

Physics-constrained two-stage reconstruction of contact pressure and real contact support in rough elastic contact

Ilkwang Jang^{1†}, Kisung Kim^{1†}, Hyeonggeun Jo¹, Yong Hoon Jang¹✉

¹ School of Mechanical Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

[†]Ilkwang Jang and Kisung Kim contributed equally to this work and share first authorship.

✉ Corresponding author. E-mail: JYH@YONSEI.AC.KR

Received: June 5, 2026; Revised: September 24, 2026; Accepted: October 23, 2026

© The Author(s) 2026.

Abstract: In rough tribological interfaces, the real contact area and local contact-pressure distribution are essential descriptors of asperity-scale load transfer and provide input fields for analyses of friction, wear, contact fatigue, and thermal or electrical contact resistance. Repeated high-resolution reconstruction of these fields from surface topography remains computationally demanding. This study proposes a physics-constrained two-stage deep-learning framework that separates latent pressure regression from contact-support prediction and combines the two outputs through hard gating. Hard gating enforces non-negative pressure and exactly zero pressure outside predicted support, while a Focal root-mean-square error (RMSE) loss emphasizes sparse high-pressure pixels. Separate load-specific models were trained at 30, 60, and 90 N. A controlled critic-

on/off ablation at 256×256 and 30 N showed comparable reconstruction accuracy with and without adversarial training, indicating that the dominant performance gain arises from the two-stage formulation and hard gating rather than from the adversarial component itself. On synthetic RMD-generated rough-surface datasets, the model achieved a mean Dice score of 0.9909 and a mean pixel-wise accuracy of 0.9587; at 30 N, R_U^2 was 0.9995 and the $p_{99.9}$ relative error was 0.0049. Batch-1 end-to-end inference required 9.0 ms at 128×128 resolution on an NVIDIA RTX 3090 GPU. The predictions preserved prescribed load, contact-area trends, and dependence on RMS slope. An auxiliary test on axisymmetric oblate spheroidal profiles assessed smooth circular single-patch reconstruction, with the numerical fields checked against independent full-profile contact theory. On measured sandpaper-finished steel topographies without fine-tuning, mean Dice and coefficient of determination were 0.940 and 0.970 against FFT-BEM numerical references. Within the prescribed-load, frictionless linear-elastic regime, the framework provides a physics-constrained surrogate for rapid reconstruction of FFT-BEM-consistent contact-support and pressure fields.

Keywords: Rough-surface contact mechanics; Contact pressure distribution; Real contact area; Physics-constrained deep learning; Two-stage reconstruction; FFT-BEM surrogate

1 Introduction

Elastic contact between rough surfaces is a fundamental physical phenomenon governing interfacial behavior in a wide range of engineering systems, including tribology, materials science, geotechnics, and microelectronic packaging. In real interfaces, contact does not occur over the entire nominal area, but is formed locally only at a subset of microscopic asperity junctions [1–4]. The resulting contact area and contact pressure distribution provide important input descriptors for analyses of friction, wear, contact fatigue, and electrical and thermal contact resistance [5–8]. Rapid and accurate reconstruction of the true contact region and contact-pressure field from a given surface topography

and loading condition is therefore of central importance for performance prediction, reliability assessment, and lifetime-oriented design of mechanical systems.

In tribological contacts, local normal pressure is not only a mechanical output but also a coupling variable linking surface topography to frictional shear strength, wear initiation, contact fatigue, flash temperature, and interfacial transport. Therefore, a surrogate that reconstructs both where real contact occurs and the local pressure magnitude can serve as a building block for downstream tribological models, even when tangential sliding, lubrication, or wear evolution are not explicitly simulated.

Direct experimental measurement of subsurface pressure and stress fields at an interface remains challenging [9–11], and numerical reproduction of these fields is also computationally demanding. The boundary element method (BEM) based on the linear elastic half-space assumption, together with FFT-based acceleration (FFT-BEM), can provide high accuracy, but computational and memory costs increase rapidly as surface resolution increases or as multiscale roughness must be resolved simultaneously [12–16]. Although a single FFT-BEM analysis is not by itself the target of replacement, repeated reference generation and evaluation become costly when many surface topographies must be solved at the same prescribed condition. The computational motivation for the surrogate is therefore an amortized many-query workflow, in which the offline data-generation and training costs are distributed over repeated evaluations, rather than an isolated one-off contact analysis.

The present study targets a class of repeated fixed-condition contact analyses rather than a universal load-parameterized contact solver. Such fixed-condition analyses arise, for example, in statistical evaluation of many surface realizations under a prescribed preload, manufacturing-variability assessment, surface-design screening, and uncertainty quantification for mechanically or electrically loaded interfaces. In these applications, the material properties and nominal load often remain fixed while a large number of measured or generated surface topographies must be evaluated. A load-

specific surrogate can therefore retain practical value even when the applied load is not treated as a continuous network input.

Against this background, deep-learning-based approaches, including convolutional neural networks (CNNs), have recently attracted attention as an alternative means of predicting contact quantities directly from surface topography. Prior studies have reported encouraging results for aggregate quantities such as total real contact area or mean pressure [17–19], and some recent work has also addressed contact localization [20]. For many engineering applications, however, the more critical requirement is the high-resolution reconstruction of both where contact occurs and how pressure is distributed within those local contact patches. In this respect, rough-surface contact differs fundamentally from standard image-regression problems: the contact region is highly sparse relative to the full surface, the physical state changes discontinuously across the contact boundary, and the most consequential high-pressure peaks are concentrated in a small number of localized regions. As a consequence, single-regression models tend to oversmooth the spatial structure or to underpredict the key contact zones and local pressure peaks because of data imbalance [20, 21]. Recent discrete rough-contact analyses further show that asperity interactions can affect not only macroscopic contact quantities such as real contact area, but also the microscale pressure distribution, motivating field-resolved pressure prediction rather than scalar-only contact modeling [22].

This challenge is not unique to contact mechanics: physics-informed and physics-constrained surrogate modeling has gained attention more broadly as a means of integrating explicit physical constraints into data-driven prediction frameworks for mechanical systems [23–25]. These examples suggest that the value of a data-driven contact surrogate lies not merely in prediction accuracy, but in how the underlying contact-mechanics constraints are embedded in the model structure. The present study is positioned within this stream: it proposes a framework that encodes a key constraint of frictionless rough-surface contact, the zero-pressure condition in non-contact regions, directly in the network output architecture, rather than relying on soft penalties or post-processing corrections.

The central gap, therefore, is that existing data-driven models have not yet satisfactorily combined two requirements that are both essential for rough-surface contact prediction: structural enforcement of zero pressure in non-contact regions and reliable recovery of sparse local high-pressure features.

A more fundamental limitation is that the governing constraints of contact mechanics are often not adequately embedded in the prediction structure itself. In frictionless normal contact, the pressure must vanish exactly in non-contact regions [26, 27], yet many existing data-driven approaches do not guarantee this condition structurally at the final output stage. Penalty terms added to the loss function may reduce the degree of violation, but they do not ensure satisfaction of the constraint itself [28, 29]. Surrogate models for rough-surface contact should therefore be designed not merely to reduce average prediction error, but to encode the essential physical constraints of the contact problem directly in the output structure.

To address these limitations, this study proposes a physics-constrained two-stage reconstruction framework that separates latent pressure regression from contact-support prediction and combines the two outputs through hard gating. The framework predicts a latent pressure field and a contact-support probability map in independent branches, then applies a non-negative projection and binary support mask to eliminate pressure outside predicted contact. This output structure targets nonphysical pressure leakage without claiming to solve the full complementarity problem. A Focal RMSE loss gives greater weight to sparse high-pressure pixels than to the extensive low-pressure background. The resulting framework is intended to improve both contact-support localization and pressure-amplitude reconstruction within contact patches.

The main contributions of this study are as follows. First, a physically meaningful task decomposition separates rough-contact reconstruction into latent pressure-field regression and explicit contact-support segmentation. Second, element-wise hard gating imposes non-negative pressure and exactly zero pressure outside the predicted support at the final output stage. Third, a Focal RMSE loss emphasizes the sparse high-pressure pixels within contact patches. Fourth, the

framework is assessed through support, pressure, load, spectral, extreme-pressure, grid-convergence, and complementarity diagnostics within a prescribed-load, frictionless linear-elastic regime. WGAN-GP is retained as auxiliary regularization in the archived implementation and is tested by a controlled critic-on/off ablation; it is not treated as the defining source of the physical constraint.

The remainder of this paper is organized as follows. Section 2 describes the physics-based dataset generation procedure, the hard-constrained two-stage reconstruction framework, the loss functions, and the evaluation metrics. Section 3 presents the quantitative performance, physical consistency, ablation results, and generalization behavior. Section 4 discusses the significance of the proposed approach, its range of applicability, its limitations, and possible directions for future extension. Section 5 concludes the paper.

2 Physics-based rough-contact data and structurally constrained pressure reconstruction

As discussed in the Introduction, two failure modes of existing data-driven contact surrogates motivate the present design: (i) the lack of structural enforcement of the zero-pressure condition outside the real-contact region, and (ii) the underprediction of sparse high-pressure peaks within contact patches. To mitigate these limitations, this study combines a physics-based dataset generated by high-fidelity numerical analysis with a hard-constrained two-stage reconstruction framework. This section first describes the generation and preprocessing of physically valid training data, then presents the two independently parameterized branches and the hard-gated output structure. Finally, evaluation metrics are defined at three levels: contact-support reconstruction, pressure regression within the contact support, and consistency with aggregate physical quantities and theory.

2.1 Physics-based data generation and preprocessing

Fig. 1 summarizes the provenance, generation, and use of the three datasets considered in this study.

The first branch comprises computationally generated RMD rough topographies and the

corresponding FFT-BEM contact-support and pressure fields at prescribed loads of 30, 60, and 90 N.

These data were used for load-specific training, validation, and testing.

The second branch comprises an auxiliary ellipsoidal-contact dataset generated from axisymmetric oblate spheroidal profiles at a prescribed load of 30 N. FFT-BEM provides the contact-support and pressure labels. An independent full-profile axisymmetric elastic half-space solution is used to assess these labels and the network predictions; the theoretical parameters are determined from the input geometry, prescribed load and elastic properties.

The third branch comprises experimentally measured sandpaper-finished steel topographies. These measured surfaces were leveled and resampled to the 128×128 model resolution. Contact-support and pressure reference fields were then computed using FFT-BEM under the same material properties, boundary conditions, and 30 N load used for the corresponding synthetic cases. The measured topographies were supplied directly to the pretrained 30 N model without additional training or fine-tuning.

A. Synthetic rough-contact branch



B. Auxiliary ellipsoidal-contact branch



*Independent theory: input geometry, prescribed load and elasticity.

C. Measured-topography transfer branch



Counts: 30 N archived train / validation / test inventory; see Section 2.1.

■ Measured ■ Generated / archived ■ FFT-BEM reference ■ Model use / evaluation

Fig. 1 Data provenance and use of the three datasets. (A) RMD-generated rough topographies and FFT-BEM solutions at 30, 60 and 90 N, with load-specific model evaluation. Counts and splits shown describe the 30 N dataset. (B) Axisymmetric oblate spheroidal profiles and FFT-BEM labels at 30 N; the independent full-profile theory assesses the BEM labels and network predictions. (C) Measured sandpaper-finished steel

topographies, preprocessing, FFT-BEM reference calculation and application of the pretrained 30 N model without fine-tuning. Only branch C topographies were measured experimentally; all contact-support and pressure reference fields were computed by FFT-BEM.

Only the surface geometry in the third branch was measured experimentally. The reference contact-support and contact-pressure fields in all three branches were obtained numerically using FFT-BEM. The color coding in Fig. 1 distinguishes experimentally measured quantities, computationally generated or stored quantities, numerically computed FFT-BEM references, and neural-network training or prediction. This distinction is particularly important for the measured-topography branch, in which the surface geometry is experimental but the reference contact-support and pressure fields are numerical.

To enable the deep-learning model to learn the input–output structure of contact mechanics without becoming overly dependent on a specific physical scale or material property, a physically valid high-quality dataset and a systematic preprocessing strategy are required. In this study, three-dimensional isotropic synthetic rough surfaces were generated using a diamond-square implementation of the Random Midpoint Displacement (RMD) algorithm, motivated by the self-similar multiscale characteristics of real machined surfaces [3, 4, 30–32]. At each refinement level, the displacement increments were sampled from Gaussian random variables. No height clipping, rank-order transformation, nonlinear height mapping, or prescribed adjustment of skewness and kurtosis was applied. The generated surfaces should therefore be regarded as approximately Gaussian finite realizations rather than as surfaces constrained to have an exactly Gaussian height distribution. The fractal dimension D was sampled to reflect experimentally observed ranges for polished and lapped surfaces, yielding a total of 50,000 rough surfaces in the 30 N dataset. Load-specific rough-contact evaluations were also performed at 60 and 90 N. Specifically, polished surfaces were sampled

uniformly over $D \in [2.17, 2.28]$, and lapped surfaces over $D \in [2.21, 2.29]$. The resulting arithmetic average roughness R_a ranged from approximately 0.036 to 0.744 μm .

The parameter D used in the generation algorithm is referred to hereafter as the nominal RMD fractal-dimension parameter. It is an input to the surface-generation procedure and is not interpreted as an independently measured fractal dimension. Owing to the finite grid, wavelength cutoffs, and the limited wavenumber interval available for spectral fitting, the effective exponent estimated from the radial power spectral density need not coincide exactly with the nominal input parameter.

The statistical nature of the generated rough surfaces was assessed using the RMS roughness, skewness, and normalized fourth moment:

$$S_q = \left[\frac{1}{N} \sum_{i=1}^N (h_i - \bar{h})^2 \right]^{\frac{1}{2}},$$

$$S_{sk} = \frac{N^{-1} \sum_{i=1}^N (h_i - \bar{h})^3}{S_q^3}, S_{ku} = \frac{N^{-1} \sum_{i=1}^N (h_i - \bar{h})^4}{S_q^4} \quad (1)$$

For an ideal Gaussian distribution, $S_{sk} = 0$ and $S_{ku} = 3$. Radially averaged power spectral densities and representative normal Q-Q plots were additionally examined.

Across the 5,000 test surfaces, the mean height skewness was $S_{sk} = -0.004 \pm 0.187$, and the normalized fourth moment was $S_{ku} = 2.920 \pm 0.260$ (Fig. 2). The mean values are therefore close to the Gaussian reference values of zero and three, respectively. Representative Q-Q plots showed broad agreement with a Gaussian distribution, with finite-realization deviations mainly in the tails. No additional spectral filter or detrending operation was applied after surface generation; consequently, the effective short-wavelength cutoff is the grid Nyquist limit. With $\Delta x = L/N$, the corresponding quantities are $q_{N_y} = \pi/\Delta x$ and $\lambda_{N_y} = 2\Delta x$. For the 128×128 grid, $\Delta x = 7.8125 \mu\text{m}$, $q_{N_y} = 0.402 \text{ rad}/\mu\text{m}$, and $\lambda_{N_y} = 15.625 \mu\text{m}$; for the 256×256 grid, they are $3.90625 \mu\text{m}$, $0.804 \text{ rad}/\mu\text{m}$, and $7.8125 \mu\text{m}$, respectively. The ensemble-averaged spectrum over 5,000 surfaces retains

its approximate power-law behavior up to the effective Nyquist cutoff, without an additional resolved-band roll-off.

These observations support describing the synthetic rough-surface family as approximately Gaussian and multiscale. No deliberately skewed, plateau-like, grooved, or otherwise non-Gaussian rough-surface family was included in training.

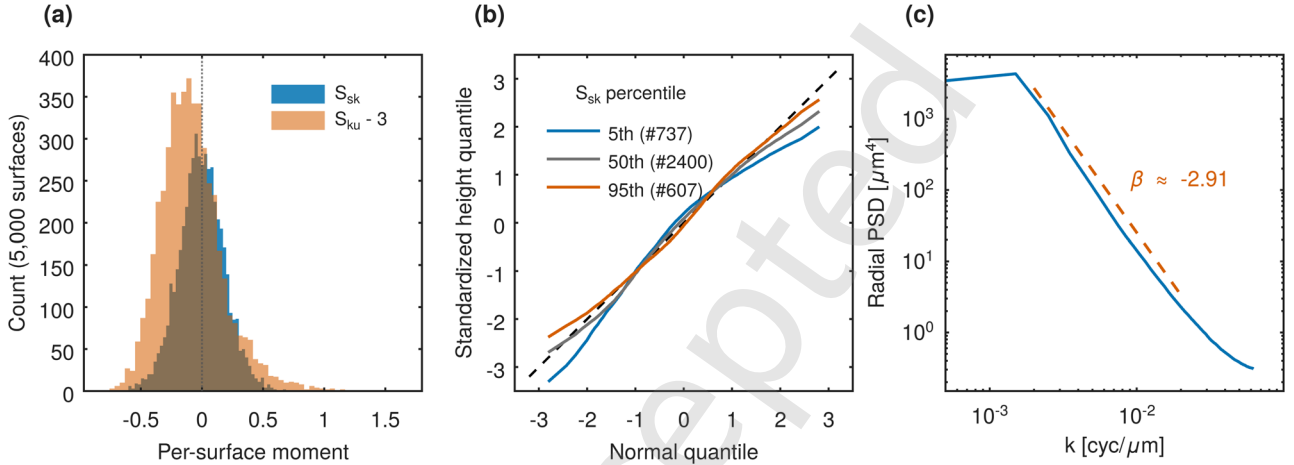


Fig. 2 Statistical characterization of the RMD-generated rough surfaces: (a) distributions of per-surface height skewness and normalized fourth moment for 5,000 test surfaces, (b) representative normal Q-Q plots selected from different skewness percentiles, and (c) a representative radially averaged power spectral density. The dashed line in panel (c) indicates an approximate power-law fit over the resolved intermediate-wavenumber band.

To obtain the ground-truth contact-pressure field and contact region corresponding to each generated surface, frictionless normal contact analysis was performed under the elastic half-space assumption. The contact problem was modeled as normal contact between a rough surface and an ideal flat, and the material was assumed to be an isotropic linear elastic solid (structural steel, $E = 200$ GPa, $\nu = 0.3$). The contact solution is defined together with the complementarity conditions [27, 33]. Let $g_0(\mathbf{x})$ denote the initial gap and Δ the rigid-body approach. Then the gap field is given by

$$g(\mathbf{x}) = g_0(\mathbf{x}) + u(\mathbf{x}) - \Delta \quad (2)$$

and the normal-contact constraints are written as

$$p(\mathbf{x}) \geq 0, \quad (3)$$

$$g(\mathbf{x}) \geq 0, \quad (4)$$

$$p(\mathbf{x}) g(\mathbf{x}) = 0. \quad (5)$$

That is, in the contact region, the pressure is positive and the gap is zero, whereas in the non-contact region, the gap is positive and the pressure is zero. The two-stage hard gating introduced in this study does not solve the full complementarity problem explicitly at the solver level; rather, it structurally enforces the key condition of zero pressure in non-contact regions in the final predicted field.

For large-scale dataset generation, an influence-coefficient-based convolutional formulation was used on the discretized grid and accelerated by the Fast Fourier Transform (FFT) [12–14]. The use of an influence-coefficient based half-space formulation is consistent with BEM interpretations of rough-surface contact mechanics [15]. Continuous in-plane position is denoted by $\mathbf{x} = (x, y)$, and continuous height, pressure, gap, and normal-displacement fields are written as $h(\mathbf{x})$, $p(\mathbf{x})$, $g(\mathbf{x})$, and $u(\mathbf{x})$, respectively. Bold symbols denote their sampled arrays, whereas i and j denote grid-cell indices and $\Delta A = \Delta x \Delta y$ is the cell area. The discrete displacement at cell i is

$$u_i = \sum_j K_{i-j} p_j \Delta A, \quad (6)$$

where K_{i-j} is the cell-averaged Green-function influence coefficient from pressure cell j to displacement cell i . The index difference is wrapped periodically in the FFT convolution. The self-influence coefficient is obtained from the same finite-cell integration used by the reference solver, normalized consistently with the explicit ΔA factor, and the sign convention is positive pressure and positive normal displacement into the elastic half-space. Continuous wavenumber $\mathbf{q}$ is distinguished from the integer FFT mode index; the latter only labels sampled Fourier coefficients. Applying the FFT converts Eq. (6) into a multiplication in the frequency domain:

$$\hat{u}_{hl} = \hat{K}_{hl} \hat{p}_{hl} \quad (7)$$

This procedure enabled efficient computation of the discrete FFT-BEM reference fields used for model development and evaluation.

The auxiliary 30 N profiles are axisymmetric oblate spheroidal domes with height $h(r) = h_0\sqrt{1 - r^2/c^2}$ for $0 \leq r < c$ and zero stored height outside the dome. The two horizontal semi-axes equal c and the vertical semi-axis is h_0 , giving circular contact. The apex-referenced gap is $f(r) = h_0 - h(r)$, and the local apex curvature radius is $R_{apex} = c^2/h_0$. These analytical definitions do not change the stored input-height encoding. The archive contains 4,000 training, 500 validation and 500 test profiles. For the 500 test profiles, theoretical contact radii span 58.18–311.59 μm and central pressures span 0.1314–3.6854 GPa.

This auxiliary dataset was not introduced as an asperity-shape model for the rough surfaces. The rough-contact branch does not decompose a topography into spherical, conical, or pyramidal asperities; the entire discretized height map is supplied directly to the FFT-BEM solver. Instead, the ellipsoidal samples provide a controlled smooth circular single-patch morphology with an independent full-profile theoretical reference, in contrast to the irregular multi-patch fields of rough contact.

The auxiliary training and validation profiles account for 4,500 of 49,500 development files (9.09%). Table 1 lists the 30 N dataset splits. The training split contains 40,000 rough and 4,000 auxiliary profiles, and the validation split contains 5,000 rough and 500 auxiliary profiles. The separate test sets contain 5,000 rough and 500 auxiliary cases.

Table 1 The 30 N, 128×128 dataset inventory by split.

Data type	Total	Train	Validation	Test
Rough contact (30 N)	50,000	40,000	5,000	5,000
Ellipsoidal contact (30 N)	5,000	4,000	500	500

Table 2 Proposed model/dataset accounting.

Load	Grid	Data/evaluation role
30 N	128 ²	Training: 40k rough + 4k ellipsoidal; validation: 5k rough + 500 ellipsoidal; tests: 5k rough + 500 ellipsoidal
60 N	128 ²	Load-specific rough-contact evaluation; 5k held-out surfaces
90 N	128 ²	Load-specific rough-contact evaluation; 5k held-out surfaces
30 N	256 ²	40k/5k/5k rough train/validation/test; scalability and critic ablation

The dataset composition is given in Tables 1–2. In the 30 N data, the fixed height and pressure encoding scales are $g_{max} = 5.72 \mu m$ and $p_{max} = 13.98 \text{ GPa}$. The input height map is normalized using its fixed global scale:

$$\tilde{g}_0(\mathbf{x}) = \frac{g_0(\mathbf{x})}{g_{max}} \quad (8)$$

The output contact-pressure field was normalized in the same manner:

$$\tilde{p}(\mathbf{x}) = \frac{p(\mathbf{x})}{p_{max}} \quad (9)$$

After inference, the original unit system can be recovered by

$$p(\mathbf{x}) = p_{max} \tilde{p}(\mathbf{x}). \quad (10)$$

Thus, this global scaling may be interpreted as a preprocessing (preconditioning) step that improves training stability while preserving the physical interpretability of the quantities.

During training, data augmentation was performed under the assumption of isotropic surfaces by applying horizontal and vertical flips and rotations in multiples of 90° identically to the inputs and ground truth. This augmentation increases the effective number of training samples without distorting the statistical directional characteristics of the surfaces. In addition, the input height maps and output pressure fields were stored in 16-bit integer representation in order to preserve fine gradient information near contact boundaries and local peaks as much as possible.

2.2 Hard-constrained two-stage reconstruction framework

This subsection presents a hard-constrained two-stage reconstruction framework designed to address simultaneously the sparsity of the contact region, the skewed distribution of pressure values, and the zero-pressure condition in non-contact regions. The central idea is to decompose contact-pressure reconstruction into latent pressure-field regression and contact-support prediction, rather than treating it as a single continuous-value regression problem, and then to combine the two outputs through hard gating. Hereafter, the continuous initial-gap field is denoted by $g_0(\mathbf{x})$, and its sampled network input by the array $\mathbf{X} = \mathbf{g}_0$. The symbol r denotes radial distance from the input apex in the auxiliary axisymmetric benchmark, $r = \|\mathbf{x} - \mathbf{x}_0\|$.

In the proposed method, the Stage 1 generator G_1 predicts a latent pressure field P_{pot} from the input height map X :

$$P_{\text{pot}} = G_1(X). \quad (11)$$

Here, P_{pot} is interpreted as a latent field prior to the final physical pressure field, and serves to learn the spatial pattern and relative magnitude distribution of the pressure field. In parallel, the Stage 2 generator G_2 predicts a contact probability map M_{prob} from the same input:

$$M_{\text{prob}} = G_2(X). \quad (12)$$

At inference, a predefined threshold τ is applied to construct a binary contact mask:

$$M_{\text{contact}}(\mathbf{r}) = \mathbb{I}(M_{\text{prob}}(\mathbf{r}) \geq \tau), \quad (13)$$

where $\mathbb{I}(\cdot)$ is the indicator function. Because pressure cannot be physically negative, a non-negativity projection is first applied to the latent pressure field:

$$P_{\text{pot}}^+(\mathbf{r}) = \max(P_{\text{pot}}(\mathbf{r}), 0). \quad (14)$$

The final pressure field is then defined as the element-wise product of the non-negative latent pressure field and the contact mask:

$$P_{\text{final}}(\mathbf{r}) = P_{\text{pot}}^+(\mathbf{r}) \odot M_{\text{contact}}(\mathbf{r}), \quad (15)$$

where $\odot$ denotes the Hadamard product. By Eq. (15), $P_{\text{final}}(\mathbf{r}) = 0$ whenever $M_{\text{contact}}(\mathbf{r}) = 0$. The model therefore enforces zero pressure with respect to the predicted contact mask at the output level, rather than relying on a penalty term to discourage nonzero values outside contact.

This guarantee is limited to the contact regions identified by the predicted binary mask. The proposed framework does not solve the full complementarity problem by jointly computing the gap field, and full complementarity must therefore be evaluated separately through subsequent physical verification. Even so, eliminating residual pressure outside predicted contact is meaningful because it suppresses a common nonphysical artifact of single-regression models.

The prediction of a contact pressure distribution from a surface height map may be viewed as a conditional image-to-image translation problem [34]. Since this study considers a deterministic mapping under fixed load and material conditions, the exogenous noise variable z used in standard GANs was not introduced. The present network receives the surface height map as its input, whereas the applied load is not provided as a continuous conditioning variable. Accordingly, separate load-specific models were trained for prescribed normal loads of 30, 60, and 90 N. Each checkpoint reconstructs the contact-support set and pressure field associated with its own training load. The present configuration therefore does not provide interpolation to arbitrary intermediate loads or extrapolation to loads outside the training conditions. This design was adopted to isolate the effects of the two-stage decomposition, hard gating, and task-specific loss functions under controlled conditions. A continuously load-conditioned model would require the target load to be supplied explicitly, together with systematic validation on additional load levels.

The support and pressure branches use the same U-Net architecture [35] but are independently parameterized: they share no weights and have no gradient coupling. Each branch is trained against its complete spatial target, including the zero-valued background of the pressure field. The encoder down-sampling path extracts multiscale surface features hierarchically, and the skip connections preserve the high-resolution information required to recover contact boundaries and local pressure

peaks. The output head of Stage 1 uses a Tanh activation to regress the latent pressure field, whereas the output head of Stage 2 uses a Sigmoid activation to generate the contact probability map. The critic adopts a conditional PatchGAN architecture to discriminate the realism of local patches and high-frequency structure in input–output pairs. Fig. 3 summarizes the overall workflow of Stage 1 latent pressure-field regression, Stage 2 contact-mask prediction, and final pressure-field construction through hard gating. The U-Net-based backbone and task-specific output heads of the two-stage generators are listed in Table 3, and the detailed architecture of the conditional PatchGAN critic is given in Table 4.

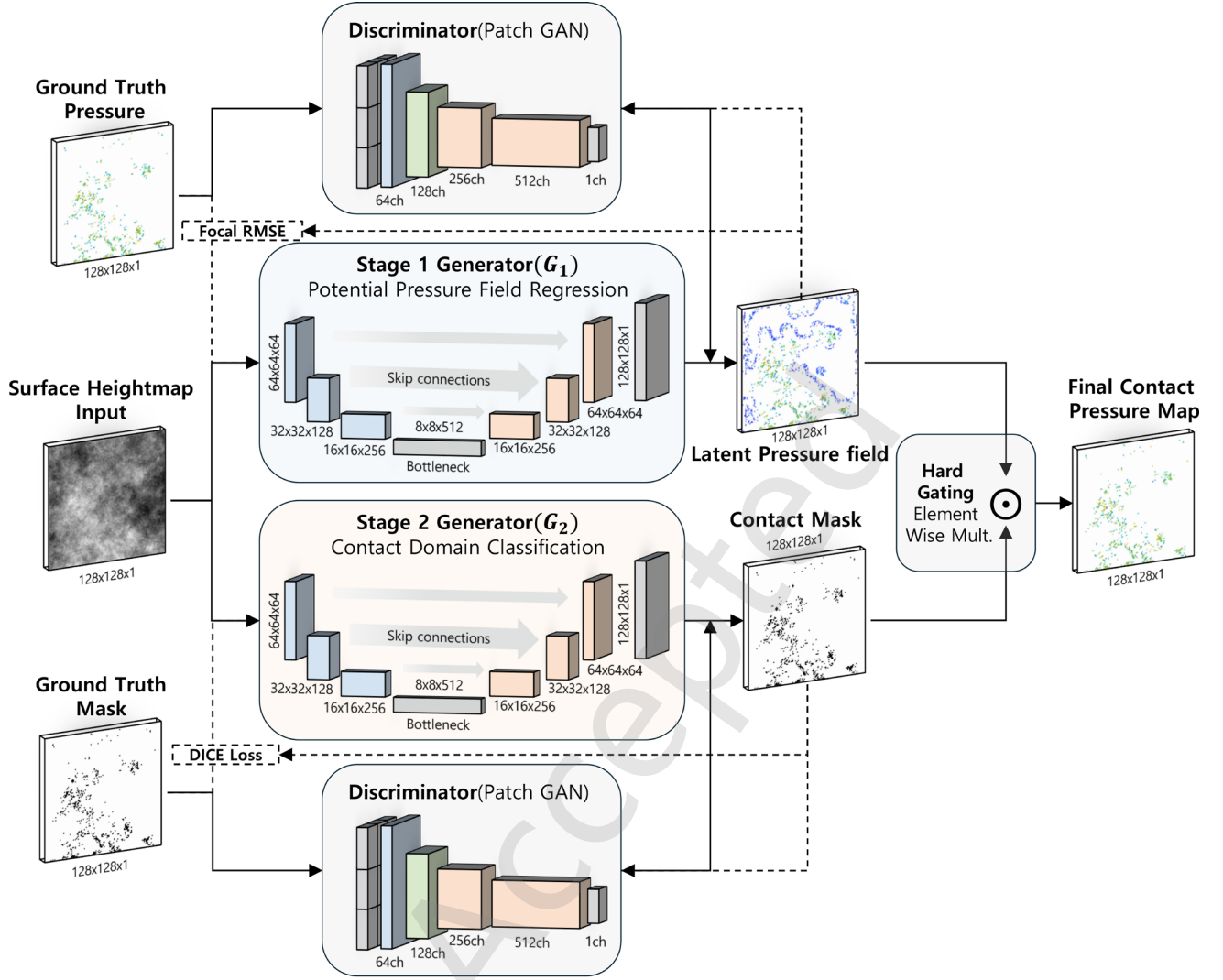


Fig. 3 Schematic of the proposed two-stage reconstruction framework. The pressure branch predicts a latent pressure field, and the support branch predicts a contact probability map that is thresholded to a binary mask. The final pressure field is obtained by element-wise gating of the non-negative latent field with the predicted mask. Each critic is auxiliary and used only during training.

Table 3 Generator architecture used by both branches. The support and pressure branches use the same U-Net architecture, independently parameterized per branch; no encoder, decoder, or output-head weights are shared.

Path	Layer	Kernel/Stride	Output Channels	Output Size
Encoder	Conv + LeakyReLU + IN	$4 \times 4 / 2$	64	64×64
	Conv + LeakyReLU + IN	$4 \times 4 / 2$	128	32×32
	Conv + LeakyReLU + IN	$4 \times 4 / 2$	256	16×16
	Conv + LeakyReLU + IN	$4 \times 4 / 2$	512	8×8

Decoder	TransConv + ReLU + IN	$4 \times 4 / 2$	256	16×16
	TransConv + ReLU + IN	$4 \times 4 / 2$	128	32×32
	TransConv + ReLU + IN	$4 \times 4 / 2$	64	64×64
Output head (G_1)	TransConv + Tanh	$4 \times 4 / 2$	1	128×128
Output head (G_2)	TransConv + Sigmoid	$4 \times 4 / 2$	1	128×128

Table 4 Conditional PatchGAN critic architecture used independently for the support and pressure branches. The input is the channel-wise concatenation of the height map and target or generated output. Each convolution uses a 4×4 kernel and padding 1; intermediate layers use InstanceNorm and LeakyReLU with slope 0.2. The last score map is averaged to obtain the scalar WGAN-GP score.

Layer	Channels	Stride	128^2 output	Operation
Input	2	—	128×128	Height plus target/generated field
Conv 1	$2 \rightarrow 64$	2	64×64	Conv, LeakyReLU
Conv 2	$64 \rightarrow 128$	1	63×63	Conv, InstanceNorm, LeakyReLU
Conv 3	$128 \rightarrow 256$	1	62×62	Conv, InstanceNorm, LeakyReLU
Conv 4	$256 \rightarrow 1$	1	61×61	Conv; spatial score map

The spatial sequence is $128 \rightarrow 64 \rightarrow 63 \rightarrow 62 \rightarrow 61$, so a 128×128 input produces a 61×61 score map. A 256×256 input produces a 125×125 score map. In both cases, the receptive field is 22×22 pixels and the critic contains 661,953 trainable parameters. WGAN-GP uses the score-map mean and $\lambda_{GP} = 10$.

Table 5 summarizes the verified training and inference specification of the primary 128×128 implementation.

Table 5 Primary 128×128 training and inference specification of the historical implementation.

Item	Verified specification
Branch relation	Independent U-Nets; architecture shared, weights not shared; no gradient coupling
Support target/loss	Full binary support field; Dice loss; smoothing 10^{-6}
Pressure target/loss	Full pressure field including zero background; Focal RMSE; $\gamma = 0.8$, $\epsilon = 10^{-6}$
Historical adversarial term	One separate conditional critic per branch; WGAN-GP; score-map mean; $\lambda_{GP} = 10$
Generator loss weights	0.5 adversarial + 0.5 supervised in the primary 128^2 implementation
Optimizer/schedule	Adam, 2×10^{-4} , $\beta = (0.5, 0.999)$; ReduceLROnPlateau factor 0.2, patience 5
Run length/batch	100 epochs per branch; batch size 1
Normalization	16-bit arrays normalized to $[0, 1]$
Integrated inference	Threshold support at 0.5; clamp pressure to non-negative 16-bit range; element-wise hard gating

All convolutional weights were initialized from $\mathcal{N}(0, 0.02^2)$, and the archived primary runs used the final-epoch checkpoints. The two branches were optimized independently according to the following procedure.

1. **Support-branch training.** Train the support U-Net on normalized height arrays and full binary support fields using Dice loss. In the historical implementation, its separate conditional critic contributes the 0.5-weighted adversarial term; the supervised-only ablation removes that term without changing the generator, data, or Dice objective.
2. **Pressure-branch training.** Independently train the pressure U-Net on normalized height arrays and full pressure fields, including zero background, using Focal RMSE with $\gamma = 0.8$ and $\epsilon = 10^{-6}$. Its historical critic is separate from the support critic, and no support-branch gradient enters this optimization.
3. **Integrated inference.** For $\mathbf{g}_0$, threshold support probability at 0.5, clamp latent pressure to the non-negative 16-bit range, and multiply the arrays element-wise.

Each stage is optimized within a common conditional WGAN-GP training framework [36]. In the general conditional setting, the critic D receives a pair consisting of the input height map X and either a real or generated output, and produces a scalar score. The critic loss is given by

$$L_D^{\text{WGAN-GP}} = \mathbb{E}_{X, \tilde{Y} \sim p_g} [D(X, \tilde{Y})] - \mathbb{E}_{X, Y \sim p_{\text{data}}} [D(X, Y)] + \lambda_{\text{gp}} \mathbb{E}_{X, \hat{Y} \sim p_{\hat{Y}}} \left(\|\nabla_{\hat{Y}} D(X, \hat{Y})\|_2 - 1 \right)^2, \quad (16)$$

where $\tilde{Y} = G(X)$ is the generated output, p_g is the generated distribution, p_{data} is the real data distribution, $p_{\hat{Y}}$ is the distribution of linearly interpolated real and generated samples, and λ_{gp} is the gradient-penalty weight. The adversarial loss of the generator is defined as

$$L_{G, \text{adv}} = -\mathbb{E}_{X, \tilde{Y} \sim p_g} [D(X, \tilde{Y})] \quad (17)$$

WGAN-GP is not required by the hard-gating formulation. The reported primary checkpoints retain their archived WGAN-GP training history, in which one conditional critic was used as auxiliary regularization for each branch. To isolate that contribution, an otherwise identical critic-free variant was trained and compared in Section 3.4; this comparison does not imply that every reported primary-result checkpoint was retrained without critics.

The Stage 1 regression problem is highly skewed, because most pixels have zero or low pressure, whereas physically important high-pressure peaks are concentrated in a small number of pixels. In addition, since wear and fatigue failure in tribological processes typically initiate near the maximum contact pressure, a Focal RMSE loss is newly introduced for the Stage 1 regression loss, inspired by the concept of focal-style hard-example reweighting [21]. This loss assigns greater weight to pixels with larger errors than standard MSE or RMSE:

$$L_{\text{Focal-RMSE}} = \sqrt{\frac{\sum_{i=1}^n (p_i - \hat{p}_i)^2 |p_i - \hat{p}_i|^\gamma}{\sum_{i=1}^n |p_i - \hat{p}_i|^\gamma + \epsilon}}, \quad (18)$$

where γ is the focusing parameter, and $\gamma = 0.8$ was used in this study. ϵ is a constant introduced for numerical stability. The weighting term $|p_i - \hat{p}_i|^\gamma$ assigns a larger learning signal to difficult

prediction regions near high-pressure peaks and contact boundaries. The pressure-branch generator loss in the primary 128×128 implementation is

$$L_{G_1} = 0.5L_{G_1,adv} + 0.5L_{\text{Focal-RMSE}} \quad (19)$$

where $L_{G_1,adv}$ is the adversarial loss obtained by applying Eqs. (16)-(17) to the Stage 1 output P_{pot} .

By contrast, the Stage 2 contact-mask prediction problem suffers from severe class imbalance, because rough-contact regions can occupy only a small fraction of the total surface area [37]. To mitigate this issue, Dice loss was used to directly optimize the overlap between the predicted mask and the true mask:

$$L_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^n m_i \hat{m}_i + \epsilon}{\sum_{i=1}^n m_i + \sum_{i=1}^n \hat{m}_i + \epsilon}, \quad (20)$$

where m_i is the true binary mask and $\hat{m}_i$ is the predicted probability. The support-branch generator loss in the primary 128×128 implementation is

$$L_{G_2} = 0.5L_{G_2,adv} + 0.5L_{\text{Dice}} \quad (21)$$

where $L_{G_2,adv}$ is the adversarial loss obtained by applying Eqs. (16)-(17) to the Stage 2 output M_{prob} .

By separating regression and classification in this manner, Stage 1 learns pressure amplitude and local peak structure within the contact region, while Stage 2 learns the geometric boundaries of contact patches more directly. Accordingly, the performance of the framework is attributed to two-stage task decomposition, hard gating, and task-specific supervised losses. Adversarial regularization is treated as an auxiliary training variant whose contribution is evaluated by the controlled critical-on/off comparison in Section 3.4.

2.3 Evaluation metrics and physical verification criteria

A single numerical score is insufficient for evaluating the proposed model. In contact problems, it is necessary to distinguish among: (i) how accurately the contact region is identified, (ii) how accurately the pressure within the contact region is reconstructed, and (iii) whether the prediction is

also consistent with aggregate physical quantities and existing theory. Accordingly, the evaluation framework in this study consists of three categories: reconstruction of contact-region geometry, pressure regression within the contact region, and verification of consistency with aggregate physical quantities and theory.

Since contact-mask prediction is a binary segmentation problem, the Dice score, which measures the overlap between the predicted and true contact regions, was adopted as the primary metric:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN}, \quad (22)$$

where A denotes the true contact region, B the predicted contact region, and TP , FP , and FN the numbers of true-positive, false-positive, and false-negative pixels, respectively. The Dice score ranges from 0 to 1, with higher values indicating better geometric reconstruction of the contact region. Because the contact region is highly sparse in the present problem, the Dice score is a much more appropriate classification metric than simple pixel accuracy [37].

Because the large number of zero values in non-contact regions can dilute global MAE and RMSE, a normalized absolute-error-based metric, pixel-wise accuracy (PWA), was defined within the contact region to properly assess pressure-field regression. Since zero pressure is structurally enforced in non-contact regions by hard gating, the essential comparison of regression performance should be conducted within the contact region. Let the evaluation set be the pixels where the reference or the predicted pressure is positive:

$$\Omega_+ = \{i \mid p_i + \hat{p}_i > 0\} \quad (23)$$

Then PWA is computed as

$$\text{PWA} = 1 - \frac{1}{|\Omega_+|} \sum_{i \in \Omega_+} \frac{|p_i - \hat{p}_i|}{p_i + \hat{p}_i} \quad (24)$$

Here p_i and $\hat{p}_i$ are the reference and predicted pressures. PWA is one minus the mean symmetric pressure difference over Ω_+ , the union of their positive-pressure supports; it equals one if both fields are zero. The denominator is positive on this set, so no stabilizing constant or clipping is required.

Dice evaluates contact-patch overlap, whereas PWA evaluates pressure agreement over this union.

This definition matches the evaluation calculations.

Finally, for the proposed model to function as a practical contact surrogate, it is necessary to verify not only pixel-level predictive accuracy but also whether the predicted fields remain physically consistent at the aggregate level. To this end, three additional physical verification criteria were introduced: total-load equilibrium, full-profile ellipsoidal-contact consistency [49, 50], and load–area consistency based on Persson theory [4, 38–40]. Total-load equilibrium assesses how closely the total load obtained by integrating the predicted pressure field matches the target load.

$$\hat{P} = \int_{\Omega} \hat{p}(\mathbf{x}) dA \approx \sum_{i \in \Omega} \hat{p}_i \Delta A \quad (25)$$

The auxiliary benchmark uses the classical axisymmetric elastic half-space solution for the complete spheroidal profile [49, 50]. It assumes frictionless, non-adhesive linear elasticity. With the same two elastic bodies, $E_1=E_2=200$ GPa and $\nu_1=\nu_2=0.3$, the combined modulus obeys $1/E^*=\sum_j(1-\nu_j^2)/E_j$ and equals 109.89 GPa. The geometry parameters c and h_0 are obtained only from the input height field. At $F=30$ N, the contact radius a is determined by the load relation:

$$\psi(x) = \frac{h_0}{c} x \operatorname{atanh}\left(\frac{x}{c}\right); \alpha = \frac{a}{c}, 0 < \alpha < 1 \quad (26a)$$

$$F = E^* h_0 c [(1 + \alpha^2) \operatorname{atanh} \alpha - \alpha] \quad (26b)$$

$$p_{ell}(r) = \frac{E^*}{\pi} \int_r^a \frac{\psi'(x)}{\sqrt{x^2 - r^2}} dx, 0 \leq r < a \quad (26c)$$

$$p_{ell}(r) = 0 (r \geq a), A_{ell} = \pi a^2 \quad (26d)$$

Here ψ is the axisymmetric profile transform and A_{ell} is the theoretical contact area. Hertz theory with $R_{apex} = c^2/h_0$ is recovered only in the local-paraboloid limit $a/c \ll 1$ [26]; the test profiles have $a/c = 0.4237\text{--}0.8553$. For p equal to the BEM or predicted field, the theory-referenced errors are $e_{p,2} = 100\|p - p_{ell}\|_2/\|p_{ell}\|_2$, $e_A = 100|A/\pi a^2 - 1|$, with both L2 norms taken over the entire original 128×128 grid. For the BEM comparison, area A is obtained from the stored FFT-BEM

contact-support mask; for the network comparison, it is obtained from the stored predicted contact-support mask. Theory is evaluated at the same points in floating point, without pressure quantization or rescaling to enforce discrete load balance. These errors are separate from DL-BEM reconstruction metrics. The full-profile solution is used within the stated half-space assumptions and the checked 1 mm square domain.

Load-area consistency examines whether the predicted real contact-area ratio, computed from the pressure field, follows the macroscopic load-area relation expected for rough elastic contact. The predicted real contact-area ratio is computed as

$$\left(\frac{\hat{A}_r}{A_{\text{nom}}} \right) = \frac{1}{A_{\text{nom}}} \int_{\Omega} \mathbb{I}(\hat{p}(\mathbf{x}) > 0) dA. \quad (27)$$

where $\hat{A}_r$ is the predicted real contact area, A_{nom} is the nominal contact area, $\hat{p}(\mathbf{x})$ is the predicted contact pressure, and $\mathbb{I}(\cdot)$ is the indicator function. For comparison, the corresponding theoretical area ratio from Persson theory is

$$\frac{A_r}{A_{\text{nom}}} = \int_0^{\infty} \Phi(p; \bar{p}, V) dp = \text{erf}\left(\frac{\bar{p}}{\sqrt{2V}}\right) \quad (28)$$

where $\Phi(p; \bar{p}, V)$ is the probability density function of the local contact pressure p , determined by the mean pressure $\bar{p}$ and the pressure variance V , and $\bar{p} = P/A_{\text{nom}}$.

The Persson-based comparison was not used as an alternative ground truth for the local pressure field. FFT-BEM remains the primary numerical reference for the contact-support geometry and pixel-resolved pressure distribution of each surface realization. The Persson relation was used only as an independent, ensemble-level consistency check of the macroscopic load-area trend.

This distinction is relevant because an accurate integrated load does not by itself guarantee an accurate real contact area. The total load may be expressed as $P = A_r \langle p \rangle_c$, where A_r is the real contact area and $\langle p \rangle_c$ is the mean pressure over the contact region. Compensating errors in these two quantities may preserve the integrated load. The Persson-based comparison therefore evaluates

whether the predicted fields retain a reasonable partition between contact area and pressure at the three evaluated loads. Exact agreement with the approximate theory is neither expected nor claimed.

To examine whether the surrogate preserves the dependence of real contact area on surface statistics, the two-dimensional RMS slope was computed as

$$S_{dq} = \left\langle \left(\frac{\partial h}{\partial x} \right)^2 + \left(\frac{\partial h}{\partial y} \right)^2 \right\rangle^{1/2} \quad (29)$$

with the RMS roughness S_q as defined in Section 2.1. The monotonic associations between A_r/A_{nom} and these surface quantities were quantified using Spearman rank-correlation coefficients. Correlations were computed separately for the FFT-BEM references and the predictions from the corresponding load-specific two-stage models at each prescribed load. The contact-area prediction error was also examined as a function of S_q and S_{dq} to identify any systematic roughness-dependent bias.

Accordingly, the evaluation metrics in this study were designed to verify performance at distinct levels. Dice quantifies the accuracy of geometric reconstruction of the contact region, and PWA quantifies the regression accuracy of the pressure field over the union of the reference and predicted contact supports. Total-load equilibrium, ellipsoidal-contact benchmark consistency, and load–area consistency serve as physical verification criteria that assess whether the predictions remain consistent with aggregate physical quantities, benchmark solutions, and established scaling relations. This multi-layered evaluation framework enables a comprehensive assessment of whether the proposed model goes beyond generating visually plausible pressure fields and faithfully reproduces the input–output mapping of contact mechanics within the validated regime.

Because the zero-pressure background can inflate a full-field correlation, pressure regression is evaluated primarily on physically relevant supports. Let $\Omega_U = \{i: p_i^{ref} > 0 \vee \hat{p}_i > 0\}$ be the union of

reference and predicted supports, and $\Omega_{GT} = \{i: p_i^{ref} > 0\}$. With $\bar{p}_i^{ref}$ denoting the reference mean over support S ,

$$R_U^2 = 1 - \frac{\sum_{i \in \Omega_U} (\hat{p}_i - p_i^{ref})^2}{\sum_{i \in \Omega_U} (p_i^{ref} - \bar{p}_{\Omega_U}^{ref})^2 + \epsilon}, \quad R_{GT}^2 = 1 - \frac{\sum_{i \in \Omega_{GT}} (\hat{p}_i - p_i^{ref})^2}{\sum_{i \in \Omega_{GT}} (p_i^{ref} - \bar{p}_{\Omega_{GT}}^{ref})^2 + \epsilon} \quad (30)$$

The former full-field Pearson coefficient is retained only as a continuity diagnostic.

Direct complementarity diagnostics are computed from predicted pressure. We reconstruct $\hat{\mathbf{u}} = \mathbf{K} * \hat{\mathbf{p}}$ and $\hat{\mathbf{g}} = \mathbf{g}_0 + \hat{\mathbf{u}} - \delta$, where δ is taken from the corresponding FFT-BEM reference solution. With $\hat{m}_i$ the predicted binary support, $\hat{\Omega}_c = \{i: \hat{m}_i = 1\}$, prescribed load W , and $[z]_+ = \max(z, 0)$,

$$\eta_{pen} = \frac{\sum_i [-\hat{g}_i]_+ \Delta A}{A_{nom} S_q}, \quad \eta_{g,c} = \frac{\|\hat{\mathbf{m}} \odot \hat{\mathbf{g}}\|_2}{\sqrt{|\hat{\Omega}_c|} S_q + \epsilon}, \quad \eta_{pg} = \frac{\sum_i [\hat{p}_i \hat{g}_i] \Delta A}{W S_q + \epsilon}. \quad (31)$$

These residuals quantify departures from complementarity.

3 Results

This section presents the performance of the proposed two-stage hard-gated framework from field-level reconstruction to aggregate physical consistency. Section 3.1 summarizes the core predictive performance. Section 3.2 examines total-load consistency, the auxiliary ellipsoidal-contact benchmark, and the ensemble-mean load-area trend at the three prescribed loads. Section 3.3 evaluates solver-consistent transfer to measured engineering topographies without fine-tuning. Section 3.4 presents the ablation study. Section 3.5 examines robustness with respect to roughness statistics, load-specific model behavior, and best- and worst-case diagnostics. Sections 3.6 and 3.7 provide supporting results on data efficiency and resolution scaling. Section 3.8 reports the fixed-spectrum grid-convergence study of the FFT-BEM reference, and Section 3.9 reports the measured computational-cost and amortization accounting.

3.1 Performance overview

Fig. 4 presents a representative qualitative prediction for a test case. Even under complex multi-asperity contact conditions, the proposed model reproduces the spatial distribution of contact patches and local high-pressure regions with close visual agreement to the reference solution. In particular, the combination of the predicted contact mask and latent pressure field through hard gating yields exactly zero pressure in predicted non-contact regions at the final output stage. Fig. 4 therefore indicates that the model reproduces both the spatial distribution of the contact regions and the localized pressure variation, rather than merely achieving visual similarity to the reference solution.

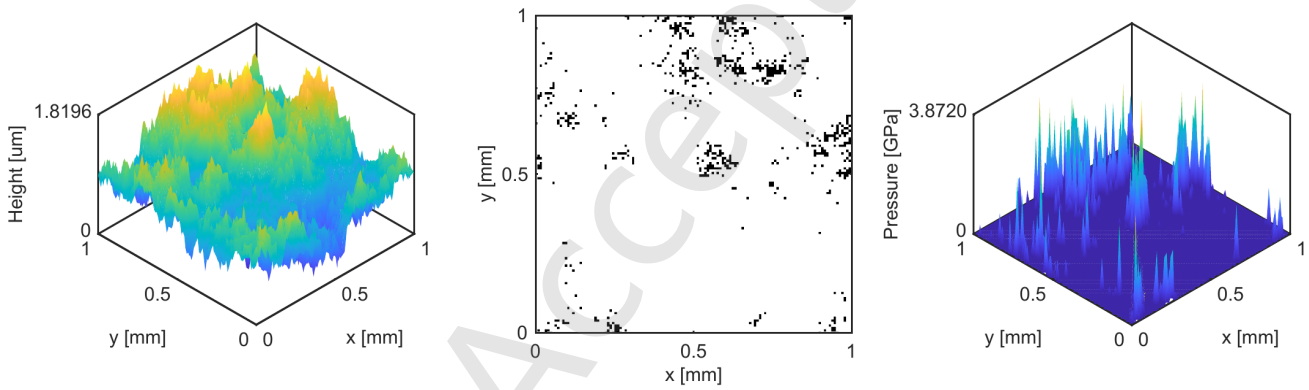


Fig. 4 Representative prediction of the integrated 30 N model at 128×128 resolution for a held-out rough-contact test case: (a) three-dimensional rough-surface topography, (b) predicted contact-support distribution, and (c) predicted contact-pressure field. The hard-gated two-stage architecture enforces zero pressure outside predicted support while preserving asperity-scale pressure localization within contact patches.

Table 6 summarizes the performance of the optimal configuration, namely the two-stage model with Focal RMSE, on the 5,000-sample test set. The contact-region segmentation accuracy remained high, with a mean Dice score of 0.9909 ± 0.00042 . Pressure regression within the contact region also showed strong performance, with a mean PWA of 0.9587 ± 0.01086 . The contact-region relative L_2 error remained 0.0140 for the primary 30 N model. Because the one-pixel maximum is grid sensitive, $p_{99.9}$ is used as the primary extreme-pressure statistic: its mean relative errors against the discrete

128 × 128 FFT-BEM references were 0.0049, 0.0043, and 0.0047 at 30, 60, and 90 N, respectively.

Across the same three load-specific test sets, mean R_U^2 was 0.9995, 0.9996, and 0.9996; R_{GT}^2 had the same rounded values, and mean boundary F1 was 0.9954, 0.9963, and 0.9968. Agreement in $p_{\max}$ is retained only relative to the discrete 128 × 128 FFT-BEM reference. For continuity, the former full-field Pearson coefficient is reported parenthetically as $R = 0.9998$, but it is not the primary regression metric. These results show that the proposed model accurately reconstructs not only the geometry of the contact region but also the amplitude and spatial distribution of pressure within contact patches.

Table 6 Summary of optimal model performance on the test dataset (5,000 samples).

Metric	Value
Mean Dice Score	0.9909 ± 0.00042
Mean PWA	0.9587 ± 0.01086
Mean Contact-Region L_2 Rel. Error	0.0140
Mean Total Load Error	0.1352 N
Mean R_U^2	0.9995
Mean Boundary F1	0.9954
Mean $p_{99.9}$ Rel. Error	0.0049
Mean $p_{\max}$ Rel. Error ^a	0.0058

^a Relative to the discrete 128×128 FFT-BEM reference; not a continuum-limit claim.

These trends were maintained consistently across the three load conditions (30 N, 60 N, and 90 N), with a mean total-load error of only 0.1352 N. The interpretation of load equilibrium is examined in more detail in Section 3.2, but Table 6 already indicates that the pixel-level predictions do not deteriorate substantially when viewed from the perspective of aggregate physical quantities. In this

sense, the high Dice and PWA values extend beyond image-level similarity and correspond to a level of quantitative accuracy relevant to engineering use of the predicted contact-pressure field.

Online latency was measured on an NVIDIA RTX 3090 in FP32 with batch size 1, including both network branches, hard gating, and host-to-device/device-to-host transfer. After 50 warm-up iterations, 1,000 synchronized iterations gave median end-to-end latencies of 9.0 ms at 128×128 and 15.8 ms at 256×256 . The corresponding PNG read-inference-write loop medians were 10.3 and 19.1 ms. Full timing and offline-cost accounting are reported in Section 3.9. This is substantially faster than conventional workflows based on repeated high-fidelity FFT-BEM analyses, and indicates that the model can substantially reduce computational overhead in applications involving large-scale parametric exploration, iterative design, or real-time surrogate evaluation. Taken together, the results in Section 3.1 provide the main quantitative evidence that the proposed framework achieves both high predictive accuracy and fast inference.

3.2 Physical consistency and auxiliary ellipsoidal-contact benchmark

This subsection evaluates whether the predicted fields remain physically consistent in terms of total-load equilibrium, an ellipsoidal-contact benchmark, and macroscopic load–area scaling. For a contact-pressure surrogate to be meaningful in engineering use, high pixel-level accuracy alone is insufficient; consistency with aggregate load and benchmark contact behavior must also be examined. First, total-load consistency indicates whether the predicted pressure field remains physically reasonable when integrated over the interface. As summarized in Table 6, the optimal configuration (2-stage architecture + Focal RMSE) yielded a mean total-load error of 0.1352 N over 5,000 test samples. This corresponds to a relative error of approximately 0.15–0.45% for the training load levels of 30 N, 60 N, and 90 N. Notably, this level of load agreement was obtained without introducing any explicit load-conservation term during training. This observation indicates that the local pressure predictions remain consistent with aggregate load measures, at least within the evaluated regime.

The 30 N dataset contains 44,000 training profiles (40,000 rough and 4,000 ellipsoidal) and 5,500 validation profiles (5,000 rough and 500 ellipsoidal). Separate rough and auxiliary test sets contain 5,000 and 500 cases, respectively. The auxiliary test assesses reconstruction of the smooth single-patch morphology represented in the development data.

As shown in Table 7, the model achieved a mean Dice score of 0.9991 and a mean PWA of 0.9936 on the 500 ellipsoidal test cases against stored FFT-BEM labels, while retaining mean Dice and PWA values of 0.9909 and 0.9587 on the rough-contact test set. These results describe prediction agreement with stored labels across the two pressure-field morphologies. They should not be interpreted as evidence that individual rough-surface asperities are spherical or that the smooth-contact class was entirely unseen during training.

Table 7 Prediction performance against stored FFT-BEM references at 30 N and 128×128 resolution: 500 ellipsoidal-contact cases and 5,000 rough-contact cases.

Metric	Ellipsoidal (500)	Rough Surface (5,000)
<i>Classification Metrics (Contact Area)</i>		
Mean Dice Score	0.9991	0.9909
Dice Score Std. Dev.	0.00014	0.00042
<i>Regression Metrics (Contact Pressure)</i>		
Mean PWA	0.9936	0.9587
PWA Std. Dev.	0.00411	0.01086

Fig. 5 compares the full-profile theoretical pressure, FFT-BEM labels and network predictions for test IDs 34, 137 and 318 at 30 N. The theory comparison comprises 4,000 training, 500 validation and 500 test profiles (5,000 in total). Across these profiles, the mean full-grid BEM pressure L2 error is 0.9972%, and the mean absolute contact-area error is 1.3224%. For the 500 test cases, the corresponding BEM mean errors are 0.9967% and 1.3199%; the network mean errors are 1.1697% and 1.3154%, respectively.

The maximum BEM pressure L2 error over the 5,000 profiles is 3.8405%, and the maximum absolute contact-area error is 7.1373%. These quantities assess the BEM fields against an independent geometry-consistent theory and are distinct from the DL-BEM reconstruction metrics in Table 7.

For rough contact, the ensemble-mean real contact-area ratio increased from 0.1307 to 0.1826 and 0.2218 in the FFT-BEM references at the prescribed loads of 30, 60, and 90 N, respectively. The corresponding predictions were 0.1308, 0.1828, and 0.2220. Thus, the absolute difference between the FFT-BEM reference and prediction was at most 0.0002, or 0.02 percentage points (Table 8).

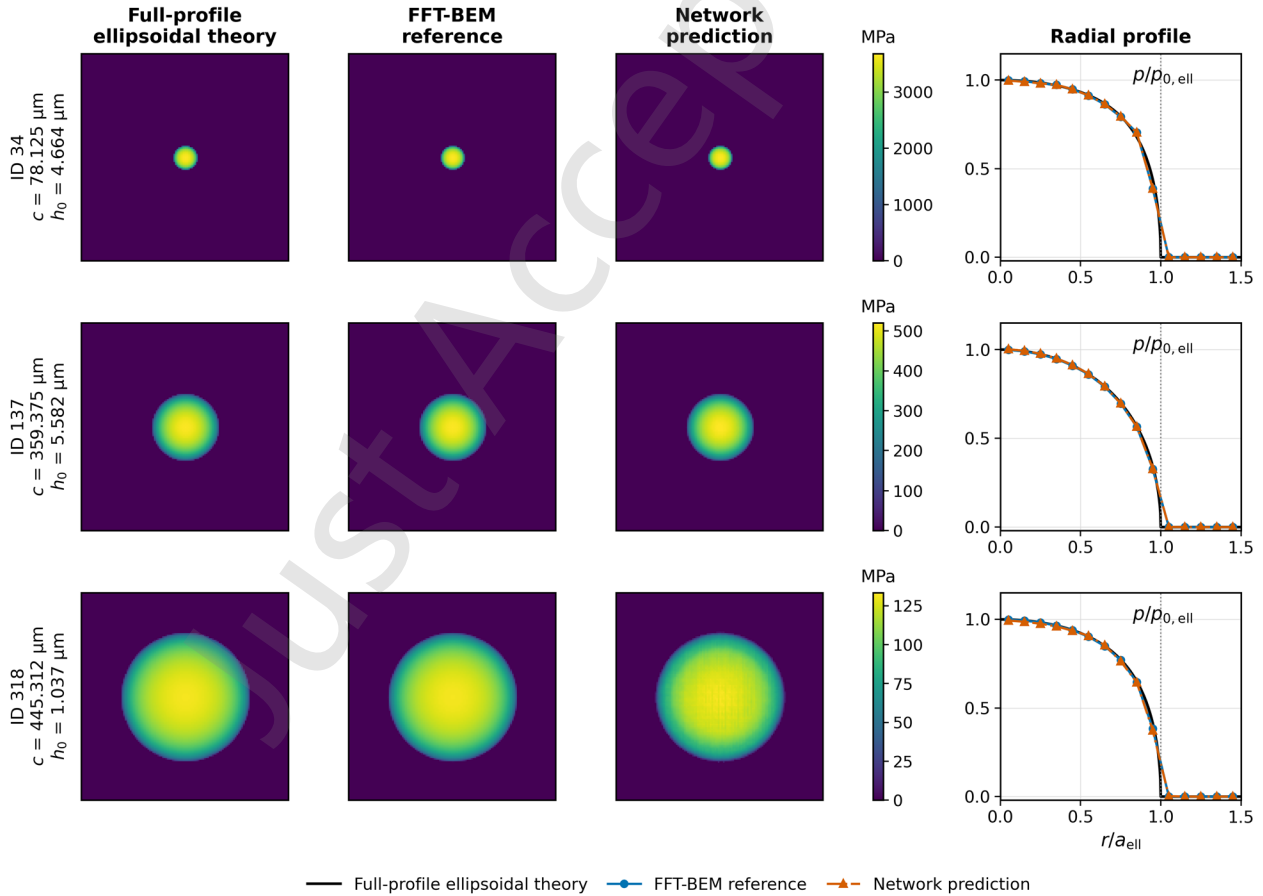


Fig. 5 Full-profile ellipsoidal theory, FFT-BEM references and network predictions at 30 N for axisymmetric oblate spheroidal profiles (test IDs 34, 137 and 318, top to bottom). Columns show theoretical pressure, FFT-BEM pressure, prediction and radial profiles. All pressure maps share coordinates and a common color scale

within each row. Every radial curve uses the same theoretical contact radius a_{ell} and central pressure $p_{0,ell}$ for that case. Numerical curves are radial-bin averages about the fixed input apex; theory is continuous. The geometry parameters c and h_0 are shown for each case.

A linear fit through the three evaluated load levels gave slopes of $1.5183 \times 10^{-3} N^{-1}$ for FFT-BEM and $1.5200 \times 10^{-3} N^{-1}$ for the predictions, corresponding to a relative slope difference of approximately 0.11%. The respective coefficients of determination were 0.9936 and 0.9935. These values indicate that the load-specific predictions preserve the near-linear increase in ensemble-mean contact area observed in FFT-BEM at the three evaluated loads.

This comparison should not be interpreted as validation of interpolation between the three load levels or extrapolation beyond 90 N. The applied load is not a network input, and each result was obtained using a checkpoint trained specifically at the corresponding load.

For comparison at the ensemble level, the Persson-based estimates of the mean real contact-area ratio were 0.1321, 0.1846, and 0.2241 at 30, 60, and 90 N, respectively. These values are slightly larger than the corresponding FFT-BEM values, as expected for an approximate statistical theory, but reproduce the same increasing load-area trend. The purpose of this comparison is not to replace the FFT-BEM reference or to claim exact theoretical agreement. Rather, it shows that spatial aggregation of the predicted fields remains consistent with the expected macroscopic rough-contact trend at the three evaluated loads.

Table 8 Ensemble-mean real contact-area ratio at 30, 60, and 90 N for the FFT-BEM references, the corresponding load-specific two-stage-model predictions at 128×128 resolution, and the Persson-based theoretical estimate. Each model is evaluated on its corresponding 5,000-case held-out rough-contact test set. The FFT-BEM and two-stage-model values show nearly identical increases across the three evaluated load levels. This comparison does not represent interpolation or extrapolation to unseen loads.

Load [N]	Mean GT	Mean Pred	Mean Persson
30	0.1307	0.1308	0.1321
60	0.1826	0.1828	0.1846
90	0.2218	0.2220	0.2241

3.3 Solver-consistent transfer to measured engineering topographies without fine-tuning

To evaluate transfer to real engineering surface geometries (Fig. 6), the model trained on synthetic rough surfaces was applied without additional fine-tuning to four experimentally measured sandpaper-finished steel topographies: P40, P320, P800, and P2000 [43, 44].

Only the three-dimensional surface topographies were measured experimentally. The contact-support and pressure fields used as references were not obtained from direct contact-pressure measurements. Instead, each measured topography was resampled to the 128×128 model resolution and analyzed using the same FFT-BEM formulation, material properties, boundary conditions, and 30 N load used for the corresponding synthetic cases. The resulting fields are therefore referred to as FFT-BEM numerical references.

The pretrained 30 N model was applied directly to the measured topographies, and its predictions were compared with these numerical references. This evaluation should be interpreted as a solver-consistent synthetic-to-measured-topography transfer test without fine-tuning, rather than as direct experimental validation of the contact-pressure field.

Whereas synthetic fractal surfaces exhibit statistically isotropic and self-affine roughness, measured sandpaper surfaces include directional scratches, heterogeneous texture, and more irregular contact boundaries.

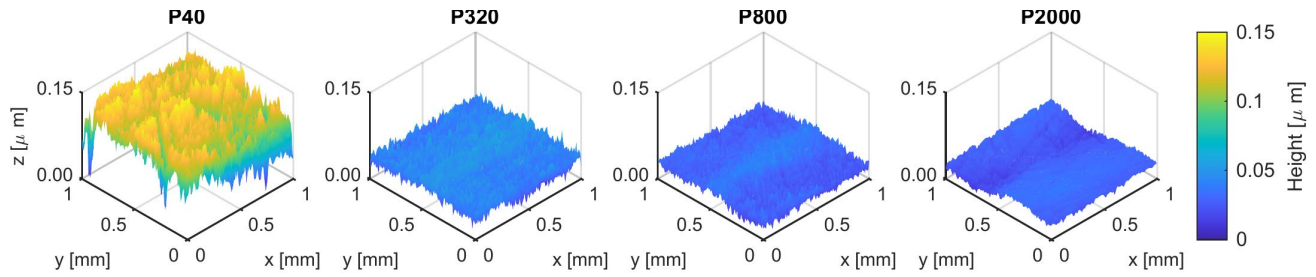


Fig. 6 3D topographies of the four sandpaper-finished steel surfaces (P40, P320, P800, P2000) used for solver-consistent transfer evaluation without fine-tuning [43, 44]. The four grit levels span coarse to fine surface morphology not seen during training. All surfaces share a common height color scale (0–0.15 μm). The spatial domain spans 1mm $\times$ 1 mm, resampled to 128 $\times$ 128 resolution for model input.

The quantitative results show that this transfer evaluation extends beyond visual similarity alone. As summarized in Table 9, the contact-region prediction performance across the four surfaces yielded Dice scores of 0.919–0.958, with a mean of 0.940, while the pressure correlation remained high, with $R^2 = 0.955$ –0.989 and a mean of 0.970. The PWA within the contact region decreased to 0.718–0.813, with a mean of 0.763, relative to the synthetic test set. This reduction is likely associated with the greater sensitivity of the absolute-error-based metric to fine boundary misalignment and directional morphological features in measured surfaces. In other words, on measured surfaces the model reproduced the contact location and overall pressure structure reasonably well, while small mismatches in irregular contact boundaries and local peak amplitudes were reflected more strongly in PWA.

Surface-wise qualitative and quantitative comparisons are shown in Fig. 7. For all four surfaces, the predicted contact regions consistently reproduced the main patches of the FFT-BEM numerical reference, and the pressure-correlation scatter plots remained tightly clustered around the identity line, yielding high R^2 values. The KDE-based comparison of pressure distributions also shows that the overall shapes of the FFT-BEM reference and predicted distributions remain broadly similar, and

that performance degradation occurs mainly near the high-pressure tail or local peak region. Surface-specific differences were also observed: the relatively rough P40 and P800 surfaces showed higher PWA, whereas the finer P2000 surface yielded the highest Dice score but the lowest PWA. This suggests that as the morphological scale becomes finer, boundary alignment of contact patches may become easier, while accurate recovery of pressure amplitude within the contact region may become more sensitive.

Taken together, these results do not replace direct experimental measurement of contact pressure, but they do provide evidence of partial transferability from synthetic rough surfaces to real measured engineering topographies under solver-consistent reference conditions.

Table 9 Solver-consistent transfer of the pretrained 30 N model (128×128 resolution) to the four measured sandpaper-finished steel topographies (P40, P320, P800, and P2000) without fine-tuning: contact-region reconstruction (Dice) and pressure-field fidelity (R^2 and PWA) relative to FFT-BEM numerical references.

Metric	P40	P320	P800	P2000	Mean
<i>Classification Metrics (Contact Area)</i>					
Dice Score	0.951	0.919	0.931	0.958	0.940
<i>Regression Metrics (Contact Pressure)</i>					
R^2	0.989	0.955	0.976	0.962	0.970
PWA	0.813	0.723	0.796	0.718	0.763

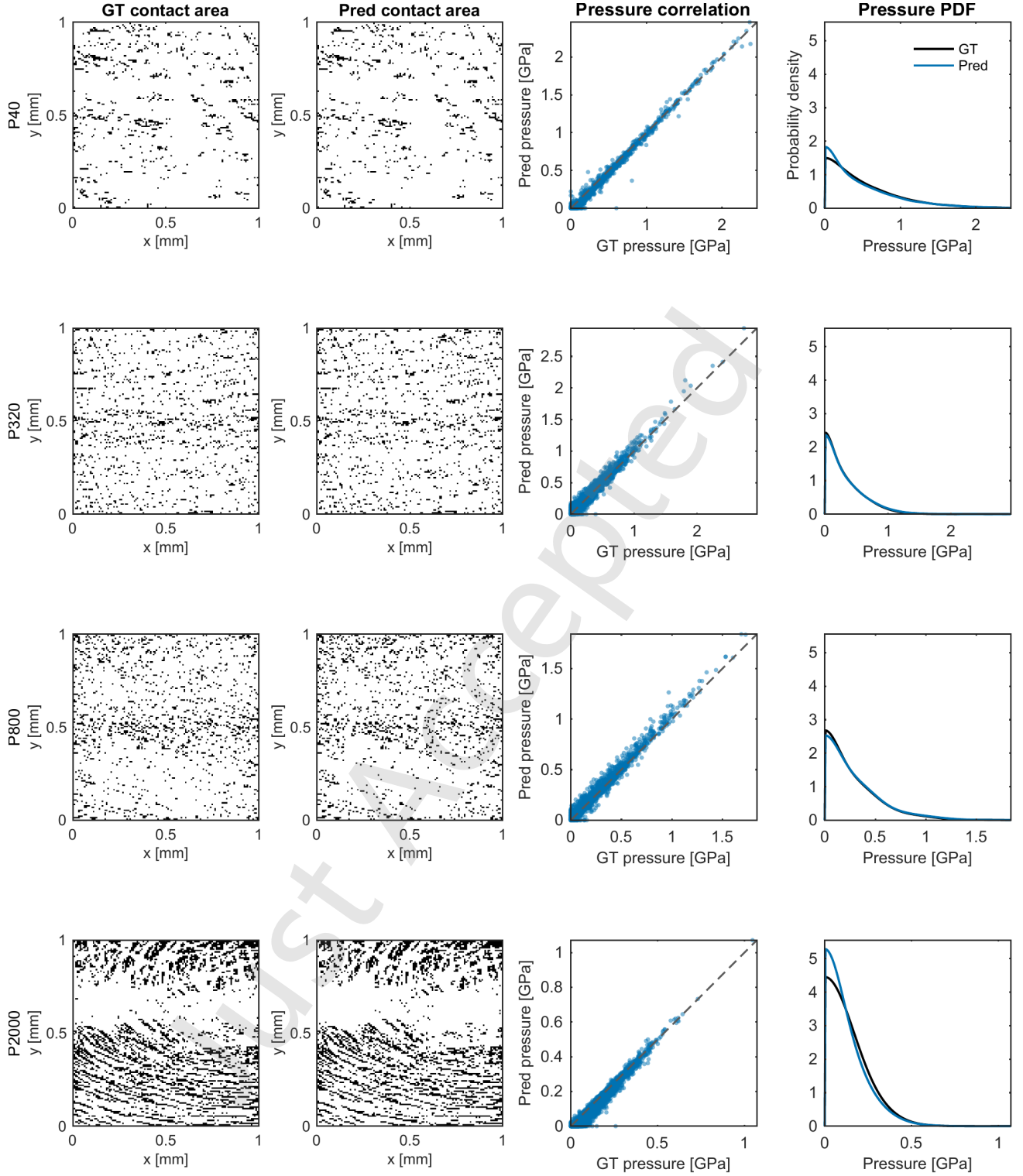


Fig. 7 Solver-consistent transfer of the pretrained 30 N model (128×128 resolution) to experimentally measured sandpaper-finished steel topographies without fine-tuning. Each row corresponds to one measured surface. The first two columns show the FFT-BEM numerical reference and predicted contact-support distributions. The third column compares the predicted pressure with the FFT-BEM numerical reference, and

the fourth compares their pressure probability-density functions. Only the surface topographies were measured experimentally; all contact-support and pressure reference fields were computed by FFT-BEM.

3.4 Ablation study

This subsection isolates the design elements responsible for the performance gains reported in Sections 3.1 and 3.2. Specifically, the contributions of the two-stage structure, hard gating, and the loss-function design are evaluated separately. To ensure fair comparison of the relative contribution of each design element, all ablation models were trained on 1,000 samples and evaluated on an independent test set.

First, to examine the effect of the two-stage structure and hard gating, a 1-stage regression model and a 2-stage model were compared under the same Focal RMSE loss. Because the 1-stage model directly regresses the pressure field without explicitly estimating the contact region, it tended to produce diffuse residual pressure over non-contact regions. By contrast, the 2-stage model learns contact-region segmentation and latent pressure-field regression separately and then combines them through hard gating, thereby enforcing zero pressure structurally in the predicted non-contact region. As a result, the mean PWA increased substantially, from 0.1336 for the 1-stage model to 0.8955 for the 2-stage model. This difference is also evident in the representative qualitative comparison shown in Fig. 8. The 1-stage model leaves broad residual responses across non-contact regions, blurring both contact-patch boundaries and local pressure structure, whereas the 2-stage model separates contact and non-contact regions more sharply and better reproduces local pressure peaks. Thus, the main advantage of the two-stage structure is not only lower average error, but also more accurate recovery of the contact-region geometry associated with the contact-pressure field.

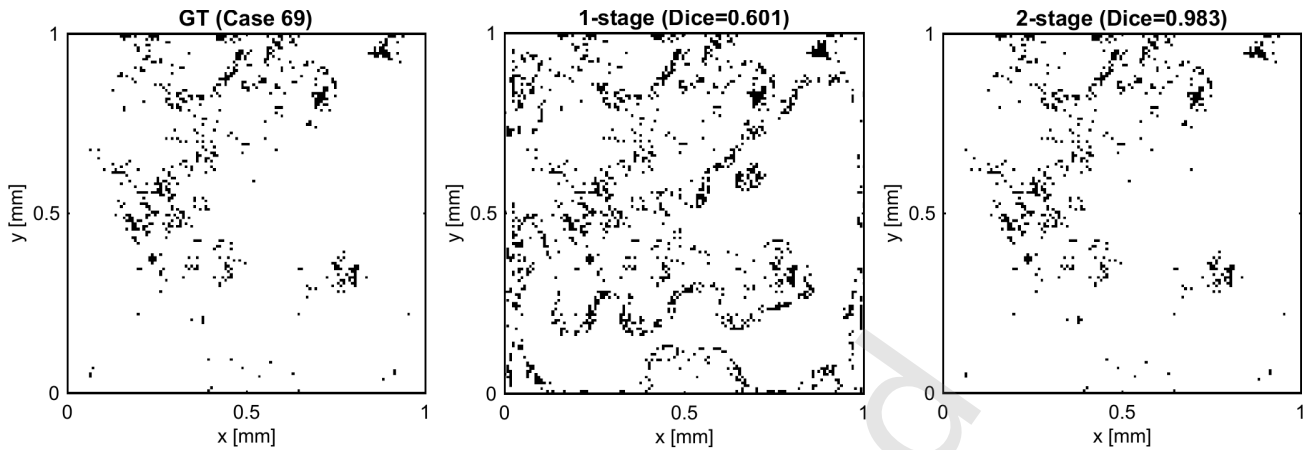


Fig. 8 Representative 30 N, 128×128 held-out rough-contact test case comparing the single-stage baseline with the two-stage model. The single-stage model (Dice = 0.601) leaves diffuse residual pressure in non-contact regions, whereas the two-stage model (Dice = 0.983) uses hard gating to localize contact boundaries and better reproduce local pressure peaks.

Next, to isolate the contribution of Focal RMSE, the 2-stage architecture was held fixed while the regression loss was varied among L1, RMSE, and Focal RMSE. As shown in Table 10, the Dice scores were similar across the three loss functions, indicating that contact-region segmentation was largely unaffected by the choice of regression loss. In contrast, the pressure prediction accuracy within the contact region, as measured by PWA, was highest for Focal RMSE (0.8955), followed by RMSE (0.8886) and L1 (0.8464). This indicates that the regression loss affects the reconstruction accuracy of the pressure field within the contact region more directly than the geometric localization of the contact region itself. The larger gain of Focal RMSE relative to L1 suggests that focal weighting provides a stronger training signal for sparse high-pressure regions with large amplitude.

Table 10 Comparison of Dice score and PWA among regression-loss variants of the independently parameterized two-stage architecture at 30 N and 128×128 resolution, based on the same 1,000-case rough-contact evaluation subset (mean $\pm$ standard deviation over three runs).

Metric	L1	RMSE	Focal RMSE
Dice	0.9770 ± 0.0011	0.9750 ± 0.0014	0.9746 ± 0.0009
PWA	0.8464 ± 0.0179	0.8886 ± 0.0005	0.8955 ± 0.0011

The advantage of Focal RMSE becomes clearer in the analysis of maximum contact pressure prediction. As reported in Table 11, Focal RMSE achieved a peak-location match ratio of 285/300, exceeding both L1 (225/300) and RMSE (279/300). The mean error at the GT peak location was 404.4 MPa, lower than that of RMSE (504.43 MPa) and L1 (1336.6 MPa). The error at the predicted peak location was also smallest for Focal RMSE, at 343.2 MPa. These results show that Focal RMSE improves the prediction of the location and magnitude of sparse high-pressure peaks while maintaining strong average performance over the full pressure field. In this sense, the modest differences in mean performance seen in Table 10 correspond to more pronounced gains in the local high-pressure regime, as shown in Table 11.

Table 11 Comparison of maximum-pressure prediction accuracy relative to the discrete 128×128 FFT-BEM reference for loss-function variants of the architecture at 30 N (300-case rough-contact test subset). The point maximum is a grid-dependent diagnostic and is not interpreted as a continuum-limit pressure.

Metric	L1	RMSE	Focal RMSE
Peak Location Match Ratio	225/300	279/300	285/300
Error at GT Peak Location [MPa]	1336.6	504.43	404.4
Error at Pred Peak Location [MPa]	1042.0	473.8	343.2

The ablation results show that hard gating improves contact-region recovery, whereas Focal RMSE improves peak-pressure regression.

Effect of adversarial training. We first performed a factorial screening at 256×256 and 30 N. The four assembled inference configurations were supervised-only (S0), mask-critic-only (S1), pressure-

critic-only (S2), and both critics (S3). Each branch condition was trained with three seeds on the same 10,000-sample subset, and all configurations were evaluated on the complete 5,000-case test set. Support metrics depend only on the support branch, so S0/S2 and S1/S3 coincide on those metrics by construction.

Table 12 Three-seed screening ablation at 256×256 and 30 N: means over the three training seeds on the complete 5,000-case test set, for models trained on the same 10,000-sample subset. σ_{seed} is the S0 seed standard deviation. The contact-support metrics depend only on the support branch, so S0/S2 and S1/S3 coincide on them by construction.

Metric	S0	S1	S2	S3	σ_{seed}
Dice	0.99094	0.99081	0.99094	0.99081	0.00016
IoU	0.98205	0.98179	0.98205	0.98179	0.00031
Boundary F1	0.99545	0.99538	0.99545	0.99538	0.00010
PWA	0.93509	0.93483	0.93543	0.93517	0.00323
R_U^2	0.99741	0.99742	0.99753	0.99754	0.00040
$p_{99.9}$ relative error	0.01169	0.01166	0.01013	0.0101	0.00140
Gradient error	0.03358	0.03360	0.03254	0.03256	0.00265
PSD error	0.00771	0.00771	0.00724	0.00724	0.00096
Absolute load error	0.00837	0.00837	0.00899	0.00899	0.00081

The screening rule was fixed before examining full-data confirmation: an adversarial branch would be retained as beneficial only if it produced a consistent improvement beyond ordinary seed variability. The mask critic did not satisfy that rule. The full-data comparison therefore contrasts supervised-only S0 with the archived both-critics S3 checkpoint on the same 5,000 test surfaces.

Table 13 Full-data comparison at 256×256 and 30 N. The last column normalizes the paired difference by the S0 seed standard deviation measured in the three-seed screening study.

Metric	S0	S3	S0-S3	Δ/σ_{seed}
Dice	0.99441	0.99385	$+5.58 \times 10^{-4}$	3.52
IoU	0.98889	0.98778	$+1.10 \times 10^{-3}$	3.54
Boundary F1	0.99721	0.99696	$+2.52 \times 10^{-4}$	2.42
PWA	0.95382	0.95483	-1.01×10^{-3}	0.31
R_U^2	0.99888	0.99897	-9.01×10^{-5}	0.23
$p_{99,9}$ relative error	0.00641	0.00628	$+1.31 \times 10^{-4}$	0.09
Gradient error	0.02198	0.02069	$+1.29 \times 10^{-3}$	0.49
PSD error	0.00459	0.00459	-2.20×10^{-6}	0.00
Absolute load error	0.00542	0.00524	$+1.80 \times 10^{-4}$	0.22

The full-data mask-critic configuration is 2.42-3.54 σ_{seed} worse than S0 on support metrics, whereas all pressure-side and related differences are within seed variability (0.00-0.49 σ_{seed}). Controlled critic training increased training time by approximately 415% and produced no material improvement. The resulting hierarchy is single-stage regression $\ll$ two-stage without critic $\approx$ two-stage with critic. Because full-data S0 and archived S3 each come from a single run, this conclusion is “no material improvement from the adversarial component,” not exact equivalence.

3.5 Robustness and scope diagnostics

The preceding subsections established the baseline predictive performance, physical consistency, solver-consistent transferability to measured topographies, and the contribution of the main design elements. For a contact surrogate to be practically useful, however, it is also necessary to evaluate whether the model mitigates structural bias relative to prior surrogates, whether its accuracy remains stable under changes in surface roughness and applied load, and under what conditions it performs best or worst. Accordingly, this subsection evaluates robustness through comparison with a prior surrogate model, the dependence of real contact area on surface roughness statistics, sensitivity

analyses with respect to roughness and load, and best- and worst-case diagnostics, all interpreted within the evaluated load and roughness regime.

3.5.1 Prior comparison

Before discussing generalization more broadly, it is useful to examine whether the structural differences introduced in this study reduce the prediction biases reported for prior deep-learning-based contact surrogates. In previous work [20], a single U-Net-based network was used to predict contact area directly, but systematic underprediction of contact area was reported because of severe class imbalance, with non-contact pixels accounting for more than 90% of the domain. That study reported a mean error rate of 12.60% and a Dice score of 0.9188.

By contrast, the two-stage hard-gated framework proposed here separates contact-region segmentation from pressure regression and combines them through hard gating, thereby reducing this bias structurally. As shown in Fig. 9, the predicted contact areas cluster closely around the identity line ($y = x$) for all load conditions (30 N, 60 N, and 90 N), with limited scatter. This indicates that the predictions remain nearly unbiased with respect to the ground truth, rather than exhibiting a systematic deviation in a particular direction.

The same trend is reflected in the quantitative performance. The optimal configuration of the present model achieved contact-region reconstruction with Dice scores above 0.99 at 128×128 resolution, representing a clear improvement over the Dice score of 0.9188 reported for the prior single-network approach. Importantly, this improvement is not attributable simply to increased network capacity, but to the combined effect of the two-stage decomposition of contact region and pressure prediction, the classification branch trained with Dice loss, and zero-pressure structural enforcement through hard gating. The primary advantage of the proposed model, therefore, lies not only in improved mean accuracy, but in the structural reduction of systematic contact-area underprediction in a strongly imbalanced contact problem.

Accordingly, the comparison in Fig. 9 should be interpreted not only as evidence of improved performance relative to a baseline, but also as an initial indication that the proposed framework suppresses a recurring bias observed in conventional single-regression models.

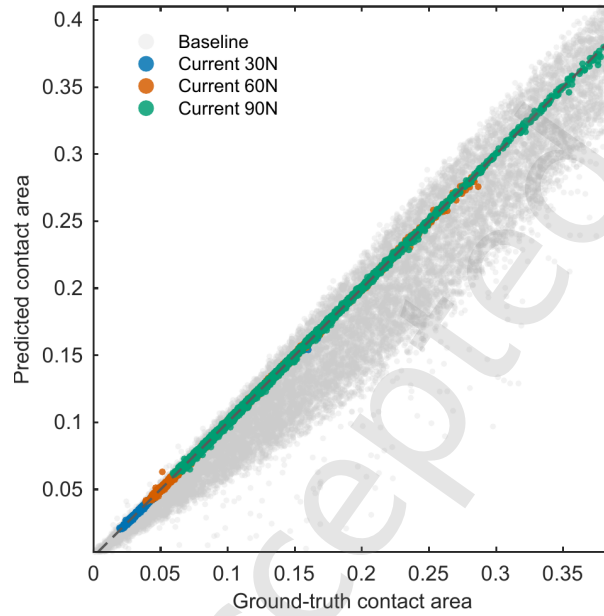


Fig. 9 Predicted versus FFT-BEM ground-truth contact area on the corresponding 5,000-case held-out rough-contact test sets for the load-specific models at 30, 60, and 90 N; all models use 128×128 grids.

3.5.2 Dependence of real contact area on surface roughness statistics

The dependence of the real contact-area ratio on the surface roughness statistics was compared between the FFT-BEM references and the two-stage-model predictions. At each prescribed load, A_r/A_{nom} showed a strong negative monotonic association with the RMS slope S_{dq} . The Spearman coefficients for FFT-BEM ranged from -0.991 to -0.996, whereas those for the predictions ranged from -0.991 to -0.995. The corresponding coefficients for the RMS roughness S_q ranged from -0.964 to -0.966 for FFT-BEM and from -0.964 to -0.965 for the predictions.

The close agreement between the FFT-BEM and prediction coefficients indicates that the surrogate preserves the ranking of real contact area with respect to surface roughness. This behavior is consistent with the low-contact-fraction scaling

$$\frac{A_r}{A_{nom}} \approx \kappa \frac{\bar{p}}{E \cdot S_{dq}} \quad (32)$$

according to which a larger RMS slope produces a smaller real contact area at fixed nominal pressure and effective modulus [2, 39, 41, 42].

No systematic dependence of the relative contact-area prediction error on S_q or S_{dq} was observed, and the mean area bias remained below 0.5% over the evaluated S_q intervals. Thus, the model reproduces not only the ensemble-mean contact area but also its mechanics-based dependence on the surface statistics within the training distribution.

3.5.3 Sensitivity to surface roughness and applied load

Table 14 and Fig. 10 show how the proposed model responds to changes in surface roughness and applied load. High average accuracy alone is not sufficient for practical engineering use; the predictive performance should also remain stable under variation of key physical parameters.

To assess sensitivity to surface roughness, the full test set was partitioned by R_a range, and the mean Dice score and mean PWA were computed for each interval (Table 14). Across the full evaluated range ($R_a = 0.0\text{--}0.7 \mu\text{m}$), the model maintained consistently high performance. Mean Dice ranged from 0.9866 to 0.9916, and mean PWA ranged from 0.9410 to 0.9626. The highest performance was observed in the more frequently represented range of $R_a = 0.1\text{--}0.3 \mu\text{m}$. A small decline was observed at larger roughness levels ($R_a > 0.4 \mu\text{m}$), but the reduction was limited. These results indicate that the model is not narrowly optimized for a specific roughness range, but maintains stable prediction accuracy across a broad roughness distribution.

Table 14 Contact-support classification and pressure-regression performance of the 30 N model at **128 × 128** resolution, segmented by R_a range over the 5,000-case held-out rough-contact test set.

R_a Range (μm)	Mean Dice	Mean PWA
0.0–0.1	0.9892	0.9410
0.1–0.2	0.9916	0.9590
0.2–0.3	0.9916	0.9626
0.3–0.4	0.9903	0.9600
0.4–0.5	0.9895	0.9569
0.5–0.6	0.9872	0.9504
0.6–0.7	0.9866	0.9481

These results also suggest that prediction difficulty is not determined solely by the absolute amplitude of roughness. If the magnitude of R_a were the dominant factor, performance would be expected to degrade monotonically with increasing roughness. Instead, the results show the highest accuracy in the intermediate roughness range, while still maintaining strong Dice and PWA values at higher roughness levels. This pattern is consistent with the possibility that prediction difficulty is influenced not only by average height amplitude, but also by geometric features such as patch fragmentation, boundary complexity, and the distribution of local peaks.

Sensitivity to load is shown in Fig. 10. For all load conditions (30 N, 60 N, 90 N), the Dice score remained above 0.99 and the PWA remained above 0.96, indicating stable predictive accuracy across the tested load range. A slight improvement in both metrics was also observed with increasing load. One possible explanation is that, as the load increases, the real contact area expands, increasing the proportion of contact pixels in the full grid and thereby partially alleviating class imbalance.

Overall, the model performance remains relatively flat with respect to changes in roughness and load, and any error increase appears gradual rather than abrupt under specific conditions.

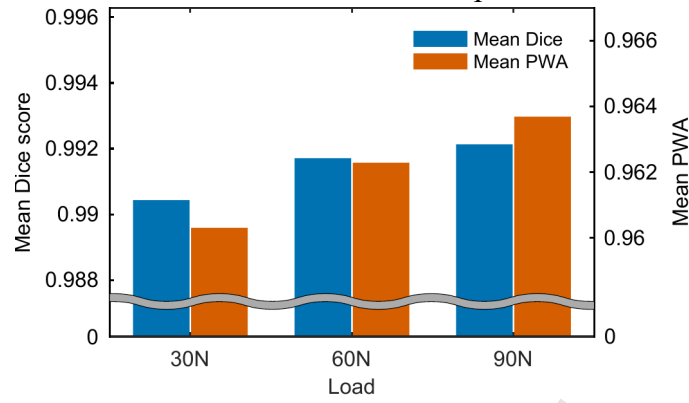


Fig. 10 Dice score and PWA on the corresponding 5,000-case held-out rough-contact test sets for the load-specific models at 30, 60, and 90 N; all models use 128×128 grids.

3.5.4 Best- and worst-case diagnostics

To understand the generalization behavior more concretely, it is useful to examine not only the mean performance but also how the best-performing and worst-performing cases differ in their error structure. In this subsection, the highest-performing and lowest-performing surfaces in the test set are compared in terms of contact-region reconstruction, pressure correlation, and the statistical distribution of pressure within the contact region, in order to diagnose both the success mode and the principal limitation of the proposed model.

The best-performing case is shown in Fig. 11. For this surface, the proposed model reproduced the contact region and pressure distribution with near-complete agreement under all three load conditions. Contact-region reconstruction yielded $\text{Dice} \geq 0.994$ at all loads, and pressure accuracy within the contact region remained high, with $\text{PWA} \geq 0.971$. In addition, the pressure correlation analysis yielded $R^2 = 1.000$ for all load conditions, indicating an essentially ideal linear correspondence between predicted and reference pressures. The qualitative comparison in Fig. 11 also shows that the locations and boundaries of the contact patches, as well as the local pressure levels within them, were reconstructed with high fidelity. The load-dependent expansion of the contact region and increase in pressure magnitude were also reproduced with close agreement to the reference solution.

The worst-performing case is shown in Fig. 12. Here, contact-region prediction was lower than the mean performance, with Dice values of 0.939 at 30 N, 0.947 at 60 N, and 0.959 at 90 N. Pressure accuracy also decreased, with PWA ranging from 0.795 to 0.871. Nevertheless, the pressure correlation remained high, with $R^2 \geq 0.991$ for all load conditions. This indicates that even in the lowest-performing case, the model preserved the overall spatial correlation structure of the pressure field reasonably well, while larger errors arose in the precise regression of the size and amplitude of local high-pressure regions. This case also showed a more pronounced performance gain with increasing load than the average trend, with PWA increasing by approximately 0.076 from 30 N to 90 N. This suggests that for geometrically complex surfaces, the stronger class imbalance at lower loads may impose a greater burden on pressure-amplitude regression.

Friction

<https://mc03.manuscriptcentral.com/friction>

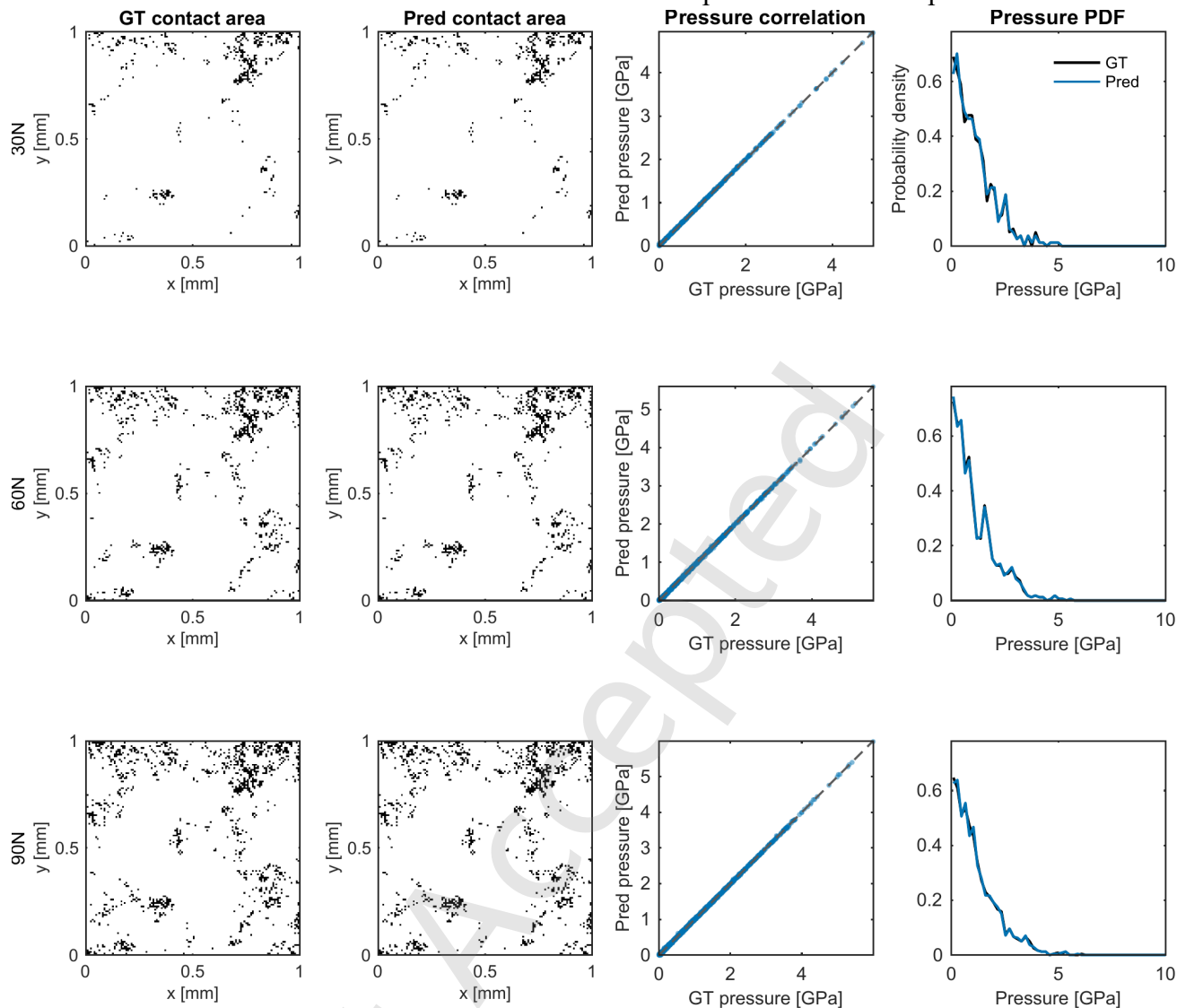


Fig. 11 Qualitative and statistical comparison for the best-performing held-out rough-contact surface using the load-specific models at 30, 60, and 90 N on 128×128 grids. Columns show the FFT-BEM ground-truth contact support, predicted contact support, contact-region pressure scatter plot, and kernel-density estimate of pressure for pixels with pressure > 0 . GT denotes the synthetic-data FFT-BEM ground truth and Pred denotes the corresponding load-specific prediction.

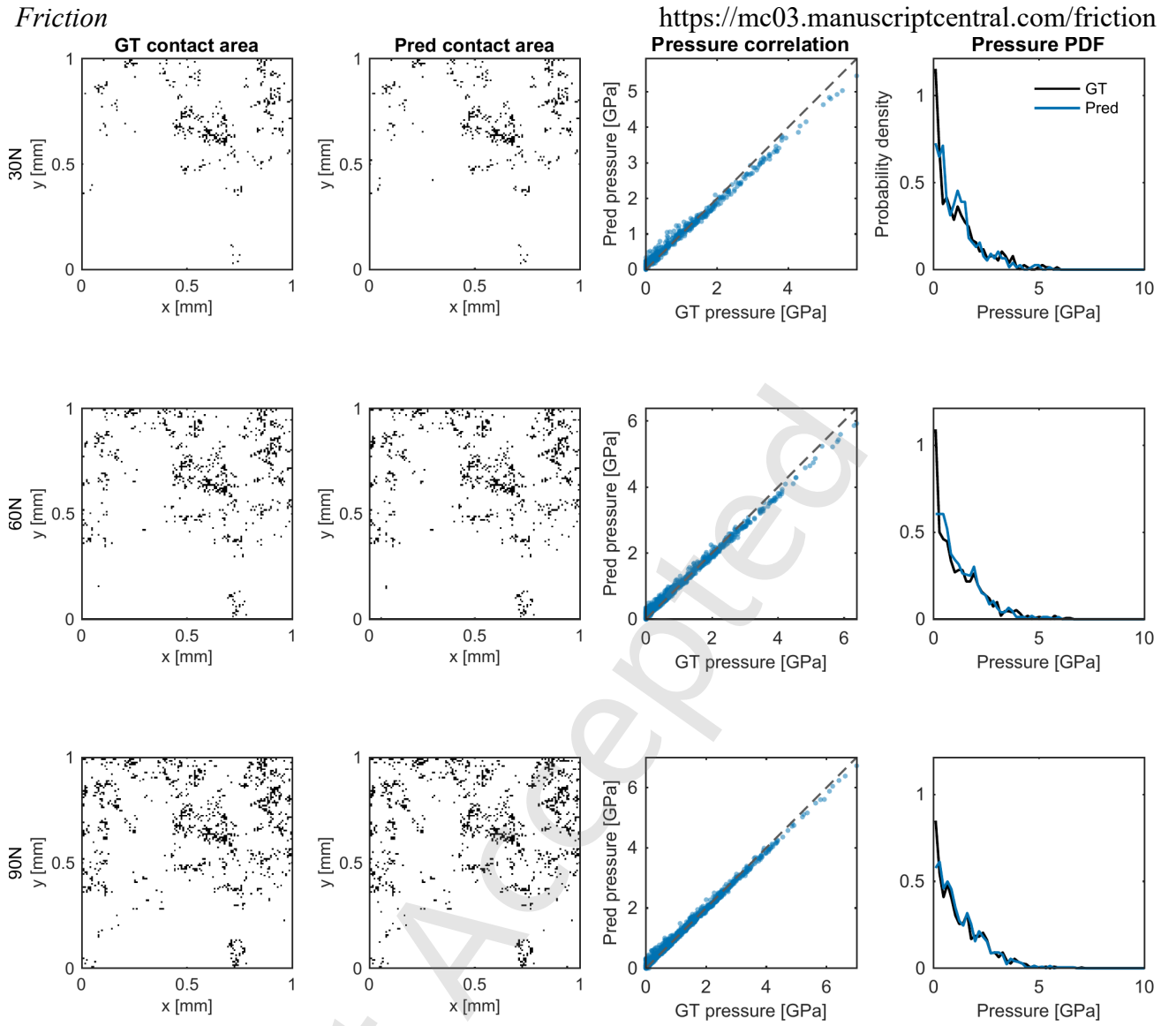


Fig. 12 Qualitative and statistical comparison for the worst-performing held-out rough-contact surface using the load-specific models at 30, 60, and 90 N on 128×128 grids. Columns show the FFT-BEM ground-truth contact support, predicted contact support, contact-region pressure scatter plot, and kernel-density estimate of pressure for pixels with pressure > 0 . GT denotes the synthetic-data FFT-BEM ground truth and Pred denotes the corresponding load-specific prediction.

This difference is also evident in the KDE-based probability density functions shown in the last column of Fig. 11 and Fig. 12. In the best-performing case, the pressure PDFs within the contact region almost overlap with the ground truth under all load conditions, reproducing not only the mean pressure level but also the high-pressure tail. In the worst-performing case, the overall distributional

shape is still preserved, but the discrepancy between GT and Pred becomes more visible near the high-pressure peak and in the tail region.

Taken together, these observations suggest that the main residual error arises less from failure to localize contact than from imperfect reconstruction of sparse high-pressure extrema.

3.6 Data efficiency

Data efficiency is a key consideration for practical deployment of surrogate frameworks in engineering workflows, where generating high-fidelity training data through numerical solvers is itself computationally expensive. Performance improved steadily as the size of the training dataset increased, although the rate of improvement was not uniform. This subsection quantifies the convergence behavior with respect to training data size in order to identify the data budget at which the proposed model reaches a practical balance between engineering accuracy and efficiency.

To this end, the training dataset size was increased stepwise from 100 to 1,000, 10,000, 20,000, and 40,000 samples, and performance was evaluated on the same 5,000-sample test set. As shown in Table 15, both contact-region segmentation and pressure regression improved monotonically with increasing data volume. With only 100 training samples, the model already achieved Dice 0.9299 and PWA 0.7845. With 1,000 samples, performance improved substantially to Dice 0.9738 and PWA 0.8962, reaching a stable baseline level of contact-region reconstruction and pressure regression. As the dataset size increased further to 10,000, 20,000, and 40,000, performance improved gradually to Dice 0.9819, 0.9894, and 0.9904, and PWA 0.9317, 0.9528, and 0.9603, respectively. These results indicate that the proposed model can achieve a stable baseline level of performance even with a relatively modest dataset, but that larger datasets are beneficial for more precise recovery of pressure amplitude within the contact region.

Table 15 Absolute prediction accuracy of two-stage architecture variants trained with the stated development-set volumes and evaluated at 30 N and 128×128 resolution on the same held-out rough-contact test split.

Dataset Volume	100	1,000	10,000	20,000	40,000
Dice	0.9299	0.9738	0.9819	0.9894	0.9904
PWA	0.7845	0.8962	0.9317	0.9528	0.9603

To interpret diminishing returns more quantitatively, the relative performance gain for each increment in dataset size was additionally computed (Table 16). In this study, practical convergence was defined as the point at which the relative improvement per increment fell below 1%. Under this criterion, the Dice score fell below a 1% improvement rate already after 1,000 samples, indicating that contact-region segmentation saturates relatively early. By contrast, PWA improved by 12.46% over 100–1,000 samples, 3.81% over 1,000–10,000 samples, and 2.21% over 10,000–20,000 samples, and decreased below 1% only over the 20,000–40,000 interval, where the gain was 0.78%. This indicates that the location and shape of the contact region can be learned relatively quickly, whereas more accurate regression of the pressure field within the contact region, especially local high-pressure structure, requires a larger data budget.

Table 16 Relative improvement per development-set volume increment for the same two-stage architecture variants evaluated at 30 N and 128×128 resolution on the same held-out rough-contact test split.

Volume difference	100–1,000	1,000–10,000	10,000–20,000	20,000–40,000
Dice [%]	4.51	0.82	0.76	0.10
PWA [%]	12.46	3.81	2.21	0.78

Thus, a useful baseline level of performance can be obtained with a relatively small dataset, but a larger data budget is required to stabilize fine pressure regression within the contact region.

3.7 Architecture scalability across grid-specific surface spectra

The 256×256 campaign extends the represented short-wavelength band by one octave relative to the 128×128 campaign and therefore constitutes a different grid-specific physical spectrum, rather than a refinement of identical surface realizations. The comparison tests whether the same two-stage hard-gated architecture can be trained on both grid-specific datasets. This comparison does not constitute a numerical convergence test of a fixed physical surface spectrum. Fixed-spectrum nested-grid behavior is examined separately in Section 3.8.

Table 17 compares the same two-stage hard-gated architecture across the two grid-specific datasets. At 256×256 , Dice was 0.9939, compared with 0.9909 at 128×128 ; PWA and R^2 were 0.9548 and 0.9991, compared with 0.9603 and 0.9998 at 128×128 . These values show that the architecture can be trained at both grid sizes. Because the two campaigns resolve different surface spectra, however, the differences cannot be attributed to grid refinement or numerical convergence and may reflect both the changed physical input band and the separate training campaign.

Accordingly, these results support architecture scalability across two grid-specific surface spectra; they do not establish mesh-refinement accuracy or continuum convergence. Fixed-spectrum numerical convergence of the FFT-BEM reference is evaluated separately in Section 3.8.

Table 17 Architecture-scalability comparison between the 30 N models trained at 128×128 and 256×256 resolution, each evaluated on its corresponding 5,000-case held-out rough-contact test set. The two datasets have grid-specific surface spectra; this table is not a fixed-spectrum numerical-convergence test.

Metric	128×128	256×256
Dice Score	0.9909	0.9939
PWA	0.9603	0.9548
R^2	0.9998	0.9991

3.8 Fixed-spectrum grid convergence of the FFT-BEM reference

A nested-grid study was performed from 30 independent 1024×1024 master spectra at 30 and 90 N. Two common physical cutoffs, $q_{N_y}^{128}$ and $0.8q_{N_y}^{128}$, were imposed before the same retained Fourier coefficients were resampled at 128^2 , 256^2 , and 512^2 ; a 20-case subset was additionally evaluated at 1024^2 . The nested coefficient identity was satisfied to 3.2×10^{-16} . For scalar Q ,

$$\epsilon_Q^{N \rightarrow 2N} = \frac{|Q_N - Q_{2N}|}{|Q_{2N}| + \epsilon}. \quad (33)$$

Table 18 Fixed-spectrum nested-grid behavior. Values are median/p95 of $\epsilon_Q^{N \rightarrow 2N}$; the **512** $\rightarrow$ **1024** column contains the 20-probe subset.

Quantity	128 $\rightarrow$ 256	256 $\rightarrow$ 512	512 $\rightarrow$ 1024
A_r/A_{nom}	15.4% / 21.3%	7.2% / 9.8%	3.6% / 4.7%
$p_{99.9}$	9.4% / 21.8%	4.4% / 30.3%	1.0% / 6.8%
Top-0.1% mean	10.0% / 23.5%	7.5% / 29.9%	12.9% / 19.7%
p_{max}	23.9% / 39.6%	28.8% / 44.4%	38.0% / 46.2%
Restricted L_2	0.49	0.31	0.18
Support Dice	0.79	0.87	0.93

The real-contact-area ratio shows approximately first-order reduction in nested-grid change, but its 15.4% median change from 128^2 to 256^2 demonstrates that the 128^2 result is not mesh independent. The $p_{99.9}$ change decreases to a 1.0% median at $512 \rightarrow 1024$, although its p95 remains 6.8%; it is therefore more stable than a one-pixel maximum without being treated as exact. By contrast, p_{max} does not converge over these grids: it increases monotonically in 106 of 120 $256 \rightarrow 512$ comparisons and has a 28.8% median change at that transition. One-pixel contact islands carry a median 2.4%, p95 7.3%, and maximum 11.4% of the load at 128^2 and are effectively absent by 512^2 , explaining why one-pixel extremes remain grid sensitive. The two cutoff choices lead to the same qualitative conclusions.

3.9 Computational cost and amortization

Neural-network latency was measured in FP32 on an NVIDIA RTX 3090 with batch size 1, including both branches, hard gating, and host-to-device/device-to-host transfer. Fifty warm-up iterations preceded 1,000 synchronized timed iterations. Median end-to-end latency was 9.0 ms at 128^2 and 15.8 ms at 256^2 ; including PNG input and output increased the medians to 10.3 and 19.1 ms. Model loading required 1.40 s and was measured separately.

FFT-BEM timings were obtained under Wine 6.0.3 with one CPU thread per solve and no more than eight simultaneous jobs. For the three 128^2 5,000-case load sets, wall-time medians were 1.45, 1.97, and 2.40 s at 30, 60, and 90 N, with median iteration counts of 25, 35, and 43. The fresh fixed-spectrum 256^2 cohort had a 22.0 s wall-time median. These measurements are hardware-, grid-, and workflow-specific.

Table 19 Measured online and offline computational-cost components.

Item	Result	Scope/qualification
NN batch-1 end-to-end	9.0 ms / 15.8 ms	128^2 / 256^2 ; RTX 3090; both branches and gating
NN file-I/O loop	10.3 ms / 19.1 ms	PNG read, inference, and PNG write
FFT-BEM at 128^2	1.45 / 1.97 / 2.40 s	30 / 60 / 90 N; 5,000 cases per load; wall-time medians
FFT-BEM at 256^2	22.0 s	Fresh fixed-spectrum cohort; wall-time median
Label generation	10.4-20.8 s/sample	256^2 historical campaign; range across recorded split cells
Primary 128^2 , 30 N training	83.23 h	41.65 h support branch + 41.58 h pressure branch
256^2 , 30 N training	116.37 h	58.05 h support branch + 58.32 h pressure branch
Controlled critic overhead	+415%	Same screening data/configuration; critics on versus off

For N evaluations at one prescribed condition, the accounting separates offline and online terms:

$$T_{\text{surrogate}} = T_{\text{data}} + T_{\text{train}} + Nt_{\text{infer}}, \quad N_{\text{BE}} = \frac{T_{\text{data}} + T_{\text{train}}}{t_{\text{BEM}} - t_{\text{NN}}} \quad (34)$$

The measured 30 N, 128^2 , batch-1 online comparison is approximately $1.45\text{ s}/9.0\text{ ms} \approx 160$ on the stated hardware, but this is a workflow-specific latency ratio rather than a universal solver speedup. Because offline accounting depends on dataset, retained labels, training configuration, and reuse horizon, no single break-even count is reported. The computational case is restricted to repeated many-query evaluation after offline costs are amortized.

4 Discussion

The principal contribution of this study lies in reformulating rough-surface contact pressure prediction not as a single continuous regression problem, but as a coupled problem consisting of latent pressure-field regression and contact-region segmentation, followed by their integration through hard gating. This design allows the zero-pressure condition in non-contact regions to be imposed directly at the output-structure level. In existing data-driven contact surrogates, small residual pressures may remain in nominally non-contact regions even when the mean prediction error is low. Such nonphysical pressure leakage can blur contact boundaries and bias the predicted contact area. By contrast, the proposed two-stage formulation suppresses non-contact outputs structurally and thereby yields a more physically consistent reconstruction of the contact-region geometry of the contact-pressure field itself.

The critic-on/off study clarifies the origin of this behavior. The physical constraint resides in the two-output structure, non-negative projection, and hard gating, not in adversarial training. WGAN-GP is an auxiliary feature of the archived implementation: removing the critics leaves pressure reconstruction within seed variability, whereas activating the mask critic slightly degrades support metrics in the full-data comparison.

The introduction of Focal RMSE complements this structural design by increasing the training emphasis on sparse local high-pressure peaks rather than the extensive low-pressure background. As a result, the framework improves not only geometric reconstruction of the contact region but also

recovery of pressure amplitude within contact patches. The significance of the present approach therefore lies not simply in improved image-level accuracy, but in combining structural constraint enforcement with loss reweighting tailored to the asymmetric distribution of the target data.

As detailed in Section 3, this design choice leads to measurable gains in both predictive performance and physical consistency. The proposed model achieved high contact-region reconstruction accuracy and pressure-field agreement on synthetic fractal surfaces, maintained aggregate physical consistency without explicit load constraints, and retained non-trivial solver-consistent transfer performance on measured sandpaper-finished steel surfaces. In particular, the fact that the model generalizes to surface morphologies not seen during training suggests that it approximates the input–output structure of contact mechanics with reasonable stability, rather than relying solely on memorization of specific surface patterns.

From a rough-contact mechanics perspective, the predicted mask identifies the real contact region, i.e., the subset of pixels where compressive pressure is allowed to exist, while the latent pressure field represents the pressure magnitude conditional on contact. Unlike end-to-end regression models that produce a pressure field without exposing any interpretable internal state, the proposed framework generates a discrete, separately inspectable binary contact mask together with the pressure field. This intermediate output serves multiple functions beyond the hard gating mechanism itself: it enables structural enforcement of the zero-pressure condition outside the predicted contact regions, it provides a separately verifiable contact-area output that can be compared against statistical rough-contact theories such as Persson theory, and it allows diagnostic separation of contact-localization errors from pressure-amplitude errors [29].

The measured 9.0 ms latency at 30 N and 128^2 applies to batch-1 inference on an RTX 3090 after model loading and does not by itself justify replacing a single FFT-BEM solve. Data generation and branch training are substantial offline costs, so practical benefit requires repeated many-query evaluation at a prescribed condition and reuse of the trained surrogate over a sufficiently large

workload. Hardware- and workflow-specific online ratios are reported separately from amortized cost, and no universal break-even count is asserted.

The diagnostic analysis in Section 3.5.4 further illustrates the engineering value of this decomposition. Even in the worst-performing cases, the spatial correlation structure of the pressure field was preserved ($R^2 \geq 0.991$), with errors concentrated in high-pressure amplitude regression rather than contact localization. This error structure is informative for engineering decision-making: it indicates that the framework's predictions are most reliable for identifying where contact occurs and least reliable for precise peak-pressure magnitude in geometrically complex configurations. Such diagnostic transparency supports risk-aware deployment, where the surrogate's outputs can be triaged by confidence level before downstream use.

The strong agreement in the $A_r/A_{nom} - S_{dq}$ correlations provides a mechanics-based assessment that is distinct from pixel-level image similarity. The RMS slope controls the dimensionless nominal pressure $\bar{p}/E^*S_{dq}$, and therefore has a direct influence on the real contact-area fraction in the low-contact regime. The ability of the surrogate to retain this relationship indicates that the predicted contact-support fields preserve an important statistical dependence of rough elastic contact within the evaluated surface distribution.

The framework remains a prescribed-load surrogate rather than a simultaneous complementarity solver. Reconstructing displacement and gap from predicted pressure gives median pressure-gap residuals $\eta_{pg} = 6.9 \times 10^{-3}$, 8.6×10^{-3} , and 1.1×10^{-2} at 30, 60, and 90 N, respectively, compared with FFT-BEM solver floors of 3.2×10^{-6} - 5.2×10^{-6} . The corresponding contact-gap residuals are $\eta_{g,c} = 0.0082, 0.010$, and 0.013 , so the RMS predicted-contact gap is approximately 0.8-1.3% of S_q .

On a matched 300-case 30 N comparison, hard-gated two-stage reconstruction reduces the median contact-gap residual from 1.18 for single-stage regression to 2.8×10^{-2} , an approximately 42-fold reduction. The penetration residual changes in the opposite direction, from 5.1×10^{-5} for the single-

stage model to 4.0×10^{-4} for the two-stage model. Reporting both directions is necessary: hard gating strongly suppresses pressure on open-gap pixels but does not enforce complete gap-pressure complementarity.

The fixed-spectrum study changes the interpretation of extreme-pressure accuracy. Agreement in p_{max} is meaningful only relative to the discrete grid used to generate the FFT-BEM label, because the one-pixel maximum remains strongly grid sensitive. The more stable $p_{99,9}$ statistic is therefore promoted for cross-grid extreme-pressure reporting, while its residual p95 change is retained as a numerical limitation. Neither statistic is presented as proof of a continuum point peak.

Separate checkpoints were trained at 30, 60, and 90 N because the applied load was not included as a continuous network input. The present study therefore establishes reconstruction performance only at these prescribed loads and does not establish interpolation to intermediate loads or extrapolation to loads substantially above 90 N. Changes in load involve the growth and coalescence of existing patches, the entry of new asperities into contact, boundary motion, and local pressure redistribution, so a checkpoint trained at one load cannot generate the contact-support evolution associated with a different load.

Despite this limitation, a load-specific surrogate remains useful for repeated evaluation of many surface realizations under a prescribed preload. Such problems occur in surface-design screening, statistical rough-contact analysis, manufacturing-variability assessment, and Monte Carlo uncertainty quantification. The main contribution of the present work is the structural combination of explicit contact-support prediction and pressure regression, rather than universal load generalization.

The current outputs also do not provide direct contact-stiffness prediction. Normal stiffness requires a load-approach relationship, whereas the network predicts only the support and pressure fields at a prescribed load. A future extension could condition the network on nominal pressure and add approach or mean separation as an additional output. More broadly, an integrated conditional model could incorporate the applied load, effective modulus, and surface statistics as continuous inputs.

The synthetic rough-surface family comprises approximately Gaussian RMD-generated multiscale topographies, alongside the separate smooth spheroidal auxiliary family. Deliberately non-Gaussian rough-surface families were not included, and systematic generalization to highly skewed, plateau-like, grooved, or engineered surface distributions remains unverified. Similarly, the present mechanical assumptions exclude adhesion, plasticity, frictional tangential loading, finite-thickness effects, and surface evolution caused by wear, as well as frictional heating, material nonlinearity, surface films, and multi-body contact [45]. The present model does not predict friction coefficients, wear rates, lubrication-film behavior, or sliding responses; rather, it reconstructs the normal contact-pressure field and the real contact region that can serve as an input descriptor for downstream tribological analyses.

Finally, the measured-topography evaluation used experimentally measured surface geometry but FFT-BEM numerical references for contact support and pressure. It therefore represents solver-consistent geometric transfer without fine-tuning, not direct experimental pressure validation. The present framework also does not address predictive uncertainty; confidence estimation, out-of-distribution detection, and identification of high-risk cases remain unresolved issues for practical deployment. Future work should combine load-conditioned modeling with controlled non-Gaussian surface families and independently measured contact-area or pressure data. Extension to multiphysics modeling, including elastic-plastic contact, frictional and thermal effects, and electrical contact resistance, will also be necessary for closer connection to realistic tribological systems. In parallel, the incorporation of uncertainty-quantification strategies, such as ensembles, Bayesian approximations, or conformal prediction [46-48], would improve the engineering reliability of the framework.

Overall, this study shows that a deep-learning framework in which a key contact-mechanics constraint is embedded in the output structure, rather than imposed only through the loss function, can achieve both high predictive accuracy and physically consistent aggregate behavior. By

producing an explicit binary contact mask and coupling it with structural enforcement of the zero-pressure condition, the framework offers a complementary surrogate that accelerates rough-contact pressure reconstruction within the validated regime. This design also provides a foundation for future extensions toward continuous conditioning, multiphysics modeling, and uncertainty-aware prediction.

5 Conclusions

This study developed a physics-constrained two-stage deep-learning framework for reconstructing real contact support and local pressure fields from rough-surface topography. Independently parameterized support and pressure branches are combined by non-negative projection and hard gating, which imposes exactly zero pressure outside the predicted support, while Focal RMSE emphasizes sparse high-pressure pixels. A controlled 256×256 , 30 N critic-on/off ablation showed no material reconstruction improvement from adversarial training; the dominant gain arises from task decomposition, hard gating, and supervised branch losses. Union-support R^2 and boundary F1 replace full-field correlation as primary local metrics, and $p_{99.9}$ replaces the one-pixel maximum as the primary extreme-pressure statistic. Any retained p_{max} agreement is explicitly limited to the discrete FFT-BEM reference. Within prescribed-load, frictionless linear elasticity, the resulting surrogate supports rapid repeated evaluation after its offline data-generation and training costs are amortized.

At the prescribed loads of 30, 60, and 90 N, the load-specific models achieved high agreement with FFT-BEM references in both contact-support geometry and pressure amplitude. The predicted fields preserved aggregate load consistency, the increase in ensemble-mean contact area across the three evaluated loads, and the strong negative dependence of real contact area on the RMS surface slope. The auxiliary axisymmetric ellipsoidal-contact test assessed smooth circular single-patch reconstruction. Against independent full-profile theory, the mean full-field pressure L2 errors over

500 test cases are 0.9967% for FFT-BEM and 1.1697% for the network, within the stated elastic half-space assumptions.

Application to experimentally measured sandpaper-finished steel topographies without fine-tuning showed partial transfer to more irregular engineering surface geometries when evaluated against FFT-BEM numerical references. Because the pressure fields were not measured experimentally, this result should be interpreted as solver-consistent transfer rather than direct experimental validation.

Within its prescribed-load, frictionless linear-elastic scope, the proposed framework provides a rapid physics-constrained surrogate for generating FFT-BEM-consistent contact-support and pressure fields. Extension to continuously varying loads, stiffness prediction, controlled non-Gaussian surfaces, and independent experimental pressure validation remains for future work.

Funding

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2026-25472952).

Author contributions

Ilkwang Jang: Conceptualization, Methodology, Writing - original draft. **Kisung Kim:** Software, Data curation, Investigation, Validation, Writing - original draft. **Hyeonggeun Jo:** Investigation, Data curation. **Yong Hoon Jang:** Supervision, Conceptualization, Project administration, Writing - review & editing.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

- [1] Greenwood J A, Williamson J B P. Contact of nominally flat surfaces. *Proc R Soc Lond A* **295**(1442): 300–319 (1966)
- [2] Bush A W, Gibson R D, Thomas T R. The elastic contact of a rough surface. *Wear* **35**(1): 87–111 (1975)
- [3] Nayak P R. Random process model of rough surfaces. *J Lubr Technol* **93**(3): 398–407 (1971)
- [4] Persson B N J. Contact mechanics for randomly rough surfaces. *Surf Sci Rep* **61**(4): 201–227 (2006)
- [5] Bowden F P, Tabor D. The area of contact between stationary and moving surfaces. *Proc R Soc Lond A* **169**(938): 391–413 (1939)
- [6] Barber J R. Bounds on the electrical resistance between contacting elastic rough bodies. *Proc R Soc A* **459**(2029): 53–66 (2003)
- [7] Bahrami M, Culham J R, Yovanovich M M, Schneider G E. Thermal contact resistance of nonconforming rough surfaces, Part 1: Contact mechanics model. *J Thermophys Heat Transf* **18**(2): 209–217 (2004)
- [8] Jang Y H, Barber J R. Effect of contact statistics on electrical contact resistance. *J Appl Phys* **94**(11): 7215–7221 (2003)
- [9] Li L T, Liang X M, Xing Y Z, Yan D, Wang G F. Measurement of real contact area for rough metal surfaces and the distinction of contribution from elasticity and plasticity. *J Tribol* **143**(7): 071501 (2021)
- [10] Woo K L, Thomas T R. Contact of rough surfaces: A review of experimental work. *Wear* **58**(2): 331–340 (1980)

- [11] Terwisscha-Dekker H, Brouwer A M, Weber B, Bonn D. Elastic contact between rough surfaces: Bridging the gap between theory and experiment. *J Mech Phys Solids* **188**: 105676 (2024)
- [12] Stanley H M, Kato T. An FFT-based method for rough surface contact. *J Tribol* **119**(3): 481–485 (1997)
- [13] Polonsky I A, Keer L M. A numerical method for solving rough contact problems based on the multi-level multi-summation and conjugate gradient techniques. *Wear* **231**(2): 206–219 (1999)
- [14] Campañá C, Müser M H. Practical Green's function approach to the simulation of elastic semi-infinite solids. *Phys Rev B* **74**(7): 075420 (2006)
- [15] Xu Y, Jackson R L. Boundary element method (BEM) applied to the rough surface contact vs. BEM in computational mechanics. *Friction* **7**(4): 359–371 (2019)
- [16] Müser M H, Dapp W B, Bugnicourt R, Sainsot P, Lesaffre N, Lubrecht T A, Persson B N J, Harris K, Bennett A, Schulze K, Rohde S, Ifju P, Sawyer W G, Angelini T, Ashtari Esfahani H, Kadkhodaei M, Akbarzadeh S, Wu J J, Vorlauffer G, Vernes A, Solhjoo S, Vakis A I, Jackson R L, Xu Y, Streater J, Rostami A, Dini D, Medina S, Carbone G, Bottiglione F, Afferrante L, Monti J, Pastewka L, Robbins M O, Greenwood J A. Meeting the contact-mechanics challenge. *Tribol Lett* **65**(4): 118 (2017)
- [17] Rapetto M P, Almqvist A, Larsson R, Lugt P M. On the influence of surface roughness on real area of contact in normal, dry, friction free, rough contact by using a neural network. *Wear* **266**(5–6): 592–595 (2009)
- [18] Kalliorinne K, Larsson R, Pérez-Ràfols F, Liwicki M, Almqvist A. Artificial neural network architecture for prediction of contact mechanical response. *Front Mech Eng* **6**: 579825 (2021)
- [19] Suman S, Prajapati D K. Predictive modeling of real contact area on rough surfaces using deep artificial neural network. *J Tribol* **147**(11): 111501 (2025)
- [20] Jang I, Jang Y H. Prediction of contact distribution on rough surfaces using deep learning algorithms. *Facta Univ Ser Mech Eng* **23**(4): 757–785 (2025)

- [21] Lin T Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2980–2988.
- [22] Yakovenko A A, Goryacheva I G. Analysis of the discrete contact characteristics based on the Greenwood-Williamson model and the localization principle. *Friction* **12**: 1042–1056 (2024)
- [23] Meng X, Bachmann M, Yang F, Rethmeier M. Porosity prediction in laser beam welding with a multimodal physics-informed machine learning framework. *Adv Eng Inform* **74**: 104611 (2026)
- [24] Sun Y, Sengupta U, Juniper M. Physics-informed deep learning for simultaneous surrogate modeling and PDE-constrained optimization of an airfoil geometry. *Comput Methods Appl Mech Eng* **411**: 116042 (2023)
- [25] Hu R, Wei P, Du X, Zhou J, Liu H, Liu G, Zhu C. Investigation of loaded contact characteristics of planetary roller screw mechanism based on influence coefficient method and machine learning. *Adv Eng Inform* **58**: 102146 (2023)
- [26] Johnson K L. Contact Mechanics. Cambridge (UK): Cambridge University Press, 1985.
- [27] Wriggers P. Computational Contact Mechanics. Berlin/Heidelberg (Germany): Springer, 2006.
- [28] Chen H, Constante Flores G E, Li C. Physics-informed neural networks with hard linear equality constraints. *Comput Chem Eng* **189**: 108764 (2024)
- [29] Beucler T, Pritchard M, Rasp S, Ott J, Baldi P, Gentine P. Enforcing analytic constraints in neural networks emulating physical systems. *Phys Rev Lett* **126**(9): 098302 (2021)
- [30] Paggi M, Barber J R. Contact conductance of rough surfaces composed of modified RMD patches. *Int J Heat Mass Transf* **54**(21–22): 4664–4672 (2011)
- [31] Pérez-Ràfols F, Almqvist A. Generating randomly rough surfaces with given height probability distribution and power spectrum. *Tribol Int* **131**: 591–604 (2019)
- [32] Ciavarella M, Demelio G, Barber J R, Jang Y H. Linear elastic contact of the Weierstrass profile. *Proc R Soc A Math Phys Eng Sci* **456**(1994): 387–405 (2000)

- [33] Laursen T A. Computational Contact and Impact Mechanics: Fundamentals of Modeling Interfacial Phenomena in Nonlinear Finite Element Analysis. Berlin/Heidelberg (Germany): Springer, 2003.
- [34] Isola P, Zhu J Y, Zhou T, Efros A A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1125–1134.
- [35] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol 9351. Cham: Springer, 2015: 234–241.
- [36] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A C. Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017: 5767–5777.
- [37] Sudre C H, Li W, Vercauteren T, Ourselin S, Cardoso M J. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Lecture Notes in Computer Science, Vol 10553. Cham: Springer, 2017: 240–248.
- [38] Persson B N J. Relation between interfacial separation and load: A general theory of contact mechanics. *Phys Rev Lett* **99**(12): 125502 (2007)
- [39] Persson B N J. Theory of rubber friction and contact mechanics. *J Chem Phys* **115**(8): 3840–3861 (2001)
- [40] Prodanov N, Dapp W B, Müser M H. On the contact area and mean gap of rough, elastic contacts: Dimensional analysis, numerical corrections, and reference data. *Tribol Lett* **53**(2): 433–448 (2014)
- [41] Hyun S, Pei L, Molinari J F, Robbins M O. Finite-element analysis of contact between elastic self-affine surfaces. *Phys Rev E* **70**(2): 026117 (2004)

- [42] Zhang X, Zhang Y, Jin Z. A semi-analytical approach to the elastic loading behaviour of rough surfaces. *Friction* **8**(5): 970–981 (2020)
- [43] Lyashenko I A. 3D surfaces of rough steel indenters. <https://doi.org/10.57703/ce-9epgb> (2024)
- [44] Lyashenko I A, Pham T H, Popov V L. Friction in adhesive contact between a rough hard indenter and smooth soft elastomer layer. *Machines* **12**(11): 754 (2024)
- [45] Ghaednia H, Wang X, Saha S, Xu Y, Sharma A, Jackson R L. A review of elastic–plastic contact mechanics. *Appl Mech Rev* **69**(6): 060804 (2017)
- [46] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017: 6402–6413.
- [47] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: *Proceedings of the 33rd International Conference on Machine Learning*, 2016: 1050–1059.
- [48] Romano Y, Patterson E, Candès E J. Conformalized quantile regression. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019: 3543–3553.
- [49] Sneddon I N. The relation between load and penetration in the axisymmetric Boussinesq problem for a punch of arbitrary profile. *Int J Eng Sci* **3**(1): 47–57 (1965)
- [50] Popov V L, Heß M, Willert E. Normal contact without adhesion. in: *handbook of contact mechanics: Exact solutions of axisymmetric contact problems*. Berlin/Heidelberg: Springer, 2019: 5–66.

Author biography



Yong Hoon Jang earned his Ph.D. in 1999 from the University of Michigan. He is currently a full professor in the School of Mechanical Engineering at Yonsei University. His research interests include friction and contact mechanics, tribological interfaces, and physics-based modeling of mechanical systems.

Just Accepted