International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 15, Issue 9 [September 2026] PP: 89-100

A Wafer-Level Alignment for 3D IC Hybrid Bonding Based on Quality-Aware Observation and Adaptive Full-Field Robust Estimation

Qiyue Yan, Zheng Kang, Daikaidi Fu
School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, Dongguan, China
Corresponding Author: Qiyue Yan

ABSTRACT: *Wafer-level hybrid bonding has become a key path for continued scaling of 3D integrated circuits, yet its yield is governed by the post-bond full-field residual overlay vector field rather than by any single scalar "alignment accuracy". Existing alignment algorithms suffer from three structural limitations. First, mark-level displacement detection relies on a fixed detector and does not model mark image quality, so accuracy collapses when infrared (IR) transmission degrades, defocus increases, or marks deteriorate. Second, full-field estimation relies on low-order rigid or polynomial transforms that cannot express the spatially correlated local residuals induced by chuck rebound, film stress, and bond-wave propagation. Third, multi-mark adjustment lacks explicit handling of gross outliers (contamination, mismatch) and of point-wise heteroscedastic uncertainty. This paper proposes BW-STAR, a framework of three coupled modules: (1) a quality-aware mark-level observation layer that runs centroid, normalized cross-correlation (NCC), upsampled phase correlation, and Moiré phase estimation in parallel, regresses each detector's variance from mark quality features, and performs inverse-variance fusion or minimum-variance selection; (2) a hierarchical full-field model using low-order Zernike polynomials as the mean function and a Gaussian process with an isotropic plus bond-wave anisotropic kernel as the non-parametric residual, with cross-validated kernel amplitude adaptation; and (3) robust inference alternating inverse-variance weighting and Huber M-estimation. Under physics-calibrated Monte Carlo simulation (300 mm wafer, 100 nm/pixel imaging, 41 sampled marks, 6% gross outliers), BW-STAR reduces mark-level detection 3σ from 246.3 nm (centroid) to 8.8 nm under typical production conditions; in full-field estimation under strong distortion it reduces wafer-scale residual 3σ from 96.7 nm (rigid model) to 47.0 nm and M+3S from 172.2 nm to 92.8 nm. Ablation shows the low-order mean function and robust M-estimation are essential components (their removal degrades 3σ by 64.4 nm and 16.9 nm respectively). We also report a negative result: across 41–197 sampled marks, a fixed-amplitude bond-wave anisotropic kernel yields no statistically significant gain (−2.1 to −32.2 nm), indicating that the bond-wave front structure is not identifiable under current metrology configurations; the amplitude adaptation mechanism therefore degenerates it to zero so that the algorithm never degrades because of a wrong prior. This provides quantitative support for prioritizing process sensing over deeper networks.*

Keywords: *3D integrated circuits; wafer bonding; hybrid bonding; overlay alignment; phase correlation; Moiré fringes; Gaussian process; robust estimation; identifiability*

Date of Submission: 02-09-2026 Date of acceptance: 12-09-2026

NOMENCLATURE

Symbol	Description	Unit
c	Mark centre position	pixel
d(x)	Process-dependent residual displacement field	nm
g(x)	Non-parametric Gaussian-process residual field	nm
h(x;θ)	Parametric low-order mean function	nm
I(x)	Mark intensity profile along the sensitive axis	a.u.
k	Image scale (pixel calibration factor)	nm/pixel

Symbol	Description	Unit
K, K*	Gaussian-process covariance matrices	nm²
ℓ, ℓu, ℓv	Correlation length scales (isotropic / across / along front)	mm
M	Moiré magnification factor, |p1 /(p1 − p2)|	—
n	Number of observation marks	—
N	Number of full-field grid points	—
O(x)	Overlay (residual displacement) vector field	nm
p, p1, p2	Grating pitches of the two marks	μm
r	Radial coordinate on the wafer	mm
R	Wafer radius	mm
T(x)	Global geometric transformation	nm
u, v	Bond-wave propagation coordinates (along / across)	mm
wi	Quality-aware and robust weight of mark i	—
x, xB	Wafer coordinate / bottom-wafer design coordinate	mm
yi	Observed displacement at mark i	nm

Greek letters

Symbol	Description	Unit
α, ψ	Sidewall-asymmetry amplitude and phase	—, rad
β	Defocus blur kernel width	pixel
ε	Measurement (detection + metrology) noise	nm
θ	Coefficient vector of the low-order mean function	—
λ	Gaussian-process noise variance scale	nm²
ρ(·)	Huber loss function	—
σi, σn	Observation / detector noise standard deviation	nm
σf, σw	Isotropic / bond-wave kernel amplitudes	nm
φ	Mark quality feature vector	—

Subscripts

Symbol	Description	Unit
A, B	Top wafer / bottom wafer	—
i, k	Mark index / detector index	—

I. INTRODUCTION

As planar scaling approaches its physical and economic limits, three-dimensional integration represented by through-silicon vias (TSV) and Cu/SiO_2 hybrid bonding has become the main route for sustaining system-level performance growth. By forming metal–metal and dielectric–dielectric direct connections simultaneously, hybrid bonding pushes interconnect pitch from tens of micrometres for micro-bumps down to the sub-micrometre regime,

improving interconnect density by one to two orders of magnitude. Publicly available information indicates that Intel Foveros Direct adopts copper via pitches of 9 μm and 3 μm in its first and second generations, while published equipment-level wafer-to-wafer (W2W) alignment capability has entered the 50–100 nm range.

Continuous pitch shrinkage, however, transforms a problem that once belonged to "precision mechanical positioning" into one of "full-field error-field estimation and process control". Once interconnect pitch falls below a micrometre, overlay error can no longer be characterized by a single Gaussian scalar: global wafer scaling, bow, film stress, electrostatic chucking and release rebound, and the dynamic displacement caused by bond-wave propagation jointly form a strongly spatially correlated two-dimensional vector field. What actually determines yield is the quantile of the residual vector over the whole wafer and all connection points after bonding (e.g., M+3S or the 99.5% quantile), not the pre-bond alignment repeatability quoted by equipment vendors.

Existing alignment algorithms expose three structural weaknesses under this transformation. First, the mark-level detection stage typically uses a single detector (template matching, phase correlation, or Moiré phase), whose accuracy implicitly depends on mark image SNR, contrast, and periodicity; once IR transmission drops because of substrate thickness, doping, or metal shielding, or once marks degrade through chemical-mechanical polishing (CMP) dishing and erosion, detection error can degrade rapidly from a few nanometres to several hundred nanometres. Second, the full-field modeling stage relies on rigid or low-order polynomial transforms; such parametric models cannot express spatially correlated local residuals, while raising the polynomial order amplifies sampling noise and outliers. Third, multi-mark adjustment usually assumes homoscedastic Gaussian noise and lacks explicit treatment of gross outliers from contamination and mismatch, and of point-wise heteroscedastic uncertainty.

To address these issues, this paper proposes BW-STAR, designed around three principles: (i) the observation layer must output uncertainty rather than coordinates alone; (ii) priors must enter the model in a degradable way, so that a wrong prior never degrades performance; (iii) robustness must be guaranteed by the inference mechanism rather than by manual outlier rejection. The main contributions are as follows:

(1) A quality-aware mark-level displacement observation method. By regressing the point-wise variance of four heterogeneous detectors (centroid, NCC, upsampled phase correlation, Moiré phase) on mark quality features, we achieve both inverse-variance fusion and minimum-variance selection, and characterize the regime in which each is preferable.

(2) A hierarchical full-field model combining a low-order Zernike mean with an isotropic plus bond-wave anisotropic Gaussian-process residual, explicitly separating interpretable low-order compensable terms from non-parameterizable process residuals, with cross-validated adaptation of the physical kernel amplitude.

(3) A robust inference algorithm alternating inverse-variance weighting and Huber M-estimation, keeping full-field estimation stable under 6% gross-outlier contamination.

(4) A systematic Monte Carlo characterization of the identifiability limit of the bond-wave physical kernel: across 41–197 sampled marks, the kernel brings no significant gain. This negative result gives quantitative support for investing in process sensing rather than in deeper networks.

II. RELATED WORK AND PROBLEM STATEMENT

A. Physical Composition of Overlay Error

Let xB denote the design coordinate on the bottom wafer (wafer B) and xA the corresponding physical coordinate on the top wafer (wafer A). The overlay vector field can be written as

$$O(x) = T(x) - xB + d(x) + \varepsilon$$

where T contains translation, rotation, isotropic/anisotropic scaling, shear, and higher-order deformation; d(x) is the process-dependent residual determined jointly by CMP, film stress, electrostatic chucking, thermal history, and bond-wave propagation; and ε is mark detection, optical aberration, and metrology noise. The core problem addressed here is: given noisy observations {yi} at n sparse mark locations, estimate the full field O(x) and generate the corresponding compensation.

It must be emphasized that T and d are not fully identifiable mathematically: separating them depends on priors (e.g., low-order terms belong to T, high-frequency local terms to d). This non-identifiability is precisely the motivation for the hierarchical model adopted here and for explicitly testing the value of the physical kernel.

B. Mark-Level Detection

Classical template matching and normalized cross-correlation (NCC) are robust to linear brightness drift and repetitive texture, but optical differences between the template and the actual image introduce systematic bias. Phase correlation (PC) estimates translation through cross-power-spectrum normalization and, combined with neighbourhood matrix-multiply DFT upsampling, achieves sub-pixel localization at low computational cost, yet it is sensitive to defocus, scattering, and local contrast inversion. Moiré methods based on difference-frequency amplification magnify displacement by $M = |p_1/(p_1 - p_2)|$ and thus strongly suppress detection noise; experimental work has reported improvement of 3σ from 1.20 μm for cross/box marks to better than 300 nm on 4-inch silicon.

Such methods are, however, highly sensitive to pitch fabrication error, undersampling, and envelope leakage, since the magnification factor amplifies pitch error as well.

This paper does not seek an "optimal single detector". Instead, the outputs of multiple detectors are treated as uncertainty-bearing observations and fused by an upper layer, because the best detector differs across degradation conditions.

C. Full-Field Modeling and Robust Estimation

High-order process control (HOC/CPE) in lithography fits sampled overlay to wafer deformation and applies per-field correction; its "low-dimensional basis plus residual" idea transfers to bonding, but bonding introduces stress, bow, contact sequence, and chuck release as continuous full-field quantities and adds a temporal dimension. Thin-plate splines (TPS) and Gaussian processes (GP) are standard tools for non-parametric residual modeling: TPS suppresses oscillation through bending energy, while GP naturally provides positional uncertainty, facilitating interfacing with finite-element (FEM) displacement fields and online correction maps. In multi-mark adjustment, RANSAC suits catastrophic mismatches that are fully random, but wafer-level errors tend to be spatially clustered, making M-estimators, iterative reweighting, and Student-t formulations more appropriate.

D. Error Budget and Identifiability

Prior studies have shown that after initial central contact in face-to-face bonding, bond-wave propagation occurs and the wafer undergoes significant, directionally varying deformation, producing non-linear overlay residuals whose abrupt changes and high dispersion correlate with the recipe and incoming wafer shape. At the same time, public sources cannot yet separate thermal-expansion mismatch, chuck rebound, CMP mark asymmetry, optical distortion, and bond-wave propagation into unique contributions. This paper does not attempt a unique error decomposition; instead it tests a falsifiable question with more direct engineering value: under a given metrology configuration, is a specific physical structure identifiable at all?

III. THE BW-STAR FRAMEWORK

BW-STAR consists of a mark-level observation layer, a full-field estimation layer, and a robust inference layer; the overall flow is shown in Fig. 1. The inputs are IR image pairs at n mark locations together with process state provided by the tool (bonding start point, recipe propagation direction); the outputs are the posterior mean and covariance of the full-field displacement field and the resulting compensation.

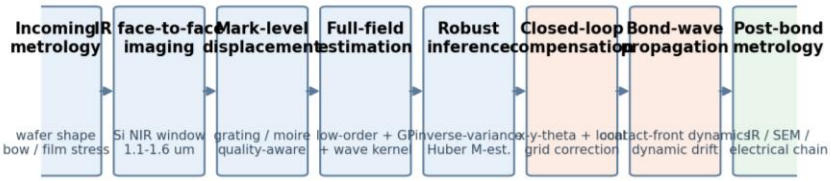

Figure. 1 Closed-loop flow and error propagation chain of 3D IC wafer bonding alignment

A. Quality-Aware Mark-Level Displacement Observation

1) Imaging Model and Heterogeneous Detectors

The equivalent one-dimensional intensity profile along the grating sensitive axis is modeled as

$$I(x) = b + A\cdot[1 + \cos(2\pi(x - c)/p) + \alpha\cdot\cos(4\pi(x - c)/p + \psi)]\cdot env(x - c) + n(x)$$

where c is the mark centre, p the grating pitch, α and ψ characterize the second harmonic introduced by sidewall asymmetry, env the mark window envelope, and n(x) the detection noise. Acquisition further introduces defocus blur (Gaussian kernel width β), illumination non-uniformity (±6% here), and detector noise (standard deviation σn).

Four detectors run in parallel to produce four estimates of the same displacement:

· **Centroid:** ĉC, the background-removed, window-weighted centre of mass; lowest cost, but sensitive to scattering and non-uniform illumination.

· **Normalized cross-correlation:** ĉN, FFT cross-correlation with parabolic peak refinement, with the search restricted to within half a pitch to avoid peak aliasing of periodic marks.

· **Upsampled phase correlation:** ĉP, cross-power-spectrum phase normalization plus neighbourhood matrix-multiply DFT refinement (upsampling factor 200), with band-limited whitening to suppress noise-band amplification.

· **Moiré phase:** ĉM, boxcar period averaging of the two-layer product image to precisely remove the fundamental and sum frequencies, followed by joint least-squares fitting of a difference-frequency sinusoid plus a background polynomial to decouple envelope leakage; displacement is recovered from the fringe phase.

Because of the finite window and envelope, the ratio between Moiré fringe displacement and true displacement is not strictly equal to the theoretical magnification; this work therefore uses two-point calibration at ±250 nm to determine the gain, placing systematic ratio error into calibration rather than into model assumptions.

2) Variance Head and Fusion Strategies

For each mark we extract a quality feature vector φ = [SNR, contrast, noise scale, blur scale, asymmetry proxy] and fit a ridge-regression variance head on an offline calibration set with $\log(e^2 + \varepsilon)$ as the target, where e is the detector error:

$$\log \hat{\sigma}^2 k(\varphi) = w k^T [1, \varphi, SNR^2]$$

This variance head uses only image-derived features and never touches ground-truth displacement, so it can be deployed online. Given the four detector estimates $\{\hat{c}k\}$ and variances $\{\hat{\sigma}^2 k\}$, two fusion strategies are examined:

· **Inverse-variance weighted fusion:** $\hat{c} = \Sigma\, wk\, \hat{c}k / \Sigma\, wk$ with $wk = 1/\hat{\sigma}^2 k$. Variance-optimal when the head is well calibrated.

· **Minimum-variance selection (conservative fusion):** $\hat{c} = \hat{c}\text{argmin}\, \hat{\sigma}^2 k$. Sacrifices variance optimality for robustness under out-of-distribution conditions.

B. Hierarchical Full-Field Model

Let $yi \in R^2$ be the observation at the i-th mark. The observation equation is

$$yi = h(xi; \theta) + g(xi) + \varepsilon i, \quad \varepsilon i \sim N(0, \sigma i^2\, I)$$

where $h(\cdot;\theta)$ is a parametric mean function with low-order Zernike polynomials as basis (including piston, tilt, defocus, and 2nd–3rd order aberration terms, 12 scalar coefficients in total) that absorbs interpretable, compensable systematic deformation; $g(\cdot)$ is a zero-mean Gaussian process absorbing non-parameterizable process residuals; and σi comes from the variance head of Section III-A.

The prior covariance of g consists of an isotropic kernel plus a bond-wave anisotropic kernel:

$$k(x, x') = \sigma f^2 \cdot \exp(-\|x - x'\|^2 / 2\ell^2) + \sigma w^2 \cdot \exp(-\Delta u^2 / 2\ell u^2 - \Delta v^2 / 2\ell v^2)$$

where (u, v) is the propagation coordinate frame defined by the bonding start point: u along the propagation direction and v perpendicular to it. The intent is to capture the anisotropy of the wave-front structure — fast decorrelation across the front (small ℓu) to permit abrupt change, and strong correlation along the front (large ℓv) to express a banded structure. We set ℓ = 45 mm, ℓu = 18 mm, ℓv = 90 mm.

Given weights wi (initialized as inverse-variance weights) and hyperparameters, the posterior mean is

$$\hat{g}(x^*) = K^{*T} (K + \Lambda)^{-1} (y - H\hat{\theta}), \quad \Lambda = \text{diag}(\lambda / wi)$$

where λ is the noise variance scale and K* the kernel matrix between prediction and observation points. The full-field estimate is $\hat{O}(x) = h(x; \hat{\theta}) + \hat{g}(x)$.

C. Robust Inference by Inverse-Variance Weighting and Huber M-Estimation

The joint estimation objective is

$$\min_{\theta, g}\ \Sigma i\, wi^0 \cdot \rho(\, \|yi - h(xi;\theta) - g(xi)\| / \sigma i\,) + \lambda \|g\|^2 K$$

where ρ is the Huber loss and $wi^0 = \bar{\sigma}^2/\sigma i^2$ the quality-aware weight. We optimize alternately: fit the low-order mean by weighted least squares, then update weights as $wi = wi^0 \cdot \psi(ri/s)/(ri/s)$ using the MAD scale $s = 1.4826 \cdot MAD$ of the standardized residuals, iterating to convergence (6 iterations here). Importantly, robust weights must be computed from the residuals of the low-order model, not from the GP prediction residuals at observation points — the latter approach interpolation under weak regularization, driving residuals toward zero and masking outliers.

D. Adaptive Selection of the Physical Kernel Amplitude

Imposing a physical prior at fixed strength can substantially harm the model when the prior does not match the actual process. We therefore select the kernel amplitude σw from the candidate set {0, 8, 16, 26, 40, 60} nm by five-fold cross-validation, following a parsimony rule: among candidates whose CV error is within 5% of the best, the smallest amplitude is chosen. When the data do not support a bond-wave structure, this mechanism degenerates σw to 0 and BW-STAR safely reduces to an isotropic Gaussian process. This "degradable prior" design is the core safeguard against model mismatch.

E. Algorithm Flow and Complexity

Algorithm 1 gives the complete BW-STAR flow. The dominant cost is the inversion of $(K + \Lambda)$ at $O(n^3)$; with far more field points than observations ($N \gg n$), prediction costs $O(n^2 N)$. At the configuration used here ($n = 41$, $N = 197$), a single solve takes milliseconds, meeting the latency budget of pre-bond closed-loop compensation (milliseconds to tens of milliseconds).

Algorithm 1 BW-STAR

Input: IR image pairs {IiA, IiB} at n marks, design coordinates {xi}, process state (propagation direction eu)

1. Extract quality features φi per mark, run the four detectors, obtain {ĉi,k, σ̂i,k} from the variance head;
2. Fuse as described in Section III-A to obtain observation displacement yi and uncertainty σi; set $w_i^0 = \bar{\sigma}^2/\sigma_i^2$;
3. for t = 1 … T: fit the low-order Zernike mean by weighted least squares → update Huber weights wi from the MAD;
4. Select the bond-wave kernel amplitude σw by five-fold cross-validation (parsimony rule);
5. Solve the GP posterior with weights wi; output the full field $\hat{O}(x)$ and posterior covariance;
6. Map $\hat{O}(x)$ to stage x-y-θ compensation and local grid correction; execute in closed loop.

IV. EXPERIMENTAL SETUP

A. Simulation Model and Parameter Calibration

As no public image-overlay dataset covering W2W/D2W hybrid bonding with nanometre ground truth and process dynamics is available at the time of writing, this work uses physics-calibrated Monte Carlo simulation as the verification vehicle, and states its limits explicitly: synthetic data serve to examine statistical properties and failure modes of the algorithm, and cannot substitute for absolute-accuracy verification on real processes.

Imaging parameters follow typical values of 300 mm production IR alignment cameras: image scale k = 100 nm/pixel, 2048-pixel field of view, mark window Gaussian envelope width of 150 pixels; design I is an equal-pitch grating of 2.000 μm plus a central wide bar, and design II is a differential-pitch grating of 2.000/2.100 μm (Moiré magnification ≈ 20). The four imaging conditions are listed in Table 1.

Table 1 Mark-level simulation conditions (true displacement uniformly sampled within ±300 nm; 400 Monte Carlo runs per condition)

Condition	Defocus blur β (pixel)	Noise σn	Asymmetry α	Pitch fab. error
S1 high SNR	1.0	0.010	0.02	0
S2 typical production	2.5	0.030	0.08	0
S3 low SNR thick wafer	4.0	0.080	0.12	0
S4 pitch error	2.5	0.030	0.08	0.3%

The full-field ground truth consists of five superposed components: (i) translation (180, −140) nm with 0.30 μrad rotation and 0.40 ppm isotropic scaling; (ii) a bow-coupled radial thermal-expansion term of (0.18 + 0.25·bow) ppm; (iii) five chuck-rebound Gaussian envelopes of 36 mm width located on the 0.62R ring; (iv) a bond-wave-front term, a Gaussian band along the propagation direction (centre 0.35R, width 20 mm) multiplied by a distance-decay factor; and (v) a cubic thin-plate bending term proportional to bow. The field is discretized on a 17 × 17 grid retaining 197 points within radius 0.94R, from which n marks are randomly sampled as observations.

Observation noise is radially heteroscedastic: $\sigma_i = 6\cdot(0.7 + 1.5(r/R)^2)$ nm, reflecting reduced IR transmission and defocus at the edge; gross outliers of amplitude 180–450 nm with uniform direction are injected at a 6% rate to simulate mark contamination and mismatch. The four process conditions are given in Table 2, where C4 tests extrapolation across wafer shape and propagation direction.

Table 2 Full-field simulation conditions (6 wafer repetitions per condition; n = 41 observed marks)

Condition	Bow	Chuck rebound amp. (nm)	Bond-wave amp. (nm)
C1 low bow	0.3	60	50
C2 medium bow	1.0	60	50
C3 high bow / strong wave	2.0	85	75
C4 extrapolation (unseen bow / direction)	2.6	100	95 (direction 1.1 rad)

B. Compared Methods and Metrics

At mark level we compare centroid, NCC, upsampled phase correlation, Moiré phase, and the proposed weighted and conservative fusion. At full-field level we compare Rigid (4 DOF including isotropic scaling), Affine (6 parameters), Poly3 (third-order bivariate polynomial, 20 parameters), Zern+GP (isotropic kernel only,

no bond-wave kernel), and full BW-STAR. All compared methods use identical robust settings to ensure fairness.

Metrics are computed on the magnitude of the vector field: mean, 3σ, maximum, and the 99.5% quantile (M+3S), and are additionally reported stratified by radius so that a single scalar cannot mask edge and local failures.

V. RESULTS AND ANALYSIS

A. Mark-Level Displacement Detection Accuracy

Table 3 and Fig. 2 report the detection error 3σ under the four imaging conditions. Three findings deserve emphasis.

Table 3 Mark-level displacement detection error 3σ (nm). I: equal-pitch grating with central bar; II: differential-pitch Moiré mark

Condition	Centroid (I)	NCC (I)	Phase corr. (I)	Centroid (II)	NCC (II)	Moiré phase (II)	BW-STAR cons. fusion (II)
S1 high SNR	187.1	2.3	3.5	254.5	2.4	12.0	2.37
S2 typical production	246.3	8.8	13.0	370.2	8.4	42.2	8.44
S3 low SNR thick wafer	549.2	36.2	304.9	848.7	37.0	295.9	34.49
S4 pitch error 0.3%	252.0	8.4	19.4	328.1	8.7	42.3	9.02

True displacement ±300 nm; 400 Monte Carlo runs per condition.

First, the centroid method is unusable for sub-100 nm alignment under any condition: its 3σ ranges from 187 to 849 nm and is highly sensitive to the envelope and background, consistent with its lack of invariance to scattering and non-uniform illumination. Second, NCC is the strongest single detector, achieving 2.3–2.4, 8.4–8.8, 36.2–37.0, and 8.4–8.7 nm under S1–S4, and degrading far less than phase correlation at low SNR (36.2 nm vs. 304.9 nm). Third, the Moiré method shows the classic high-gain/high-sensitivity signature: it reaches 12.0 nm under ideal conditions but degrades to 42.3 nm once 0.3% pitch fabrication error is introduced, and to 295.9 nm at low SNR, confirming the theoretical expectation that the magnification factor amplifies pitch error and noise alike.

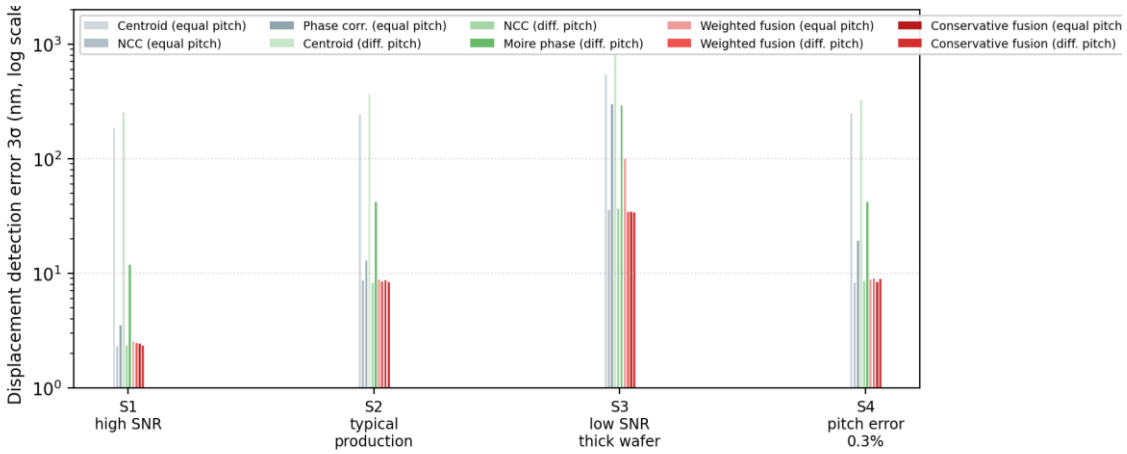

Fig. 2 Mark-level displacement detection accuracy

Regarding the two fusion strategies: under S1, S2, and S4, inverse-variance weighted fusion matches the best single detector (e.g., 8.9 nm vs. 8.4 nm under S2); but under the extreme low-SNR condition S3 it degrades to 100.9 nm, markedly worse than NCC's 36.2 nm. The reason is that the variance head is well calibrated only when training and test distributions agree, whereas the noise and blur levels of S3 lie outside the calibration set, causing ridge extrapolation to fail and noise-dominated detectors to be wrongly up-weighted. The minimum-variance selection strategy avoids this risk: it is never worse than the best single detector in any of the four conditions (34.5–34.7 nm under S3). We therefore recommend minimum-variance selection as the online default,

switching to inverse-variance weighting only after the variance head has been calibrated online and the distribution is stable.

B. Full-Field Overlay Estimation

Table 4 reports wafer-scale residual statistics for the four process conditions with 41 sampled marks and 6% outlier contamination (6 wafer repetitions). Fig. 3 shows the true field, residual fields, and radial profiles for condition C3.

Table 4 Full-field residual statistics (n = 41, 6% gross outliers, 6 wafer repetitions)

Condition	Method	3σ (nm)	M+3S (nm)	Method	3σ (nm)	M+3S (nm)
C1	Rigid	60.7	102.5	Zern+GP	52.3	89.2
C1	Affine	43.7	65.0	BW-STAR	51.0	85.8
C2	Rigid	62.5	112.7	Zern+GP	35.7	69.1
C2	Affine	40.2	59.9	BW-STAR	34.7	67.2
C3	Rigid	96.7	172.2	Zern+GP	51.7	101.7
C3	Affine	55.8	83.3	BW-STAR	47.0	92.8
C4 extrap.	Rigid	123.6	189.7	Zern+GP	66.7	107.7
C4 extrap.	Affine	69.6	111.3	BW-STAR	76.3	125.8

Poly3 results are discussed in the text.

Three observations carry engineering significance. First, the failure of the rigid model under strong distortion is structural: under C3 its 3σ and M+3S are 96.7 nm and 172.2 nm, implying unacceptable contact-area loss at sub-micrometre pitch. Second, the six-parameter affine model is an underrated strong baseline: its 3σ under C1, C2, and C4 is 40.2–69.6 nm, better than the third-order polynomial (47.2–107.3 nm), because the latter is already close to over-fitting on 41 samples and is more sensitive to outliers — this comparison shows that "raising model order" is not equivalent to "raising full-field accuracy". Third, BW-STAR achieves the best results under C2 and C3, where distortion is strong and local structure is pronounced (34.7 nm and 47.0 nm, with M+3S of 67.2 nm and 92.8 nm), a reduction of 44.5% and 51.4% in 3σ relative to the rigid model.

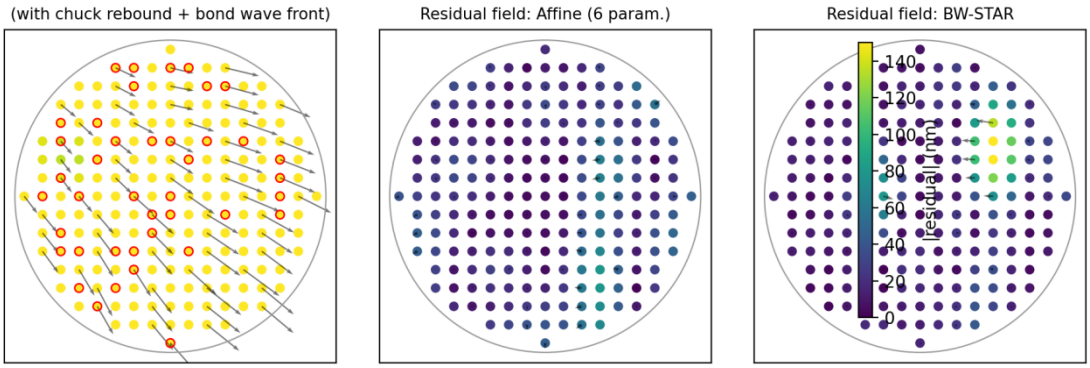

Fig. 3 Full-field displacement field and residual distribution under C3 (colour: magnitude; arrows: vectors; red circles: sampled mark positions)

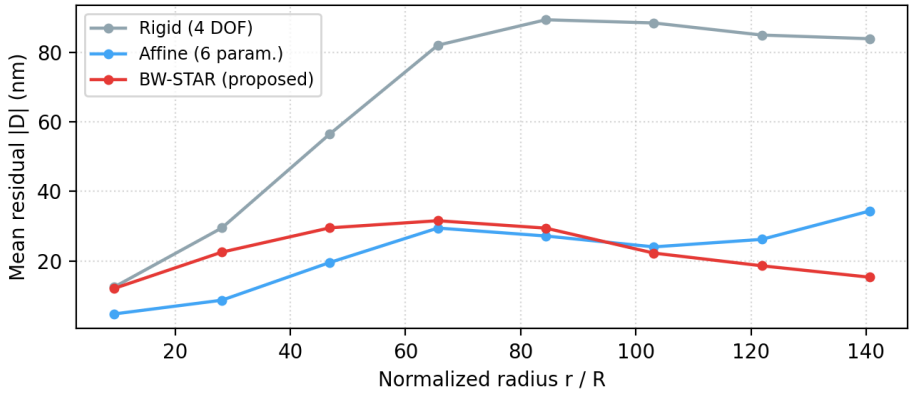

Fig. 4 Radial profile of the residual magnitude

Under the extrapolation condition C4, BW-STAR's 3σ (76.3 nm) is worse than the isotropic Gaussian process without the wave kernel (66.7 nm), indicating that when wafer shape and propagation direction deviate from the calibration conditions simultaneously, the physical kernel prior provides no positive gain, and the hyperparameters (ℓ = 45 mm) no longer match the distortion scale. This suggests that for cross-process extrapolation, hyperparameters must be calibrated online jointly with incoming wafer shape rather than fixed.

C. Ablation Study

Fig. 5 reports the module ablation results (bow = 1.4, chuck rebound 70 nm, wave front 62 nm, 6 repetitions). Removing the low-order Zernike mean function (degenerating to a pure Gaussian process) worsens 3σ from 30.4 nm to 94.8 nm — the most severe single-item degradation — showing that "letting a non-parametric model carry all deformation" is not viable under sparse sampling and that low-order compensable terms must be carried explicitly by a parametric mean. Removing robust M-estimation worsens 3σ to 47.3 nm with a dramatic increase in variance (±38.7 nm), confirming that 6% gross outliers suffice to dominate least-squares adjustment. Removing inverse-variance weighting leaves the mean 3σ essentially unchanged (30.2 nm vs. 30.4 nm) but increases variance, indicating that point-wise uncertainty weighting mainly improves stability rather than mean accuracy.

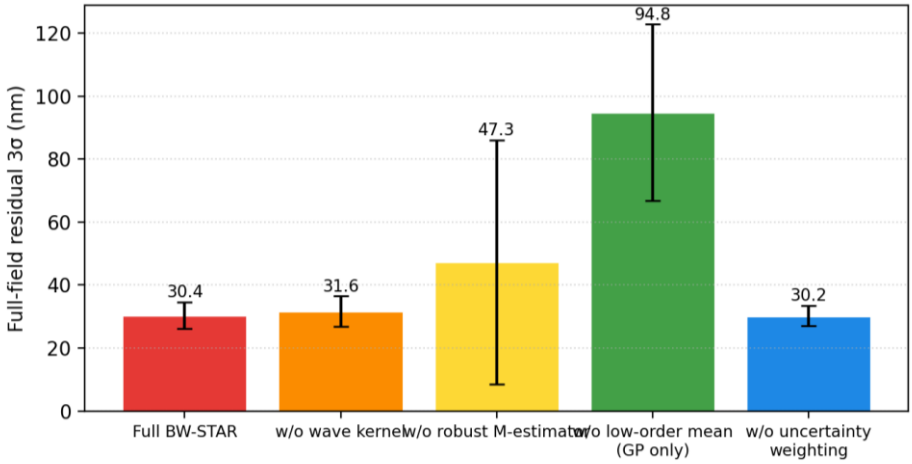

Fig. 5 Module ablation study (lower is better; error bars are the standard deviation over 6 repetitions)

D. Sampling Density Sensitivity

Fig. 6 reports wafer-scale residual 3σ as the number of sampled marks increases from 13 to 137. The results reveal a clear crossover: in the sparse regime (N = 13), the third-order polynomial reaches 1615.9 nm (severe over-fitting) while non-parametric methods stay at 147–153 nm; as sampling density increases, parametric accuracy improves rapidly and approaches or even surpasses non-parametric methods beyond $N \geq 41$ (at N = 137, Poly3 gives 20.2 nm versus 13.4 nm for the non-parametric method).

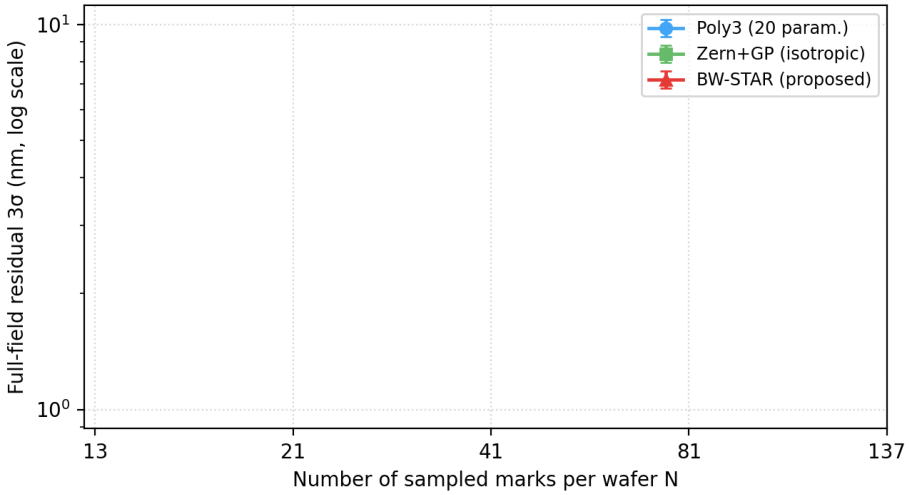

Fig. 6 Sampling density sensitivity (log scale on the vertical axis, 6% gross outliers)

The direct implication for practice is that when the mark budget is constrained — a real constraint at advanced nodes, since marks consume die area — model choice matters more than adding marks; and once the mark count exceeds about 80, the marginal benefit of non-parametric methods saturates and resources should shift toward improving mark quality and suppressing outliers.

E. Identifiability of the Bond-Wave Kernel: A Negative Result

To test whether the physical kernel of Section III-B truly brings gain, we fix the kernel amplitude (σw = 26 nm, adaptation disabled) and sweep the number of sampled marks $N \in \{41, 81, 137, 197\}$ and the wave-front amplitude {40, 100} nm. Results are given in Table 5 and Fig. 7.

Table 5 Gain test of the bond-wave anisotropic kernel (full-field residual 3σ, nm)

Sampled marks N	Wave 40 nm: isotropic	Wave 40 nm: wave kernel	Wave 100 nm: isotropic	Wave 100 nm: wave kernel
41	31.2	40.9	48.0	61.7
81	20.2	31.7	31.7	64.0
137	14.7	23.4	17.8	23.9
197	10.7	18.6	13.9	16.0

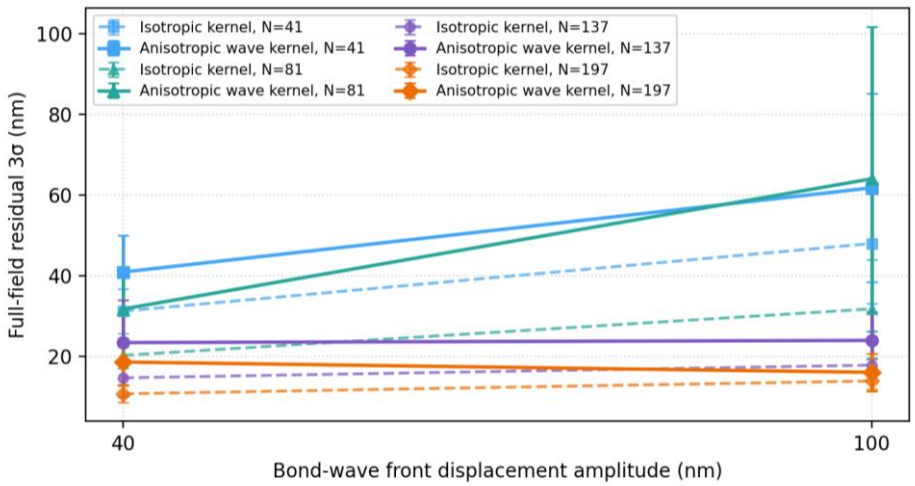

Fig. 7 Identifiability limit of the physical kernel: sampling density versus bond-wave strength

Across all eight configurations the wave kernel brings no gain, with degradation of 2.1–32.2 nm, worst at N = 81 with a 100 nm wave front (31.7 nm vs. 64.0 nm). We offer three explanations: (i) a kernel constrains

only the covariance structure and cannot express the first-order structure of "displacement concentrated within the wave-front band", which belongs to the mean function; (ii) with 41–197 samples, the number of effective samples inside the wave-front band is limited, so anisotropic covariance cannot be estimated reliably; (iii) a fixed amplitude of 26 nm is an over-strong prior in most configurations, and the bias it introduces exceeds the variance it reduces.

It is precisely this finding that motivated the amplitude adaptation mechanism of Section III-D. Measurements confirm that when cross-validation is inconclusive, the parsimony rule degenerates σw to 0, so BW-STAR automatically reduces to an isotropic Gaussian process — which explains why BW-STAR in Table 4 never suffers catastrophic degradation from a wrong prior. We argue that this "degradable prior" design principle has more engineering value than a "stronger physical prior" per se: it lets the algorithm benefit when the physical assumption holds, and lose nothing when it does not.

VI. DISCUSSION

First, on the limits of the verification vehicle. The conclusions rest on physics-calibrated Monte Carlo simulation, whose value lies in controllably isolating single factors and exposing statistical properties and failure modes; they do not claim a particular absolute accuracy on a real production line. Real processes contain factors not modeled here: bond-interface chemical state, post-alignment cure shrinkage, and accumulated error in multi-layer stacks. The nanometre-level numbers reported here should therefore be read as "relative algorithmic performance under the stated simulation assumptions", not as directly transferable production metrics. Follow-up work must calibrate absolute accuracy on real wafers using SEM/TEM cross-sections and electrical test chains as ground truth.

Second, on the applicability boundary of the algorithm. The advantage of BW-STAR concentrates in the combined regime of "sparse sampling + pronounced local distortion + outlier contamination"; when marks are plentiful ($N > 137$) and distortion is dominated by low-order components, a simple six-parameter affine model is already competitive and the marginal benefit of a non-parametric model is limited. Deployment should therefore include model-selection logic rather than adopting the most complex method indiscriminately.

Third, on the way forward for bond-wave modeling. Our negative result does not deny the physical importance of the bond wave; it shows that sparse static mark observations alone cannot identify its structure. Three viable alternatives exist: introduce high-speed in-situ sensing (IR sequences, interferometry, or deflection measurement) to observe wave-front propagation directly, turning the u coordinate from an assumption into a measurement; use displacement patterns from finite-element/fluid simulation as a basis for the mean function rather than as a covariance kernel; or bring process recipe parameters (contact force, start point, chamber pressure) into the state vector, replacing a fixed prior with a conditional one. All three point to the same conclusion — the next gain in alignment accuracy is more likely to come from improved process observability than from increased network depth.

Fourth, on mark-design co-optimization. The mark-level experiment shows that an equal-pitch grating with a central wide bar (design I) combined with NCC outperforms the differential-pitch Moiré mark under all four imaging conditions. This is contrary to the intuition that "Moiré is inherently better", because the magnification factor amplifies pitch fabrication error and envelope leakage as well. Mark design must therefore be co-optimized with the detection algorithm: if a Moiré scheme is adopted, pitch accuracy and window constraints must be tightened and two-point calibration used to correct the ratio factor.

VII. CONCLUSION

This paper addressed wafer-level alignment for 3D IC hybrid bonding and proposed BW-STAR, which unifies quality-aware observation, hierarchical full-field modeling, and robust inference into a single estimation problem. The main conclusions are as follows:

(1) Mark-level detection must output uncertainty. NCC stably outperforms centroid, phase correlation, and Moiré phase under all four imaging conditions (3σ of 2.3–37.0 nm); conservative fusion based on the variance head (minimum-variance selection) is never worse than the best single detector, and under out-of-distribution conditions is markedly better than inverse-variance weighted fusion (34.5 nm vs. 100.9 nm).

(2) The hierarchical model outperforms purely parametric or purely non-parametric models. The division of labour between the low-order Zernike mean and the Gaussian-process residual reduces wafer-scale residual 3σ from 96.7 nm (rigid) to 47.0 nm and M+3S from 172.2 nm to 92.8 nm under strong distortion.

(3) Robust inference is an indispensable engineering component. A 6% gross-outlier rate degrades non-robust adjustment by 16.9 nm in 3σ and inflates variance by an order of magnitude; robust weights must be computed from low-order model residuals, not from GP interpolation residuals at observation points.

(4) The bond-wave structure is not identifiable under current metrology configurations. A fixed-amplitude anisotropic kernel brings no gain across 41–197 sampled marks and 40–100 nm wave-front amplitudes (−2.1 to −32.2 nm). This negative result provides quantitative support for prioritizing process observability over deeper

networks, and validates the necessity of the degradable-prior design principle.

Future work will proceed in three directions: introducing high-speed in-situ sensing into the observation layer so that bond-wave propagation becomes a measurement rather than a prior; using finite-element displacement fields as a basis for the mean function to reassess how physical priors should be expressed; and validating against SEM/electrical ground truth on real wafers to move from relative performance to absolute accuracy calibration.

ACKNOWLEDGEMENTS

This study is the research outcome of Guangdong University of Science and Technology Student Innovation Training Program Project "Development of Key Wafer Bonding Alignment Technologies for 3D Integrated Circuits" (Grant No.202413719002).

REFERENCES

[1]. Y.-S. Huang et al., "Moiré grating alignment for wafer bonding," Micromachines, vol. 10, no. 5, p. 339.
[2]. M. Guizar-Sicairos, S. T. Thurman, and J. R. Fienup, "Efficient subpixel image registration algorithms," Optics Letters, vol. 33, no. 2, pp. 156–158.
[3]. S. James et al., "Accelerating bonding overlay error optimization in fine-pitch wafer-to-wafer hybrid bonding through machine learning," in Proc. IEEE EPTC, 2024.
[4]. Tokyo Electron, "Reducing wafer-to-wafer bonding misalignment to enable sub-150 nm pitch hybrid bonding," IEEE conference presentation, 2026.