Defect detection in magnetic systems using U-Net and statistical measures

Ross Knapman [1,2,*] Atreya Majumdar [1] Nasim Bazazzadeh [1,3] Kübra Nur Kalkan [1] Katharina Ollefs [1,4] Oliver Gutfleisch [5] and Karin Everschor-Sitte [1]

[1] *Faculty of Physics and Center for Nanointegration Duisburg-Essen (CENIDE), University of Duisburg-Essen, 47057 Duisburg, Germany*

[2] *Institute of Mechanics, Faculty of Engineering, University of Duisburg-Essen, 45141 Essen, Germany*

[3] *Radiology Department, Massachusetts General Hospital, 175 Cambridge Street, Boston, Massachusetts 02114, USA*

[4] *Kirchhoff Institute for Physics, University of Heidelberg, 69120 Heidelberg, Germany*

[5] *Functional Materials, Institute of Materials Science, Technical University of Darmstadt, 64287 Darmstadt, Germany*

(Received 30 March 2026; revised 10 June 2026; accepted 24 July 2026; published 24 August 2026)

Local material inhomogeneities can strongly influence magnetization dynamics and macroscopic magnetic properties, yet detecting such defects from magnetic imaging data remains challenging when thermal fluctuations and experimental noise obscure static contrast. Here, we investigate defect detection in strongly fluctuating magnetization regimes, where signatures of inhomogeneities largely average out in time-resolved measurements. Using finite-temperature micromagnetic simulations with randomly distributed defects and material parameters representative of $Ni_{80}Fe_{20}$, we compute per-pixel temporal mean, temporal standard deviation, and latent entropy and use them as inputs for U-Net-based semantic segmentation models. We find that the most effective descriptor depends on the noise level and, importantly, that robust detection requires training data that reflect the expected noise statistics. These results provide practical guidance toward designing noise-robust defect-detection workflows in magnetic imaging.

DOI: 10.1103/czsp-gnmy

I. INTRODUCTION

Local material inhomogeneities can strongly influence magnetization dynamics in bulk and thin-film systems. Small local variations in exchange interaction, magnetic anisotropy, or easy-axis orientation can modify domain nucleation and domain-wall motion, thereby affecting macroscopic magnetic properties such as coercivity and hysteresis [1–10]. Accurately locating and characterizing such defects is therefore important for both fundamental studies and technological applications.

Magnetic imaging techniques provide spatially resolved information on magnetization behavior, but subtle inhomogeneities can nevertheless be difficult to detect. In many situations, particularly at finite temperature or in strongly fluctuating systems, the magnetization evolves rapidly in time, so that defect-related signatures may largely average out in time-averaged magnetization images.

*Contact author: ross.knapman@uni-due.de

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

As a result, conventional static indicators such as domain-wall width or persistent contrast features may not reliably reveal the underlying inhomogeneities. Experimental noise can further obscure weak signals in the data [11–14]. Recently, machine-learning approaches, especially U-Net-style convolutional neural networks, have demonstrated promising capabilities for identifying spatial features in magnetic systems from imaging data [15–17].

In this work, we mainly investigate defect detection in highly fluctuating regimes, where static contrast is weak, and information about defects can instead be extracted from magnetization dynamics. We develop and evaluate a combined approach that integrates statistical measures of temporal behavior with semantic segmentation. From finite-temperature micromagnetic simulations containing random defects, we compute per-pixel temporal mean, temporal standard deviation, and latent entropy—a physics-motivated measure of transition stochasticity or predictability inferred from temporally ordered state transitions [13,14]. Each measure is used separately as input for training U-Net segmentation models. For concreteness, the simulations employ material parameters representative of $Ni_{80}Fe_{20}$. By focusing on strongly fluctuating data, the detection task relies on robust dynamical signatures rather than on static structural features.

We evaluate segmentation performance across magnetization components and under different levels of postprocessed noise applied to the simulated magnetization time series, using additive and multiplicative noise models that represent common experimental noise contributions.

For the simulated magnetization dynamics considered here, we find that the temporal mean, temporal standard deviation, and latent entropy allow for defect detection at low noise levels. As the noise increases, however, we find that typically the temporal standard deviation and latent entropy emerge as the most effective measures. The temporal standard deviation primarily captures the Gaussian fluctuations of the magnetization dynamics, whereas the latent entropy is sensitive to more general stochastic transition behavior [13] and therefore goes beyond purely Gaussian statistics. For the considered magnetization dynamics, however, this additional sensitivity plays only a minor role, leading to comparable performance of the two measures, with the temporal standard deviation performing slightly better at lower noise levels. Importantly, we also observe that robustness under moderate noise improves strongly when the training data reflect the noise statistics of the target data.

The present work provides a reproducible workflow for detecting localized magnetic inhomogeneities in thermally fluctuating magnetization data, where defects may not be visible in individual magnetization images. The workflow links finite-temperature micromagnetic simulations, temporal feature extraction, and U-Net-based segmentation to compare temporal descriptors and support future extensions toward more realistic defect models and experimental datasets.

The remainder of the article is organized as follows. In Sec. II, we present the approach, consisting of the statistical measures and training procedures based on the synthetic data obtained from micromagnetic simulations. Sections III and IV present the main segmentation results and discuss their behavior under varying noise conditions.

II. COMPUTATIONAL WORKFLOW FOR DEFECT DETECTION

For automated defect detection, we followed a three-step workflow: (1) run finite-temperature micromagnetic simulations of magnetic films with randomly distributed defect regions to generate synthetic magnetization dynamics data; (2) extract scalar measures (temporal mean μ, temporal standard deviation σ, and latent entropy LE) from each pixel time series; and (3) train U-Net models to automatically segment defects from the resulting two-dimensional (2D) measure maps.

A schematic overview of the full processing pipeline is shown in Fig. 1; the following subsections describe these three steps.

A. Micromagnetic simulations

Synthetic magnetization dynamics data were generated using finite-temperature micromagnetic simulations of a thin ferromagnetic film with in-plane uniaxial anisotropy and randomly distributed

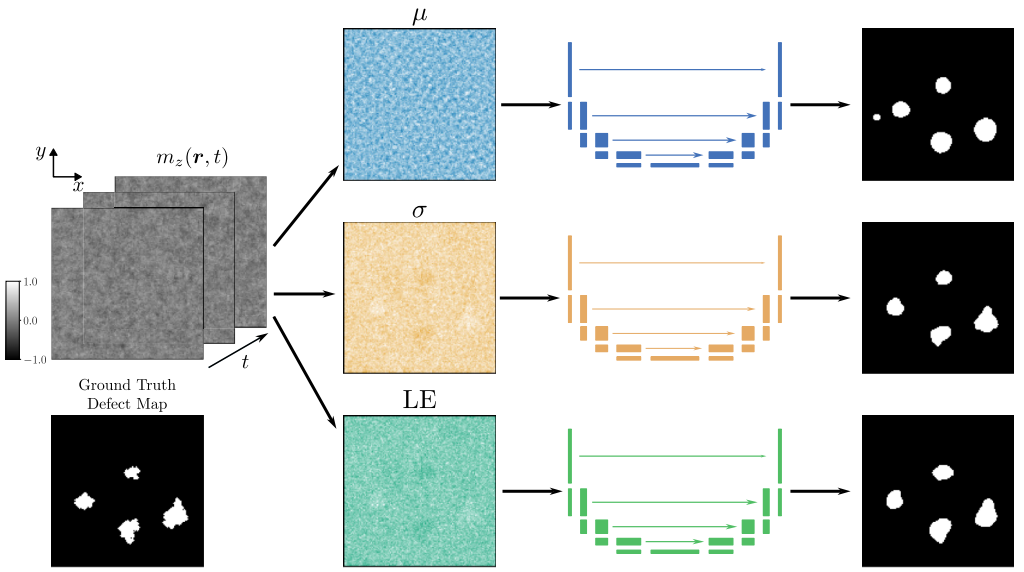


FIG. 1. Schematic overview of the data processing pipeline. Highly fluctuating time-series magnetization fields obtained from finite-temperature micromagnetic simulations are processed on a per-pixel basis to compute three statistical measures: the temporal mean μ (top, blue), temporal standard deviation σ (center, yellow), and latent entropy LE (bottom, green). Each measure is then provided separately as a single-channel input to a U-Net trained to detect defect regions. For the example shown, the standard deviation and latent entropy perform best.

impurities at approximately room temperature. The magnetization dynamics were modeled by the stochastic Landau-Lifshitz-Gilbert equation

$$\frac{\mathrm{d}\boldsymbol{m}}{\mathrm{d}t} = -\gamma\, \boldsymbol{m} \times (\boldsymbol{H}_{\mathrm{eff}} + \boldsymbol{H}_{\mathrm{th}}) + \alpha\, \boldsymbol{m} \times \frac{\mathrm{d}\boldsymbol{m}}{\mathrm{d}t}, \tag{1}$$

where $\boldsymbol{m}(\boldsymbol{r},t)$ is the unit magnetization vector with Cartesian components (m_x, m_y, m_z), γ the gyromagnetic ratio, α the Gilbert damping parameter, and $\boldsymbol{H}_{\mathrm{th}}$ a Langevin thermal field representing fluctuations at temperature T [18–20]. In this work, we report results for the in-plane component m_x and the out-of-plane component m_z. The component m_y is omitted due to the normalization condition $|\boldsymbol{m}| = 1$. The effective magnetic field $\boldsymbol{H}_{\mathrm{eff}} = -(\mu_0 M_{\mathrm{s}})^{-1}\, \delta E/\delta \boldsymbol{m}$ is obtained from a micromagnetic energy density functional with spatially varying exchange and uniaxial anisotropy, allowing us to distinguish between the background material and localized defect regions,

$$E = \int \left[A(\boldsymbol{r})(\nabla \boldsymbol{m})^2 + K(\boldsymbol{r})(1 - (\boldsymbol{m} \cdot \hat{\boldsymbol{u}}(\boldsymbol{r}))^2) - \frac{\mu_0}{2} M_{\mathrm{s}}\, \boldsymbol{m} \cdot \boldsymbol{H}_{\mathrm{d}} \right] \mathrm{d}V, \tag{2}$$

where $A(\boldsymbol{r})$ is the exchange stiffness, $K(\boldsymbol{r})$ the uniaxial anisotropy constant, $\hat{\boldsymbol{u}}(\boldsymbol{r})$ the local easy-axis direction, and $\boldsymbol{H}_{\mathrm{d}}$ the demagnetizing field. Simulations were performed on a $128 \times 128 \times 1$ grid with cell dimensions (2 nm, 2 nm, 10 nm) using `MuMax3` [21].

Background material parameters were set to values representative of $Ni_{80}Fe_{20}$: $A = 1.0 \times 10^{-11}\,\mathrm{J\,m^{-1}}$, $K = 1.5 \times 10^{2}\,\mathrm{J\,m^{-3}}$, and everywhere, $M_{\mathrm{s}} = 8.4 \times 10^{5}\,\mathrm{A\,m^{-1}}$, damping $\alpha = 0.01$ [22], $T = 300\,\mathrm{K}$, with the background easy axis aligned along the x direction, i.e., $\hat{\boldsymbol{u}} = \hat{\boldsymbol{x}}$.

Four nonoverlapping inhomogeneity regions were placed at random positions for each simulation run. Each inhomogeneity was generated on the 128×128 grid as an irregular, simply connected blob with a target size on the order of a few cells. Within inhomogeneity regions, exchange stiffness and anisotropy were scaled relative to the background with fixed factors: Two inhomogeneity regions were set to 1.2 times the background value and two to 0.8 times the background value in

every simulation (i.e., this scaling pattern was not randomized between runs). The in-plane easy-axis orientation was randomly perturbed around $\hat{\boldsymbol{x}}$ by drawing angles from a Gaussian distribution with a 20° standard deviation, while the background anisotropy remained along $\hat{\boldsymbol{x}}$.

The full dataset consisted of 2048 independent simulations, with ground-truth defect masks constructed directly from the known defect regions in each simulation. Initial magnetization was uniformly aligned along the easy axis x. Each simulation extended over 1 ns with the magnetization output every 10 ps (100 frames). From each simulation, we then derived per-pixel summary measures from the magnetization time series.

B. Measures and latent entropy

To capture different aspects of local dynamics, we computed per-pixel summary measures from magnetization time series: temporal mean μ, temporal standard deviation σ, and latent entropy LE [13,14]. To assess robustness to experimental noise, we synthetically perturbed each magnetization-component time series (here, m_x or m_z) before computing the measures. We considered two standard zero-mean noise models with variance σ_n^2: additive Gaussian noise,

$$\tilde{m}_i = m_i + n, \quad n \sim \mathcal{N}\big(0, \sigma_n^2\big), \tag{3}$$

and multiplicative speckle noise,

$$\tilde{m}_i = m_i(1 + n), \quad n \sim \mathcal{N}\big(0, \sigma_n^2\big). \tag{4}$$

The measures μ, σ, and LE were then recomputed from the noisy series. The latent entropy [13] was computed for each pixel time series (m_x or m_z). The signal was discretized into $n_{\text{bins}} = 5$ states, and the latent entropy was evaluated following the procedure described in Ref. [13].

C. U-Net architecture and model training

For semantic segmentation of defect regions, we used a U-Net architecture [23] trained on single-channel measure maps (μ, σ, or LE) for each magnetization component (m_x, m_z). Separate models were trained for each measure/component combination, and robustness was tested using both clean and noise-matched training conditions. Full details of the architecture, optimization setup, data splitting, normalization, and model selection are provided in the Appendix.

III. RESULTS

We evaluated the ability of a U-Net to segment defect regions from single-channel inputs derived from the magnetization time series on the held-out test set. Representative predictions are shown in Fig. 2 and quantitative performance is summarized in Fig. 3, which compares models trained on clean data with models trained using matching noisy data under clean, additive Gaussian, and multiplicative speckle evaluation conditions.

A. Baseline performance on clean data

On clean (noise-free) inputs, Fig. 2 already shows a strong component dependence in visual contrast. For m_x, defect regions are visible in the mean map, while for m_z the mean map shows much weaker visual contrast between defect and background regions. This qualitative difference is consistent with the quantitative Dice scores in Fig. 3. For the in-plane component m_x, the temporal mean provides the strongest single-channel signal (baseline Dice $\approx$0.89), followed by the temporal standard deviation ($\approx$0.85) and latent entropy ($\approx$0.81). For the out-of-plane component m_z, the temporal mean is markedly less informative (baseline Dice $\approx$0.65), whereas the temporal standard deviation ($\approx$0.84) and latent entropy ($\approx$0.80) remain competitive.

Across all measures, the predicted masks in Fig. 2 also tend to have smoother, more rounded boundaries than the jagged pixel-level contours in the ground-truth masks. This can be partially

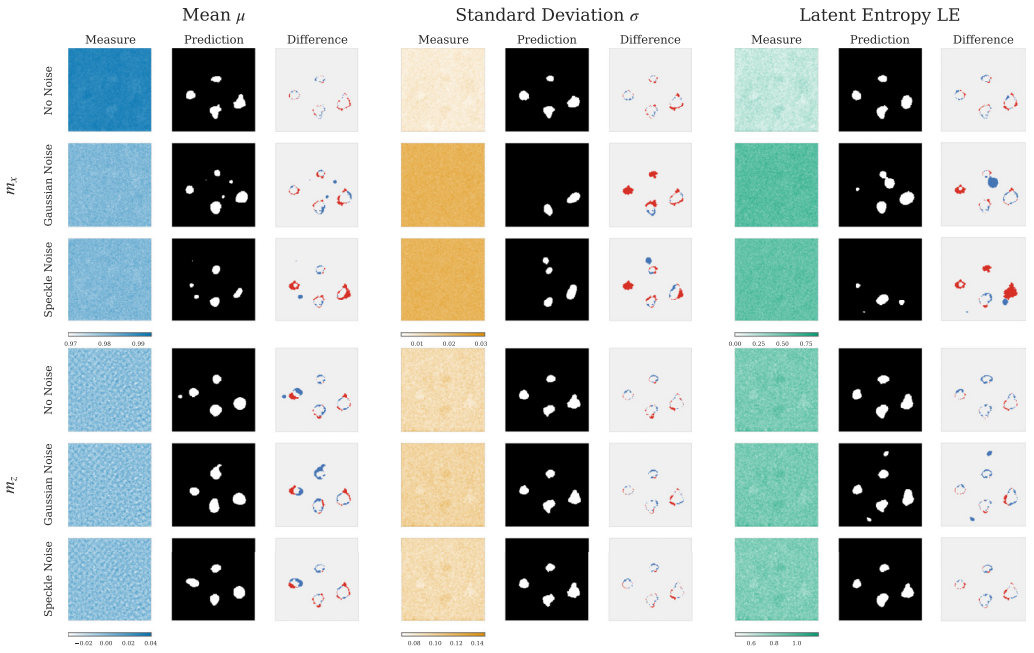


FIG. 2. Qualitative prediction overview for defect segmentation using the three input measures: temporal mean (μ), temporal standard deviation (σ), and latent entropy (LE). Columns are grouped by measure, and within each group show (from left to right) the input measure map, the U-Net prediction, and the prediction difference relative to the ground truth shown in Fig. 1 (false positives/negatives highlighted in blue/red, respectively). Rows are organized by magnetization component (m_x top block, m_z bottom block) and noise condition (no noise, additive Gaussian noise, and multiplicative speckle noise). Noisy examples are evaluated using models trained with matching noise statistics.

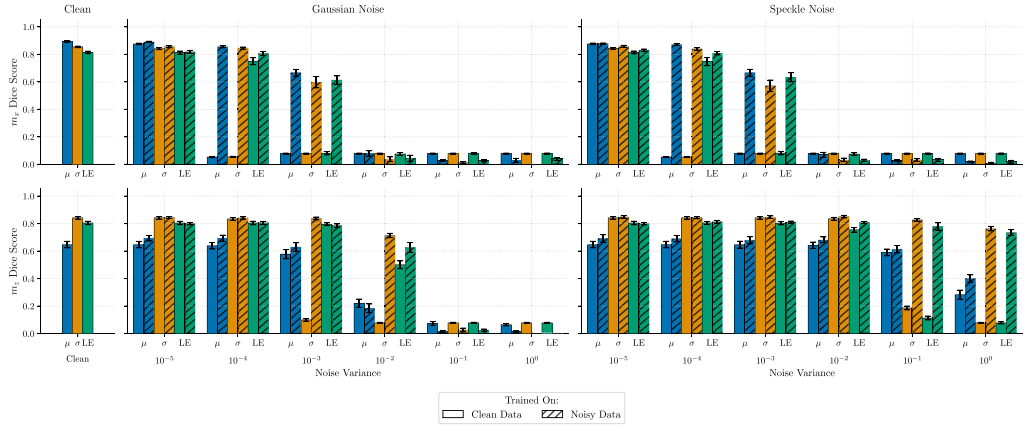


FIG. 3. Test-set Dice coefficients for defect segmentation from the three single-channel inputs (μ, σ, LE). Top row: m_x; bottom row: m_z. The left column reports clean-data performance, while the center and right columns show performance vs noise variance for additive Gaussian and multiplicative speckle noise, respectively. For each measure and noise level, results are shown for models trained on clean data and for models trained with matching noisy data, as indicated in the legend.

explained by the static magnetization having a characteristic length scale $\sqrt{A/K}$, which results in smooth transitions between the defect regions and background. Furthermore, this behavior is expected for U-Net-style segmentation: The encoder-decoder pathway with pooling and upsampling favors spatially coherent regions, and optimization of pixelwise cross-entropy and Dice overlap rewards correct area overlap more than reproducing high-frequency boundary roughness. After thresholding soft output probabilities, small one-pixel protrusions in the labels are therefore often suppressed, leading to slightly smoother predicted defect contours [24,25].

B. Robustness under additive (Gaussian) noise

Under weak Gaussian noise (variance 10^{-5}), Dice scores remain close to their clean baselines for all measures (Fig. 3), and noise-augmented training provides only marginal improvements. At moderate noise levels (variance 10^{-4} to 10^{-3}), the benefit of matched noise augmentation becomes pronounced: Models trained only on clean data fail (Dice $\lesssim 0.1$ for several measures), whereas noise-trained models retain high accuracy. For example, for m_x at variance 10^{-4}, noise-trained models achieve Dice $\approx$0.85 (mean) and $\approx$0.85 (standard deviation), compared with $\approx$0.05 for the corresponding clean-trained models. For m_z, the standard-deviation input is particularly robust when trained with matching noise (Dice $\approx$0.84 at variance 10^{-4} and $\approx$0.84 at variance 10^{-3}). Latent entropy benefits similarly from noise-aware training at these noise levels, but remains below the best-performing measure for each component.

At large Gaussian noise variance ($\gtrsim 10^{-2}$), performance degrades substantially even for noise-trained models, with Dice scores dropping below $\sim$0.1 for most measure/component combinations. Representative noisy examples at variance 10^{-3} are shown in Fig. 2, where the input measures are strongly distorted and the predicted masks show missed defects and false positives, consistent with the sharp deterioration in quantitative scores. At the highest noise levels, the two training strategies fail in qualitatively different ways, which explains the nonzero Dice values visible in Fig. 3 for several clean-trained results (for both Gaussian and speckle noise). Clean-trained models often collapse to near-all-defect predictions (the defect class assigned to most pixels). Because the true defect mask occupies only a limited but nonzero area, this near-constant prediction can still yield a finite Dice overlap. In contrast, models trained with noisy data at these extreme noise levels more often collapse to near-all-nondefect predictions (defects mostly missed), for which the Dice score tends toward zero. Thus, the residual Dice of clean-trained models at very high noise should not be interpreted as better segmentation quality, but as a different failure mode.

C. Robustness under multiplicative (speckle) noise

The qualitative trends under speckle noise are similar at low-to-moderate variance, with matched noise augmentation again being critical once the noise level exceeds the weak-noise regime (Fig. 3). However, compared to additive Gaussian noise, multiplicative speckle noise is better tolerated for certain inputs and components. In particular, for m_z the standard-deviation and latent-entropy inputs remain robust over a wider range of speckle variances when trained with matching noise (Dice remaining $\gtrsim$0.73 even at variance 1.0), whereas m_x performance drops more rapidly at high speckle variance. This component dependence suggests that the out-of-plane fluctuations captured by m_z contain a comparatively stable signature of inhomogeneity regions under multiplicative intensity distortions. Within the high-variance tail for m_x, latent entropy can also edge out the mean and standard-deviation inputs (e.g., speckle variances 0.1 and 1.0), though all measures are weak in that regime.

IV. DISCUSSION

The results highlight three main messages for defect detection in magnetic imaging. First, the most informative input measure depends on the specific magnetization dynamics and the associated fluctuations, which are determined by the physical model and may manifest differently across

magnetization components. For the dynamics considered here, the temporal mean of m_x produced the strongest defect contrast, whereas for m_z the temporal standard deviation was more informative. This indicates that feature design should be selected for the specific signal channel and contrast mechanism, rather than assumed to transfer unchanged between magnetization components.

Second, robustness is determined primarily by how well training data match inference conditions. Models trained only on clean inputs performed well on clean test data, but degraded strongly under moderate noise. In contrast, training with matched noise preserved high Dice scores over a broad range of Gaussian and speckle perturbations. For future applications to measured data, this supports two complementary strategies: incorporating representative noise during training and updating pretrained models with domain-adaptation or transfer-learning steps when measurement conditions shift. One possible interpretation of the particularly fast drop for clean-trained m_x models is that they may rely more heavily on high-contrast intensity cues that are prominent in clean m_x maps, whereas m_z models are forced to use subtler structure already in the clean case. Under added noise, those dominant m_x cues may be disrupted first, causing a sharper loss in segmentation quality.

Third, latent-entropy maps are a useful complementary descriptor. They were competitive with standard-deviation maps on clean data and remained particularly robust for m_z under speckle noise, but they were not consistently superior across all regimes for the magnetization dynamics considered here.

For future applications to experimental data, our results suggest the following strategy. The temporal mean can be preferred when strong in-plane–like contrast is present (analogous to m_x), whereas the temporal standard deviation or latent entropy is more suitable when the out-of-plane-like contrast is weaker (analogous to m_z). As a default noise model, we recommend additive Gaussian noise, since additive-noise descriptions are commonly employed in magnetic imaging acquisition and denoising workflows, where detector-related noise often represents a key practical limitation [11,12]. Rather than relying on training with clean data only, the noise level should be estimated from background regions, and the model should be trained or fine-tuned using data with matching noise statistics, for example, obtained from a short calibration dataset.

Beyond these central findings, this work provides a controlled benchmark that can be extended systematically. We deliberately used synthetic defect distributions, a single geometry, and a restricted range of defect morphologies and parameter variations to enable reproducible comparisons across measures and noise models. Likewise, the single-channel training design isolated the contribution of each descriptor. These controlled choices provide a clear basis for future investigations that include more realistic defect landscapes, additional material inhomogeneities, different spatial scales, and experimental time-resolved magnetic imaging data. To support such extensions by the community, the accompanying implementation provides reusable tools for data generation, temporal feature extraction, neural-network training, and evaluation [26]. A natural next step is multichannel segmentation that combines mean, standard deviation, and latent entropy as joint inputs, which may further improve accuracy and robustness.

These results provide guidance for designing noise-aware defect-detection workflows and illustrate how dynamical magnetization signatures can support future approaches for identifying material inhomogeneities in magnetic imaging data.

ACKNOWLEDGMENTS

We thank Illia Horenko, Omer Fetai, Philipp Gessler, and Heiko Wende for discussions. We acknowledge funding from the German Research Foundation (DFG) Project No. 405553726 (CRC/TRR 270, projects B12, A03, and B01) and Project No. 505561633 in the TOROID project cofunded by the French National Research Agency ANR under Contract No. ANR-22-CE92-0032. A.M. acknowledges support from the Joachim Herz Foundation through the Add-on Fellowship for Interdisciplinary Science and Transfer. AI-assisted language editing was used.

R.K. performed the computational work, developed the methodology, implemented the code, and generated the figures. R.K. and K.E.-S. wrote the manuscript. N.B. contributed to the initial

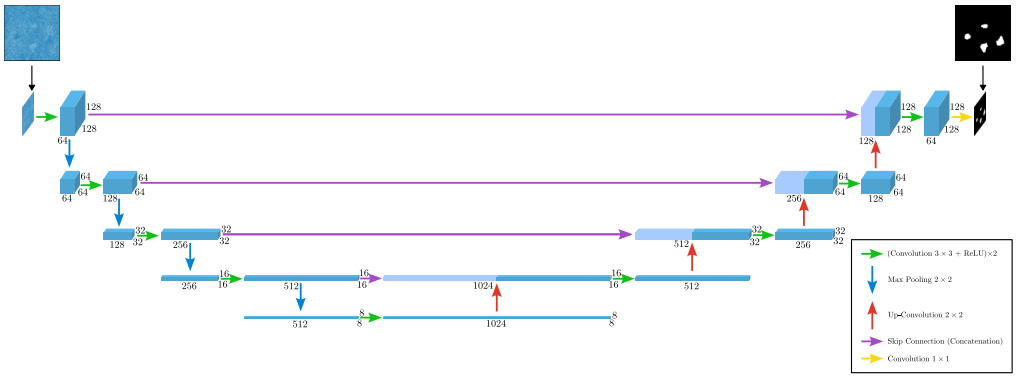


FIG. 4. U-Net architecture used for defect segmentation. The network consists of a four-level encoder-decoder structure with feature sizes 64–128–256–512 and a 1024-channel bottleneck. Each block contains two 3×3 convolutional layers with ReLU activation. Downsampling is performed using 2×2 max pooling, and upsampling using 2×2 transposed convolutions. Skip connections concatenate encoder feature maps with the corresponding decoder levels. A final 1×1 convolution maps the features to two output channels (defect and nondefect logits). Numbers indicate feature-channel counts at each resolution level.

development of the computational framework and methodology. A.M. contributed to the conceptual development of the study and provided scientific guidance. K.N.K., K.O., and O.G. gave scientific input and provided feedback. K.E.-S. contributed to the methodology, led and supervised the project, provided scientific guidance, and acquired funding. All authors reviewed and approved the final manuscript.

All authors declare no financial or nonfinancial competing interests.

DATA AVAILABILITY

All code required to generate the synthetic data, perform the analysis, train and evaluate the models, and reproduce the results presented in this paper is provided in the supplementary software archive [26].

APPENDIX: U-NET ARCHITECTURE AND MODEL TRAINING

For pixel-level segmentation of defect regions, we employed a U-Net convolutional neural network architecture, originally developed for biomedical image segmentation [23]. The network comprises an encoder-decoder structure with skip connections between corresponding layers. Our implementation uses four resolution levels with feature sizes 64–128–256–512 and a bottleneck of 1024 channels. Each block consists of two 3×3 convolutions with ReLU activations; downsampling is performed by 2×2 max pooling and upsampling by transposed convolutions. The architecture is illustrated in Fig. 4.

Each model received a single input channel (one measure map per magnetization component) and output two channels representing defect and nondefect logits. We trained separate models for each combination of measure $\{\mu, \sigma, \mathrm{LE}\}$ and component $\{m_x, m_z\}$.

Training used the Adam optimizer with a learning rate of 10^{-4}, batch size 32, and 500 epochs. Dropout regularization (spatial/channelwise) with probability $p = 0.1$ was inserted inside each double-convolution block after the first Conv + ReLU and before the second Conv + ReLU; this was applied in all encoder blocks, the bottleneck, and all decoder blocks. The training conditions were chosen after having tested different hyperparameters and regularization methods. The dataset was split into training/validation/test subsets in a 0.70/0.15/0.15 ratio, ensuring reproducible splits across experiments.

Inputs were normalized separately for each measure/component combination using the mean and standard deviation computed on the noise-free training subset; these statistics were then applied unchanged to all noise conditions. The loss function was the pixelwise cross-entropy between predicted logits and the binary ground-truth labels. Model performance was assessed using the validation Dice coefficient, and the checkpoint with the best validation Dice score was used for all test-set evaluations.

Model performance is quantified by the Dice coefficient, which ranges from 0 (no overlap between predicted and ground-truth masks) to 1 (perfect overlap). This configuration enables the network to learn spatial signatures of magnetic inhomogeneities from both clean and noise-degraded inputs.

[1] M. Katter, J. Weber, W. Assmus, P. Schrey, and W. Rodewald, A new model for the coercivity mechanism of $Sm_2(Co,Fe,Cu,Zr)_{17}$ magnets, IEEE Trans. Magn. **32**, 4815 (1996).

[2] G.-P. Zheng and M. Li, Influence of impurities on dynamic hysteresis of magnetization reversal, Phys. Rev. B **66**, 054406 (2002).

[3] P. Kühn, Y. Yang, G. Chen, S. N. Foster, H. Egger, and B.-X. Xu, Microstructural influence on hysteresis and eddy-current losses of additively manufactured electrical steel: A multiphysical approach, Comput. Mater. Sci. **263**, 114408 (2026).

[4] H. Chen, F. Yun, J. Qu, Y. Li, Z. Cheng, R. Fang, Z. Ye, S. P. Ringer, and R. Zheng, Coercivity degradation caused by inhomogeneous grain boundaries in sintered Nd-Fe-B permanent magnets, Phys. Rev. Mater. **2**, 054404 (2018).

[5] Y. Yang, T. D. Oyedeji, X. Zhou, K. Albe, and B.-X. Xu, Tailoring magnetic hysteresis of additive manufactured Fe-Ni permalloy via multiphysics-multiscale simulations of process-property relationships, npj Comput. Mater. **9**, 103 (2023).

[6] Y. Yang, P. Kühn, M. Fathidoost, and B.-X. Xu, Micromagnetics and multiscale hysteresis simulations of permanent magnets, COMPEL **42**, 993 (2023).

[7] F. Scheibel, T. Gottschall, A. Taubel, M. Fries, K. P. Skokov, A. Terwey, W. Keune, K. Ollefs, H. Wende, M. Farle, M. Acet, O. Gutfleisch, and M. E. Gruner, Hysteresis design of magnetocaloric materials—From basic mechanisms to applications, Energy Technol. **6**, 1397 (2018).

[8] O. Gutfleisch, T. Gottschall, M. Fries, D. Benke, I. Radulov, K. P. Skokov, H. Wende, M. Gruner, M. Acet, P. Entel, and M. Farle, Mastering hysteresis in magnetocaloric materials, Philos. Trans. R. Soc. A **374**, 20150308 (2016).

[9] H. Sepehri-Amin, L. Liu, T. Ohkubo, M. Yano, T. Shoji, A. Kato, T. Schrefl, and K. Hono, Microstructure and temperature dependent of coercivity of hot-deformed Nd–Fe–B magnets diffusion processed with Pr–Cu alloy, Acta Mater. **99**, 297 (2015).

[10] J. Lyubina, R. Schäfer, N. Martin, L. Schultz, and O. Gutfleisch, Novel design of $La(Fe, Si)_{13}$ alloys towards high magnetic refrigeration performance, Adv. Mater. **22**, 3735 (2010).

[11] C. S. Arnold, M. Dunlavy, and D. Venus, Magnetic susceptibility measurements of ultrathin films using the surface magneto-optic Kerr effect: Optimization of the signal-to-noise ratio, Rev. Sci. Instrum. **68**, 4212 (1997).

[12] M. Kopp, S. Harmeling, G. Schütz, B. Schölkopf, and M. Fähnle, Towards denoising XMCD movies of fast magnetization dynamics using extended Kalman filter, Ultramicroscopy **148**, 115 (2015).

[13] I. Horenko, D. Rodrigues, T. O'Kane, and K. Everschor-Sitte, Scalable computational measures for entropic detection of latent relations and their applications to magnetic imaging, Commun. Appl. Math. Comput. Sci. **16**, 267 (2021).

[14] D. R. Rodrigues, K. Everschor-Sitte, S. Gerber, and I. Horenko, A deeper look into natural sciences with physics-based and data-driven measures, iScience **24**, 102171 (2021).

[15] X. Cao, B. Chen, and W. He, Unsupervised defect segmentation of magnetic tile based on attention enhanced flexible U-Net, IEEE Trans. Instrum. Meas. **71**, 1 (2022).

[16] I. Labrie-Boulay, T. B. Winkler, D. Franzen, A. Romanova, H. Fangohr, and M. Kläui, Machine-learning-based detection of spin structures, Phys. Rev. Appl. **21**, 014014 (2024).

[17] A. R. C. McCray, A. Bender, A. Petford-Long, and C. Phatak, Simulation-trained machine learning models for Lorentz transmission electron microscopy, APL Mach. Learn. **2**, 026120 (2024).

[18] W. F. Brown, *Micromagnetics*, Interscience Tracts on Physics and Astronomy (Interscience Publishers, New York, 1963), Vol. 18.

[19] A. Lyberatos, D. V. Berkov, and R. W. Chantrell, A method for the numerical simulation of the thermal magnetization fluctuations in micromagnetics, J. Phys.: Condens. Matter **5**, 8911 (1993).

[20] J. Leliaert, J. Mulkers, J. De Clercq, A. Coene, M. Dvornik, and B. Van Waeyenberge, Adaptively time stepping the stochastic Landau-Lifshitz-Gilbert equation at nonzero temperature: Implementation and validation in MuMax^3, AIP Adv. **7**, 125010 (2017).

[21] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, The design and verification of MuMax3, AIP Adv. **4**, 107133 (2014).

[22] J. M. D. Coey, Micromagnetism, domains and hysteresis, in *Magnetism and Magnetic Materials* (Cambridge University Press, Cambridge, UK, 2010), Chap. 7, pp. 231–263.

[23] O. Ronneberger, P. Fischer, and T. Brox, U-Net convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18* (Springer, Cham, 2015), pp. 234–241.

[24] H. Zunair and A. Ben Hamza, Sharp U-Net: Depthwise convolutional network for biomedical image segmentation, Comput. Biol. Med. **136**, 104699 (2021).

[25] A. Kamath, J. Willmann, N. Andratschke, and M. Reyes, The impact of U-Net architecture choices and skip connections on the robustness of segmentation across texture variations, Comput. Biol. Med. **197**, 111056 (2025).

[26] R. Knapman, A. Majumdar, N. Bazazzadeh, K. N. Kalkan, K. J. Ollefs, O. Gutfleisch, and K. Everschor-Sitte, Supplementary code for "Defect detection in magnetic systems using U-Net and statistical measures", Zenodo, 2026, doi:10.5281/zenodo.21809287.