Interlaboratory Round-Robin Study of Errors and Best Practices for Scanning Electron Microscope Digital Image Correlation

Tucker V, Vermeij T, Montgomery C, Stinville JC, Liu M, Pollock T, Sottos N, Hoefnagels J, LePage W

PUBLISHED AS

Tucker V, Vermeij T, Montgomery C, Stinville JC, Liu M, Pollock T, Sottos N, Hoefnagels J, LePage W. "Interlaboratory Round-Robin Study of Errors and Best Practices for Scanning Electron Microscope Digital Image Correlation." Experimental Mechanics (2026).

https://doi.org/10.1007/s11340-026-01347-6

Interlaboratory Round-Robin Study of Errors and Best Practices for Scanning Electron Microscope Digital Image Correlation

Victoria Tucker[1,2], Tijmen Vermeij[3,4], Christopher Montgomery[5], Jean-Charles Stinville[5,6], Mengying Liu[7,8], Tresa Pollock[6], Nancy Sottos[5], Johan Hoefnagels[3], and William LePage[1†]

[1]University of Tulsa, Department of Mechanical Engineering
[2]Purdue University, School of Materials Engineering
[3]Eindhoven University of Technology, Department of Mechanical Engineering
[4]Empa Thun, Laboratory for Mechanics of Materials & Nanostructures
[5]University of Illinois at Urbana-Champaign, Department of Materials Science and Engineering
[6]University of California, Santa Barbara, Materials Department
[7]Texas A&M University, Department of Materials Science and Engineering
[8]Washington and Lee University, Physics and Engineering Department
[†]Correspondence: lepage@utulsa.edu

Abstract

Background: Scanning electron microscope digital image correlation (SEM-DIC), also referred to as high-resolution digital image correlation (HR-DIC), is a powerful method to measure displacements and strains at the micro to nanoscale. While SEM-DIC is gaining popularity, the nature of various errors in SEM-DIC is not clear.

Objective: To advance the practice, this work seeks a more extensive and cohesive understanding of the types and causes of errors in SEM-DIC. Previous works have identified SEM imaging errors that are problematic for DIC, including spatial distortions, drift distortions, noise, stabilization errors, and line jumps.

Methods: The work presented here outlines a unique, interlaboratory, round-robin format to investigate how beam and hardware settings can optimize SEM-DIC image quality. Characterization using multiple SEMs of different vendors, models, and types (e.g., electrostatic and electromagnetic columns) helps identify the types, magnitudes, and causes of SEM-DIC errors, as well as the optimum conditions to reduce such errors.

Results: As part of the iDICs round-robin challenge, over 2,000 total SEM images, which are available as an open-access dataset, were analyzed in this study, comprising images from five SEMs, two sample patterning types, and two magnifications. To quantify SEM-DIC errors, this work developed new error metrics to quantify the severity of line jumps, left-edge smear, horizontal skewing, vertical skewing, and translational drift.

Conclusions: This round-robin study found that most SEM-DIC experiments should utilize high beam voltages (e.g., 30 kV), fast dwell times (e.g., 1 µs), low to moderate beam currents (e.g., less than 1 nA), no line averaging/integration, and cropping of the left-most portion of images.

Keywords: electron microscopy, digital image correlation, mechanical behavior, round robin

1 Introduction

Digital image correlation (DIC) is useful for characterizing the mechanical behavior of materials through local measurements of displacement[1–3]. While DIC is typically applied at the millimeter scale using optical cameras, it can utilize images from any length scale that can be appropriately imaged. At the microscale, scanning electron microscope digital image correlation (SEM-DIC), also referred to as high-resolution DIC, has emerged as a powerful technique[4–6], as described in detail in a recent review paper[7]. However, compared to optical systems, scanning electron microscopes (SEMs) introduce additional imaging errors that can degrade the quality of SEM-DIC results[8–14]. For example, SEM images are time-dependent. While the pixels in an optical image can typically be assumed to come from the same point in time,[1] the pixels in an SEM image are inherently captured at different times due to the scanning nature of an SEM. Along with errors from imaging sources, additional errors can originate from the DIC analysis algorithms itself. For example, when comparing two identical images and images with known ground-truth displacements, such errors can be identified, as procedures in other round-robin studies on DIC analysis algorithms have shown[15,16]. In this work, the focus is entirely on errors from the SEM imaging process, but we acknowledge that other errors are present in the measurements from the DIC analysis algorithms itself.

Understanding the types and sources of error in SEM images is crucial for reliable SEM-DIC measurements. Previous works have identified several error types in SEM-DIC, encompassing spatial distortions, drift distortions, noise, line jumps, and lens preparation errors. The following paragraphs describe these error types as identified by previous works[8–11,17–22].

[1] There are exceptions to this, such as rolling-shutter CMOS sensors.

Spatial distortions are position-dependent asymmetries analogous to barreling distortions in optical imaging systems[8,11,17]. These distortions come from imperfect electromagnetic lenses. Practically, these distortions manifest as skewing or stretching imperfections of the image field when referenced to a precise, gridded sample, such as a TEM grid[18]. Equipment positioning and settings, such as improper lens alignment, stigmation, and focus, contribute to these distortions[8]. As spatial distortions are position-dependent, they become apparent during sample displacement or translation. Importantly, these distortions are relatively time-independent and generally do not appear in static or non-translated images taken in repetition. Finally, spatial distortions manifest in both the vertical and horizontal displacement fields.

Drift distortions are time-dependent errors that manifest as a translation complex curve error over a period of time[10,11,13]. Since drift distortions occur gradually, they become apparent when capturing static or non-translated images in repetition. These distortions are typically attributed to charging of the SEM stage or sample[10,18], which may result in stage motion or heating[10]. Drift distortions can accumulate in both the horizontal and vertical displacement fields[9,10]. However, for normal horizontal scanning, drift distortions only lead to a stretch in the vertical direction and a shear in the horizontal direction, while the amplitude of this stretch and shear only varies in the vertical direction[13].

Noise is a random, pixel-wise discrepancy in brightness[10,19]. In SEM images, sources of noise could include external electromagnetic interference, external vibration of the SEM, as well as variations during signal conversion of the analog voltage from an Everhart-Thornley detector to a digital signal for the image.

Stabilization errors occur when the electron detector does not have sufficient time to stabilize after moving to the location of a new pixel (Figure 1a) or a new line (Figure 1b). It is a pixel-wise and line-wise error of which the magnitude is hypothesized to be constant over the image on average[20]. Stabilization errors result from inconsistencies in the beam/detector control loop. The detector response time, and consequently stabilization error, depends on voltage, dwell time, electrostatic versus electromagnetic column type, and magnification[20].

Line jumps are scanning errors that occur line-wise, i.e., across the vertical displacement field (Figure 1b). Although similar to line-wise stabilization errors, it has been speculated that line jumps could result from dust particles in the SEM column charging and deflecting the beam[21]. This would indicate that, unlike stabilization errors related to detector response times, line jumps would be unaffected by beam and scan settings.

Left-edge smear (LES) is an error that manifests as horizontal and vertical displacement errors on the far left side of the image. Presumably, LES is a specific stabilization error that occurs as the beam jumps from the right side of the previous row of pixels and begins scanning a new line of pixels (Figure 1b). This stabilization issue may be related to timing inconsistencies (Figure 1c). LES is distinct from line jump errors and more general stabilization errors, while a line jump error creates erroneous displacements across the full image width, and general stabilization errors are only pixel-wise, the LES is confined to the left-most $\approx 5\%$ of the image. Therefore, LES is commonly addressed among the community by removing the left-most region of the images (often $\approx 10\%$ of the image width) from the DIC analysis.

Lens preparation errors (LPE) result from imperfect setup of the SEM optical system prior to imaging, including focus, stigmator correction, and aperture alignment. These errors introduce spatially varying intensity and geometric distortions that can bias SEM-DIC measurements. Unlike scanning-related errors, LPE originate from residual electron-optical aberrations and persist across the entire image. Although automated alignment methods are an active area of research, images in this study were acquired by trained microscopists using conventional manual preparation[23,24]. This approach reflects current typical SEM-DIC practice, and thus includes the level of lens preparation error commonly encountered in experimental measurements.

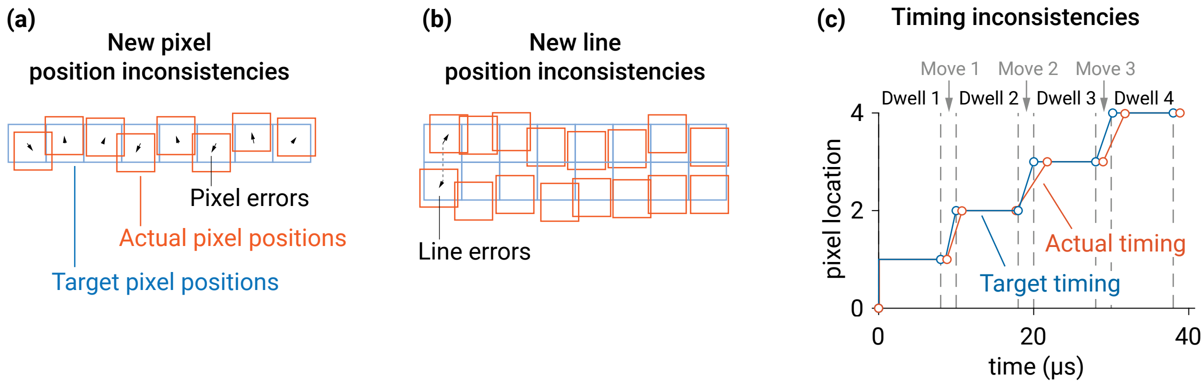

Figure 1: Stabilization scanning errors in SEM images (a) new pixel position inconsistencies (b) new line position inconsistencies (c) timing inconsistencies.

Due to errors in SEM images, SEM-DIC results typically require correction through post-processing procedures[9,10,25] or capturing with specific error-reducing scan techniques[20]. In either case, when capturing an SEM image, several settings influence the image quality. These settings include beam voltage, beam current, dwell time, and magnification. Optimizing these SEM settings often involves a balance. For instance, while higher beam current reduces noise, it can also amplify charging, leading to drift distortion. Alongside SEM settings, the quality of SEM images is also influenced by the SEM hardware itself. SEMs can utilize either electrostatic or electromagnetic deflectors, and the faster response time of electrostatic deflectors may reduce time-dependent system stabilization errors[20].

Previous works have studied the relationships between SEM settings and SEM-DIC errors. Table 1 provides a summary of the conditions examined in the literature, along with the relationships between these conditions and the various error types mentioned above.

	Drift distortion	Spatial distortion	Noise	Line jumps	Stabilization errors
Deflector Type	†	†	†	†	Electrostatic columns deflect the beam faster, resulting in fewer time-dependent stabilization errors compared to electromagnetic deflectors [20].
Beam Voltage	Higher beam voltages can increase sample charging, which can lead to drift distortions [18,22].	†	†	†	Lower voltages reduce stabilization errors [20].
Beam Current	Higher beam currents can increase sample charging, which can lead to drift distortions [22].	†	Higher currents reduce noise by increasing the signal to noise ratio [10,22].	†	†
Dwell Times	Long dwell times can lead to charging, which can lead to drift distortions [20].	†	Long dwell times increase the signal-to-noise ratio [10,22].	†	Longer dwell times reduce the presence of stabilization errors [20].
Magnification	Lower magnifications (wider fields of view) significantly reduce drift distortions [10].	Higher magnifications (smaller fields of view) significantly reduce spatial distortions [10].	†	†	†
Line averaging	†	†	†	Bias strain is removed through line averaging [26]	†
Image Integration	Image integration (frame averaging) reduces drift distortions, but also removes the possibility to perform drift correction [19,20].	†	Noise is reduced through image integration [10].	†	†

Table 1: Summary of literature on SEM-DIC conditions and error types.
† No relationship found in literature

There are various beam and hardware settings that can enhance image quality, but the connection between beam settings and error reduction has not been extensively characterized across multiple SEMs and samples. Additionally, although this study focuses on conventional SEM, the general methodology to distinguish time-dependent (drift) and position-dependent (scan/lens) distortions would extend to other raster-scanned electron imaging modes; however, the dominant error sources and optimal operating conditions may differ substantially in TSEM, ESEM, and STEM. Therefore, the goal of this work is to better understand and optimize hardware and scan settings for enhancing SEM-DIC quality. Future analyses of the images from this study, which are publicly available, can be utilized to further optimize SEM-DIC results.

2 Methods

The relationships between errors and conditions provided in prior studies (Table 1) guided the procedure for this round-robin study. For example, concerning different SEM types, electrostatic column SEMs have been found to provide better temporal accuracy and precision by deflecting the beam faster[20]. Based on this finding, both electrostatic and electromagnetic SEMs were utilized in this study. Additionally, multiple currents and voltages were used to identify the balance and severity of errors concerning overall SEM-DIC image quality and beam conditions. In this way, the study could assess the balance between high currents that increase signal-to-noise ratio but can also introduce drift[18,22]. Noise (reduced by long dwell times) and drift distortions (reduced by short dwell times) suggested dwell time as a parameter to include in this study[20]. Thus, multiple dwell times were studied, and line averaging was also considered. Finally, since magnification relates to drift and spatial distortions, with more drift expected at high magnification and more spatial distortion expected at low magnification, samples were imaged at two different magnifications[10].

2.1 Samples

The study utilized two sets of samples for imaging at different magnifications and for providing consistent samples to examine with the five different SEMs. Both sets of samples used diced Si wafer substrates (10×10mm wide, 500 µm thick; Ted Pella), but the two sets had different speckle patterning methods (Figure 2). For low magnifications, samples of self-assembled Au nanoparticles for a 200 µm horizontal field width (HFW) were created using the method from Kammers and Daly[10]. Additionally, for high magnifications, samples of Ag-sputtered nanodroplets for a 20 µm HFW were created using the method from Montgomery, Koohbor, and Sottos[27]. It is important to note that in the main study, sample types cannot be deconvoluted from magnification. Section 3.7 and the Supplementary Information further outline the specific effect of magnification and conductivity on SEM-DIC error.

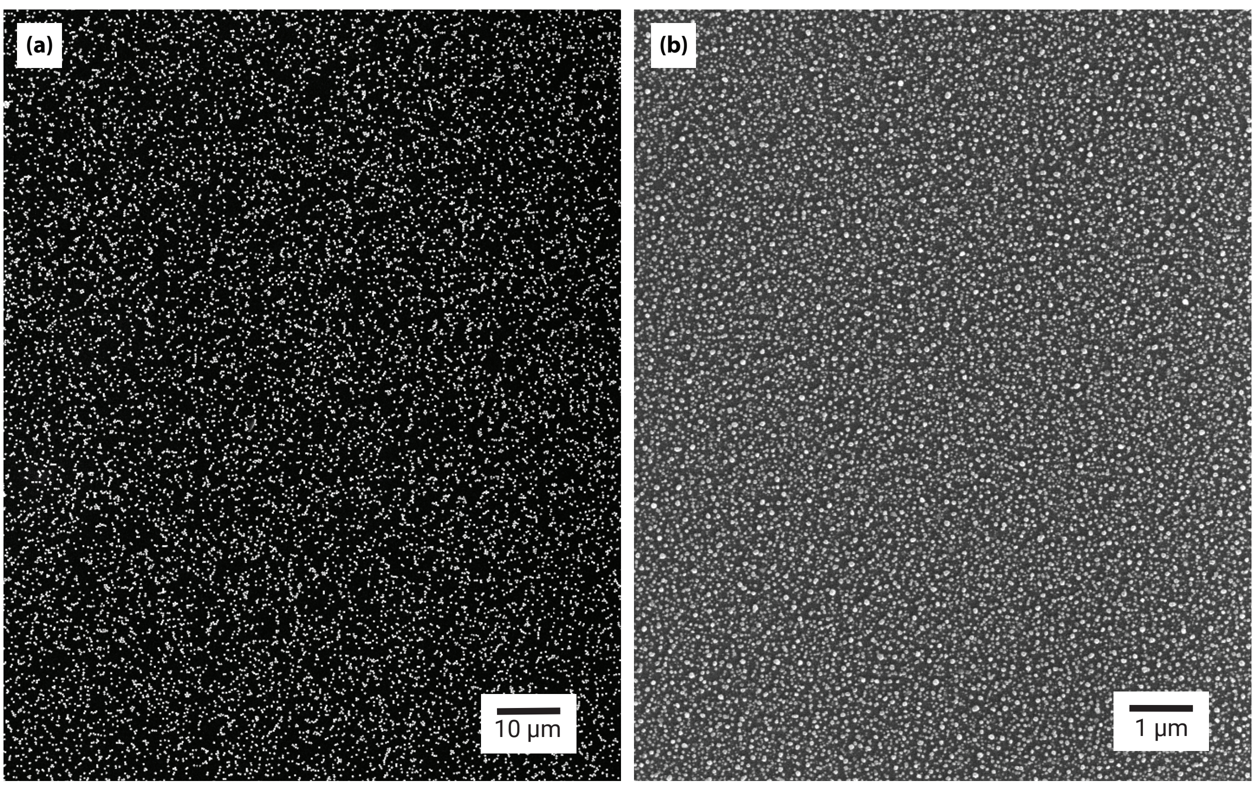

Figure 2: SEM image of (a) low-magnification self-assembled Au nanoparticle coated sample (suited for 200 µm HFW)[10] and (b) high-magnification Ag sputtered nanodroplet sample (suited for 20 µm HFW)[27]

The Ag and Au samples differ due to the necessity of optimizing speckle size (at least 3 px per speckle)[28,29]. The Ag samples have good speckle density and good electronic conductivity (due to the sputtered metal base layers). However, the speckle sizes are inconsistent and often too small, which increases aliasing (subset misidentification due to the lack of contrast between speckled and non-speckled regions caused by high speckle density)[28,30]. Alternatively, the Au nanoparticle samples feature uniform speckle sizes large enough to avoid aliasing. However, the Au samples have sparse speckles and presumably lower conductivity due to the lack of a metallic base layer[29]. Another key difference here is that the electron beam interaction volume size becomes relatively more significant at higher magnifications, regardless of the speckle pattern used. Since secondary electrons originate from an interaction volume near the surface, rather than a precise point, finer structures needed for higher resolution tend to exhibit reduced contrast due to signal bleed. One way to mitigate this effect is by lowering the beam voltage or current to shrink the interaction volume, though this may negatively impact image quality in other ways. However, it is hypothesized that this issue is more dependent on magnification than on the choice of speckle pattern. Finally, it has also been shown that pattern bias can play a key role in DIC measurement; however, this was not explored in our study[30].

To apply Au nanoparticles (AuNPs), a self-assembling method was followed, as described by Kammers and Daly[10], with the following modifications: instead of producing AuNPs with the Frens method, nanoparticles were purchased (250 nm diameter; Alfa Aesar Gold nanoparticles supplied in 0.1 mg/mL sodium citrate with stabilizer). Additionally, hydroxylation was not executed, as the samples naturally contained sufficient hydroxyl groups to achieve a dense AuNP surface coverage[31]. The substrates were silanized, rinsed with a mixture of methanol and deionized water, and then submerged in a colloid with the AuNPs for 4 hours, following the method from prior work[32].

The Ag nanodroplet samples were fabricated with a thin film deposition and reconfiguration method[27]. Thin films were sputtered on the substrate sample surfaces using a magnetron sputtering system[27]. A 3 nm thick Ti adhesion layer was deposited onto the substrate, followed by a 56 nm Au layer. Then, a second 20 nm Ti adhesion layer was added, and finally, an Ag layer was deposited. In addition to adhesion for the Ag, the base layers ensure that the same pattern is formed for the same conditions across a variety of different substrates. Reconfiguration was performed at ambient temperature by exposing the surface to a 1 wt% NaCl solution. The pattern size was controlled by adjusting the thickness of the Ag layer, while the pattern morphology was regulated by varying the exposure time to the NaCl solution.

2.2 Experimental Procedure

Image sets were collected on five different SEMs (outlined in Table 2) over a range of accelerating voltages, beam currents, and dwell times. Au samples I, II, IV, and V were used across two different microscopes. Each microscope had a different microscopist, and each microscopist followed the procedure illustrated in Figure 3 and listed fully in the Supplementary Information. All images in the main study were captured with the standard in-chamber Everhart-Thornley secondary electron detectors. For SEM-4, images from the backscattered electron (BSE) detector were also captured and analyzed outside of the main study (see Section 3.7 for details). Scanning was controlled using the integrated microscope scan components (not an external scan controller). Working distances (WDs) were kept constant throughout the sample image set. Microscopists selected a WD of 5 mm, 10 mm or 15 mm, based on whichever WD was closest to the one they would typically select for an in-situ SEM-DIC experiment. Prior to imaging, each SEM was aligned with respect to the source/gun tilt, objective/lens alignment, and astigmatism correction.

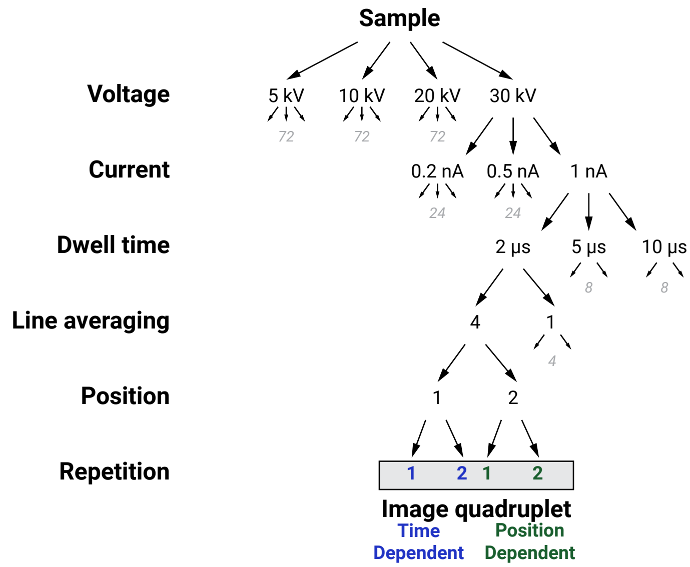

Figure 3: Visual representation of procedure scanning and beam conditions throughout the procedure. Grey numbers represent how many image-quad images stem from each set of arrows, highlighting that each sample includes 288 image-quad images before the additional location reference images.

Name	Microscope	Au image set	Au sample	Ag image set	Ag sample	Microscope model	Microscope vendor	Column type	Resolution (px)
SEM1-F, SEM1-H	1	F	I	H	IV	Quanta	TFS/FEI	Electromag.	3072 × 2048
SEM2-B, SEM2-K	2	B	II	K	V	FERA3	Tescan	Electromag.	3072 × 2048
SEM3-O, SEM3-P	3	O	I	P	IV	Versa 3D	TFS/FEI	Electromag.	3072 × 2048
SEM4-G, SEM4-M	4	G	III	M	VI	MIRA3	Tescan	Electromag.	3072 × 3072
SEM5-Q, SEM5-N	5	Q	II	N	V	Helios	TFS/FEI	Electrostatic	4096 × 3536

Table 2: Microscopes used in the present study.

The procedure involved capturing images with various beam voltages, currents, dwell times, and line averaging. The ordering of these conditions in the image acquisition sequence was randomized for each image set to limit bias from time-dependent errors that may result from images being captured over several hours. For example, SEM chamber pressure often decreases over a number of hours as images are collected. This decrease in chamber pressure (and thus, fewer molecules in the chamber to interfere with electrons) may lower SEM-DIC error.

Beam voltages of 5, 10, 20, and 30 kV and beam currents of 200, 500, and 1000 pA were prescribed in the experimental procedure. For SEM-3, the beam currents as specified in the procedure were not available. In this case, the nearest beam currents were selected based on spot size measured with a Faraday cup. For images taken with SEM-1, the beam current could not be prescribed. Instead, particular spot sizes were selected. In all cases, the various image conditions (beam voltages, currents, dwell times, line averaging, HFW, and working distance) were extracted from the image metadata, to avoid misnamed conditions. Furthermore, for SEM-5, beam current was varied across each location in order to achieve greater variation. Beam current values ranged from approximately 100 to 1200 pA.

Dwell times of 2 µs, 5 µs, and 10 µs were specified in the procedure. In the case that the specified dwell time was not available, the microscopists were instructed to choose the most comparable values. For analysis, data were grouped by 'ShortDwell': $1 \le t < 3$ µs, 'MidDwell': $3 \le t < 6$ µs, and 'LongDwell': $6 \le t \le 10$ µs.

Four images were captured for each beam condition (voltage, current) and scanning scheme (dwell time and line averaging). This group of four images formed an image quadruplet: "quad" (Figure 4). Each image quad was correlated independently, with the first image in the quad as the reference image and the other three images in the quad as the deformed images. In the absence of SEM errors, there would be zero displacement in the second image in the quadruplet, and a uniform displacement field for the third and fourth images in the quad (the translated image pair). The repeated images (images 2 and 4 in each quad) served to capture the time-dependent (TD) nature of drift distortions, while the translated images (images 3 and 4 in each quad) captured the position-dependent (PD) nature of spatial distortions and left-edge smear. All four images in a quad may manifest noise, line jumps, and stabilization errors. Prior to capturing each quad, magnification was adjusted and approached from a lower magnification for consistent engagement of objective lenses[20]. Additionally, microscopists were instructed to adjust brightness and contrast with a built-in auto-adjustment prior to capturing each quad.

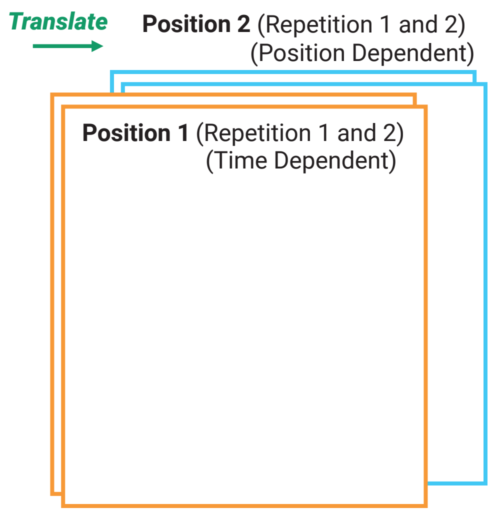

Figure 4: Image Quadruplet "Quad": static image pair (Position 1; orange), translated image pair (Position 2; blue)

Next, images 1 and 2 were captured as a static pair in an initial position on the sample. Afterward, the SEM stage was translated horizontally to the right (same direction as the beam scanning path) by 10% of the horizontal field width (HFW). The images from the AuNP samples had 200 µm HFW and translations of 20 µm. The images from the Ag sputtered nanodroplet samples had 20 µm HFW and translations of 2 µm. After translation, images 3 and 4 were captured.

Once the first quadruplet image set for a particular beam condition was complete, scanning schemes (dwell time and line averaging settings) were changed. The SEM stage was returned to the pre-translated (reference image) position. Next, the procedure for capturing an image quad was repeated three times (varying dwell time with each image quad). After each group of four image quadruplets (16 images) was completed, the stage was translated to a new region of the sample. This translation to a new region of the sample limited the degree of carbon contamination and thus loss of contrast that can occur by repeatedly imaging a region. Two images were taken at this new location to serve as a comparison for speckle pattern quality, using the same settings (dwell time, line averaging, beam current, and beam voltage) as the first image in the quadruplet. This process was repeated with the next beam conditions at the new location for a total of 12 locations. Ultimately, 312 images were taken on each sample (Figure 3). The exception to this is the image set for SEM-2. For this SEM, a limited number of dwell times were available, resulting in a total of 216 images per image set. Across all SEMs and image sets, in total, over one thousand images were analyzed for each sample type.

2.3 Correlation

A summary of the imaging and correlation details is provided in Table S1.[2] The image sets were correlated with a commercial code (Vic 2D-6, Correlated Solutions). Drift correction was not performed in any way. For the correlation, we employed low-pass image filtering, Gaussian-weighted subsets, optimized 8-tap interpolation, and a correlation criterion of zero-normalized sum of squared differences. After the correlation, the data were exported as MATLAB MAT files for analysis.

To determine appropriate subset and step sizes, an initial subset size study was performed. Two quads were randomly selected from each of the two image pixel widths (one selected from SEM-5, with 4096 px image width, and another selected from SEM-3 with 3072 px image width). For each of these two quads, 21 different correlations were performed with subset sizes ranging from 25 to 125 px. The fraction of successful DIC points were examined with respect to subset size (Figure 5). Ultimately, a 3.5 µm subset size for the 200 µm HFW Au samples was chosen. For the Ag samples with 20 µm HFW, a subset size of 0.35 µm was chosen. Since the images varied in pixel width (due to differences in SEM capabilities), the subset and step sizes were measured in nanometers, such that subset sizes from different microscopes were matched in physical space (instead of pixel space). However, we note that the subset sizes chosen correspond to 71 px for SEM-5 and 55 px for SEMs 1-4. These values were chosen by selecting the minimum subset size where the fraction of successful DIC points had converged to $\approx 100\%$. A step

[2]This table was adapted for SEM-DIC from the reporting requirements suggested in the DIC Good Practices Guide from the International DIC Society[33].

size of 1/8 of the subset size was chosen to ensure appropriate spatial resolution[33,34].

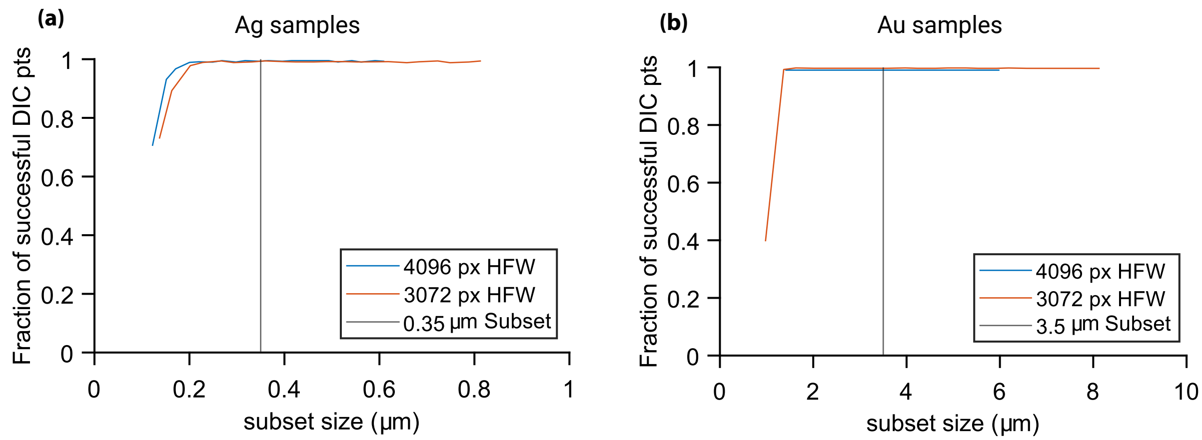

Figure 5: Subset size study results

2.4 Data Analysis and Visualization

The data were analyzed to quantify image quality (on the raw SEM images) and DIC errors (from the correlated SEM-DIC results). Image quality metrics were used to compare image sets, sample types, speckle patterns, and microscopes. Error metrics were used to identify trends between SEM-DIC error types (drift, line jumps, *etc.*) and beam or hardware conditions. Furthermore, bivariate (Pearson Coefficient) and multivariate regressions (MVR) were used to quantify the relationships between the various error metrics and imaging conditions.

2.4.1 Image quality metrics

The different magnification samples required different speckle patterns in order to achieve appropriate speckle sizes. While both samples roughly maintained the 3 pixel per speckle threshold for optimizing SEM-DIC speckle pattern quality[28], the patterns are different in their speckle density, size, brightness, contrast, and conductivity. It should be noted that both brightness and contrast are influenced by the speckle pattern and magnification. At higher magnifications, the same interaction volume becomes relatively larger. To assess whether pattern quality varied across samples and could influence SEM-DIC error results, the mean intensity gradient (MIG) was calculated and compared, in addition to the "correlation confidence interval" (“sigma”), for each location reference image pair. Here, “sigma” refers to the Correlated Solutions (CSI) software correlation residual metric, which is analogous to the correlation residual (matching error), sometimes also referred to as an epipolar error, in other DIC software packages.

MIG is often used as an assessment metric for DIC speckle pattern quality, and higher MIG values correspond to a better speckle pattern[35]. MIG is an image-wide metric that assesses the speckle pattern quality by quantifying the degree of local contrast through a spatial gradient. MIG (δ_f) is calculated by

$$\delta_f = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \left| \nabla f(a_{ij}) \right| \tag{1}$$

where W and H are the image pixel dimensions and

$$\nabla f(a_{ij}) = \sqrt{\left(f_x(a_{ij})\right)^2 + \left(f_y(a_{ij})\right)^2} \tag{2}$$

is the modulus of the local intensity at each pixel, a_{ij}. Following prior work[35], a Prewitt operator was used to compute the intensity derivative in both the horizontal (f_x) and vertical (f_y) directions. After the MIG of the image was calculated, this metric was used to compare the pattern quality of samples and locations.

Along with the MIG to assess speckle pattern quality, the overall correlation quality was measured with the correlation confidence interval “sigma.”[3] Sigma provides the 95% confidence interval (in pixel units) of the location of the deformed position for each individual subset. Theoretically, all errors (speckle pattern quality, scanning errors, lens errors, *etc.*) may increase the sigma value. Therefore, the average sigma provides an overall measure of the “goodness” of the correlation.

[3] We chose to denote correlation confidence interval as “sigma” instead of σ to avoid confusion with stress.

2.4.2 Error Metrics

Error metrics were used to quantify the relationships between SEM-DIC error types and SEM conditions. These metrics were calculated using the horizontal and vertical positions (x and y) and horizontal and vertical displacements (u and v) of the correlated image. All displacement-based error metrics in this section are reported in physical displacement units (nm); pixel-wise DIC displacements were converted to nm using the calibrated image scale before calculating the metrics. Table 3 outlines the error metric characteristics and the error type. The image within the image quad analyzed (Figure 4) is denoted by the Position and Repetition columns. For some error metrics, the left 10% of the image was cropped, as indicated by the "Crop?" column to remove the LES from the given error metric. Also, for some error metrics, the displacement fields were normalized by subtracting the mean or median of the field, as indicated by the "Deviation" column. This is particularly relevant for Position 2 Repetition 1 images because the subtracted mean or median displacement field includes the rigid-body translation magnitude such that the reported values represent only residual error. The displacement components used in the error metric calculation are denoted by the u and v columns, where u is horizontal displacement and v is vertical displacement.

Error Metric	**Error type**	**Pos.**	**Rep.**	**Crop?**	**Deviation**	u	v
MAD (TD)	Time-dependent robust overall displacement	1	2	×	✓	✓	✓
MAD (PD)	Position-dependent robust overall displacement	2	1	×	✓	✓	✓
Max FD	Outlier-influenced overall displacement	1	2	×	×	✓	✓
Mean D	Immediate drift distortion	1	2	✓	×	✓	✓
Column Error	Left-edge smear (LES)	2	1	×	✓	✓	✓
U_{error} (TD)	Time-dependent horizontal skewing	1	2	✓	×	✓	×
U_{error} (PD)	Position-dependent horizontal skewing	2	1	✓	✓	✓	×
V_{error} (TD)	Time-dependent vertical skewing	1	2	✓	×	×	✓
V_{error} (PD)	Position-dependent vertical skewing	2	1	✓	✓	×	✓
$\varepsilon_{yy,\text{max}}$	Line jump severity	1	2	✓	×	×	✓
$\varepsilon_{yy,\text{rows}>1\%}$	Line jump quantity	1	2	✓	×	×	✓

Table 3: Error metrics for each SEM-DIC error type.

Median absolute deviation (MAD) is the median of the absolute values of the residuals from the median of the data. It provides an overall measurement of displacement error for a rigid-body translation by analyzing both the horizontal and vertical displacement fields. Specifically, it takes the vector magnitude of the horizontal and vertical displacement fields, and then subtracts the median magnitude from it. For each displacement measurement location i (i.e., each pixel/subset value in the displacement field), the displacement magnitude is defined as

$$d_i = \sqrt{u_i^2 + v_i^2}, \quad i = 1, \ldots, n \tag{3}$$

$$\tilde{d} = \mathrm{median}\left(d_i\right) \tag{4}$$

$$\mathrm{MAD} = \mathrm{median}\left(\left|d_i - \tilde{d}\right|\right), \tag{5}$$

where u_i and v_i are the horizontal and vertical displacement components, respectively, and n is the total number of displacement measurement locations in the field.

MAD is used in this study as a robust measure of the overall error of the data set because it is not influenced by outliers. MAD values were calculated for Position 1- Repetition 2 images and Position 2- Repetition 1 images within the image quad to differentiate position-dependent (PD) and time-dependent (TD) errors. In analyzing images both before and after translation, MAD can capture all of the error types.

Maximum False Displacement (Max FD) is the maximum vector magnitude of the horizontal and vertical displacement fields for the second image in a quad, with

$$\text{Max FD} = \max d_i. \tag{6}$$

This metric captures the maximum displacement in the image, meaning it is sensitive to outliers (unlike MAD).

Mean D is the mean of the vector magnitude of displacement of the images, calculated as

$$d_i = \sqrt{u_i^2 + v_i^2}, \quad i = 1, \ldots, n \tag{7}$$

$$\text{Mean D} = \text{mean}(d_i) \tag{8}$$

Mean D is evaluated on the pre-translated image in the quad to capture the time dependence of drift.

Column Error is the range of column mean difference from the mean for the total displacement field. This metric is evaluated on the translated image (Position 2-Repetition 1) due to the position dependence (PD) of the left-edge smear. For each DIC column, the field mean ($\bar{d}$) is subtracted from the displacement magnitude field ($d = \sqrt{u^2+v^2}$), and then the column error is the range of the column-wise means of $\left|d-\bar{d}\right|$. This metric relates to the LES, as shown in Figure 6.

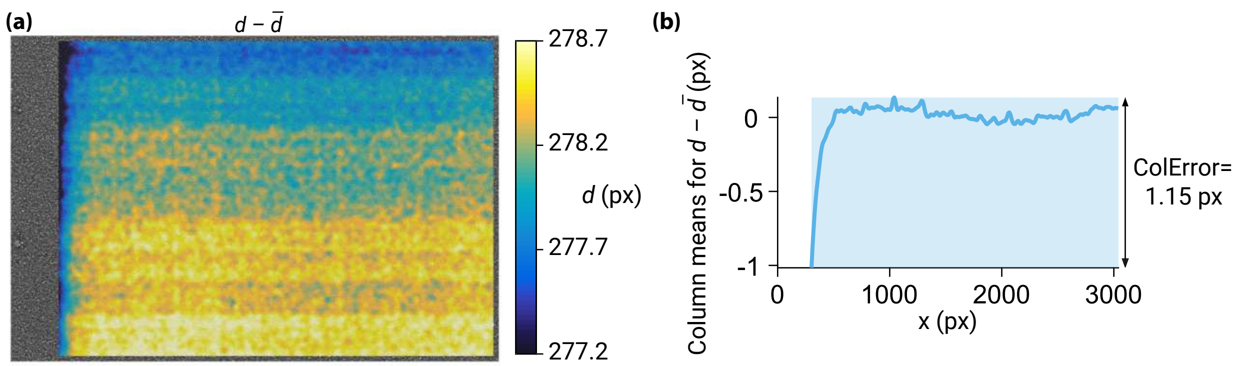

Figure 6: Representation of column error for a single image including (a) total displacement field and (b) calculated column error.

U_{error} is the range of column error difference from the mean of the cropped horizontal displacement (u) field. For every horizontal displacement field DIC point, the field mean is subtracted, and then the column means of those differences are taken. The overall image U_{error} is the range of the column means. Images before and after translation are analyzed with this metric. Post translation images are cropped to avoid the effect of the LES. This metric is designed to capture horizontal skewing across the image as a scalar value (neglecting the LES). Therefore, the time-dependent (Position 1 Repetition 2, TD) analysis of this error captures drift distortion in the horizontal field while the position-dependent (Position 2- Repetition 1, PD) analysis captures spatial distortion in the horizontal field.

V_{error} is the range of row error difference from the mean of the cropped vertical displacement (v) field. For every vertical displacement field DIC point, the field mean is subtracted, and then the row means of those differences are taken (Figure 7). The overall image V_{error} is the range of the row means. Both pre- and post-translation images are analyzed with this metric. Post-translation images are cropped to avoid the effect of the LES. This metric is designed to capture line jumps and vertical skewing (spatial and drift distortions) across the image as a scalar value, and was based on the work by Maraghechi et al.[21]. Therefore, the time-dependent (Position 1 Repetition 2, TD) analysis of this error captures drift distortion in the vertical field while the position-dependent (Position 2- Repetition 1, PD) analysis captures spatial distortion in the vertical field.

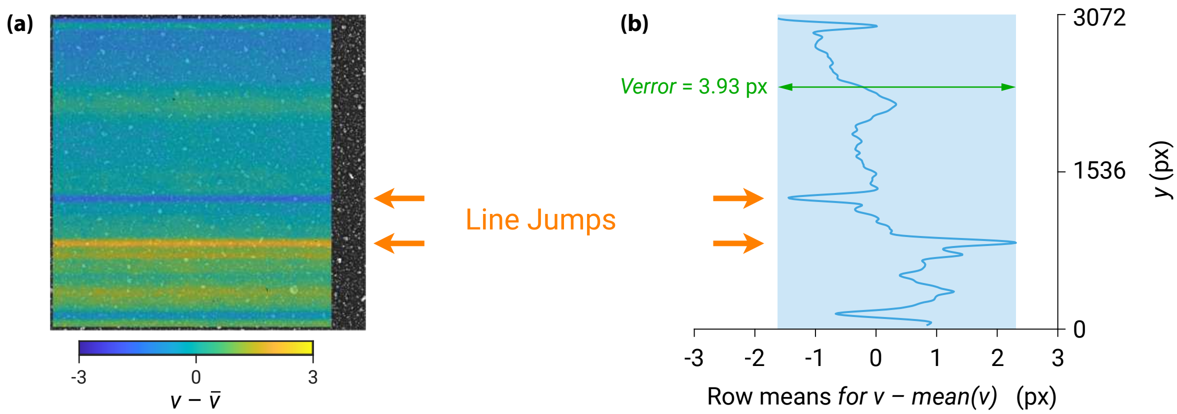

Figure 7: Illustration of the V_{error} with (a) the vertical displacement field minus the mean of the vertical displacement field and (b) the calculated row error. The image shown is Position 2- Repetition 2 from sample M at 30 kV, 1 nA, and 10 µs with no line or frame averaging.

$\varepsilon_{yy,\text{max}}$ is the maximum row-wise mean of the cropped ε_{yy} strain field. (Strains are calculated according to the parameters in Table S1.) This vertical strain metric differentiates the magnitude of line jump error from overall vertical skewing captured by V_{error}. Scanning errors are present throughout the image quad. Thus, $\varepsilon_{yy,\max}$ is measured on the pre-translated image in the quad in order to maximize the effective correlated area.

$\varepsilon_{yy,\text{rows}>1\%}$ captures the quantity of line jumps above 1% strain by indicating what fraction of lines in the image show a strain greater than 1%. Like $\varepsilon_{yy,\max}$, this metric is measured on the pre-translated, cropped image in the quad to maximize the effective correlated data area and negate the effect of the LES. Justification of the use of the 1% threshold is provided in Section S4.

2.4.3 Regressions and Main Effect Plots

Bivariate and multivariate regressions were used to quantify relationships between error metrics and beam conditions. For bivariate relationships, the Pearson correlation coefficient (r) was selected, as it is appropriate for evaluating linear relationships between continuous numerical variables (in contrast to metrics designed for categorical or non-numeric data, such as Cramér's V). For multivariate relationships, a multivariate linear regression (MVR) was performed.

The Pearson correlation coefficient (r) quantifies the strength and direction of a linear relationship between two variables by normalizing their covariance by the product of their standard deviations, enabling direct comparison of correlation strength across different parameter–error pairs. Values of r range from -1 to $+1$, where $r = +1$ indicates a perfect positive linear relationship, $r = -1$ indicates a perfect negative linear relationship, and $r \approx 0$ indicates little to no linear correlation. The statistical significance of each correlation was assessed using the associated p-value, which represents the probability of observing a correlation of equal or greater magnitude under the null hypothesis of no linear relationship between the variables; correlations were considered statistically significant for $p < 0.05$.

Multivariate linear regression was used to measure the weighted effect of each condition on a single error metric. The MATLAB regression learner application was used to determine the combined effect of each condition on each error metric. Within the application, beam current, beam voltage, dwell time, line averaging, and sample type were defined as the predictors. The application was set to use 75% of the data for training, with the remaining 25% held out for validation. Each error metric response was fit to a linear regression model.

Multivariate regression data is commonly presented in "main effect plots." For illustrative purposes, a main effect plot is shown below in Figure 8.

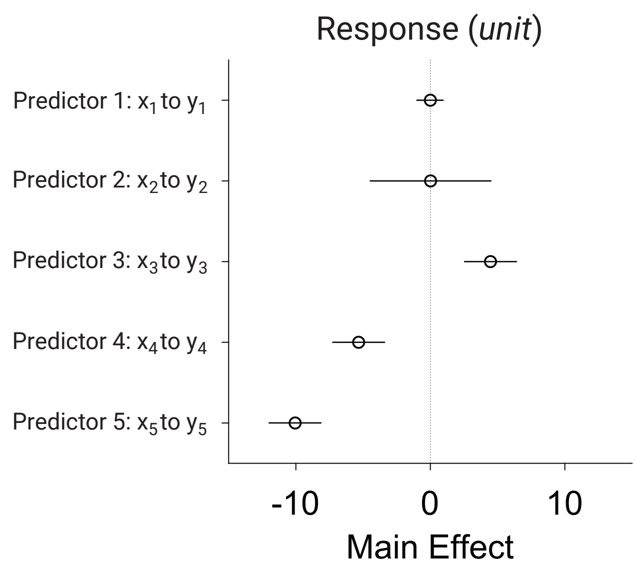

Figure 8: Example Main Effect plot showing the learned regression relationship between predictors and response.

These plots measure the effect to which each predictor increases or decreases the response. For the purpose of this study, the predictors are the image set conditions: beam voltage, beam current, dwell time, line averaging, and sample type (paired with magnification). Within the main effects plot, each predictor has its own measured range, the length of the line for each predictor represents the 95% confidence interval. As seen from Figure 8, predictor 2 has the greatest amount of uncertainty based on the length of the 95% confidence interval line. The point for each predictor represents the predicted response. For the multivariate regression analysis of the round robin, the main effect is presented as nanometer displacement (the unit of each error metric). Positive main effect points represent an increasing effect across the range of the predictor. Likewise, a negative main effect point represents a decreasing effect across the range of the predictor. Points that rest on the $x = 0$ line have no analyzed effect on the predictor (Figure 8 - Predictor 1 & 2). According to the main effect plot in Figure 8, Predictor 5 has the largest effect on the response. Specifically, Predictor 5 shows a decrease of 10 units for the bounds of its property (x_1 to y_1). Likewise, for Predictor 4, there is a decrease of 5 units attributed to varying Predictor 4 from its bounds of x_4 to y_4. Alternatively, across the range of x_3 to y_3 there is an increase of 5 units. With these main effects plots, we are able to visualize and determine which imaging conditions drive the responses for a particular error metric.

2.4.4 Box-and-Whisker Plots

Box plots are used extensively in this study to visualize the data and extract trends. Each box in the plot is characterized by a central mark representing the median. The lower and upper boundaries of the box correspond to the 25th and 75th percentiles, respectively. The whiskers extend to encompass the data points that are not outliers, while any outliers are individually marked with a '$\circ$' symbol on the plot. Outliers are defined as values that are more than 1.5 times the interquartile region (IQR) away from the top or bottom of the box. The IQR is the span of the box, between the 25th and 75th percentiles.

3 Results

3.1 Comparison of round-robin samples

The round-robin study captured images across five SEMs with the previously described beam/imaging parameters, as well as matching samples of two types (Ag and Au). While each sample type was fabricated in a batch and beam conditions were controlled, differences among the samples were found. Variations were found in the sample patterns, along with differences from image to image. These differences presumably arose from variations in lens adjustments (alignment and focus), noise, and brightness/contrast. For example, Figure 9 shows the differences in MIG and sigma for 1 µm portions of images from each SEM, captured for Au and Ag samples under the same beam conditions (20 kV, 1000 pA, 5 µs and no line averaging).

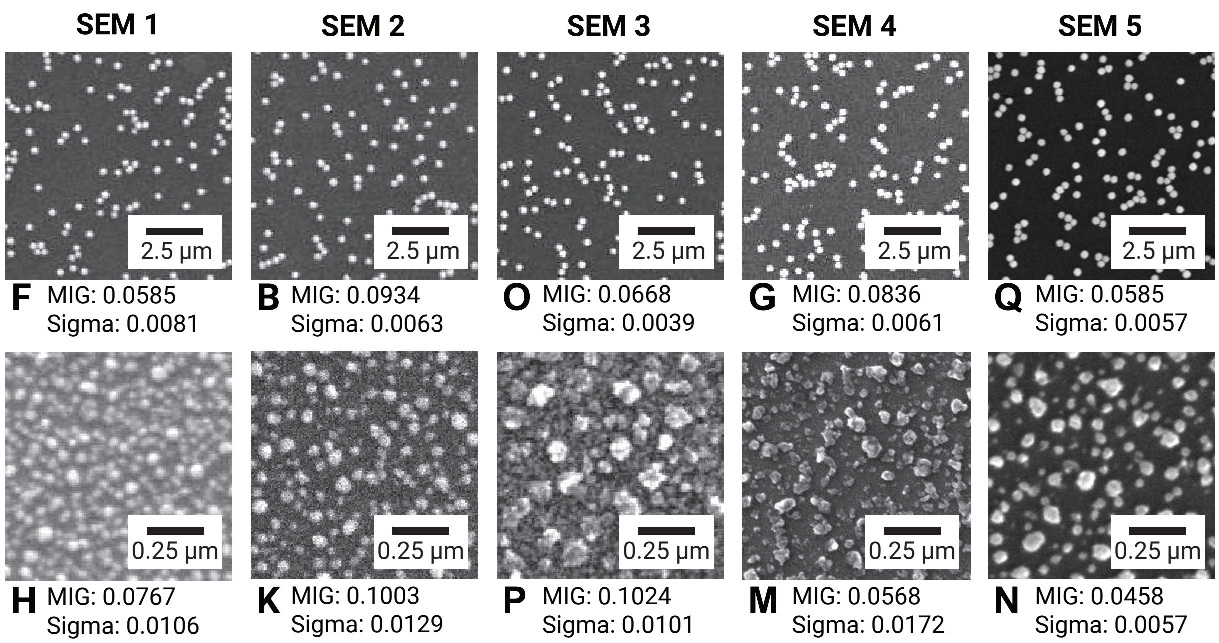

Figure 9: Portions of images from each of the five SEMs are shown for the Position 1, Repetition 2 images of all Au (top) and Ag (bottom) samples at the following conditions: 20 kV, 1000 pA, 5 µs, without line averaging.

Despite sample-to-sample variations, the pattern quality was reasonably consistent to permit an aggregate analysis of the round-robin results. Figure 10 outlines the average MIG and sigma values for the 12 reference images from each image set. The MIG and sigma from each image set are plotted from the 12 different reference images, taken at the start of each new location (as detailed in the round-robin procedure in Section 2.2). Ag samples (black boxes) and Au samples (blue boxes) maintained median MIG values between 0.05 and 0.10. Repeated samples across SEMs are denoted by color outline (e.g., the same Au sample was used for both image set F from SEM 1 and image set O from SEM 3). The median sigma remained near 0.005 px for the Au samples, and between 0.01 and 0.02 px for the Ag samples.

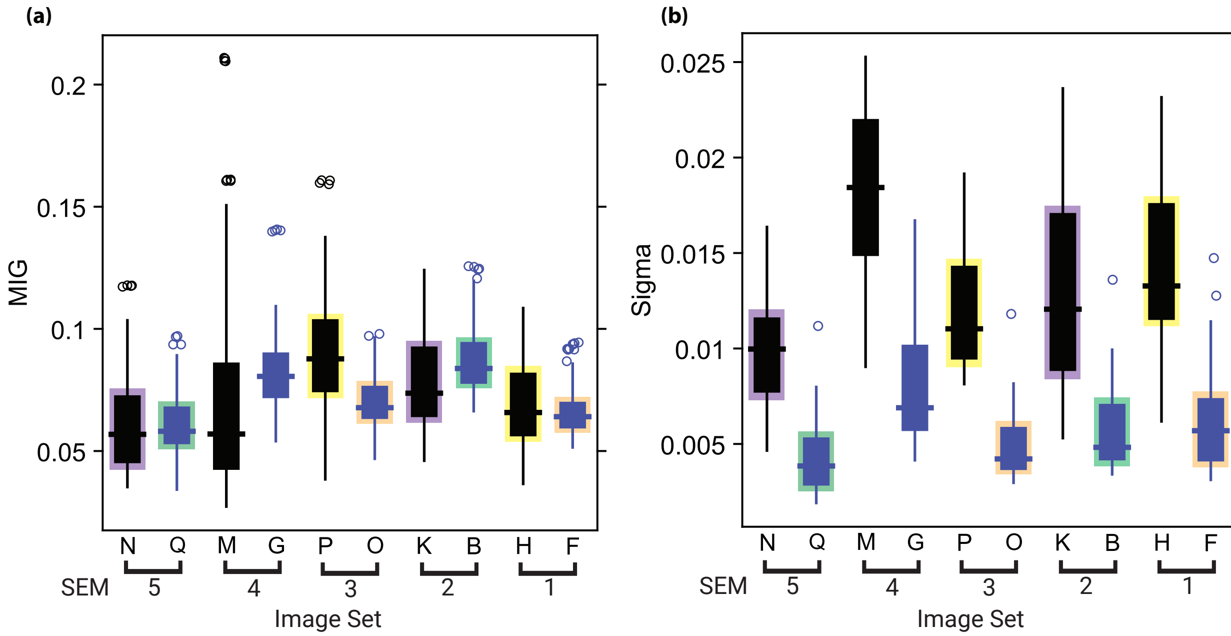

Figure 10: Average MIG and sigma of all reference images differentiated by sample type (Ag-Black, Au-Blue), sample repeated across SEMs (color outline), and microscope (SEM number).

While there are image-wise differences (Figure 9), the quantitative differences among samples and SEMs are reasonably consistent, with no clear outliers (Figure 10). Overall, the round-robin data set is representative of SEM-DIC results from samples patterned with the two different techniques used here. In SEM-DIC applications, it is not expected or reasonable to assume that variations (such as user-to-user lens preparation differences) will be absent. Therefore, the round-robin results are analyzed in aggregate, not in an effort to identify optimum individual microscopes, but rather to find and assess the overall trends across SEMs and beam conditions for advancing the practice of SEM-DIC.

3.2 Error Magnitude Comparison

An analysis was performed to determine the relative importance of each error type on SEM-DIC results, broken down by the Ag and Au samples. In the main round-robin study, the high-magnification (Ag) samples correspond to the 20 µm HFW condition, while the low-magnification (Au) samples correspond to the 200 µm HFW condition;

therefore, magnification, patterning method, conductivity, and sample type are coupled in the aggregate comparisons.

Figure 11 outlines the results for the Ag samples before translating (Position 1 Repetition 2, TD) and after translating (Position 2 Repetition 1, PD). For the Ag samples, the largest error type was Max FD (an outlier-influenced metric of the overall image quality), contributing to a displacement error of about 20 nm. After Max FD, the second largest error type for the Ag samples was translational drift distortion (measured by Mean D), at about 15 nm. The third largest error contribution was from the LES (measured by Col Error), at about 10 nm. Vertical skewing and line jumps (measured through V_{error}) were near 10 nm error before and after translating. Additionally, displacement from horizontal skewing, measured by U_{error}, was small (less than 2 nm) in the Ag images.

For additional context, the number of outliers identified by the boxplot quartile criterion (1.5×IQR) was extracted directly from the plotted distributions. Across the Ag datasets shown here, each error metric contains 25, 31, 27, 18, 34, 14, 19, 18, and 43 outliers, respectively, out of 336 total data points, providing perspective on the fraction of measurements that fall outside the interquartile range.

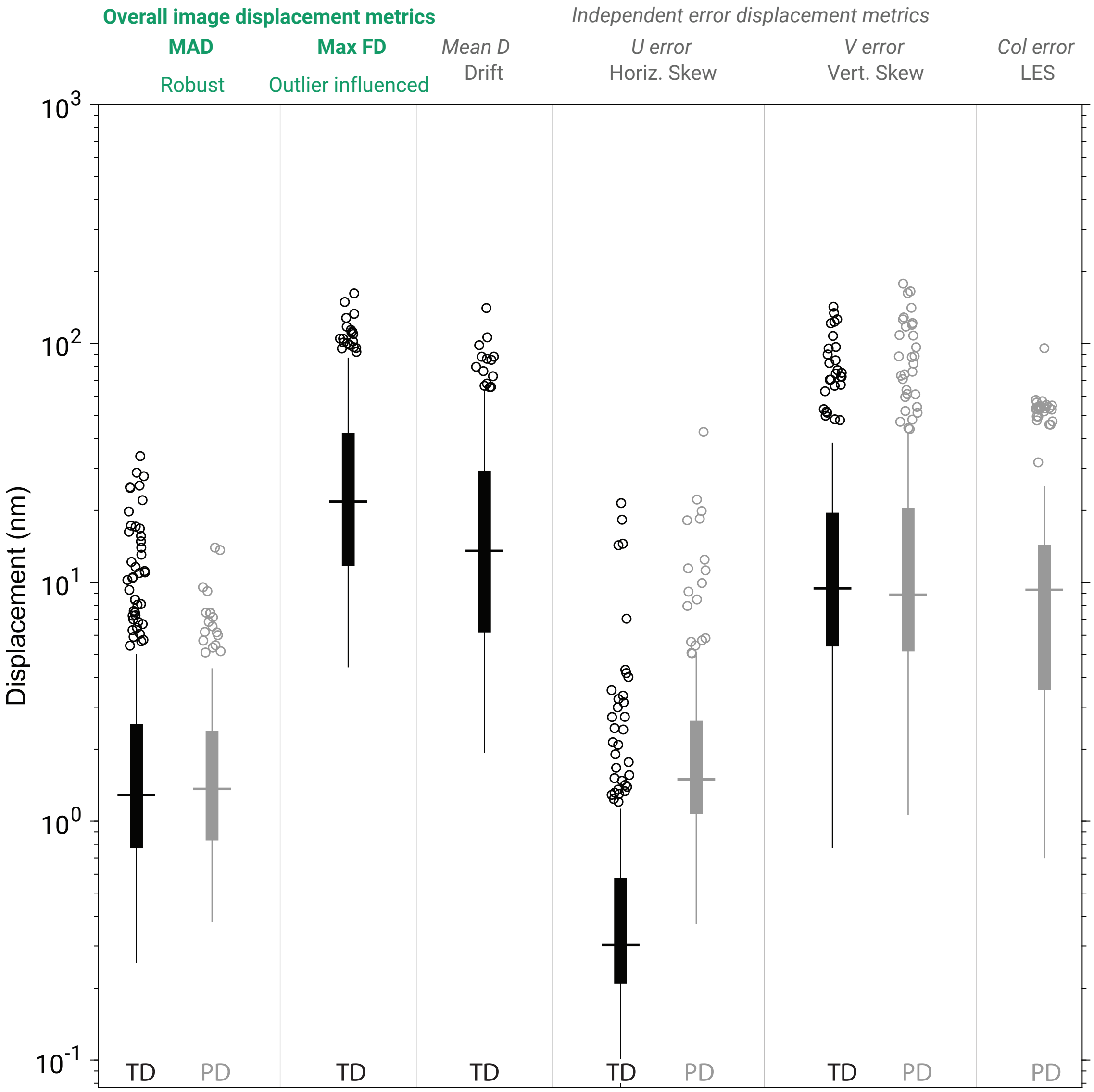

Figure 11: Quantitative displacement measurements captured across all beam and hardware conditions with each developed error metric for the high magnification Ag samples. Black boxplots represent time-dependent error metrics (Position 1 Repetition 2, TD), while grey boxplots represent position-dependent error metrics (Position 2- Repetition 1, PD). Overall image displacement metrics probe all error types, while independent error displacement metrics capture drift distortion in both the horizontal and vertical displacement fields (Mean D), drift distortions in the horizontal and vertical displacement fields independently (U_{error} (TD), and V_{error} (TD)), spatial distortions in the horizontal and vertical displacement fields independently (U_{error} (PD), and V_{error} (PD)), and the left-edge smear (LES, Column Error).

For the Au samples (Figure 12), the left-edge smear (LES, Col Error) contributed most to the SEM-DIC error (about 100 nm). Another large error type for the Au samples was horizontal skewing (U_{error}) in the translated images (about 15 nm), in contrast with the small U_{error} in the Ag samples. In a similar way as the Ag images, line jumps and vertical skewing (V_{error}) were relatively consistent throughout the image quad.

Comparing the Mean D with U_{error} (TD) and V_{error} (TD) in both Figures 11 and 12, we observe that while the Mean D captures drift distortions in both the horizontal and vertical displacement fields, the magnitudes of the individual components (u and v) differ significantly. Similarly, when comparing U_{error} (PD) and V_{error} (PD), the magnitudes of u and v are again shown to differ. Due to the rastering nature of the SEM, it is not expected that spatial and drift distortions in the horizontal and vertical fields will behave the same under different beam and scanning settings. Therefore, in the remainder of this study, we refer to the distortions in the horizontal and vertical fields independently as 'horizontal skewing' (encompassing both spatial and drift distortions in the horizontal displacement field) and 'vertical skewing' (encompassing both spatial and drift distortions in the vertical displacement field).

Across the Au datasets, each error metric contains 23, 9, 15, 22, 23, 1, 7, 12, and 23 outliers, respectively, out of 336 total data points, providing perspective on the fraction of measurements that fall outside the interquartile range.

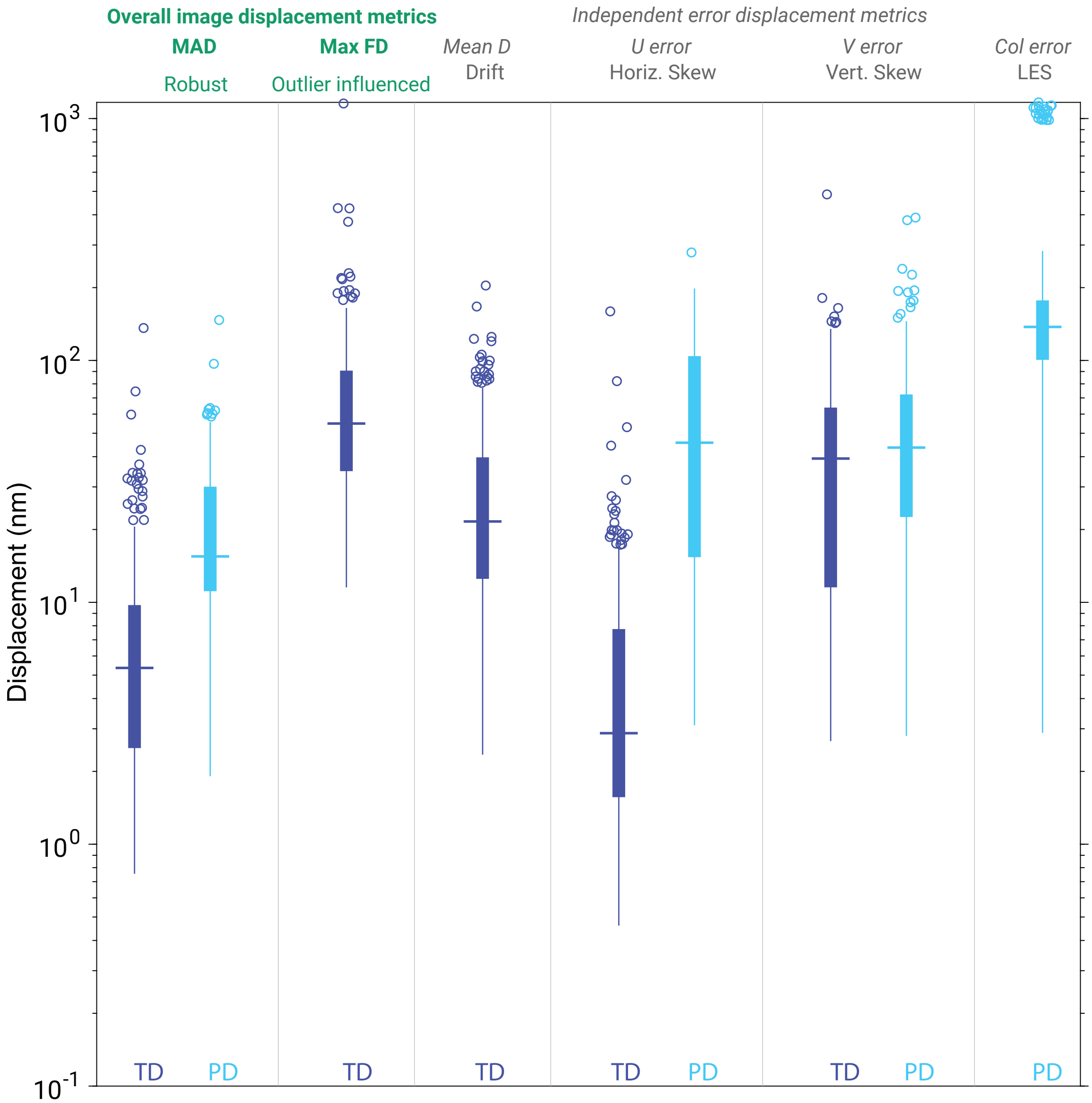

Figure 12: Quantitative displacement measurements captured across all beam and hardware conditions with each developed error metric for the low magnification Au samples. Dark blue boxplots represent time-dependent error metrics (Position 1 Repetition 2, TD), while light blue boxplots represent position-dependent error metrics (Position 2- Repetition 1, PD). Overall image displacement metrics probe all error types, while independent error displacement metrics capture drift distortion in both the horizontal and vertical displacement fields (Mean D), drift distortions in the horizontal and vertical displacement fields independently (U_{error} (TD), and V_{error} (TD)), spatial distortions in the horizontal and vertical displacement fields independently (U_{error} (PD), and V_{error} (PD)), and the left-edge smear (LES, Column Error).

3.3 The Left-edge smear

The left-edge smear (LES) was captured from the Column Error for the uncropped images at Position 2 Repetition 1 (PD). Both the vertical and horizontal displacement fields are encompassed within Col Error. The round-robin experiment parameters designate certain conditions as predictors for multivariate regression, including beam voltage, current, dwell time, line averaging, and sample type (which incorporates magnification and general sample differences such as conductivity and pattern type). In Figure 13, the results of the multivariate regression analysis for LES are outlined.

The analysis reveals that beam voltage, beam current, and line averaging have minimal influence on LES. Notably, the combined variables of sample type and magnification emerged as the primary factor affecting LES. It is demonstrated that Au samples generally exhibited an error about 240 nm greater than their Ag counterparts. Beyond sample type, the multivariate regression results indicate that longer dwell times result in an average reduction of LES by about 180 nm across the aggregate model, with Pearson values of $r_{Ag} = -0.46$ and $r_{Au} = -0.39$. However, because LES magnitude is strongly coupled to HFW/magnification, this dwell-time effect should be interpreted alongside the substantially larger LES observed in the low-magnification (Au) images. There were no statistically significant correlations between LES and either beam voltage or beam current (all with $p \geq 0.05$ for both Ag and Au samples).

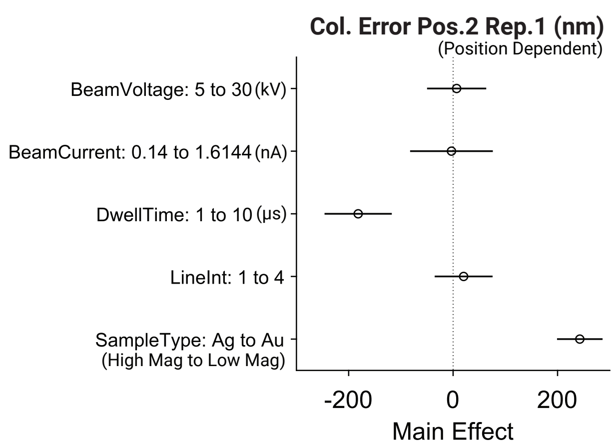

Figure 13: Column error at Position 2 Repetition 1 (LES) with the multivariate regression main effect results.

The correlation between dwell time, sample type/magnification, and LES is further illustrated in Figure 14. There is a clear decrease in LES with longer dwell times. The magnitude of Col Error is about an order of magnitude higher in the Au samples, as the HFW for the Au samples is also an order of magnitude larger. While the effects of sample type and magnification cannot be decoupled from the round-robin data, the impact of magnification on LES is addressed later in Section 3.7.3.

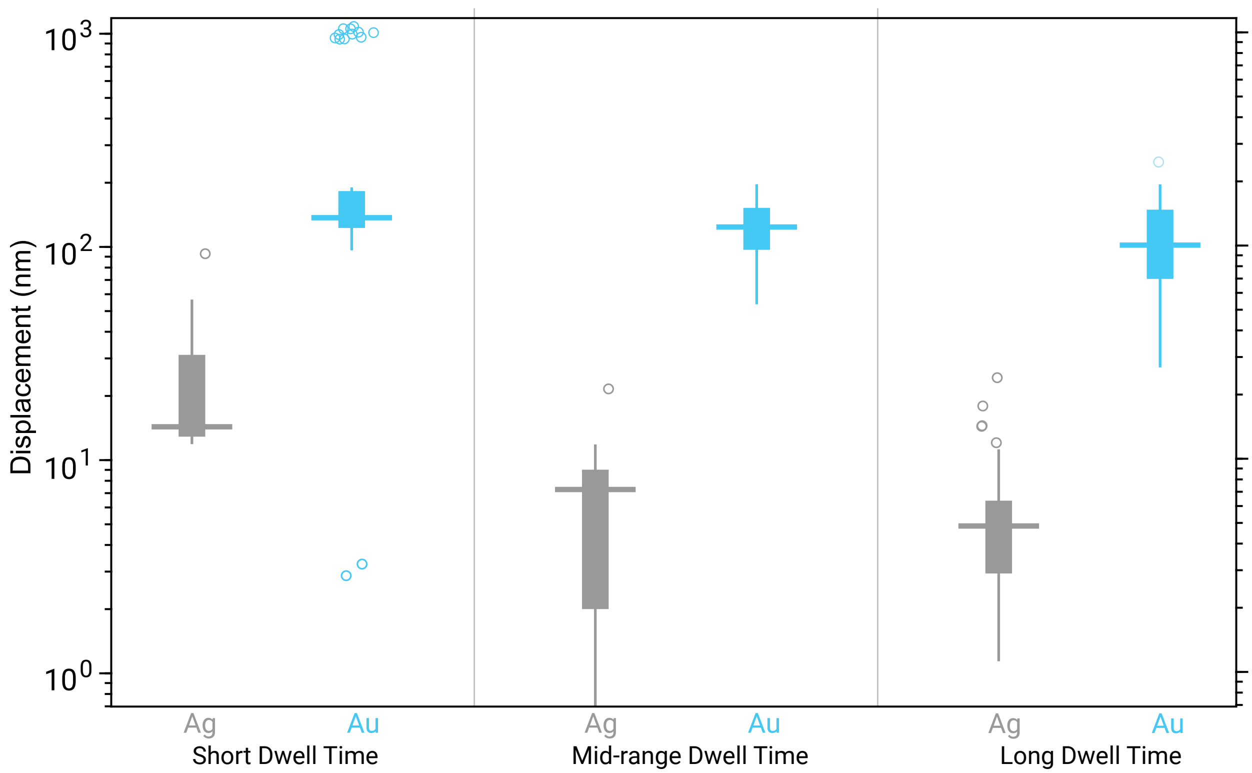

Figure 14: Column error at Position 2 Repetition 1 (LES) vs. Dwell time. Data are grouped into short ($< 3\,\mu$s), mid-range (3–6 μs), and long ($> 6\,\mu$s) dwell-time regimes.

3.4 Short-term drift distortion

Figure 15 outlines the multivariate regression results for drift distortions over a short time span (minutes). According to these results, drift distortion increases the error by up to 20 nm for long dwell times (10 µs). Additionally, line averaging is shown to elevate drift distortions by approximately 11.5 nm over 4 line averagings.

Literature has pointed out that high current and voltage settings can significantly increase drift distortions[10,18]. However, the current findings suggest that drift over short time scales is more heavily influenced by sample type and magnification, line averaging, and dwell time. The relatively small Pearson values for beam voltage ($r_{Ag} = -0.04$) and beam current ($r_{Ag} = -0.08$) suggest that beam conditions do not have a significant effect on the Ag sample type (p ≥ 0.05). Conversely, for the Au samples, it is observed that higher beam currents increase drift distortion ($r_{Au} = 0.18$). However, as in the Ag samples, beam voltage does not show a statistically significant correlation with drift ($r_{Au} = -0.14$, $p \geq 0.05$). The disparity in Pearson values across sample types for the beam conditions aligns with the multivariate regression result shown for sample type.

Quantitatively, the regression model (Figure 15) estimates that low-magnification Au samples generally have 7.0 nm more drift distortion compared to high-magnification Ag samples. In absolute displacement, this means that drift is worse at low magnifications, which contradicts previous work that has shown drift to be worse at high magnifications[10]. However, the low-magnification images have an order of magnitude larger field of view, so the Mean D error of ≈ 20 nm in the Au samples (Figure 12) is relatively low compared to the ≈ 13 nm average Mean D error in the Ag samples (Figure 11). Additional studies (Sections 3.7.2 and 3.7.3, and SI: Conductivity) delve into the drift/sample type and drift/magnification relationships more thoroughly.

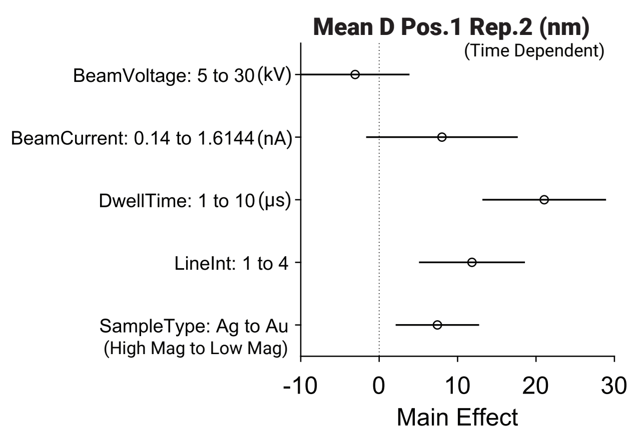

Figure 15: Mean D Position 1- Repetition 2 (Short-Term Drift Distortion) multivariate regression main effect results

Dwell time emerges as the most influential factor on short-term drift, contributing to an increase in drift distortions by an average of 20.7 nm across the 1–10 µs range. Additionally, dwell time exhibits a linear correlation for both sample sets, as evidenced by Pearson values of $r_{Ag} = 0.29$ and $r_{Au} = 0.29$ ($p < 0.05$).

The relationship between drift distortion and dwell time is further illustrated in Figure 16. Across an aggregate data set encompassing all conditions and magnifications, an evident trend emerges: as dwell time increases, drift displacement also increases. This finding aligns with existing literature (Table 1). Since drift distortions are time-dependent and charging tends to accumulate with longer beam and sample contact, the prolonged dwell times contribute to the amplification of this error.

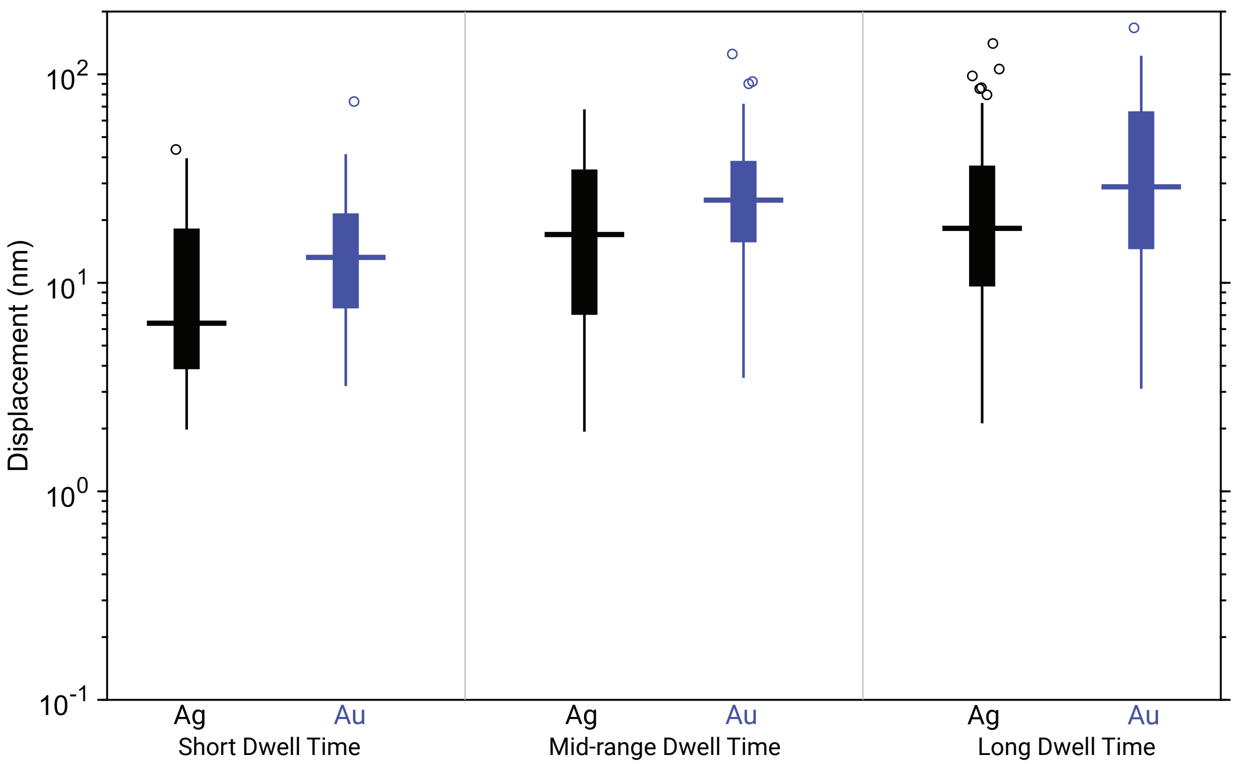

Figure 16: Mean D Position 1- Repetition 2 (Short-Term Drift Distortion) vs. Dwell time. Data are grouped into short ($< 3\,\mu s$), mid-range (3–$6\,\mu s$), and long ($> 6\,\mu s$) dwell-time regimes.

3.5 Horizontal and Vertical Skewing (time-dependent drift and position-dependent spatial distortions)

In SEM-DIC literature, both spatial and drift distortions are defined in some capacity as a skewing in both the horizontal and vertical displacement fields. While drift distortions may manifest as a rigid body displacement, these errors also have a time-dependent skewing characteristic[10]. Alternatively, the literature compares spatial distortions to optical barreling[8]. In our results, drift distortions are captured in the second image of each image quad (Position 1 Repetition 2, TD), while spatial distortions are captured in the third image of each quad (Position 2 Repetition 1, PD).

In this section of the study, both components of vertical and horizontal error metrics are analyzed independently to better understand SEM error. Because of the raster nature of the SEM, it should not be assumed that these errors manifest from the same conditions and are therefore addressed independently here. For vertical skewing, V_{error} was established. It captures the overall error in the vertical displacement field (including vertical stretching and line jump errors), and it is measured for both the pre-translated image (Position 1 Repetition 2, TD) and the post-translated image (Position 2 Repetition 1, PD). Comparably, U_{error} measures horizontal skewing in the time-dependent case (Position 1 Repetition 2, TD), which probes drift distortions, and in the position-dependent case (Position 2 Repetition 1, PD), which probes spatial distortions. Related to common SEM-DIC practice, the left 10% of each image is cropped to negate the effect of the LES in the U_{error} and V_{error} analyses.

3.5.1 Horizontal Skewing

Based on the multivariate results presented in Figure 17, it is evident that sample type is the primary contributor to horizontal skewing. It is important to note that samples differ in both patterning technique and imaging magnification. Concerning sample type, the horizontal skewing error increases by 5.8 nm for Au without translation (drift distortion, TD) and by 64 nm after translation (spatial distortion, PD). In contrast, for Position 1-

Repetition 2 (TD) voltage ($r_{Ag} = -0.16$, $r_{Au} = -0.06$), current ($r_{Ag} = -0.13$, $r_{Au} = 0.07$) and scanning conditions (dwell time: $r_{Ag} = 0.10$, $r_{Au} = 0.10$), the multivariate regression does not reveal significant trends ($p \geq 0.05$). Similarly, for Position 2- Repetition 1 (PD) voltage ($r_{Ag} = 0.11$), beam current ($r_{Ag} = -0.10$, $r_{Au} = 0.10$), and scanning conditions (dwell time: $r_{Ag} = 0.03$, $r_{Au} = -0.06$), the multivariate regression does not reveal highly significant trends, with the exception of a slight correlation in beam voltage ($r_{Au} = -0.22$, $p < 0.05$).

These Pearson values are in accordance with the MVR results showing very little influence from any condition with the exception of sample type.

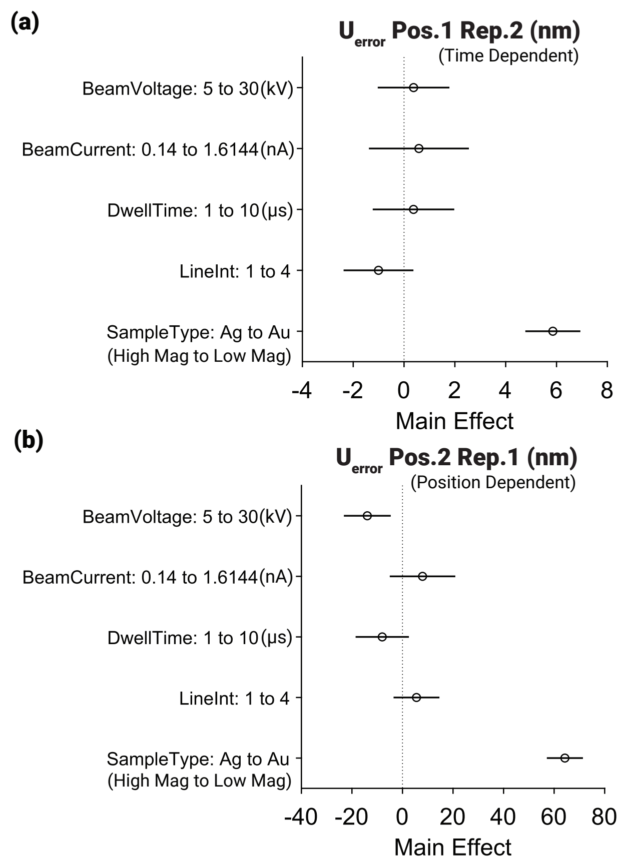

Figure 17: U_{error} Position 1- Repetition 2 (TD horizontal skewing) and U_{error} Position 2- Repetition 1 (PD horizontal skewing) multivariate linear regression results

In Figure 18, the relationship between beam voltage and sample type is compiled for the conditions where Pearson results indicated a slight degree of correlation. As suggested by the Pearson values, the regression trend for Au samples with increasing voltage ($r_{Au} = -0.22$) is more pronounced than the increasing error observed in the Ag samples ($r_{Ag} = 0.11$). At present, the reason for the opposite responses of U_{error} (PD) to beam voltage in the Au and Ag image sets is unclear. However, this comparison underscores the necessity of analyzing Pearson values in tandem with multivariate regression results.

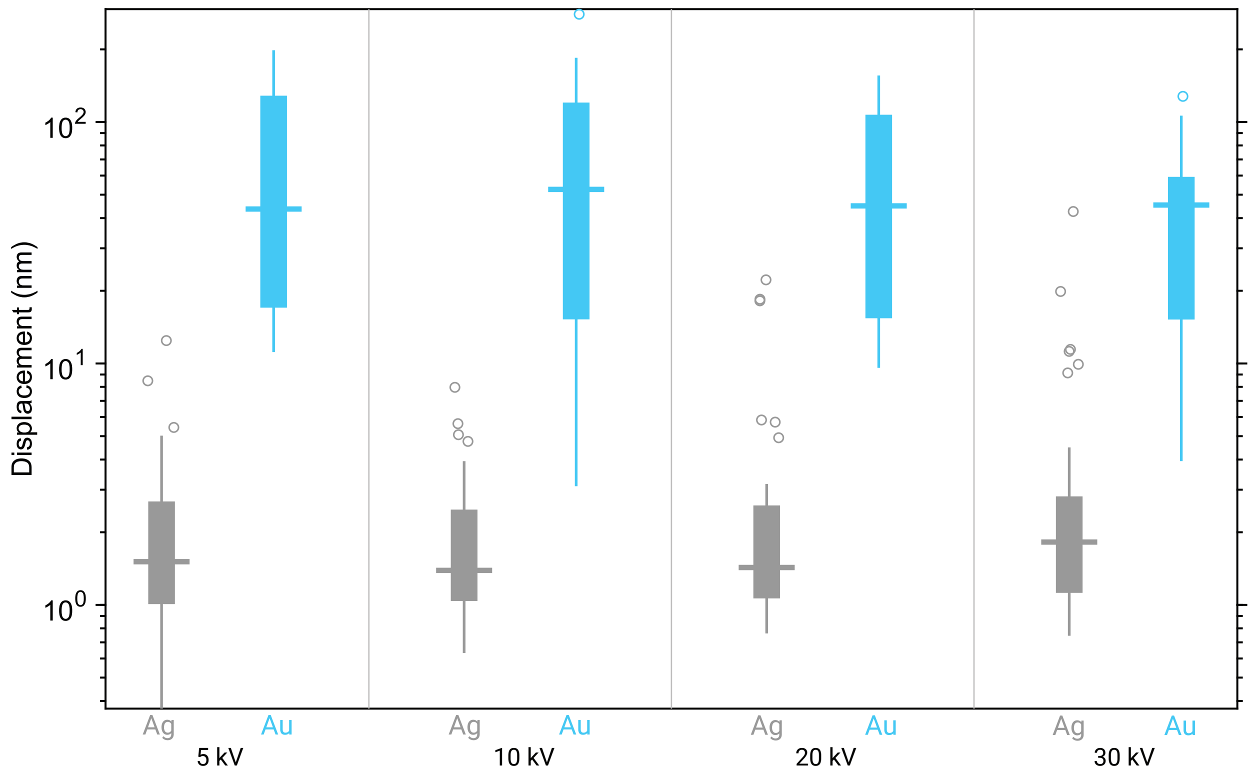

Figure 18: U_{error} Position 2 Repetition 1 (PD, spatial distortion) vs. beam voltage

In conclusion, both the multivariate regression model and Pearson values indicate that beam and scanning conditions have limited influence on horizontal skewing. The predominant factor shaping horizontal skewing is sample type, with magnification playing a coupled role. To disentangle sample type from magnification, a more in-depth exploration of this relationship is undertaken in Section 3.7.3, SI: Conductivity.

3.5.2 Vertical Skewing

The main effect plots for multivariate regression in V_{error} for Position 1 Repetition 2 (TD) and Position 2 Repetition 1 (PD) (Figure 19) exhibit almost identical predictor influences.

In both cases, dwell time, sample type, and line averaging are highlighted as having the most substantial impact.

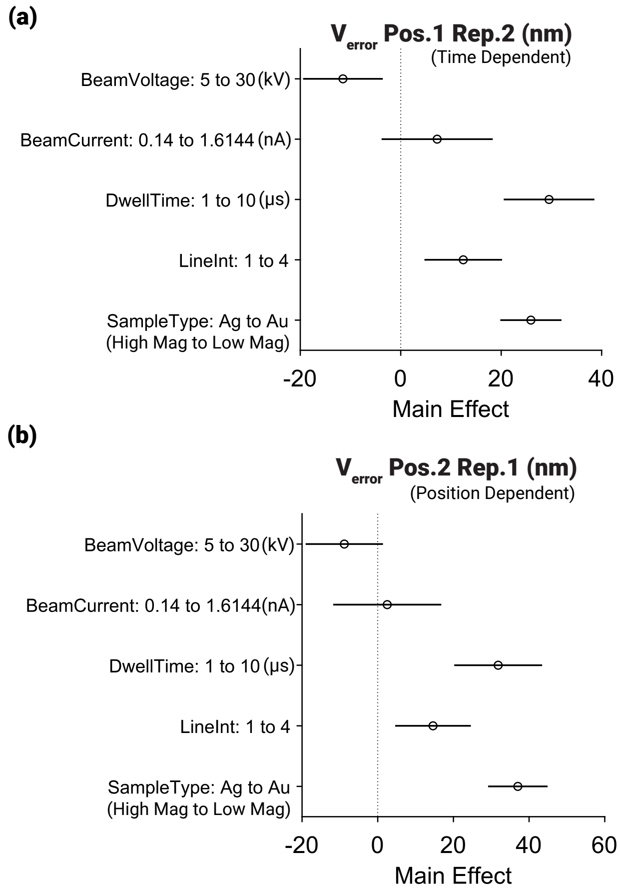

Figure 19: V_{error} Position 1- Repetition 2 (time-dependent vertical skewing) and V_{error} Position 2- Repetition 1 (position-dependent vertical skewing) multivariate linear regression results

For both Position 1 and 2, Pearson values indicate no significant correlation for the relationship between beam voltage and V_{error} in the Ag samples (Position 1: $r_{Ag} = -0.10$, Position 2: $r_{Ag} = 0.03$ ($p \geq 0.05$)). Conversely, for the Au samples, Pearson values suggest a degree of correlation between beam voltage and V_{error} (Position 1: $r_{Au} = -0.21$, Position 2: $r_{Au} = -0.25$ $p < 0.05$). The voltage results for Au samples are detailed in Figure 20.

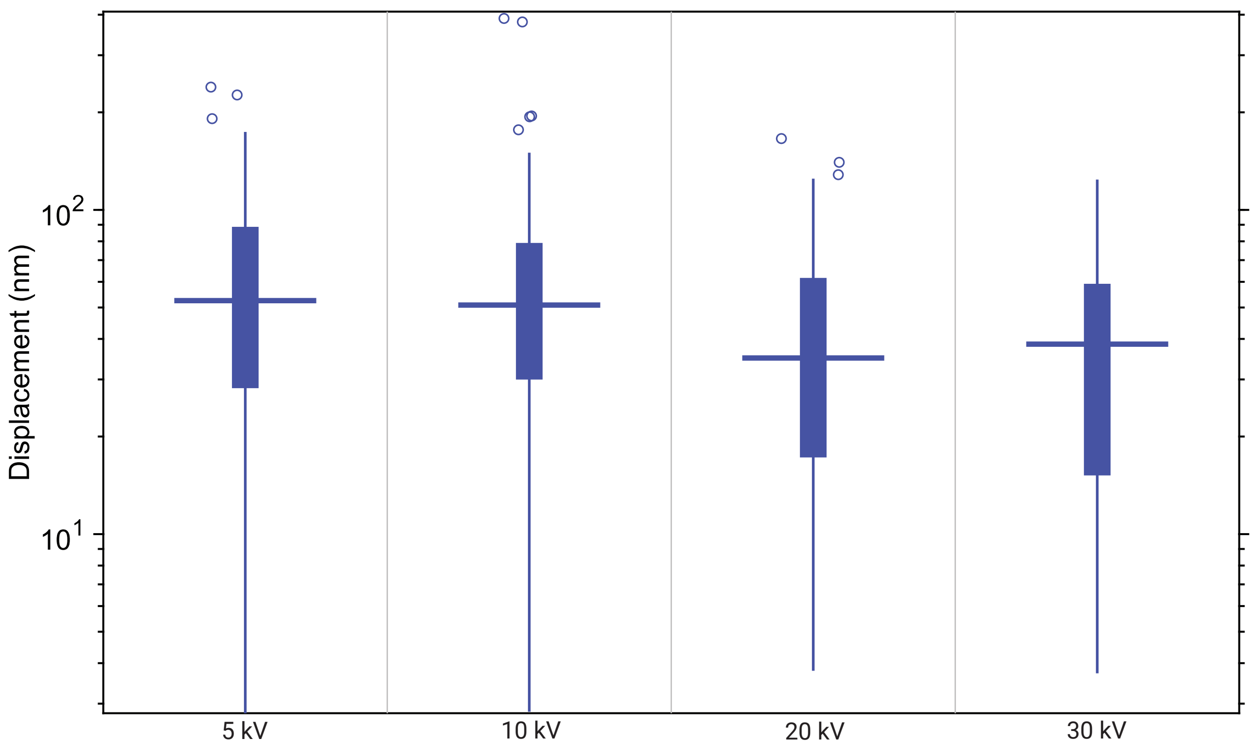

Figure 20: Au Sample V_{error} aggregate Position 2 Repetition 1 and Position 1 Repetition 2 vs. Voltage

This difference between sample types is also reflected by sample type being one of the largest predictors of V_{error} (along with dwell time), as shown in Figure 19. It is worth noting that Ag and Au samples vary in both image magnification and conductivity. Further details on the effects of magnification and sample conductivity are outlined in Section 3.7.4, SI: Conductivity.

Based on the MVR and Pearson correlation coefficient results, it appears that beam current is unlikely to significantly influence V_{error} for both the Au and Ag image sets (Position 1: $r_{Au} = -0.01$, $r_{Ag} = 0.13$; Position 2: $r_{Au} = 0.01$, $r_{Ag} = 0.06$). Overall, the reduction in the magnitude of V_{error} from both beam voltage and beam current is negligible compared to the impact of dwell time and sample type. Alternatively, for both Positions 1 and 2, line averaging is observed to increase V_{error}. Finally, apart from sample type, dwell time emerges as the primary influence of V_{error} for both Positions and samples. Pearson values of $r_{Ag} = 0.38$ and $r_{Au} = 0.19$ were calculated to quantify the correlation between Position 1 V_{error} and dwell time. Similarly, Pearson values of $r_{Ag} = 0.22$ and $r_{Au} = 0.32$ were calculated for Position 2. Pearson values relating longer dwell times to increased vertical skewing and line jumps are statistically significant across both positions and sample types ($p < 0.05$). Figure 21 illustrates a linear correlation between longer dwell times increasing V_{error} for both Position 1 and Position 2, across both Ag and Au samples.

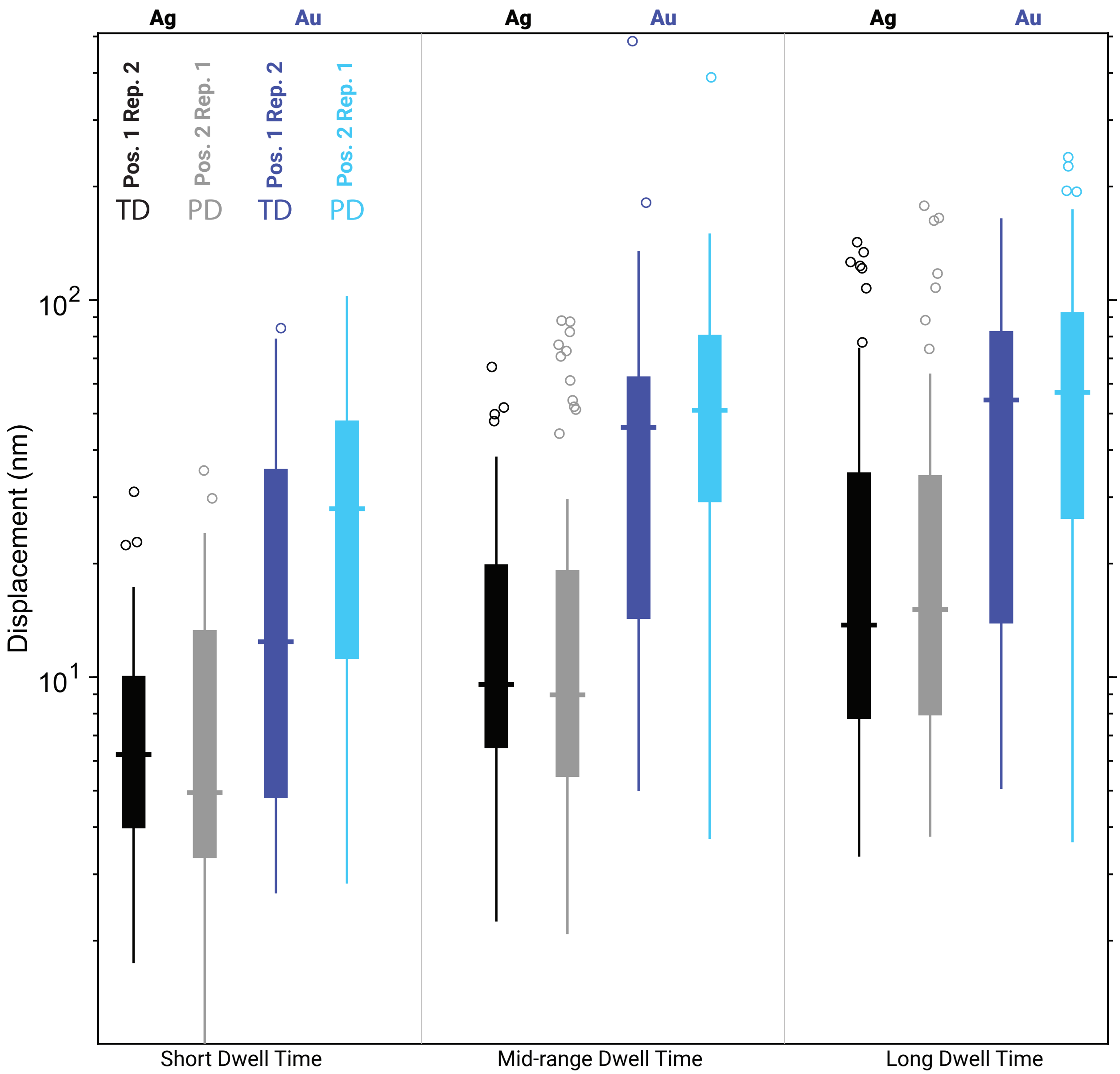

Figure 21: V_{error} Position 1- Repetition 2 (time-dependent, TD) and Position 2-Repetition 1 (position-dependent, PD) vs. Dwell time. Illustrating that in both time and position-dependent cases, V_{error} increases with long dwell times. Data are grouped into short ($< 3\,\mu$s), mid-range (3–6 μs), and long ($> 6\,\mu$s) dwell-time regimes.

Figure 22 shows the effect of dwell time on the vertical displacement fields of Position 1, Repetition 2, 20 kV, 500 pA images for all image sets. For most samples, it is clear that as dwell time increases, the variations in vertical displacements increase.

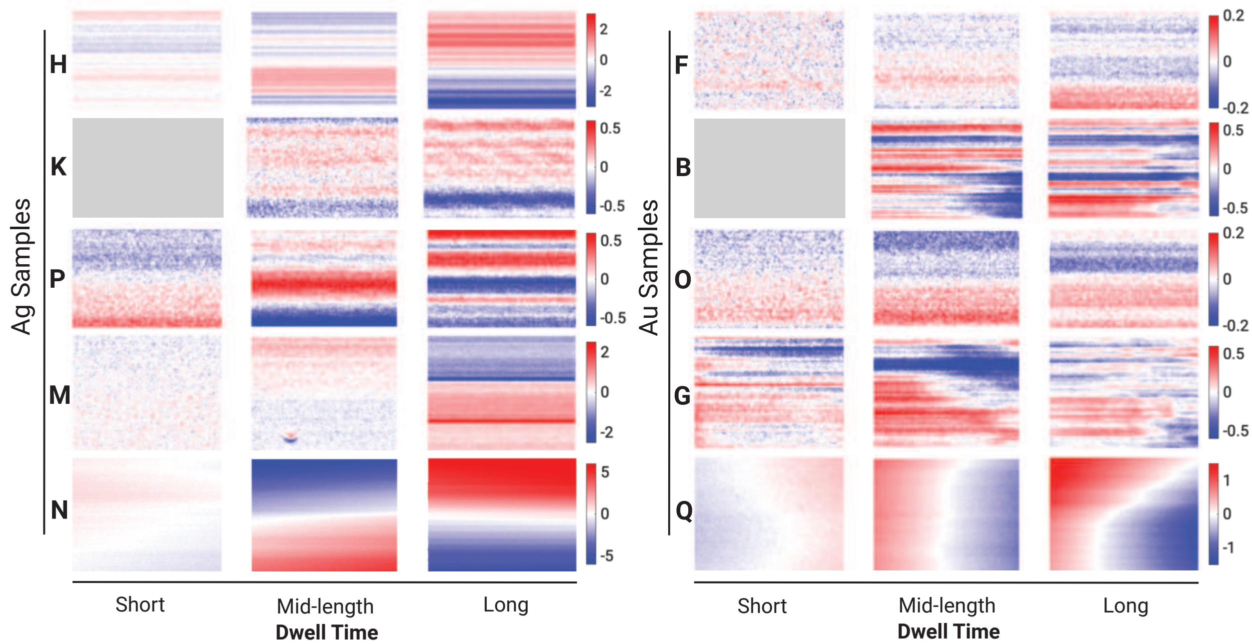

Figure 22: Vertical displacement field map of Position 1- Repetition 2 (time-dependent, TD), 20 kV, 500 pA, for all image sets differentiated by dwell time range classes. Color bars representing the vertical displacement (nm) are normalized across each image set (represented by each unique sample letter). Data are grouped into short ($< 3\,\mu$s), mid-range (3–6$\,\mu$s), and long ($> 6\,\mu$s) dwell-time regimes.

3.6 Line Jumps

As detailed in Section 2.4.2, two error metrics were employed to quantify line jump errors: (1) $\varepsilon_{yy,\max}$, which represents the magnitude of the largest line jump error, and (2) $\varepsilon_{yy,\text{rows}>1\%}$, indicating the percentage of rows with ε_{yy} above 1% strain. These metrics provide a more nuanced depiction of the intensity and quantity of line jumps compared to V_{error}.

In Figure 23, the multivariate linear regression analysis for $\varepsilon_{yy,\max}$ and $\varepsilon_{yy,\text{rows}>1\%}$ is outlined. Both error metrics are influenced almost identically by each condition.

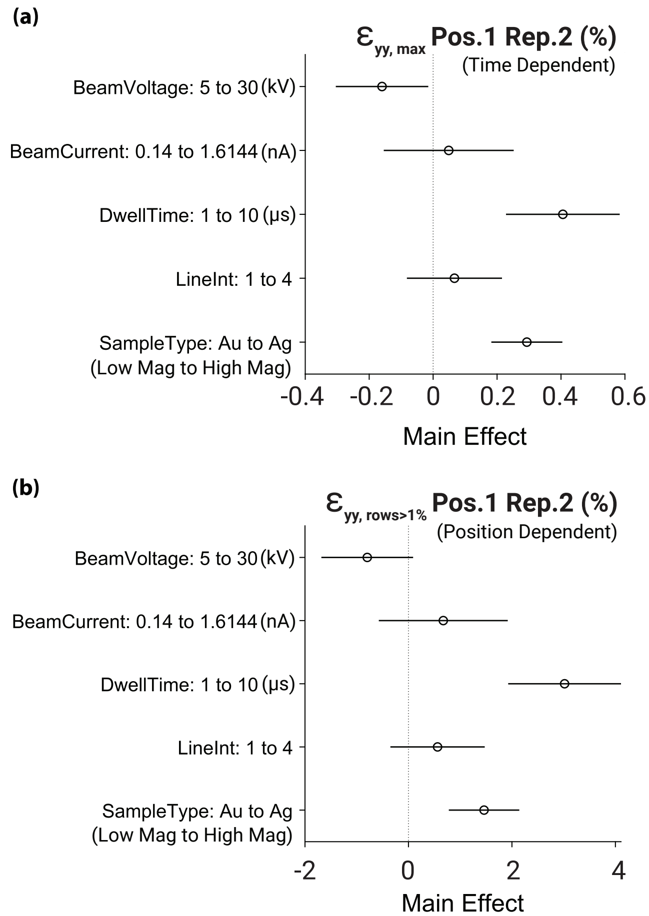

Figure 23: (a) $\varepsilon_{yy,\max}$ Position 1- Repetition 2 multivariate regression main effect results (b) Percent rows greater than 1% ($\varepsilon_{yy,\text{rows}>1\%}$) multivariate regression main effect results

Dwell time stands out as the most influential factor affecting line jump error in both the MVR analysis and Pearson values. Over dwell times ranging from 1 to 10 μs, the maximum line jump strain increases by 0.4% in $\varepsilon_{yy,\max}$. While this may seem like a small strain, it represents the entire row. Additionally, the similarity in the multivariate regression results for $\varepsilon_{yy,\max}$ and $\varepsilon_{yy,\text{rows}>1\%}$ indicates that the significance of line jumps also increases with dwell time. Therefore, dwell time has a significant compounding effect on line jump error, resulting in a 3.0% increase in $\varepsilon_{yy,\text{rows}>1\%}$ at long dwell times compared to short dwell times (Figure 24).

Apart from dwell time, sample type or magnification exhibits a notable impact on line jump error (Figure 23). The multivariate regression results reveal that the high-magnification Ag samples experience a 0.2% increase in $\varepsilon_{yy,\max}$ and a 1.4% increase in $\varepsilon_{yy,\text{rows}>1\%}$. Pearson values also underscore the dependence of sample type on line jump error when specifically looking at the sample-dwell time influence ($p < 0.05$). For the high-magnification images of the Ag samples, significant Pearson values of $r_{Ag} = 0.33$ and $r_{Ag} = 0.31$ were calculated for $\varepsilon_{yy,\max}$ and $\varepsilon_{yy,\text{rows}>1\%}$, respectively. In contrast, Pearson values indicating almost no correlation were calculated for the low-magnification images of the Au samples: $r_{Au} = 0.07$ for $\varepsilon_{yy,\max}$ and $r_{Au} = 0.05$ for $\varepsilon_{yy,\text{rows}>1\%}$.

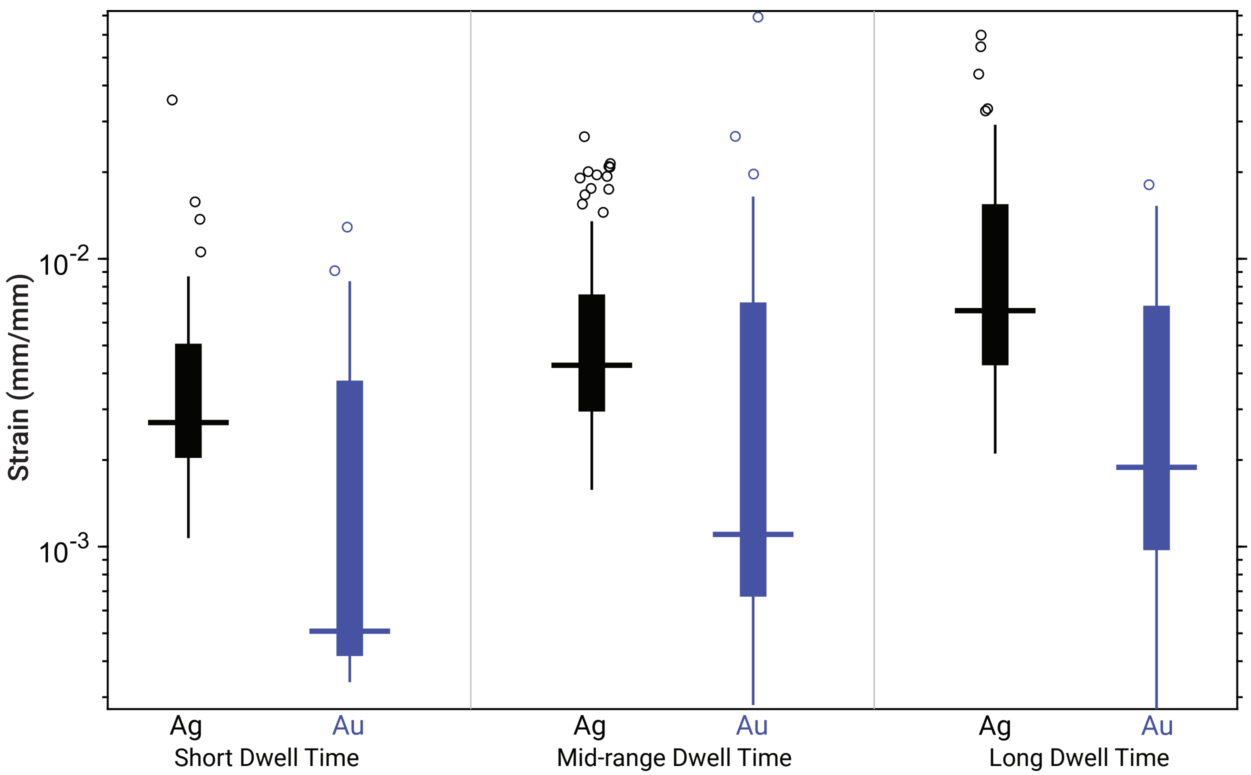

Figure 24: $\varepsilon_{yy,\max}$ for Position 1, Repetition 2, shown as a function of dwell time. Data are grouped into short (< 3 μs), mid-range (3–6 μs), and long (> 6 μs) dwell-time regimes.

To provide a more detailed illustration of the relationships between dwell time and sample type, ε_{yy} strain maps for Position 1, Repetition 2, 20 kV, 500 pA are presented across various dwell time settings in Figure 25. In the Ag samples, a clear trend emerges as dwell time increases, showcasing an evident rise in both the quantity and magnitude of line jumps in the ε_{yy} fields.

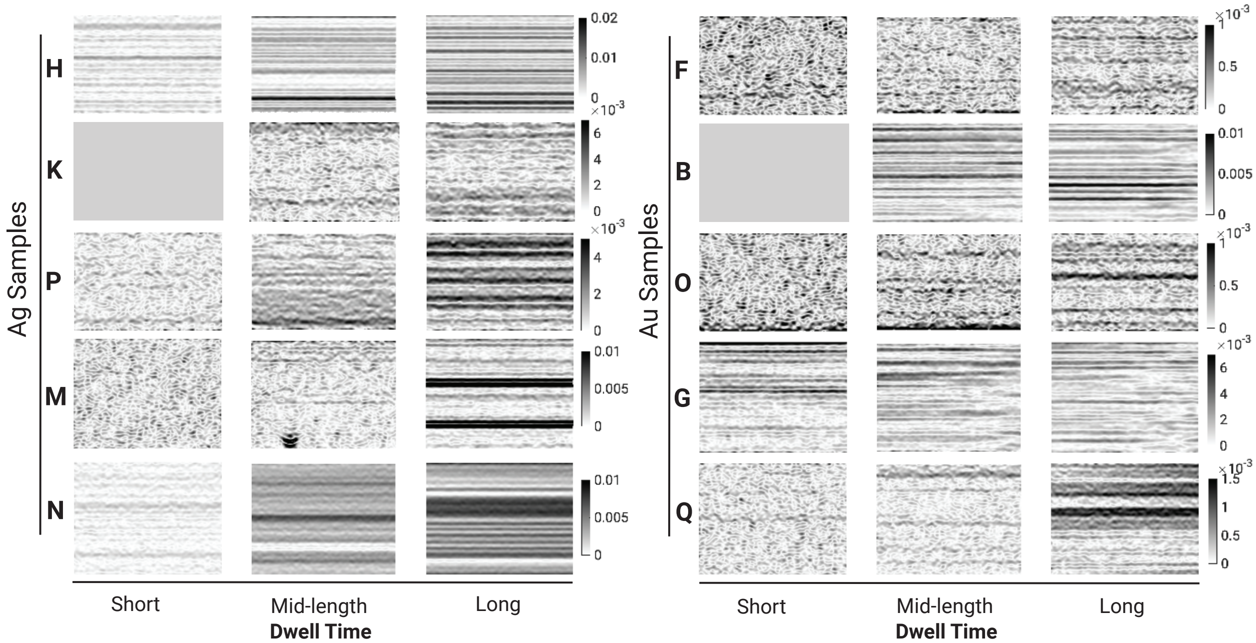

Figure 25: ε_{yy} strain field map of Position 1- Repetition 2, 20 kV, 500 pA, for all image sets differentiated by dwell time range classes. Color bars representing the strain (mm/mm) are normalized across each image set (represented by each unique sample letter). Data are grouped into short ($< 3\,\mu$s), mid-range (3–6 μs), and long ($> 6\,\mu$s) dwell-time regimes.

The analysis indicates that beam current did not exert a significant influence on line jumps, with $p \geq 0.05$ ($\varepsilon_{yy,\max}$ for $r_{Ag} = 0.00$ and $r_{Au} = 0.06$; $\varepsilon_{yy,\text{rows}>1\%}$ for $r_{Ag} = 0.11$ and $r_{Au} = -0.05$). However, higher beam voltages were found to reduce the quantity and magnitude of line jump errors for Au and Ag samples. Significant negative Pearson values for $\varepsilon_{yy,\max}$ and rows over 1% ε_{yy} ($\varepsilon_{yy,\text{rows}>1\%}$) were observed ($r_{Au} = -0.25$, $r_{Au} = -0.18$) and ($r_{Ag} = -0.16$, $r_{Ag} = -0.16$), respectively. Both the multivariate regression results and the negative Pearson values highlight the trend of reduced line jump errors with higher beam voltages, as depicted in Figure 26.

It is noteworthy that our finding of higher voltages reducing line jump errors, at least for the Au samples at low magnification, contrasts with prior work (Table 1). The exploration of sample conductivity in SI: Conductivity is intended to further understand errors attributable to charging.

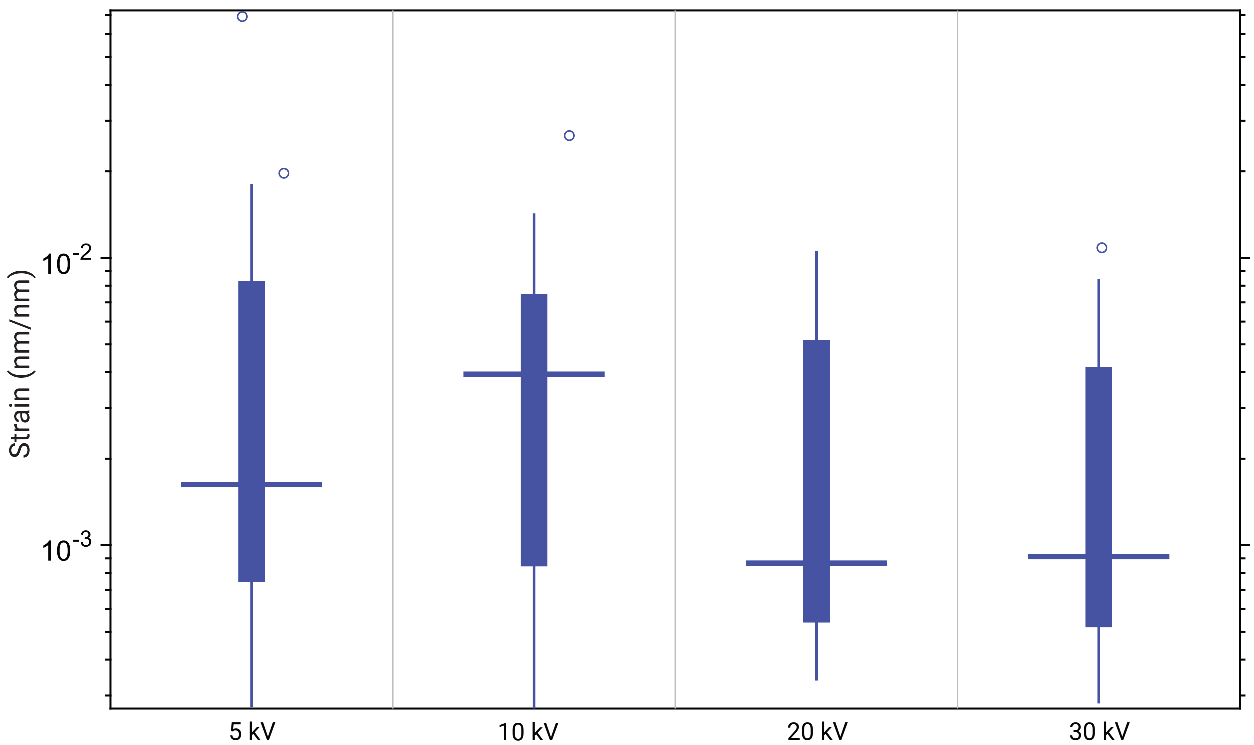

Figure 26: $\varepsilon_{yy,\max}$ Position 1- Repetition 2 (magnitude of line jumps) vs. voltage for Au sample

3.7 Additional studies with individual SEMs

In addition to the round-robin study detailed above, five additional relationships were explored on individual SEMs. First, detector influence was elucidated using SEM-4. The secondary electron detector (SE) was compared to the backscattered electron detector (BSE). Each image location, beam conditions, and scanning conditions were concurrently investigated while capturing both BSE and SE images simultaneously. Additionally, SEM-5 was used for 4 individual studies: a 24 hour drift distortion, horizontal field width, investigation of working distance, and sample conductivity (SI). In contrast to the main study, multivariate regression of these smaller datasets employed cross-validation to protect against overfitting, rather than holdout validation, which was used for the study of the large data set.

3.7.1 Detector

Detector type was not a parameter explored in the round-robin, as the Everhart-Thornley SE detector is typically used for SEM-DIC. Beyond the main round-robin study, this section compares the presence of SEM-DIC error across images taken at the same time and location on sample M (Ag, high magnification), using both backscattered electron (BSE) and secondary electron (SE) detectors simultaneously. We note that the BSE detector used was a standard retractable BSE detector, with a single solid-state diode. Multivariate linear regression results show that while detector type influences SEM-DIC error, there are many other conditions that have a more dramatic effect on overall error reduction (represented by MAD in Figure 27). Across all of the study parameters (i.e. main effect plot predictors), detector type was found to have a greater effect on MAD error only in the case of line averaging. It is shown, however, that MAD Position 1 Repetition 2, and MAD Position 2 Repetition 1, were both increased slightly by the use of the BSE detector. Specifically, the non-translated robust overall image displacement (MAD Position 1 Repetition 2) was increased by the BSE detector 0.092 nm with a 95% confidence interval of -0.25 to 0.43 nm. The translated overall image displacement error (MAD Position 2 Repetition 1) increased with the BSE detector by 0.11 nm with a 95% confidence interval of -0.42 to 0.64 nm.

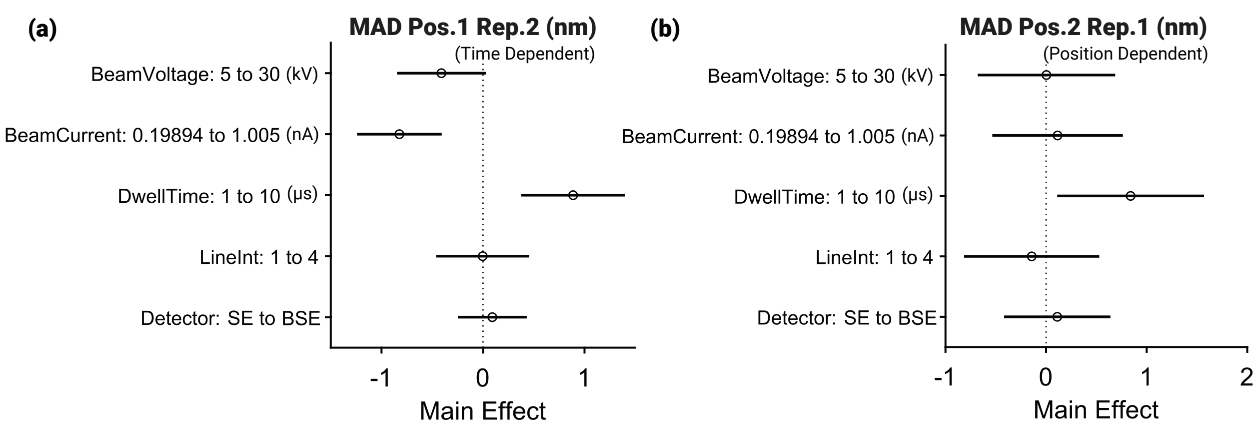

Figure 27: Overall image displacement multivariate regression results including detector type.

Table 4 outlines the relative influence of detector type on horizontal skewing, vertical skewing, rigid body drift distortion, magnitude of line jump error, and quantity of line jumps. In all error metric cases, the SE detector was seen to reduce error. In this study, error is typically reduced by roughly a nanometer or less with the exception of max false displacement. However, when compared to the effect beam and scanning conditions have on the multivariate regression results, it is clear that the SE detector only marginally reduces SEM-DIC error.

Error Metric	**Range**	**Predicted Response (nm)**
MAD Position 1 Repetition 2	SE to BSE	0.092
MAD Position 2 Repetition 1	SE to BSE	0.111
Max FD Position 1 Repetition 2	SE to BSE	**4.433**
Mean D Position 1 Repetition 2	SE to BSE	0.093
U_{error} Position 1 Repetition 2	SE to BSE	**0.231**
U_{error} Position 2 Repetition 1	SE to BSE	**0.984**
V_{error} Position 1 Repetition 2	SE to BSE	0.010
V_{error} Position 2 Repetition 1	SE to BSE	0.118
$\varepsilon_{yy,\text{max}}$ Position 1 Repetition 2	SE to BSE	0.000
$\varepsilon_{yy,\text{rows}>1\%}$ Position 1 Repetition 2	SE to BSE	0.002

Table 4: Multivariate regression error metric relation to prediction of the effect of detector type.

The comprehensive results of the BSE versus SE images presented in Figure 28. Notably, Position 1, Repetition 2 U_{error} exhibits the most significant difference in error. This aligns with the multivariate regression results, indicating an increase in U_{error} in the BSE images by 0.231 nm and 0.984 nm for Position 1 and Position 2 images, respectively. No discernible impact is observed in the cases of Mean D error, Column error, and V_{error}.

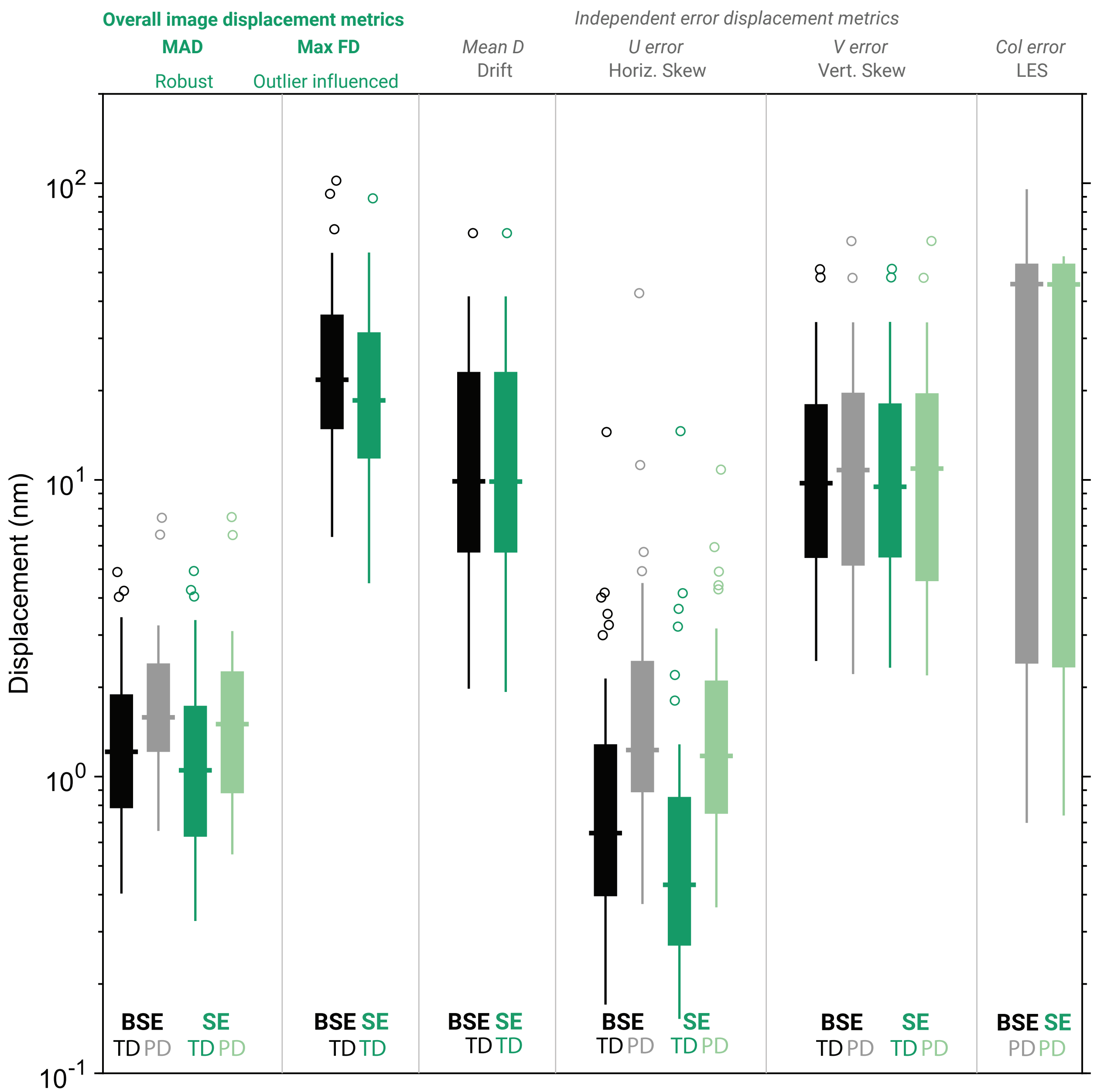

Figure 28: Displacement results including detector type comparing the backscattered electron detector (BSE) to the secondary electron detector (SE).

3.7.2 Long duration drift distortion

While the main study addressed short-term drift distortion, this additional study explores drift over long time scales. Based on the main study results, beam voltage and current did not seem to influence short-term drift. This conclusion was contradictory with previous literature conclusions (Table 1). Thus, parameters of the current study were developed based on previous literature findings that drift is worse at long dwell times and high magnifications[10,20]. To further investigate this relationship, a 24 hour time scale drift study was performed. This study specifically compares the effect of drift distortion over both high-power (30 kV, 1.4 nA) and low-power beam conditions (5 kV, 0.17 nA). Images were taken with a HFW of 20 µm and a 10 µs dwell time; conditions that are consistent with the main study's longest dwell time and highest magnification.

Before capturing images for the 24 hour drift study, the typical lens alignment, focus, and adjustments to the stigmator were carried out. Image quads were taken hourly. Prior to capturing the image quad, the beam was refocused and the stigmator was adjusted if necessary. Similar to the main study, the path dependence of magnification was accommodated by approaching the target magnification from a lower magnification[20,22].

Figure 29 shows that the low-power beam conditions experienced much less drift compared to the high-power beam condition image set. The maximum drift displacement for the high beam condition image set was 3,688 nm whereas the total drift displacement for the low beam condition was 568 nm. Additionally, the drift velocity for the high-power image set was calculated as 129 nm/hr, and the low-power image set velocity was 18.7 nm/hr.

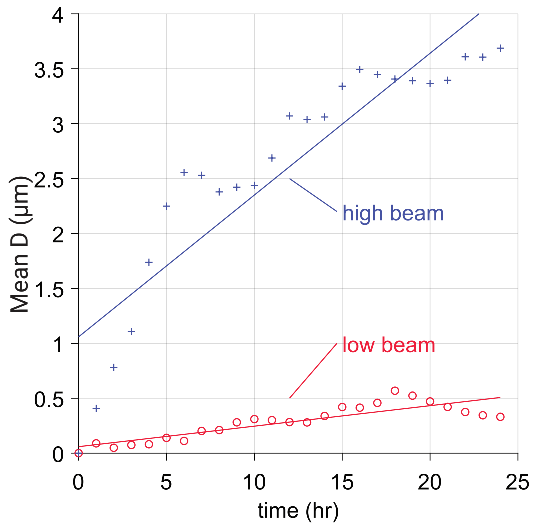

Figure 29: Rigid Body displacement over 24 hours and line of best fit used to determine drift velocity.

3.7.3 Horizontal Field Width (Magnification)

Similar to the sputter coating study, this horizontal field width study was proposed based on the main study multivariate regression results outlining sample type as a main effect. Since both magnification and patterning varied across samples, this study explores magnification independently of sample type. A unique sample was constructed for this study interlayering the 250 nm Au nanoparticles with 20 nm Au nanoparticles. The sample was imaged at HFWs ranging from 5 µm to 330 µm. Figure 30 shows the sample at the HFW bounds of the study.

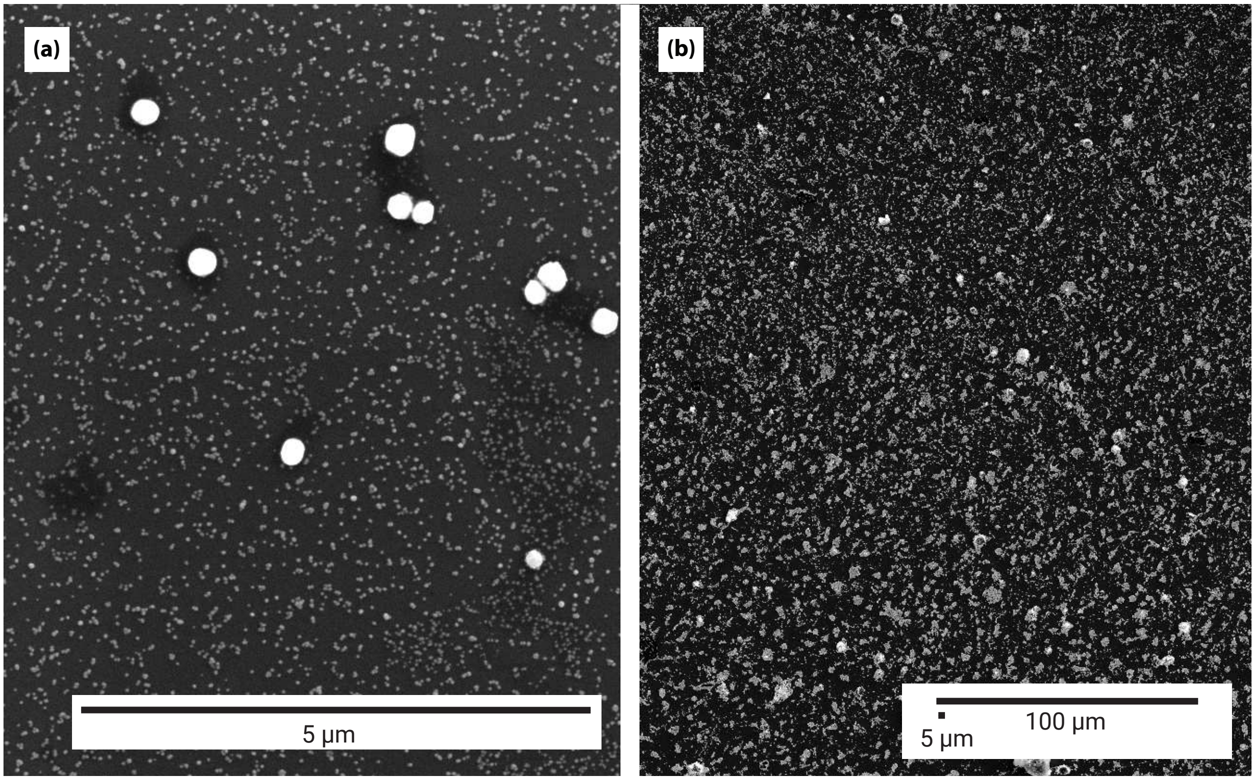

Figure 30: Image of HFW study sample: 250 nm Au nanoparticles interlayered with 20 nm Au nanoparticles. Left: HFW 5 µm; Right: 330 µm

Upon initial examination, there was only a marginal linear trend overall in image error for increasing the HFW from 5–330 µm (Figure 31). However, because of the limited data set and targeted error analysis approach, relatively small changes in the data represent high correlation Pearson values. Additionally, the sample speckle size and pattern were not well controlled within the standard SEM-DIC subset size and pixel quality guidelines[28]. Table 5 displays the calculated Pearson values. Pearson values with high degree of correlation are represented in bold and plotted in Figure 32.

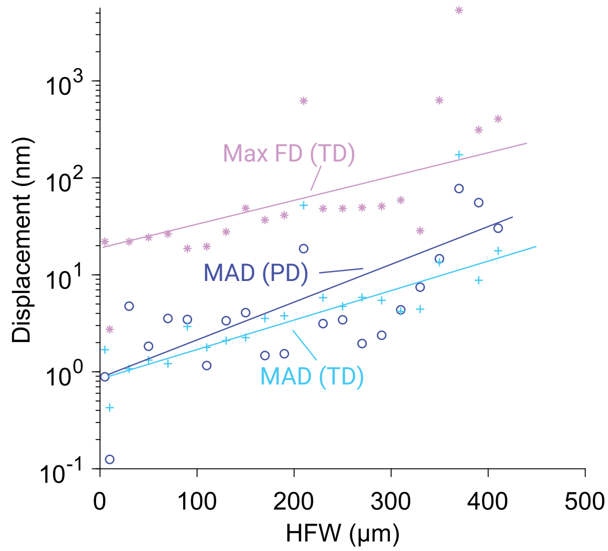

Figure 31: Overall image error displacement metrics (MAD (time-dependent, TD), MAD (position-dependent, PD), Max FD time-dependent, TD) vs. HFW

Error Metric	Pearson Value	Mean Error Mag. (nm)	Stdev. Error Mag. (nm)
MAD Position 1 Repetition 2	0.6039	11.1610	19.5469
MAD Position 2 Repetition 1	0.3368	14.4420	37.0725
Max FD Position 1 Repetition 2	0.5297	358.9752	1131.2
Mean D Position 1 Repetition 2	0.5034	290.8645	985.2255
U_{error} Position 1 Repetition 2	0.1678	16.3024	56.6874
U_{error} Position 2 Repetition 1	**0.7546**	23.8954	70.7966
V_{error} Position 1 Repetition 2	0.6430	74.5952	173.4358
V_{error} Position 2 Repetition 1	0.5530	213.7807	658.7579
Column Error Position 2 Repetition 1	**0.9977**	85.6530	81.5443
$\varepsilon_{yy,max}$ Position 1 Repetition 2	-0.5162	0.41%	0.75%

Table 5: Error metric relationship Pearson values based on magnification. Additionally mean and standard deviation of error is recorded.

Based on the Pearson values and error magnitude comparison (mean and standard deviation), it is clear that magnification greatly impacts SEM-DIC error. Not only do many error metrics have Pearson values greater than 0.5, indicating a moderate correlation, but the mean error corresponds to a high degree of displacement. Generally, low magnification (high HFW) images have more error than high magnification (low

HFW) images. However, the opposite trend is observed for line jump error ($\varepsilon_{yy,\max}$ Position 1 Repetition 2).

Figure 32 outlines a distinct linear relationship between HFW and Column Error as well as HFW and horizontal skewing in the Position 2 image. The magnitude of these errors changes with magnification. Through this observation, it can be hypothesized that LES and horizontal skewing is an inherent error within the SEM hardware and can be reduced at small HFWs (high magnification).

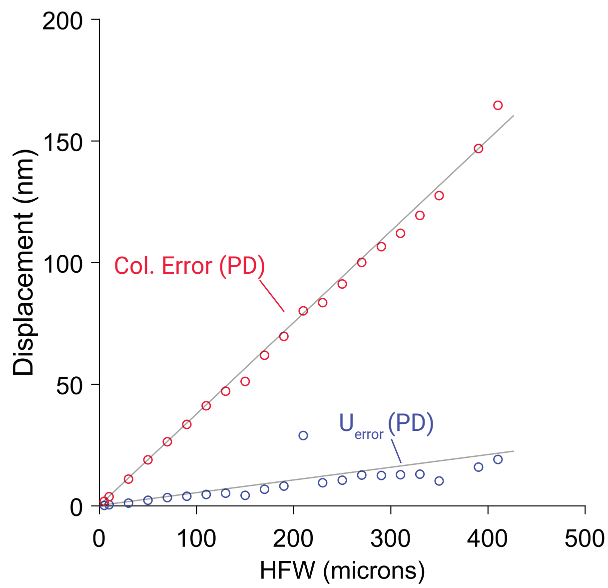

Figure 32: Position-dependent (PD) errors: column Error (LES) and U_{error} (horizontal skewing) vs. HFW

3.7.4 Working Distance

The main study procedure instructed each microscopist to choose a working distance of 5 mm, 10 mm or 15 mm based on whichever WD was closest to the value they would typically select for an in-situ SEM-DIC experiment. Ultimately, the main study consisted of 4 image sets with a working distance of 5 mm (G, M, O, P) and 6 image sets with a working distance of 10 mm (B, F, H, K, N, Q). This additional study was proposed to introduce working distance as a study condition. Image sets were taken at seven different working distances. These values were chosen based on the parameters of the SEM stage z axis bounds. The main study's mid-range beam conditions were used in the study (20 kV, 500 pA) with short dwell times. Image quads were taken at each WD after a typical alignment procedure was performed (focus, lens alignment, and adjusting stigmator).

Due to the nature of this individual image set, it is difficult to confidently conclude correlation. Table 6 outlines the Pearson values for error metrics as they relate to WD. Similar to the HFW study, Pearson values from 0 to 0.3 represent a low degree of correlation, values from 0.3 to 0.7 represent a moderate degree of correlation, and values from 0.7 to 1.0 represent a high degree of correlation. Pearson results are outlined in Table 6.

Error Metric	**Pearson Value**	**Mean Error Mag. (nm)**	**Std. Error Mag. (nm)**
MAD Position 1 Repetition 2	**0.8305**	2.4460	0.7560
MAD Position 2 Repetition 1	-0.1468	4.6584	0.6770
Max FD Position 1 Repetition 2	0.2633	41.8472	12.1643
Mean D Position 1 Repetition 2	0.1780	26.7001	13.1418
U_{error} Position 1 Repetition 2	-0.0576	2.4340	0.4185
U_{error} Position 2 Repetition 1	-0.5522	14.1883	0.8778
V_{error} Position 1 Repetition 2	**0.7327**	12.3565	5.4116
V_{error} Position 2 Repetition 1	0.5781	18.7672	9.3505
Column Error Position 2 Repetition 1	**-0.8317**	136.7063	1.5449
$\varepsilon_{yy,\text{max}}$ Position 1 Repetition 2	**0.8218**	0.72%	0.41%

Table 6: Error metric relationship Pearson values based on working distance. Additionally, absolute values of mean and standard deviation of error are given.

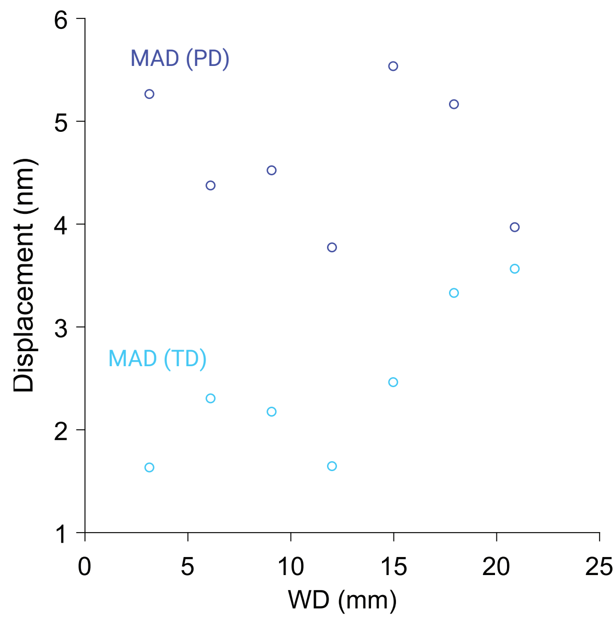

Figure 33: Overall image displacement error (MAD Position 1 Repetition 2, MAD Position 2 Repetition 1, Max FD Position 1 Repetition 2) vs. working distance

This image set shows possible correlation between the LES and working distance and $\varepsilon_{yy,\max}$ Position 1 Repetition 2. It should be noted, however, that the magnitude of displacement error attributed to working distance and these conditions is extremely small. This indicates that working distance is not likely a large contributing factor to SEM-DIC error. However, the trends observed between working distance and different error types can aid to elucidate hypothesis of errors. It is likely that errors which are strongly related to working distance have some component of either charging or inherent lens hardware error. Alternatively, since lenses are designed to have distinct focal points, if an error is highly altered by working distance, this could indicate inherent lens hardware error.

4 Discussion

4.1 Aggregate understanding from various samples and SEMs

The round-robin format of this work is a unique advancement for the field to develop understanding of SEM-DIC errors. Nonetheless, the format opens the door for error from variations in SEMs, samples, and microscopists. The analysis does not explicitly compare individual SEMs, but rather aims to identify and evaluate trends in the data across various beam conditions. Based on the comparisons of MIG and sigma values (Section 3.1), the samples and image sets effectively fulfill these objectives. It was generally shown that the high magnification Ag samples had a higher correlation confidence (sigma) compared to the low magnification Au samples. Alternatively, MIG (related to pattern quality) was highly comparable across sample types and samples. Further sources of error and mitigation strategies are discussed in the following sections.

4.2 Largest sources of error for different sample types/magnifications

The comparison of error types in the Ag and Au samples reveals distinct patterns in their performance. For the Ag samples (higher magnification), the order of errors from largest to smallest is led by Max FD (Figure 11), emphasizing a significant discrepancy in the maximum feature displacement. This is followed by Mean D, highlighting notable errors in the drift displacement.

In contrast, the Au samples exhibit a different error distribution. For the Au samples (lower magnification), the predominant error type is column error (Figure 12), indicating a substantial LES at lower magnifications. Max FD follows closely, suggesting a significant variation in maximum feature displacement. These distinctions in the prominent error types underscore variations in the two sets of samples, providing valuable insights into their respective characteristics and potential areas for improvement or optimization in the SEM-DIC analysis. Conducting a specific analysis of errors allows for a detailed examination of the response to different beam conditions, as outlined below. While a more in-depth error-specific analysis is to follow, this initial assessment holds significance as it provides a concise overview.

4.3 Left-edge smear increases as magnification decreases

The analysis of the left-edge smear (LES), measured through column error, revealed that beam voltage, current, and line averaging had minimal to no impact on this error. In contrast, both dwell time and sample type demonstrated a significant effect in reducing the error. Notably, Au samples exhibited considerably higher errors compared to their Ag counterparts. However, a Pearson analysis, conducted independently for Ag and Au samples, indicated that regardless of sample type, longer dwell times consistently reduced column error in both cases. This leads to the hypothesis that the LES incorporates an element of stabilization error, where the timing of the initiation of the new line to raster is delayed. Recall, based on the error magnitude comparison previously discussed, that the LES was the greatest error contributor in the Au image set and the fourth greatest contributor in the Ag image set (Figures 11 and 12). Thus, error reduction (calculated from multivariate regression) of roughly 181 nm by using longer dwell times (10 µs) is highly significant in reducing overall SEM-DIC error.

Furthermore, considering the sample type as a contributing factor to error, it is hypothesized that either magnification or conductivity plays a role in diminishing the LES. This hypothesis was substantiated in the additional study on magnification, where it was demonstrated that the LES is directly correlated with image magnification, exhibiting a Pearson value exceeding 0.99. In other words, when magnification decreases, LES increases linearly. In contrast to the robust relationship established between magnification and the LES, the relationships identified with sample conductivity and the LES were observed to be marginal. It was also observed (Section 3.7.4) that as working distance increased, the LES decreased. In combination with the LES and magnification relationship, this suggests that the LES may also have a contribution from hardware-related factors. This finding adds to our existing hypothesis that the LES is influenced by stabilization error, particularly in relation to dwell time.

Based on these results, the most practical recommendation for mitigating the impact of LES in SEM-DIC analysis is to crop the left-most portion of the image affected by LES. In this study, and in common SEM-DIC practice, a 10% crop was used as a practical threshold; however, this value was a conservative data processing choice rather than an optimized cropping percentage. This strategy proves particularly beneficial when conducting imaging at low magnifications and employing extended dwell times.

4.4 Short-term drift is best mitigated by fast dwell times

The examination of short-term drift (assessed by Mean D) uncovered a hierarchy of contributing conditions in terms of their impact. Dwell time emerged as the most influential factor, followed by line averaging, current, sample type, and voltage, in descending order of importance. Prolonged dwell times, elevated currents, and multiple line averagings were associated with increased error, while lower voltages were correlated with reduced error. It is important to note that, although our analysis revealed a trend of lower voltages contributing to error reduction and higher currents contributing to increases in error, these observed trends were relatively modest, as shown by a sizable 95% confidence interval. The degree of uncertainty associated with the impact of these conditions on error emphasizes the need for careful interpretation and consideration of other influencing factors. For example, while dwell time exhibited a consistent influence on both the Ag and Au samples, noteworthy distinctions were observed concerning the impact of beam voltage and current. Specifically, the Au samples demonstrated a more pronounced error based on variations in beam voltage and current compared to their Ag counterparts. When analyzing the effect of beam voltage and current independently of sample type (Pearson values), it was shown that for the Ag samples, there was almost no correlation between beam current and voltage, and drift error. This differential response highlights the material-dependent nature of certain influencing factors on short-term drift, emphasizing the need for tailored considerations based on the specific sample composition.

For this reason, a long time scale drift study was performed with an emphasis on outlining the differences in high beam (voltage and current) and low beam conditions, and the effect sample conductivity may have on drift distortion. It was shown that high beam conditions greatly increased the drift velocity over a 24 hour study. For the high beam conditions, drift velocity was calculated as 129 nm/hr while the low beam condition drift velocity was calculated as 18.7 nm/hr. This highlights the profound impact of beam power on drift distortion over prolonged duration. Furthermore, the study revealed $\approx 40\%$ reduction in drift when the sample was sputter-coated (Figure S6), indicating that enhanced sample conductivity significantly diminishes drift distortions during extended experiments. Contrary to this conclusion, it was shown that higher substrate resistivity led to a reduction in drift error during immediate image repetitions. This unexpected outcome is likely due to differences between short- and long-duration drift errors. Given the unexpected nature of this hypothesis, which contradicts conventional expectations and previous literature, it is advisable to further explore and validate this trend, potentially through additional rounds of experimentation or collaborative verification. Beyond the realm of sample characteristics, such as type and conductivity, as well as beam voltage and current considerations, practical strategies to minimize drift distortion include using short dwell times and avoiding line averaging.

4.5 Horizontal skewing is primarily a time-independent error, while vertical skewing can manifest strongly as both time- and position-dependent

In this work, skewing was intentionally categorized into horizontal and vertical displacement, each examined independently (U_{error} and V_{error}). This strategic division stems from the recognition that, given the nature of SEM imaging where each pixel row is rastered individually, making assumptions about the uniform influence of beam conditions on both vertical and horizontal components of error may not be warranted. Indeed, the analysis revealed distinct and varied effects of beam conditions on vertical skewing compared to horizontal skewing, underscoring the need for a nuanced understanding of these responses to different factors. Additionally, recall that Position 1

Repetition 2 images were examined to identify the time-dependent characteristics of errors. Conversely, Position 2 Repetition 1 images were examined to identify the position-dependent aspects of errors. This dual approach allowed for a comprehensive understanding of how errors evolve over time and how they manifest in different positions, contributing to a nuanced interpretation of the overall error landscape.

Analyzing the horizontal skewing error (U_{error}) under both time-dependent and position-dependent conditions, it becomes evident that sample type emerges as the predominant contributing factor. Notably, there is a substantial degree of position dependence in horizontal skewing, illustrated by the maximum MVR error magnitudes of 5.9 nm for U_{error} Position 1 Repetition 2 and 65 nm for U_{error} Position 2 Repetition 1. In the context of time-dependent horizontal skewing (Position 1 Repetition 2) in Ag samples, the influence of high beam voltage and high beam current on decreasing error is marginal. However, these relationships lack significance due to their low Pearson correlation values (-0.16 and -0.14) and minimal degree of change (less than 1 nm). As such, it is concluded that the primary determinant of horizontal skewing appears to be sample type, which encompasses factors like magnification or conductivity. This conclusion is further corroborated by the analysis of Position 2 Repetition 1, where Pearson values for beam and scanning conditions indicate low degrees of correlation, reinforcing the notion that sample type remains the primary contributing factor. The synthesis of these results leads to the assertion that horizontal skewing is a position-dependent, time-independent error profoundly influenced by sample type. Subsequent validation in the horizontal field study (Section 3.7.4) directly links position-dependent horizontal skewing to image magnification (Pearson Value $r = 0.75$), revealing that decreasing magnification corresponds to an increase in horizontal skewing error. Additionally, the study establishes an influence of working distance on position-dependent horizontal skewing (Pearson Value $r = -0.55$), indicating that increased working distance correlates with decreased position-dependent horizontal skewing. Mitigation strategies for horizontal skewing may include optimizing conditions such as large working distances, high magnifications, and the use of the SE detector. Moreover, it is crucial to recall that horizontal skewing was identified as having the lowest contribution to SEM-DIC error. Therefore, while efforts can be directed toward minimizing this error, it might be more beneficial to prioritize optimizing beam conditions to address other error types with more significant contributions.

Analyzing the vertical skewing error (V_{error}) under both time-dependent and position-dependent conditions, it becomes evident that sample type, dwell time and line averaging emerge as the predominant contributing factors. Unlike horizontal skewing, which was shown to be predominantly a time-independent error, vertical skewing exhibits large magnitudes of error in both the time-dependent and position-dependent conditions.

Based on the results of this study, it is shown that vertical skewing remains the least understood: it seems to have both time and position, conductivity, and magnification dependence. Generally, in SEM-DIC practice, it is recommended that short dwell times and no line averaging be used to reduce vertical skewing. Recall that V_{error} encompasses both vertical skewing and the impact of line jump errors. While vertical skewing cannot be deconvoluted from line jumps in this metric, a more thorough analysis of beam and scanning conditions that impact line jumps specifically is outlined below.

4.6 Line jump errors are tied to sample conductivity, dwell time, and charging

The analysis of both line jump error magnitude ($\varepsilon_{yy,\text{max}}$) and frequency ($\varepsilon_{yy,\text{rows}>1\%}$) revealed that line jump errors are influenced by both beam and scanning conditions. Specifically, line jump errors increase with line averaging, long dwell times, high beam currents, and low beam voltages. Perhaps the most illuminating result from this MVR line jump error analysis, however, is how different sample types behaved with respect to increasing dwell time. For the case of the Ag samples, Pearson values of 0.334 and 0.3081 were extracted for line jump magnitude and line jump frequency, respectively. Alternatively, for the Au samples Pearson values were less than 0.08 in both cases. This disparity underscored the necessity for further investigations specifically addressing sample conductivity and magnification.

Accordingly, first, a substrate conductivity study was proposed, wherein three full image sets were taken on three unique samples with different levels of substrate conductivity. Results demonstrated that, for all error metrics excluding line jump errors, errors decreased as substrate resistivity increased. Conversely, for line jump errors, errors exhibited an upward trend with increasing substrate resistivity. This observation supports the hypothesis that line jump errors may stem from speckle or surface contaminant charging. Furthermore, the horizontal field width (HFW) study revealed a statistically significant relationship ($r = 0.8218$) between line jump magnitude and HFW. This finding further substantiates the hypothesis that line jump errors may arise from speckle or surface contaminant charging, as a larger HFW corresponds to a higher imaged area, greater number of speckles, and an increased probability of surface deformities. Practical strategies to mitigate line jump errors include the application of a highly conductive pattern for speckling samples whenever feasible. Alternatively, users should employ low beam currents and high voltages during imaging to reduce both the magnitude and quantity of line jump errors.

5 Conclusions

The SEM and sample conditions investigated in this study reveal various trends among errors and imaging conditions. Furthermore, hypotheses regarding the sources of error are elucidated. Table 7 summarizes the conditions, beam settings, and line scanning conditions determined in this study, along with hypotheses regarding the causes of errors, ultimately providing insights into mitigating each error. We also note that the appendix highlights a colorized table of the main study Pearson values. This round-robin investigation concludes that optimal conditions for SEM-DIC experiments involve high beam voltages (e.g., 30 kV), fast dwell times (e.g., 1 µs), low to moderate beam currents (e.g., less than 1 nA), and the exclusion of line averaging/integration. The study further disentangles the error contributions from specific elements of charging, distinguishing between sample charging and speckle charging. To refine the understanding of charging effects, the study proposes targeted experiments for future research. A new hypothesis is put forth, suggesting that line jump errors may be linked to speckle charging or surface deformities. Additionally, the LES is hypothesized to result from stabilization/timing errors. It is recommended that users interested in utilizing SEM-DIC should carefully review Table 7 and the resulting recommended practices. This work also suggests potential new studies to further advance the investigation of SEM-DIC error, aiming to enhance the overall practice. In addition, the images analyzed here originate from the iDICs round-robin challenge and are available as an open-access dataset, providing a transparent benchmark to support broader community efforts in understanding, comparing, and optimizing in-situ SEM-DIC measurements.

	Left-edge smear (LES)	**Immediate Drift Distortion (IDD)**	**Long Drift Distortion (LDD)**	**Horizontal Skewing TD (HSTD)**	**Horizontal Skewing PD (HSPD)**	**Vertical Skewing TD (VSTD)**	**Vertical Skewing PD (VSPD)**	**Line Jumps (LJ)**
Beam Voltage (V)	†	†	V $\propto$ LDD	△	V $\propto$ 1/HSPD	V $\propto$ 1/VSTD	△	V $\propto$ 1/LJ
Beam Current (I)	†	△	I $\propto$ LDD	†	†	†	†	†
Dwell Time (DT)	DT $\propto$ 1/LES	DT $\propto$ IDD	×	†	†	DT $\propto$ VSTD	DT $\propto$ VSPD	DT $\propto$ LJ
Line integration (LI)	†	LI $\propto$ IDD	×	†	†	LI $\propto$ VSTD	LI $\propto$ VSPD	†
Detector	†	†	×	SE <BSE	†	†	†	†
Magnification (Mag)	Mag $\propto$ LES	Mag $\propto$ IDD	×	△	Mag $\propto$ HSPD	Mag $\propto$ VSTD	Mag $\propto$ VSPD	Mag $\propto$ 1/LJ
Working Distance (WD)	WD $\propto$ 1/LES	△	×	†	△	WD $\propto$ VSTD	WD $\propto$ VSPD	WD $\propto$ LJ
Cause of Error	Stabilization error (exaggerated by magnification)	Stage Charging	Charging	Position of SE detector	Hardware errors	Charging, hardware error, stabilization error	Stage charging, hardware error, stabilization error	Speckle or contaminant charging
Means to Negate	Crop left most 10% or increase dwell time	Short dwell times, no line averaging, low magnification	Low voltage and current	Use SE over BSE	High voltage, low magnification	High voltage, short dwell time, low magnification	Short dwell time, low magnification	High voltage, short dwell time, high magnification

Table 7: Summary of the influence of beam and scanning parameters on SEM-DIC errors (TD: time-dependent, PD: position-dependent, SE: secondary electron, BSE: Backscattered electron).
$\propto$ Significant relationship found; symbol describes proportionality of condition and error
† No relationship found (based on multivariate regressions (MVR) and Pearson values)
△ Small relationship outlined (see paper for details)
× Not outlined in this study

5.1 Recommended Practices

- Voltage: In general, higher voltages reduce the largest sources of error.
- Current: For short experiments (minutes to hours), current does not greatly influence SEM-DIC error, but low currents are recommended for experiments spanning multiple hours to days.
- Dwell time: Shorter dwell times (about 1 μs) are preferred to reduce multiple error types.
- Line averaging: It is also recommended that line averaging is not used. In all cases, it was shown to increase SEM-DIC error.
- Cropping: When a standard typewriter/rastering pattern is deployed, it is recommended to crop the left-most 10% of the image to mitigate LES.
- Conductivity: Higher sample conductivity reduces the largest sources of error.
- Working distance: Working distances between 5 and 15 mm appear to all be suitable, with no clear trend in error magnitude among them.

5.2 Future work

While this study provided insights into various factors contributing to SEM-DIC error, it also raised new questions about the underlying causes of these errors. Certain errors like horizontal and vertical skewing and short-term drift may be linked to stage charging, and further investigation is warranted. A dedicated study centered on charging could offer valuable insights into whether specific charging components, such as stage or sample charging, demonstrate distinct correlations with particular errors. Furthermore, it is recommended to conduct additional analysis to explore the impact of sample contamination and, potentially, the influence of specific patterning techniques on line jump errors. Additionally, the behavior of vertical skewing (excluding the effect of line jumps) under various beam conditions is still not thoroughly understood. Overall, future work should determine better ways to decouple these error types using alternative error metrics.

Acknowledgments

The authors are very grateful to Zhe Chen, Samantha Daly, Phillip Reu, Michael Demkowicz, Marc van Maris, and William Lenthe for thoughtful discussions. Additionally, funding sources per author are provided below. VT and WL acknowledge the National Science Foundation Major Research Instrumentation (Grant Number #0923461) and The University of Tulsa. TV and JH funding from the "UNFAIL" project, under project number S17012b in the framework of the Partnership Program of the Materials innovation institute M2i and the Netherlands Organization for Scientific Research. CM and NS acknowledge grant No. FA9550-12-1-0445 to the Center of Excellence on Integrated Materials Modeling (CEIMM) at Johns Hopkins University, awarded by the AFOSR/ Aerospace Materials for Extreme Environments Program (Program Manager Ali Sayir) and AFRL/RX (Monitors C. Woodward and C. Przybyla). This work was carried out in part in the Frederick Seitz Materials Research Laboratory Central Research Facilities, University of Illinois. JCS and TM acknowledge the National Science Foundation Major Research Instrumentation (Grant Number #2320030) and the UCSB Microscopy and Microanalysis Facility. ML acknowledges Texas A&M University.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found in the online version.

Data Availability

The as-received TIF images, correlated DIC (in Vic2D format) results and MATLAB data analysis scripts are published in an open access data repository (https://idics.org/challenge). With these data, researchers may compare the performance of their SEM and further assess SEM-DIC errors as post-processing techniques.

References

[1] Sutton M, Orteu J, Schreier H (2009) Image Correlation for Shape, Motion, and Deformation Measurements. Springer, New York, NY, doi:10.1007/978-0-387-78747-3

[2] Pan B (2013) Bias error reduction of digital image correlation using Gaussian pre-filtering. Optics and Lasers in Engineering 51:1161–1167, doi:10.1016/j.optlaseng.2013.04.009

[3] Delaire F, Raphanel JL, Rey C (2000), PLASTIC HETEROGENEITIES OF A COPPER MULTICRYSTAL DEFORMED IN UNIAXIAL TENSION: EXPERIMENTAL STUDY AND FINITE ELEMENT SIMULATIONS

[4] Lagattu F, Bridier F, Villechaise P, Brillaud J (2006) In-plane strain measurements on a microscopic scale by coupling digital image correlation and an in situ SEM technique. Materials Characterization 56(1):10–18, doi:10.1016/j.matchar.2005.08.004

[5] Stinville J, Callahan P, Charpagne M, Echlin M, Valle V, Pollock T (2020) Direct measurements of slip irreversibility in a nickel-based superalloy using high resolution digital image correlation. Acta Materialia 186:172–189, doi:10.1016/j.actamat.2019.12.009

[6] Di Gioacchino F, Quinta da Fonseca J (2015) An experimental study of the polycrystalline plasticity of austenitic stainless steel. International Journal of Plasticity 74:92–109, doi:10.1016/j.ijplas.2015.05.012

[7] Briffod F, Edwards T, da Fonseca JQ, Stinville JC, Texier D, Vermeij T (2026) Understanding strain localization in metallic materials: a review of high-resolution digital image correlation and related techniques. Science and Technology of Advanced Materials 27(1):2630488

[8] Sutton MA, Li N, Joy DC, Reynolds AP, Li X (2007) Scanning electron microscopy for quantitative small and large deformation measurements Part I: SEM imaging at magnifications from 200 to 10,000. Experimental Mechanics doi:10.1007/s11340-007-9042-z

[9] Sutton MA, Li N, Garcia D, Cornille N, Orteu JJ, McNeill SR, Schreier HW, Li X, Reynolds AP (2007) Scanning electron microscopy for quantitative small and large deformation measurements Part II: Experimental validation for magnifications from 200 to 10,000. Experimental Mechanics 47, doi:10.1007/s11340-007-9041-0

[10] Kammers A, Daly S (2013) Digital Image Correlation under Scanning Electron Microscopy: Methodology and Validation. Experimental Mechanics 53(9):1743–1761, doi:10.1007/s11340-013-9782-x

[11] Weidner A, Biermann H (2021) Review on Strain Localization Phenomena Studied by High-Resolution Digital Image Correlation. Advanced Engineering Materials 23(4), doi:10.1002/adem.202001409

[12] Mansilla C, Ocelík V, De Hosson JTM (2014) A New Methodology to Analyze Instabilities in SEM Imaging. Microscopy and Microanalysis 20(6):1625–1637, doi:10.1017/s1431927614013282

[13] Maraghechi S, Hoefnagels JPM, Peerlings RHJ, Rokoš O, Geers MGD (2019) Correction of Scanning Electron Microscope Imaging Artifacts in a Novel Digital Image Correlation Framework. Experimental Mechanics 59(4):489–516, doi:10.1007/s11340-018-00469-w

[14] Postek MT, Vladár AE (2013) Does your SEM Really tell the truth?—how would you know? Part 1. Scanning 35(6):355–361, doi:10.1002/sca.21075

[15] Reu PL, Toussaint E, Jones E, Bruck HA, Iadicola M, Balcaen R, Turner DZ, Siebert T, Lava P, Simonsen M (2018) DIC challenge: developing images and guidelines for evaluating accuracy and resolution of 2D analyses. Experimental Mechanics 58:1067–1099

[16] Reu PL, Blaysat B, Andò E, Bhattacharya K, Couture C, Couty V, Deb D, Fayad SS, Iadicola MA, Jaminion S, et al. (2022) DIC challenge 2.0: developing images and guidelines for evaluating accuracy and resolution of 2D analyses: focus on the metrological efficiency indicator. Experimental Mechanics 62(4):639–654

[17] Schreier HW, Braasch JR, Sutton MA (2000) Systematic errors in digital image correlation caused by intensity interpolation. Optical Engineering 39:2915 – 2921, doi:10.1117/1.1314593

[18] Stinville J, Echlin M, Texier D, Bridier F, Bocher P, Pollock T (2015) Sub-Grain Scale Digital Image Correlation by Electron Microscopy for Polycrystalline Materials during Elastic and Plastic Deformation. Experimental Mechanics 197–216, doi:10.1007/s11340-015-0083-4

[19] Gioacchino FD, da Fonseca JQ (2013) Plastic Strain Mapping with Sub-micron Resolution Using Digital Image Correlation. Experimental Mechanics 53:743–754, doi:10.1007/s11340-012-9685-2

[20] Lenthe WC, Stinville J, Echlin MP, Chen Z, Daly SH, Pollock TM (2018) Advanced Detector Signal Acquisition and Electron Beam Scanning for High Resolution SEM Imaging. Ultramicroscopy

[21] Maraghechi S, Hoefnagels JP, Peerlings RH, Geers MG (2018) Correction of scan line shift artifacts in scanning electron microscopy: An extended digital image correlation framework. Ultramicroscopy 187:144–163, doi:10.1016/j.ultramic.2018.01.002

[22] Mello AW, Book TA, Nicolas A, Otto SE, Gilpin CJ, Sangid MD (2017) Distortion Correction Protocol for Digital Image Correlation after Scanning Electron Microscopy: Emphasis on Long Duration and Ex-Situ Experiments. Experimental Mechanics 57(9):1395–1409, doi:10.1007/s11340-017-0303-1

[23] Cho Y, Cho J, Park J, Wang J, Jeong S, Lee J, Hwang Y, Kim J, Yu J, Chung H, et al. (2024) Automatic beam optimization method for scanning electron microscopy based on electron beam Kernel estimation. Communications Engineering 3(1):82

[24] Lee W, Nam HS, Kim YG, Kim YJ, Lee JH, Yoo H (2021) Robust autofocusing for scanning electron microscopy based on a dual deep learning network. Scientific reports 11(1):20933

[25] Vermeij T, Verstijnen J, Ramirez y Cantador T, Blaysat B, Neggers J, Hoefnagels J (2022) A nanomechanical testing framework yielding front&rear-sided, high-resolution, microstructure-correlated SEM-DIC strain fields. Experimental Mechanics 62:1625–1646

[26] Montgomery CB (2018) Multiscale characterization of carbon fiber-reinforced epoxy composites. Ph.D. thesis, University of Illinois at Urbana-Champaign

[27] Montgomery CB, Koohbor B, Sottos NR (2019) A Robust Patterning Technique for Electron Microscopy-Based Digital Image Correlation at Sub-Micron Resolutions. Experimental Mechanics 59(7):1063–1073, doi:10.1007/s11340-019-00487-2

[28] Reu P (2014) All about speckles: Aliasing. Experimental Techniques doi:10.1111/ext.12111

[29] Bornert M, Brémand F, Doumalin P, Dupré JC, Fazzini M, Grédiac M, Hild F, Mistou S, Molimard J, Orteu JJ, Robert L, Surrel Y, Vacher P, Wattrisse B (2009) Assessment of digital image correlation measurement errors: Methodology and results. Experimental Mechanics 49:353–370, doi:10.1007/s11340-008-9204-7

[30] Fayad SS, Seidl DT, Reu PL (2019) Spatial DIC Errors due to Pattern-Induced Bias and Grey Level Discretization. Experimental Mechanics 60(2):249–263, doi:10.1007/s11340-019-00553-9

[31] G Frens (1973) Controlled Nucleation for the Regulation of the Particle Size in Monodisperse Gold Suspensions. Nature Physical Science doi:10.1038/242117a0

[32] LePage WS, Shaw JA, Daly SH (2020) Effects of texture on the functional and structural fatigue of a NiTi shape memory alloy. International Journal of Solids and Structures doi:10.1016/j.ijsolstr.2020.09.022

[33] Jones E, Iadicola M (2018) A Good Practices Guide for Digital Image Correlation. International Digital Image Correlation Society

[34] Vermeij T, Hoefnagels J (2021) Plasticity, localization, and damage in ferritic-pearlitic steel studied by nanoscale digital image correlation. Scripta Materialia 208

[35] Pan B, Lu Z, Xie H (2010) Mean intensity gradient: An effective global parameter for quality assessment of the speckle patterns used in digital image correlation. Optics and Lasers in Engineering doi:10.1016/j.optlaseng.2009.08.010

Supplementary Information

S1 Round-robin procedure

The following procedure was distributed to the round-robin participants for use during image acquisition.

S1.1 Sample information

There are two sets of samples for the SEM-DIC round-robin. Each participant should receive one sample from each set, so that each participant has two samples.

- Both sets of samples are 10×10 mm Si wafer chips, but the two sets have different speckle patterning methods.
  - The first set (samples A through G) has 300 nm Au nanoparticles in a sparse monolayer. They were produced in 2017 by Will LePage at University of Michigan.
  - The second set (samples H through N) has 30 nm Ag sputtered nanodroplets in a dense monolayer. They were produced in 2018 by Chris Montgomery at University of Illinois, Urbana-Champaign.
- The "Samples" sheet of the round-robin spreadsheet details each sample and the participant.
  - The first set of samples (A through G) was distributed to participants at the iDICs 2017 and SEM 2018 annual conferences. The second set of samples (H through N) will be distributed to participants by postal mail from Will LePage. If a sample has already been analyzed before by a previous participant, it will be assigned another sample letter in sequence down the alphabet.
- The two sets of samples will have different horizontal field widths (HFW) and translation distances for their corresponding SEM images.
  - Samples A through G and sample O must have HFW of 200 µm and

translations of 20 µm

- Samples H through N and sample P must have HFW of 20 µm and translations of 2 µm

S1.2 Experiment procedure

The following procedure was followed by microscopists to gather the round-robin data. The procedure is presented below verbatim as the procedure shared with the microscopists.

Detector

The images should be captured with the SEMs standard Everhart-Thornley secondary electron detector.

Image formats

All images should be saved as uncompressed TIF format at 16-bit depth with a pixel size nearest to 3072 pixels wide by 2048 pixels tall. The minimum pixel count in the horizontal field width is 3000 pixels (higher is okay). If the SEM cannot save at 16-bit depth, then the maximum bit depth should be used (e.g. some SEMs can only save up to 14-bit depth).

Preparation

1. Load sample onto stage and pump the chamber.
2. Navigate towards a working distance (WD) that is appropriate for your SEM and that you would typically use for SEM-DIC. For example, many SEMs have optimum imaging at WD=5 mm, while others are used at WD=15 mm to accommodate the extra size of an in-situ tension/compression stage. Please select one of three options for your image set's WD: 5 mm, 10 mm, or 15 mm. Pick just one of these three WD settings (not all three), and this should be the WD for all of your images with that SEM.

3. Perform the typical alignment procedure that would be performed before capturing an important image set. This does not need to be the full service-level alignment that is often performed monthly for the SEM, but should be a user-level alignment that will optimize the SEM performance. Different SEMs have different alignment procedures and terminology, but broadly this user-level alignment should include source/gun tilt, objective/lens alignment, and astigmatism correction.

4. With many SEMs, the beam current is not an option that can be controlled, but is rather only a value that is measured. If your SEM cannot set the beam current, then use a Faraday cup (most SEM stages have this as a small hole in the main stage area) to find the approximate beam "intensity" setting that gets closest to the desired beam current.

Image acquisition

1. At each beam condition (voltage and current) and scanning scheme (dwell time and integration settings), 4 images will be captured. These four images are called the "image quadruplet." Images 1 and 2 in the quadruplet will be in a reference position, and then images 3 and 4 in the quadruplet will be in a translated position.

2. Set beam conditions to the appropriate values for the current image quadruplet.
   (a) Refer to the ImageNameConvention spreadsheet for the list of images that should be captured on the sample.
   (b) If your SEM is unable to achieve exactly the conditions specified (e.g. if the nearest dwell time available on your SEM is 3.2 µs instead of 5 µs), then use the condition that your SEM can achieve and update the image names accordingly, but do not change the precision of the number.
   (c) The analysis script uses the image names as the inputs for the data for voltage, current, dwell time, etc., so it is important that the image names reflect the actual conditions.

(d) For example, if the SEM dwell time will be 3.2 µs instead of 5 µs, an example image name from the
spreadsheet is: `B_20kV_0200 pA_05us_1Line_1Frame_loc02_pos01_rep02`

(e) For the SEM that cannot dwell at 5 µs, though, change the image name to:
`B_20kV_0200pA_03us_1Line_1Frame_loc02_pos01_rep02`

(f) Or alternatively, change it to the following, but just keep the number of characters in your image names consistent:
`B_20kV_0200pA_032us_1Line_1Frame_loc02_pos01_rep02`

3. Perform an alignment for this current image quadruplet's conditions. This alignment should include source/gun tilt, objective/lens alignment, and astigmatism correction, and the same alignment procedure must be repeated for each beam condition.

4. Record the time and chamber pressure (if this information is not captured in image metadata).

5. Adjust magnification, going from a lower to higher magnification.

   (a) Background: there is asymmetric engagement of SEM lenses

   (b) The "asymmetric engagement of additional lenses" (or path dependence in coil activation/deactivation) was recently discussed by Mello, et al., 2017, and Lenthe, et al., 2018. From Lenthe, et al. 2018: "Because of the asymmetry, particular care is required to study the influence of magnifications when lens engagement boundaries are crossed [Mello, et al.]. Images collected in this work were always approached from lower magnification when a boundary was crossed (if a relay was audible) to avoid the effect."

   (c) The SEM-DIC Challenge will follow the method from Lenthe, et al. 2018, and approach from lower magnification when a boundary is crossed (if a relay is audible) to avoid this effect.

(d) To account for this path dependence in magnification, the images should be captured after going from a lower magnification up to the higher magnification.

(e) Set the correct HFW for the current sample (for samples A through G, HFW = 200 µm; for samples H through N, HFW = 20 µm).

6. Adjust brightness and contrast with a built-in auto-adjustment.

7. Capture images 1 and 2 (the reference image pair) in the current quadruplet.

8. Translate the stage horizontally in the direction that the beam scans during one line of the image (typically the beam scans left to right, so move the center of the image to the right). The translation distance is 20 µm for samples A through G, and 2 µm for samples H through N.

9. Capture images 3 and 4 (the deformed images) in the current quadruplet.

10. Change to the next dwell time and line averaging settings for the current beam condition.

11. Translate the stage horizontally back to the previous position before Step 7 (above), in the opposite direction that the beam scans during one line of the image. For samples A through G, the translation distance is 20 µm. For samples H through N, the translation distance is 2 µm.

12. Return to step 3 and begin the next image quadruplet for the current beam condition.

13. After all four of the image quadruplets are completed for these beam voltage and current settings (16 images, with 4 images in each set of 4 different dwell time and line averaging settings), translate the stage to a new region of the speckle pattern that is subjectively similar to the previous. The distance between the previous and new region of the speckle pattern should be at least one HFW. This translation is to avoid carbon contamination and loss of contrast on the same region of the sample.

14. Change the dwell time and line averaging to match that specified in the first image at the new location by the ImageNameConvention spreadsheet.

15. Take two images at the new Location. These images will serve as a comparison for the speckle pattern quality between the locations on the speckle pattern.

16. Return to step 1 and proceed with the next beam conditions.

S2 DIC details

Table S1: Summary of the hardware and software parameters for digital image correlation (DIC) measurements.

Parameter	**Value**
SEM	See Table 2
Field of view	Either 20 µm (Ag sputtered nanodroplet samples) or 200 µm (Au nanoparticle samples)
Working distances	5, 10, or 15 mm (selected by microscopist, closest to WD they would typically use for SEM-DIC)
Dwell time	Varied between 1 and 10 µs
Patterning technique	Either Ag sputtered nanodroplets[27] or Au nanoparticle samples[10]
Approximate pattern feature size	About 5 to 50 nm (Ag nanodroplets) or about 250 nm (Au nanoparticles)
Software	Vic2D 6 (Correlated Solutions, Inc.)
Image filtering	Low-pass filter
Subset size	0.35 µm (Ag nanodroplets) and 3.5 µm (Au nanoparticles)
Step size	1/8 of the subset size
Subset shape function	Affine
Strain formulation	Finite Lagrangian
Strain filtering	Gaussian window of 5×5 DIC data points
Maximum consistency threshold	0.02 px
Maximum confidence interval	0.05 px
Maximum matchability threshold	0.1 px

S3 Pearson's correlation coefficient

To assess linear relationships between acquisition parameters and SEM-DIC error metrics, Pearson's correlation coefficients (r) were computed for each parameter–metric pair. Correlations were evaluated separately for Ag and Au samples.

The resulting correlations are summarized in Figure S1. The numerical value in each cell reports Pearson's r, indicating the strength and direction of the linear relationship between the acquisition parameter (row) and the error metric (column). Cell color encodes the associated p-value, providing a visual indication of statistical significance across the parameter space. Red denotes non-significant correlations, while yellow-to-green indicates statistically significant correlations ($p < 0.05$).

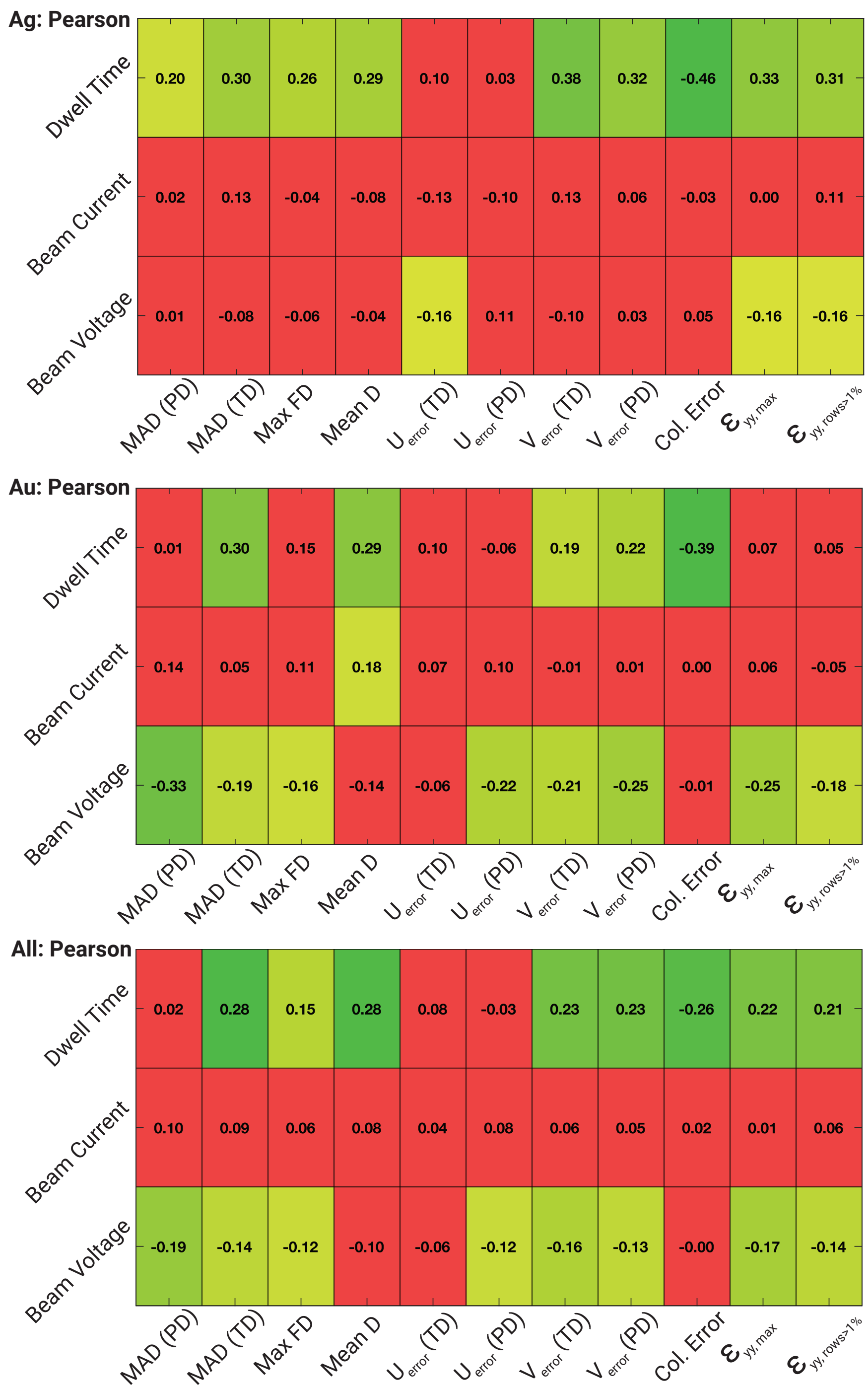

Figure S1: Pearson correlation analysis between SEM acquisition parameters and SEM-DIC error metrics for (a) Ag and (b) Au samples. Values shown in each cell correspond to Pearson's correlation coefficient (r). Cell color indicates the statistical significance of the correlation based on the associated p-value: red denotes non-significant correlations ($p \geq 0.05$), while yellow-to-green indicates statistically significant correlations ($p < 0.05$), with increasing green intensity corresponding to smaller p-values (greater statistical significance).

S4 Justification for the 1% line-jump error threshold

We would like to clarify two points regarding how the line-jump metrics are intended to be interpreted. Two complementary quantities are used to assess scanline artifacts: $\varepsilon_{yy,\mathrm{max}}$ and $\varepsilon_{yy,\mathrm{rows}>1\%}$. The metric $\varepsilon_{yy,\mathrm{max}}$ is computed for every image and provides a measure of the overall magnitude of strain excursions across the dataset. In contrast, $\varepsilon_{yy,\mathrm{rows}>1\%}$ is a more targeted diagnostic that captures the spatial extent of scanline artifacts by reporting the fraction of scanlines in an image that exceed a 1% strain threshold.

In images that do not contain scanline artifacts, no scanlines exceed this threshold, and $\varepsilon_{yy,\mathrm{rows}>1\%}$ is therefore zero (within the resolution of the distribution). This behavior is expected and intentional. The metric is not designed to distinguish among artifact-free images, but rather to flag and quantify those images that exhibit pronounced line jumps.

Figure S1 is included to illustrate this intended operating regime, not to characterize variability across the entire image population. As shown, 80% of images fall into the first histogram bin of $\varepsilon_{yy,\mathrm{rows}>1\%}$, corresponding to values near zero (0–0.0106). These images contain no scanlines exceeding the 1% strain threshold and are therefore considered nominal with respect to line-jump artifacts.

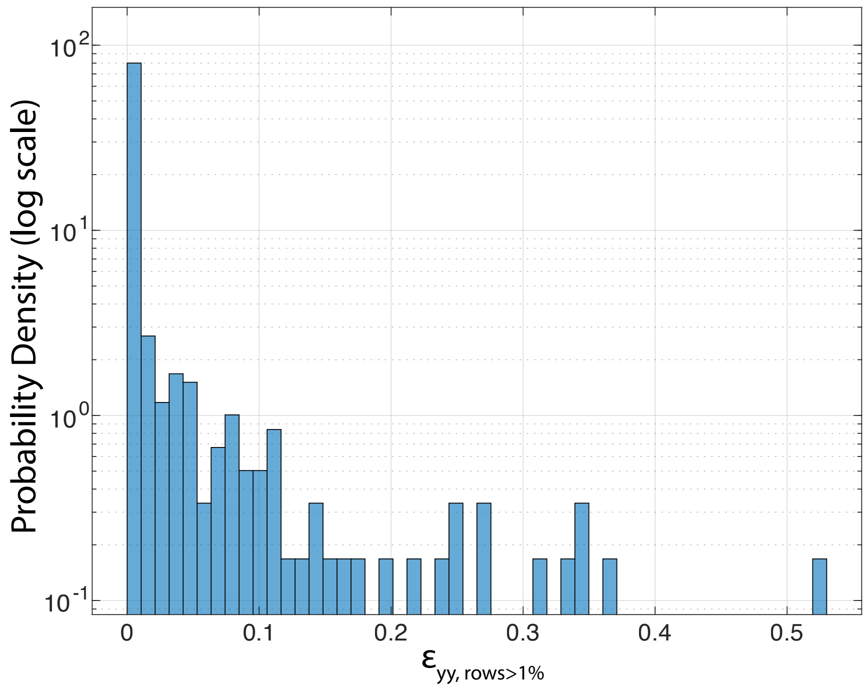

Figure S1: Distribution of $\varepsilon_{yy,\text{rows>1\%}}$ across all images. Approximately 80% of images fall within the first histogram bin, corresponding to nominal behavior with no scanlines exceeding the 1% strain threshold. The remaining $\sim$ 20% of images populate a long tail, indicating the presence of scanline discontinuities. While $\varepsilon_{yy,\text{rows>1\%}}$ is intentionally sparse for nominal images, it provides a quantitative measure of line-jump severity for this subset of affected frames.

Accordingly, the purpose of $\varepsilon_{yy,\text{rows>1\%}}$ is to isolate and analyze the minority of images affected by severe scanline discontinuities. Analysis of this metric therefore focuses on the remaining $\sim$ 20% of images in the tail of the distribution, where it provides a physically meaningful measure of the fraction of scanlines impacted by line jumps.

Population-wide comparisons across all images, including those without scanline artifacts, are instead performed using $\varepsilon_{yy,\text{max}}$, which remains sensitive to overall strain excursions regardless of their origin. The two metrics thus serve complementary roles: $\varepsilon_{yy,\text{max}}$ for global comparison across the full dataset, and $\varepsilon_{yy,\text{rows>1\%}}$ for targeted assessment of severe scanline corruption.

S5 Conductivity

Charging of the sample is likely to increase SEM-DIC errors. To better understand the influence of charging on specific error types, both substrate conductivity and sample conductivity were examined. Substrate conductivity was varied with three different Si wafer resistances. Sample conductivity was varied with pre- and post-sputter-coated images.

Substrate conductivity

Two additional samples were created in order to study varying substrate conductivity on SEM-DIC error. The two Si wafers, one of 10^{-4} $k\Omega$·cm and one of 10 kΩ·cm, were patterned with Au nanoparticles in accordance with the main study sample procedure[31]. Entire image sets were taken on both additional samples with the procedure outlined in the experimental procedure section of the main study. The main study sample taken on SEM-5 was also used in this analysis. The substrate conductivity of the main study samples was 0.1 $k\Omega$·cm.

These multivariate linear regression results show substrate conductivity to have the third greatest effect on the overall image error (MAD Position 1 Repetition 2, MAD Position 2 Repetition 1) (Figure S2). Substrate conductivity has a greater effect on these metrics than beam current and voltage, yet not as great of an effect compared to line averaging and dwell time.

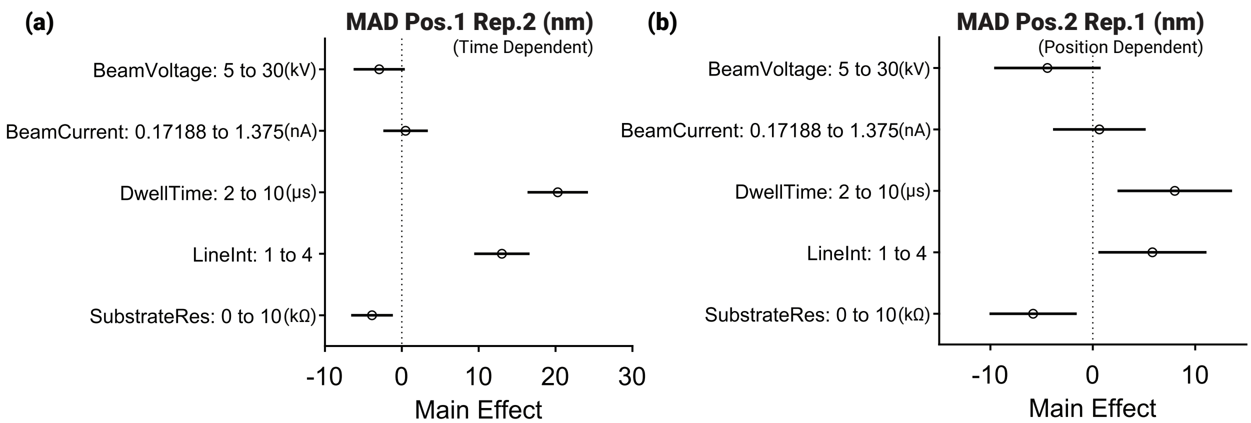

Figure S2: Multivariate regression results of the overall image displacement metrics including substrate resistance.

Table S2 outlines the multivariate regression calculated effect of substrate conductivity on each error metric: short-term drift (Mean D), the LES (Column Error), horizontal and vertical skewing (U_{error} and V_{error}), and line jumps. It is shown that substrate conductivity has very little effect on V_{error} and Position 1 U_{error}, as seen by the Pearson values. Alternatively, MAD Position 1 Repetition 2, MAD Position 2 Repetition 1, Max FD Position 1 Repetition 2, Mean D, and U_{error} Position 2 Repetition 1 are all shown to reduce error with the high resistance substrate. $\varepsilon_{yy,\max}$ and $\varepsilon_{yy,\text{rows}>1\%}$ were shown to be reduced at low resistance substrates.

Error Metric	Range	Predicted Response	Pearson Value
MAD Position 1 Repetition 2	0-10 kΩ	-3.877	**-0.2004**
MAD Position 2 Repetition 1	0-10 kΩ	-5.829	**-0.2158**
Max FD Position 1 Repetition 2	0-10 kΩ	-54.739	**-0.231**
Mean D Position 1 Repetition 2	0-10 kΩ	-45.956	**-0.2125**
U_{error} Position 1 Repetition 2	0-10 kΩ	-0.945	-0.0501
U_{error} Position 2 Repetition 1	0-10 kΩ	-2.518	**-0.2502**
V_{error} Position 1 Repetition 2	0-10 kΩ	-0.686	-0.0274
V_{error} Position 2 Repetition 1	0-10 kΩ	-25.655	-0.0935
$\varepsilon_{yy,\max}$ Position 1 Repetition 2	0-10 kΩ	0.561	**0.229**
$\varepsilon_{yy,\text{rows}>1\%}$ Position 1 Repetition 2	0-10 kΩ	1.344	0.1028

Table S2: Relationship between error metric, responses predicted from multivariate regression, and Pearson values for image sets with different substrate conductivity.

This result is contrary to what was expected. Literature establishes that sample charging is a hypothesis for many SEM-DIC errors. However, in this analysis, error is reduced when greater sample charging should occur (high resistance substrate: low conductivity). The cause for this unexpected trend is not known at this time. However, it may be due to different aspects of the SEM charging (i.e. stage charging vs. sample charging vs. chamber charging). Future work includes investigating this trend more thoroughly.

Unlike the other error metrics, line jumps (magnitude measured by $\varepsilon_{yy,\mathrm{max}}$ and quantity measured by $\varepsilon_{yy,\mathrm{rows}>1\%}$) increase with the reduction of substrate conductivity.

Sputter coating

The main study used both Au and Ag patterned samples imaged at different magnifications. It should be noted that in using these different patterning methods, the samples had different resistances. It was proposed that sputter coating the samples could aid in differentiating the effect of magnification versus sample resistance. For this study, samples were sputter coated with gold palladium (Au-Pd) for 3 minutes.

The minimum and maximum current, voltage, and dwell time conditions from the main study were used in this analysis. Dwell times of 10 µs and 2 µs, beam voltages of 5 kV and 30 kV, and currents of 200 pA and 1000 pA were used. Therefore, 8 image quads were taken pre-sputter coating and post-sputter coating on both the Ag and Au samples. This study was unique from the main study as all images were taken at the same location throughout the experiment.

Figure S3 outlines the robust overall image displacement of the pre- and post-sputter-coated samples. It can be seen that sputter coating has little effect on overall SEM-DIC error when compared to dwell time. Additionally, multivariate regression results show Position 1 Repetition 2 overall image error (MAD) to increase after sputter coating, whereas Position 2 Repetition 1 error decreases. Regardless of this conclusion, the wide 95% confidence interval shown in each of these plots should be noted (−2.91 to 5.60 for

Position 1 Repetition 2 and −0.914 to 3.28, for Position 2 Repetition 1). In both robust overall image error (MAD) main effect plots, it is seen that sputter coating has a larger effect on the error than sample type. MAD Position 1 Repetition 2 is calculated to increase error in the Ag samples by 0.494 nm where the post sputter-coated sample increases error by 1.35 nm. Alternatively, for MAD Position 2 Repetition 1, the Ag samples decrease error by 0.708 nm and the post sputter-coated samples decrease error by 1.18 nm. Based on these results, it is not likely that sample patterning conductivity greatly influences overall SEM-DIC results. We therefore hypothesize that the error in the main study due to sample type is more greatly influenced by magnification and pattern quality rather than sample conductivity. A deeper study on magnification is provided in Section 3.7.3.

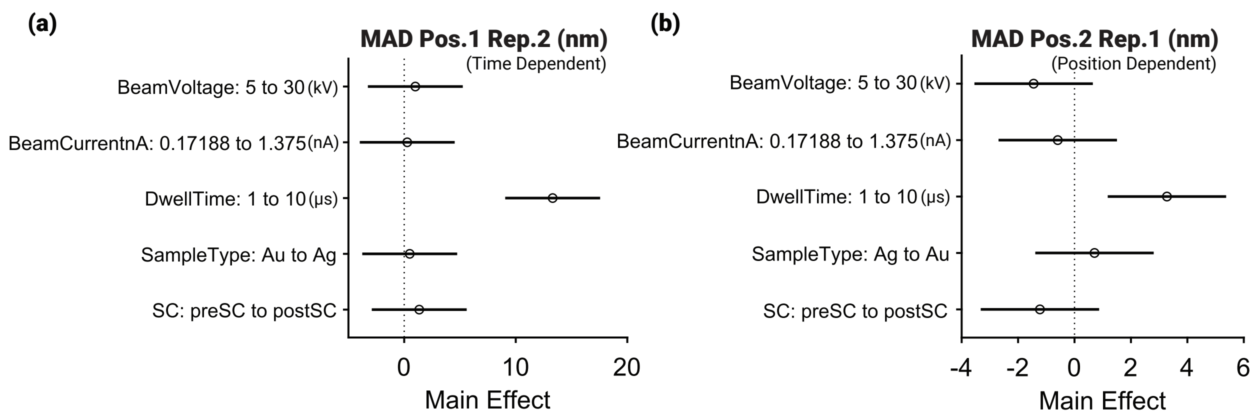

Figure S3: Multivariate regression results of the MAD including sputter coating as a predictor

Figures S4 and S5 outline the trends observed across all error metrics for the pre- and post-sputter-coated samples. For the Au samples, Figure S4, it is shown that error is reduced across MAD Position 2 Repetition 1, Max FD, Mean D, U_{error}, V_{error} Position 2 Repetition 1, and Column Error. For Position 1 Repetition 2 MAD and V_{error}, sputter coating actually increases the error. Alternatively, for the Ag samples, MAD Position 1 Repetition 2, U_{error} Position 1 Repetition 2, and V_{error} Position 1 Repetition 2 are seen to decrease after sputter coating. MAD Position 2 Repetition 1, Max FD, Mean D, U_{error} Position 2 Repetition 1, and V_{error} Position 2 Repetition 1, and Column Error increase the error after sputter coating.

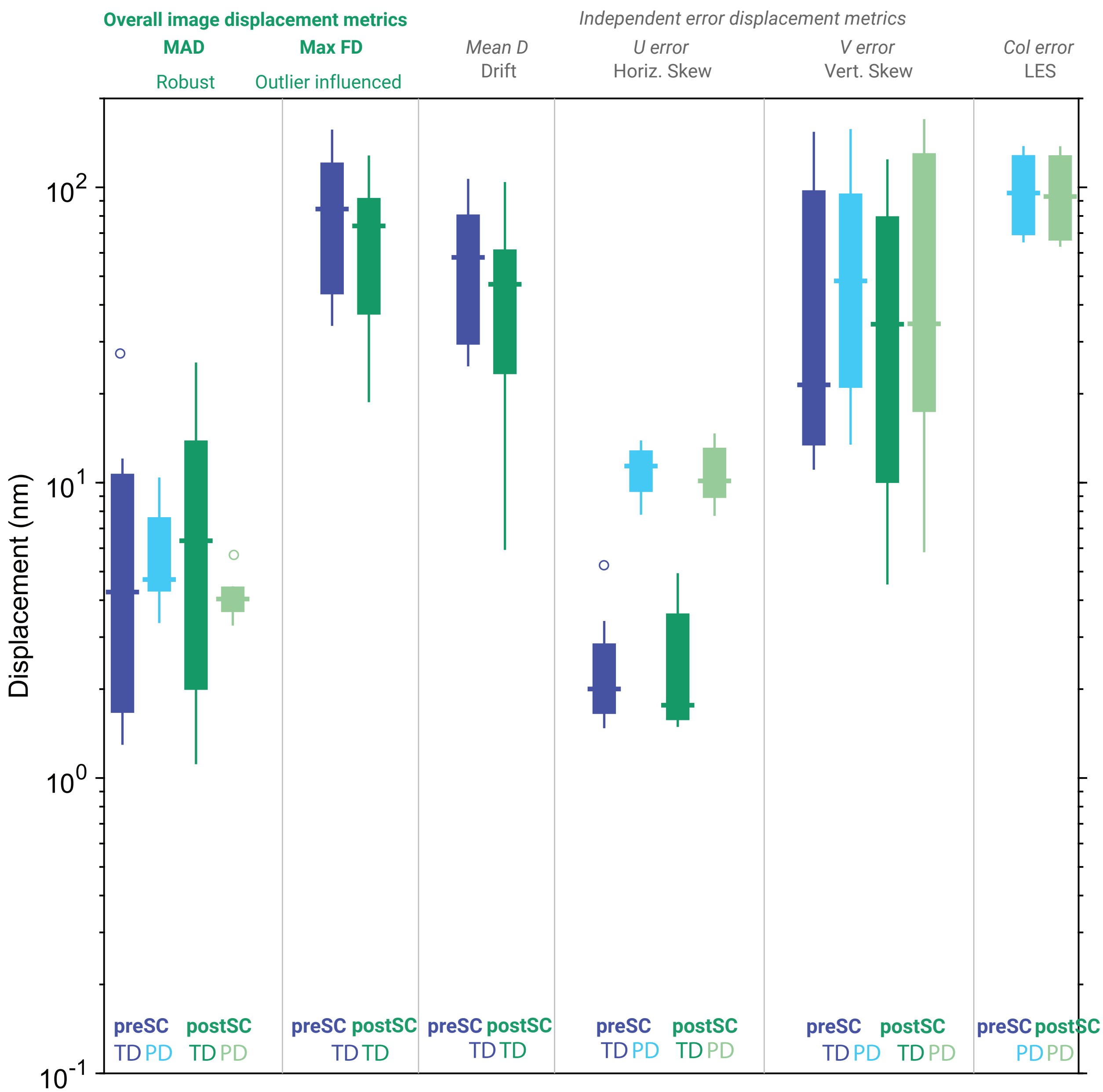

Figure S4: MAD Position 1 Repetition 2 (time-dependent, TD), MAD Position 2 Repetition 1 (position-dependent, PD), and Max FD error comparing the pre-sputter-coated images to the post-sputter-coated images for the Au samples.

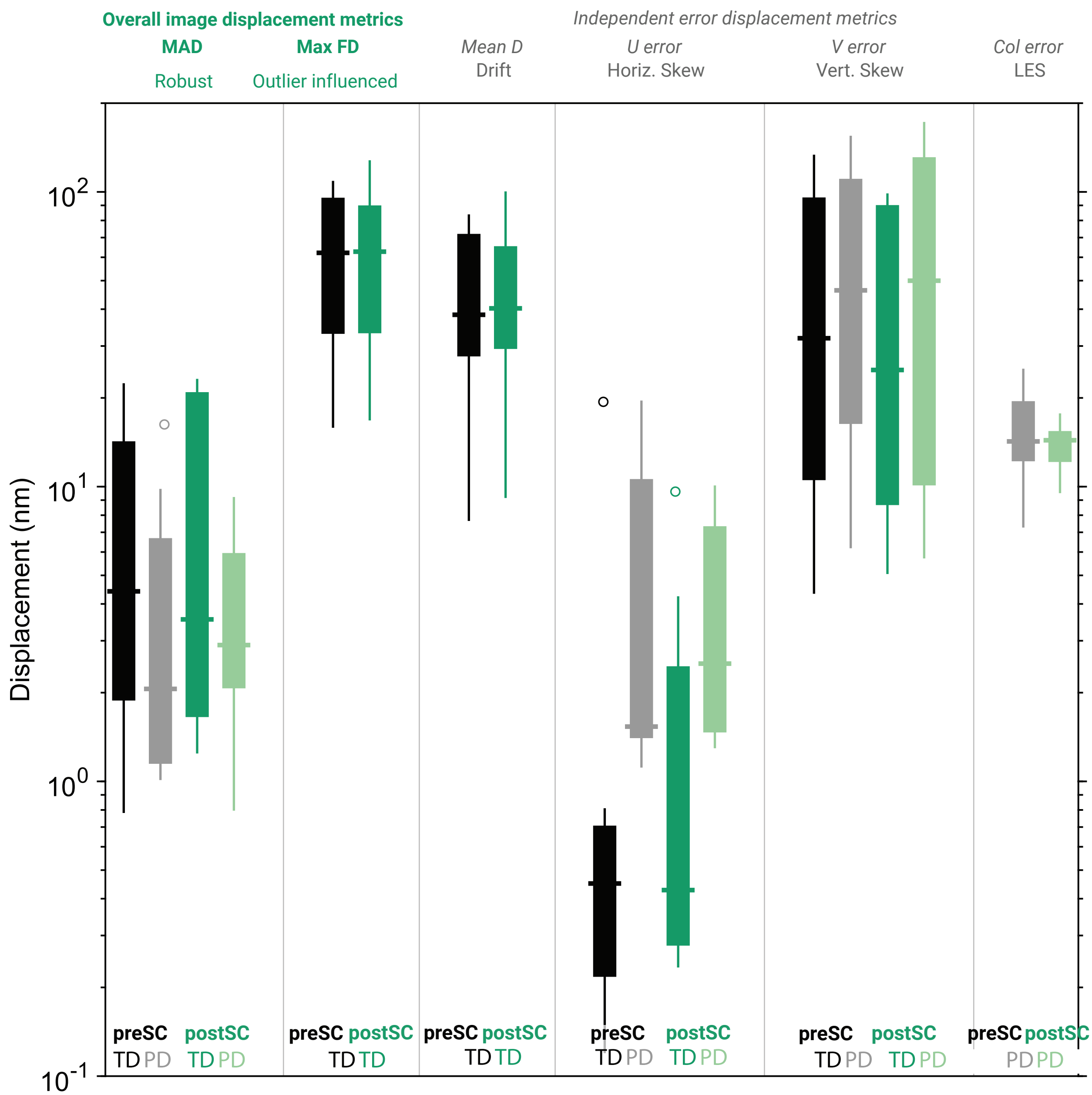

Figure S5: MAD Position 1 Repetition 2, MAD Position 2 Repetition 1, and Max FD error comparing the pre-sputter-coated images to the post-sputter-coated images for the Ag samples.

To capture the long time scale drift translation, the Ag pre- and post-sputter-coated sample was used for a 12 hour drift analysis. High-power conditions were used and the long time scale drift procedure was followed. Figure S6 outlines the results of the pre- and post-sputter-coated samples aligned with the high- and low-power condition drift study.

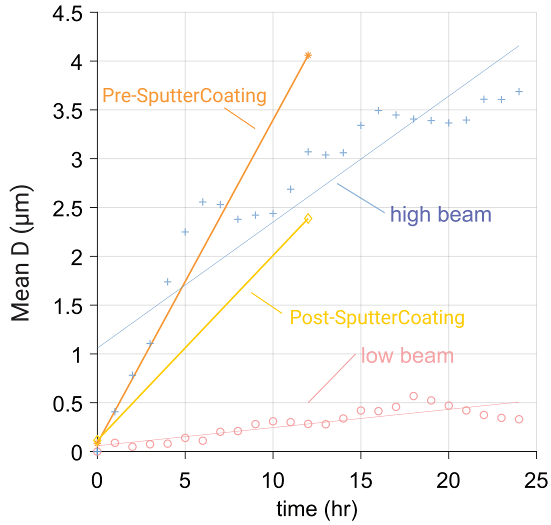

Figure S6: Rigid-body displacement of 24 hour drift study compared to results of shorter time scale sputter-coated sample study.

Drift velocity was calculated to be 338.3 nm/hr for the pre-sputter-coated sample and 199.0 nm/hr for the post-sputter-coated sample. Ultimately, these data show that while increasing the conductivity of the sample only marginally reduces errors other than drift, drift over several hours is reduced by approximately forty percent. However, it should be noted that a sputter-coated film may crack under loading and interfere with SEM-DIC. While the MVR results show some influence of sputter coating, the large confidence interval suggests that conductivity only marginally influences overall image error. We therefore hypothesize that the error in the main study due to sample type is more greatly influenced by magnification or pattern shape rather than sample conductivity. Future efforts could quantify and compare the overall effect of magnification and sample type on error with multiple SEMs and samples for a more thorough perspective.