. IEEE, 2024. 1, 2
- [19] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international conference on computer vision*, pages 118–126, 2015. 2
- [20] Veyis Solak, Sultan Aldirmaz-Colak, and Ahmet Serbes. Fast and efficient 2-D and K-D DFT-based sinusoidal frequency estimation. *IEEE Transactions on Signal Processing*, 70:5087–5101, 2022. 4
- [21] Zongliang Wu, Chengshuai Yang, Xiongfei Su, and Xin Yuan. Adaptive deep PnP algorithm for video snapshot compressive imaging. *International Journal of Computer Vision*, 131(7):1662–1679, 2023. 1
- [22] Kaihua Zhang, Lei Zhang, Kin-Man Lam, and David Zhang. A level set approach to image segmentation with intensity inhomogeneity. *IEEE transactions on cybernetics*, 46(2):546–557, 2015. 2
- [23] Linfang Zheng, Tze Ho Elden Tse, Chen Wang, Yinghan Sun, Hua Chen, Ales Leonardis, Wei Zhang, and Hyung Jin Chang. GeoReF: Geometric alignment across shape variation for category-level object pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10693–10703, 2024. 2
- [24] Xiang Zhu, Scott Cohen, Stephen Schiller, and Peyman Milanfar. Estimating spatially varying defocus blur from a single image. *IEEE Transactions on image processing*, 22(12):4879–4891, 2013. 4
- [25] Zijian Zhu, Chenyang Zhao, and Yueping Xi. Micro-vision super-resolution restoration and positioning based on ultra-precision machining topography guidance. *IEEE Transactions on Instrumentation and Measurement*, 2024. 1, 2