Supplementary Materials for

Finger-level tactile intelligence toward robotic dexterous operation via afferent encoding and neuronal multiplexing

Jin Ma[1,5], Mengqi Fang[1,5], Mengli Sui[2,5], Jiayi Ding[1], Xinshi Yang[1], Jie Cao[3], Zhijie Zhang[1], Shuo Wang[1], Long Cheng[1], Ming Wang[3*], Xinge Yu[4*], Yu Wang[1*]

[5]These authors contributed equally to this works.

[*]Corresponding author. E-mail: wang_ming@fudan.edu.cn; xingeyu@cityu.edu.hk; yu.wang@ia.ac.cn

This file includes:

Supplementary Text 1 to 3

Supplementary Table 1 to 3

Supplementary Figures 1 to 10

Legends for Supplementary Videos 1 to 4

Supplementary Text:

Supplementary Text 1: Design and fabrication of the visuotactile system

(1) Design of the visuotactile system HydroPalm

The visuotactile system, named HydroPalm, is designed for dual-environment operation in both terrestrial and underwater conditions. The exploded view of the HydroPalm assembly is illustrated in Figure S1. Structurally, HydroPalm comprises four functional modules: a gel layer module, an illumination module, an imaging module, and a waterproof module. The gel layer module consists of a hemispherical acrylic dome coated with an optically transparent elastic gel layer, within which retroreflective markers are embedded in a concentric lattice pattern. This configuration enables high-resolution deformation tracking for tactile shape reconstruction. The illumination module integrates a ring-shaped LED strip and an aperture to generate uniform and diffused lighting across the dome surface, minimizing specular reflection and ensuring consistent visual feedback in turbid water. The imaging module includes a binocular camera pair for stereoscopic capture and an STM32 microcontroller responsible for tactile image acquisition and neuromorphic spike-train encoding. This module translates contact-induced deformation into spike-based tactile representations. The waterproof module comprises a sealed aluminum enclosure with a quartz glass window, top and bottom covers, and O-ring interfaces, providing a fully dry environment for the imaging and illumination components during submersion. Assembly procedures for each component are detailed in Supplementary Text 1 (2-5), and fabrication parameters are summarized in Supplementary Table 2.

(2) Fabrication of the gel layer module

The gel layer module consists of a hemispherical acrylic dome covered by a multilayer optically transparent gel layer. The gel layer comprises three sublayers, a protective coating, a marker pattern, and an elastomer base, which are functionally analogous to the epidermis, tactile corpuscles, and dermis of human skin, respectively. The elastomer base is fabricated using a 5-degree addition-cure silicone gel, which is cast in a custom-designed 3D-printed mold and cured at 65℃ for 6 hours under ambient pressure to form a hemispherical substrate. The marker pattern is applied using transparent water-transfer decals printed in a petal-like layout, corresponding to a planar projection of the spherical surface. Each petal-shaped segment is trimmed along its edges and carefully adhered onto the surface of the cured elastomer base, resulting in a concentric and uniformly spaced distribution of retroreflective markers over the hemispherical surface. The protective coating is then formed by pouring a 0-degree addition-cure silicone gel over the marker layer. A centrifugation step is employed to ensure uniform coverage across the hemisphere, followed by curing at 35 ℃ for 12 h under ambient pressure to form a transparent protective layer approximately 0.2 mm thick. The acrylic dome is precision-machined from transparent acrylic and mechanically joined to the top cover via a threaded interface. The gel layer is subsequently bonded to the dome surface using a thin optical adhesive, ensuring full encapsulation and optical continuity.

(3) Fabrication of illumination module

The illumination module consists of a ring-shaped LED strip and an aperture diaphragm for light collimation and glare suppression. The LED strip incorporates six evenly spaced white-light LEDs, each providing a luminous intensity of approximately 15-20 cd at a driving voltage of 4.5

V. The LEDs are arranged in a circular configuration to generate uniform illumination over the gel surface. A black resin aperture with a width of 2.0 mm is mounted above the LED strip to confine the illumination angle and prevent direct reflection toward the imaging module. The aperture ring is fabricated using precision 3D printing and fixed onto the LED housing with adhesive epoxy. This configuration ensures a homogeneous and diffused illumination field across the hemispherical gel surface while minimizing specular artifacts under underwater conditions.

(4) Fabrication of imaging module

The imaging module comprises a binocular camera and an STM32 microcontroller. The binocular camera provides synchronized stereo images at a resolution of 1280×720 pixels and a frame rate of 60 fps, which is powered at 5 V. The camera is mounted at a height of 40.2 mm above the apex of the gel layer, ensuring complete visibility of all embedded markers within its field of view. Prior to deployment, the binocular camera is calibrated using a standard checkerboard pattern to obtain intrinsic parameters under both air and underwater conditions. The calibration process is performed following the pinhole camera model, and the resulting intrinsic and distortion coefficients are summarized in Supplementary Table 3. The STM32 microcontroller is installed inside the waterproof housing and directly connected to the camera. Both components operate on a 5 V DC supply and interface via a 5-pin waterproof connector, which transmits power (+/−) and data (D+/D−) signals. Data transfer conforms to the USB communication standard, enabling real-time acquisition of stereo image sequences for spiking signal generation.

(5) Fabrication of waterproof module

The waterproof module consists of a top cover, quartz glass, waterproof housing, a bottom cover, and a waterproof connector. The upper interface is sealed by compressing the quartz glass against a fluoro-rubber (F-O-ring) gasket, ensuring watertight isolation. The top cover contains a precision-machined groove that holds the quartz glass in position and is thread-fastened onto the waterproof housing to provide mechanical rigidity and long-term sealing stability. At the lower interface, the bottom cover is coupled to the waterproof connector through a locking nut and fixed with hexagonal screws to prevent mechanical loosening during submersion. The waterproof housing encloses an internal cylindrical cavity (ϕ 40 mm, length 80 mm), which provides a dry and pressure-resistant compartment for the electronic components, including the imaging and illumination modules. All joints were further reinforced with silicone sealant to prevent micro-leakage during prolonged underwater operation.

Supplementary Text 2: Force prediction network architecture

The objective of the force prediction task is to regress a three-axis force vector $(F_\mathrm{X}, F_\mathrm{Y}, F_\mathrm{Z})$ from a spatiotemporal sequence of features extracted from the sensor's deformation. To this end, we design a deep neural network based on the Transformer encoder architecture. The principal advantage of this network is its use of a multi-head self-attention mechanism, which enables it to effectively learn the complex spatial dependencies between deformations at different regions of the sensor surface, thereby allowing for a high-fidelity reconstruction of the contact forces. The network consists of three primary components: a positional encoding layer, a multi-head attention module, and a feed-forward network layer.

(1) Positional encoding

The self-attention mechanism is inherently permutation-invariant; it does not process the order of elements in a sequence. In the context of force perception, however, the spatial location of each feature vector—corresponding to a specific region on the sensor surface—is critically important. To inject this spatial information into the model, positional encodings are added to the input features before they are passed to the attention module. We employ the standard sinusoidal positional encoding scheme using sine and cosine functions of different frequencies:

$$PE_{(\mathrm{pos,2i})} = sin(pos/10000^{\frac{2i}{d_{\mathrm{model}}}})$$

$$PE_{(\mathrm{pos,2i+1})} = cos(pos/10000^{2i/d_{\mathrm{model}}}),$$

where pos is the position of a feature in the sequence, i is the dimension index of the encoding, and d_{model} is the dimensionality of the feature vectors. This encoding provides the model with unique information regarding the relative or absolute position of each input feature.

(2) Multi-head self-attention module

This module is the computational core of the network, allowing it to weigh the importance of different parts of the input sequence. Rather than performing a single attention calculation, it projects the features into multiple "heads" in parallel, allowing attention to be performed independently in different representation subspaces.

The computation for a single attention head follows the Scaled Dot-Product Attention mechanism. The Query (Q), Key (K), and Value (V) matrices are derived from linear projections of the input feature sequence. The attention output is computed as:

$$\mathrm{Attention}(Q,K,V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}})V,$$

where d_{k} is the dimension of the key vectors. The scaling factor $\sqrt{d_{\mathrm{k}}}$ is used to counteract the effect of large dot products, which can saturate the softmax function and lead to vanishing gradients.

In Multi-Head Attention, we compute independent attention heads in parallel:

$$\mathrm{head}_{\mathrm{i}} = \mathrm{Attention}(XW_{\mathrm{i}}^{\mathrm{Q}}, XW_{\mathrm{i}}^{\mathrm{K}}, XW_{\mathrm{i}}^{\mathrm{V}}) \ ,$$

where X is the input feature sequence and $W_{\mathrm{i}}^{\mathrm{Q}}$, $W_{\mathrm{i}}^{\mathrm{K}}$, $W_{\mathrm{i}}^{\mathrm{V}}$ are the learnable projection matrices for the i -th head. This structure allows each head to attend to different feature subsets from different positions in the input sequence. In the context of force perception, for example, one head might specialize in identifying patterns of normal pressure in the central contact region, while another might learn the relationship between shear forces at the periphery and pressure at the center.

The outputs of all heads are then concatenated and passed through a final linear projection to integrate the information learned from the different subspaces:

$$\text{MultiHead}(X)\text{=Concat(head}_1\text{,...head}_\text{h})W^\text{O},$$

where W^O is the output projection matrix.

(3) Feed-forward network and output

Following the attention module, the output for each position is processed by an identical Position-wise Feed-Forward Network (FFN). This network consists of two linear layers with a ReLU activation:

$$\text{FFN}(x) = max(\,0, xW_1 + b_1)W_2 + b_2.$$

The FFN applies a further non-linear transformation to the features aggregated by the attention mechanism. Finally, the processed feature sequence is globally average-pooled and mapped by a final linear layer (the regression head) to a 3-dimensional output, representing the predicted force vector $(F_\text{X}, F_\text{Y}, F_\text{Z})$. Residual connections and layer normalization are employed throughout the network to stabilize the training process.

Supplementary Text 3: Multi-modal texture recognition network architecture

The objective of the texture recognition task is to identify object surfaces by processing two complementary streams of information. To this end, we design and implement a parallel, multi-modal deep neural network. The architecture comprises two specialized feature extraction pathways: one for processing Low-pass (LP) information and another for processing High-pass (HP) information. The features extracted from both pathways are ultimately fused to produce a highly robust classification decision regarding the surface texture. The network is composed of three primary components: an LP information feature pathway, an HP information feature pathway, and a multi-modal fusion and classification head.

(1) LP information feature pathway

This pathway is designed to extract complex dynamic patterns from sequential data. It consists of a parallel Long Short-Term Memory (LSTM) network and a 1D Convolutional Neural Network (CNN), which process the input sequence concurrently.

LSTM module: An LSTM is a type of Recurrent Neural Network (RNN) specialized in learning long-range dependencies in sequential data. This is critical for texture recognition, where surface characteristics often unfold over the course of a dynamic interaction. The core of the LSTM is its cell state and three gating mechanisms: the forget gate, the input gate, and the output gate. At a given time step t, their operations are defined as:

$$f_{\mathrm{t}} = \sigma(W_{\mathrm{f}} \cdot [h_{\mathrm{t}-1}, x_{\mathrm{t}}] + b_{\mathrm{f}})$$

$$i_{\mathrm{t}} = \sigma(W_{\mathrm{i}} \cdot [h_{\mathrm{t}-1}, x_{\mathrm{t}}] + b_{\mathrm{i}})$$

$$o_{\mathrm{t}} = \sigma(W_{\mathrm{o}} \cdot [h_{\mathrm{t}-1}, x_{\mathrm{t}}] + b_{\mathrm{o}}),$$

where x_{t} is the current input, $h_{\mathrm{t}-1}$ is the hidden state from the previous time step, σ is the sigmoid activation function, and W and b are learnable weights and biases. These gates collectively determine what information is discarded from the cell state, what new information is stored, and how the hidden state h_{t} is outputted.

1D CNN module: In parallel with the LSTM, a 1D CNN is employed to extract local, translation-invariant motifs from the sequence. The convolutional kernels slide across the 1D sequence to detect specific, recurring patterns within short temporal windows, which may correspond to fundamental physical features of the texture. The 1D convolution operation is defined as:

$$(x * k)[n] = \sum_{\mathrm{m}=0}^{\mathrm{M}-1} x[n-m]k[m] \ ,$$

where x is the input sequence and k is a kernel of length M. By stacking multiple convolutional layers, the network can learn a hierarchical representation of features, from low-level patterns (e.g., individual spikes) to higher-order combinations.

(2) HP information feature pathway

This pathway is responsible for extracting high-level representations from 2D data. We employ a pre-trained EfficientNet-B5 model as the feature extractor. The EfficientNet family of models is renowned for its exceptional balance between accuracy and computational efficiency, which is achieved through a compound scaling method that uniformly scales network depth, width, and resolution. Its core building block is the MBConv (Mobile Inverted Bottleneck Convolution), which incorporates depthwise separable convolutions and Squeeze-and-Excitation optimization to learn rich feature representations efficiently. The input data is processed through the convolutional base of the EfficientNet-B5 to produce a high-dimensional feature map, which is then globally average-pooled to form a compact feature vector.

(3) Modal fusion strategy for texture recognition task

To integrate the LP data and HP textural (vibration) data, we implement a conditional, asymmetric fusion strategy. This strategy designates the HP modality as the primary source for classification, while the LP modality provides supplementary information to resolve predictive ambiguities during the inference phase.

The fusion mechanism is triggered if and only if two specific conditions are met: (1) a discrepancy exists between the predicted classes from the two modalities, (2) the prediction confidence of the force modality falls below an adaptive threshold, denoted as τ. This threshold represents the minimum confidence required for the force-based prediction to be considered reliable.

When these conditions are satisfied, a dynamic weighted harmonic mean is employed to compute the final fused probability vector, P_{fuesd}. For each class i, the fused probability is calculated as:

$$P_{\mathrm{fuesd,i}} = \frac{w_{\mathrm{HP}}+w_{\mathrm{LP}}}{\frac{w_{\mathrm{HP}}}{P_{\mathrm{HP,i}}}+\frac{w_{\mathrm{LP}}}{P_{\mathrm{LP,i}}}},$$

where $P_{\mathrm{HP,i}}$ and $P_{\mathrm{LP,i}}$ are the predicted probabilities for class i from the HP and LP modalities, respectively. The weights w_{HP} and w_{LP} represent the contribution of each modality to the fusion, constrained by $w_{\mathrm{HP}} + w_{\mathrm{LP}} = 1$. The final classification is then determined by the class with the maximum probability in the P_{fuesd} vector.

If the triggering conditions are not met, the prediction from the primary high-frequency modality is directly used as the final output. The fusion weights (w_{HP} and w_{LP}, initialized at 0.7 and 0.3, respectively) and the confidence threshold τ are treated as learnable parameters and are optimized jointly with the network parameters throughout the training process to maximize overall classification accuracy.

Supplementary table:

Supplementary Table 1: Comparison of various robotic tactile sensors' resolutions and capabilities.

No.	Reference	Sensor Type	Spatial resolution (mm)	Texture roughness (µm)	Slip prediction	Underwater
1	[17]	High spatiotemporal resolution texture sensor	NaN	15	No	No
2	[20]	Vision-tactile fusion (RGB-D camera + XELA tactile)	NaN	50	No	No
3	[25]	Stretchable pressure sensor array with low crosstalk	5	NaN	No	No
4	[28]	Vision-based tactile sensor	0.4	NaN	No	No
5	[59]	Sparsely distributed tactile units	0.73	NaN	No	No
6	[24]	Polydimethylsiloxane (PDMS) sensor array	1	NaN	No	No
7	[49]	GelStereo tactile sensor array	1	NaN	No	No
8	[52]	Capacitive sensor	2.2	NaN	No	No
9	[53].	Doped polysilicon	1	NaN	No	No
10	[57]	Barometer pressure sensor chip (BPSC)	NaN	3000	No	No
11	[60]	Artificial neural tactile sensing system	NaN	520	No	No
12	Ours.	HydroPalm	0.8	0.258	Yes	Yes

Supplemental Table 2: Fabrication parameters of the HydroPalm.

Module	Components	Dimensional parameters	Materials and Manufacturing Methods
Gel layer module	Protective coating	ϕ54 mm × L 0.3 mm	Psycho Paint Smooth On, pouring
	Marker pattern	ϕ50 mm	Water-transfer paper, water transfer printing
	Elastomer base	ϕ54 mm × L 2.0 mm	5-degree addition-cure silicone gel, mold forming
	Acrylic dome	ϕ50 mm	Acrylic, CNC machining
Illumination module	LED strip	105 mm × 3 mm	LEDs, product
	Aperture	ϕ28 mm × L 1.5 mm	Black resin, 3D printing
Imaging module	Binocular camera	19 mm × 12 mm × 5 mm	Camera, product
	STM32	42 mm × 35 mm × 5 mm	Microcontroller, product
Waterproof module	Top cover	ϕ54 mm × L 16.5 mm	Aluminum alloy, CNC machining
	Quartz glass	ϕ41 mm × L 2 mm	Quartz glass, CNC machining
	Waterproof housing	ϕ54 mm × L 96 mm	Aluminum alloy, CNC machining
	Bottom cover	ϕ54 mm × L 5 mm	Aluminum alloy, CNC machining
	Waterproof connector	ϕ22 mm × L 40 mm	Connector, product

Supplemental Table 3: Parameters of the binocular camera in air and underwater environments.

Medium	**Air**		**Underwater**	
Camera ID	**Left**	**Right**	**Left**	**Right**
Focal length $(\boldsymbol{f}_{\mathrm{x}}, \boldsymbol{f}_{\mathrm{y}})$ (px)	644.49 644.39	642.60 642.61	1031.1 1033.4	1031.8 1030.8
Principal point $(\boldsymbol{c}_{\mathrm{x}}, \boldsymbol{c}_{\mathrm{y}})$ (px)	597.31 380.10	584.52 365.09	594.71 383.97	589.69 379.96
Skew (px)	0.00	0.00	0.003	0.001
Distortion coefficients $(\boldsymbol{k}_1, \boldsymbol{k}_2, \boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{k}_3)$	-0.0013 0.0537 -0.0008 -0.0016 -0.0819	-0.0035 0.0622 -0.0011 0.0007 -0.0915	0.5343 3.7701 0.00 0.00 -10.860	0.5916 2.9142 0.00 0.00 -7.4617
Rotation Matrix R	$\begin{bmatrix} 0.9998 & 0.0004 & 0.0054 \\ -0.0004 & 0.9999 & -0.0003 \\ -0.0054 & 0.0003 & 0.9998 \end{bmatrix}$		$\begin{bmatrix} 0.9995 & -0.0012 & 0.0310 \\ 0.0013 & 0.9999 & 0.0031 \\ -0.0310 & -0.0030 & 0.9995 \end{bmatrix}$	
Translation Vector $\boldsymbol{T}$ (mm)	$[-7.5093 \quad -0.0857 \quad -0.0566]$		$[-7.8859 \quad -0.1181 \quad -0.0276]$	

Supplementary figure:

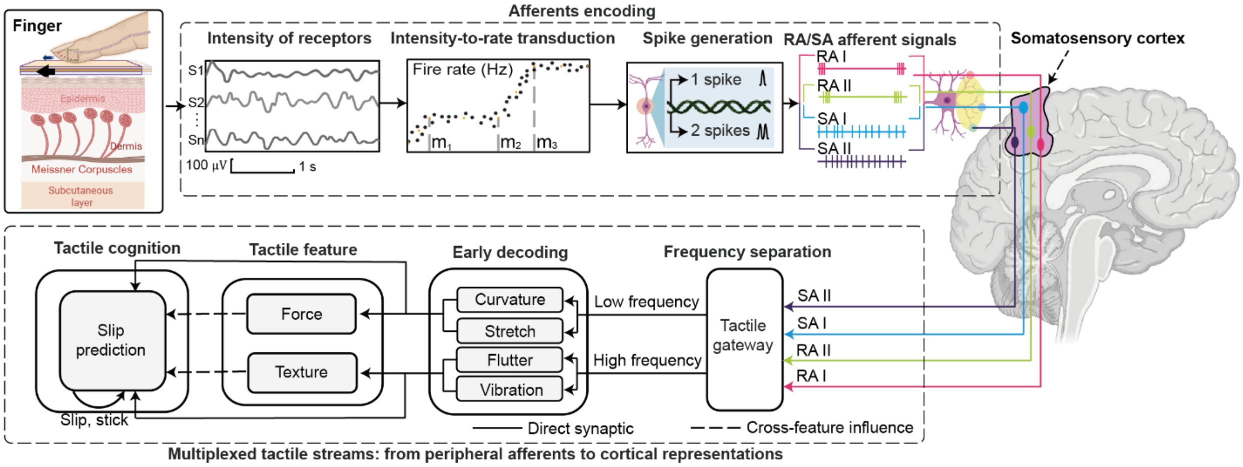


Fig. S1. Multiplexed tactile streams: from peripheral afferents to cortical representations. Stimulus intensity at the finger is converted into spike rate, generating single- or double-spike events in RA I, RA II, SA I, and SA II afferents. These multiplexed afferent streams are segregated into high-frequency and low-frequency pathways in the somatosensory cortex, where early decoding yields flutter, vibration, curvature, and stretch representations. These are integrated into force and texture features that converge on slip prediction, which is dynamically updated through direct and cross-feature influences.

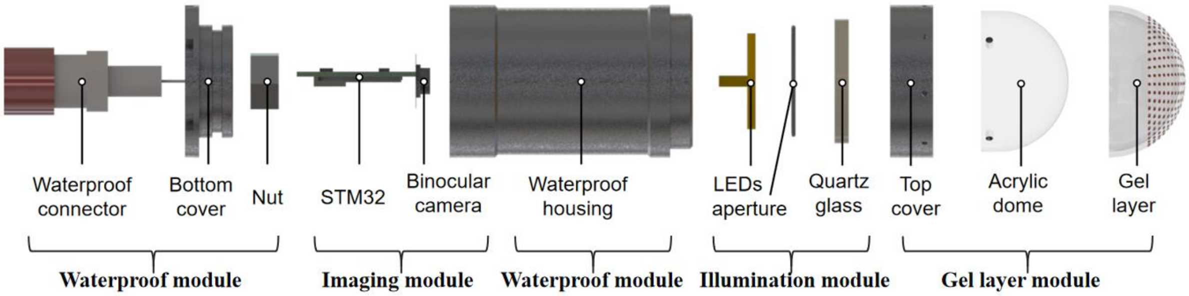


Fig. S2. Exploded view of the visuotactile sensing system. The waterproof module provides a hermetically sealed enclosure for amphibious operation and consists of a waterproof connector, a bottom cover, nuts, and the main waterproof housing. The imaging module is responsible for data acquisition and processing, containing an STM32 microcontroller and a binocular camera. The illumination module, comprising LEDs and an aperture, ensures uniform internal lighting of the contact interface. The gel layer module is the primary mechanotransduction component; it includes a top cover, a protective quartz glass window, a transparent acrylic dome, and the soft gel layer in which a marker array is embedded. The working principle is as follows: when the sensor makes contact with an object, the gel layer deforms, resulting in the displacement of the markers. The binocular camera captures these displacements, which are then processed by the STM32 to generate the raw tactile data stream.

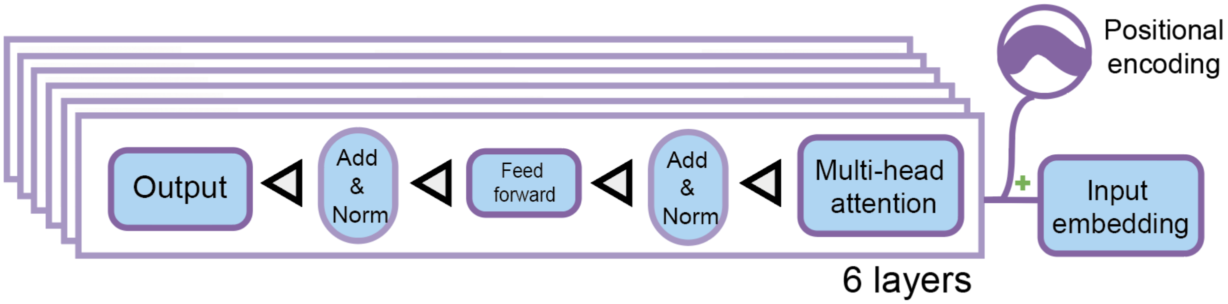


Fig. S3. Transformer network architecture for force prediction. The architecture consists of an input embedding layer, followed by 6 identical layers, and an output layer. Each of the 6 layers contains two main sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Both sub-layers are followed by a residual connection and layer normalization (Add & Norm). Positional encodings are added to the input embeddings to incorporate the order of the low-frequency force information. The output layer processes the final representation to predict force values.

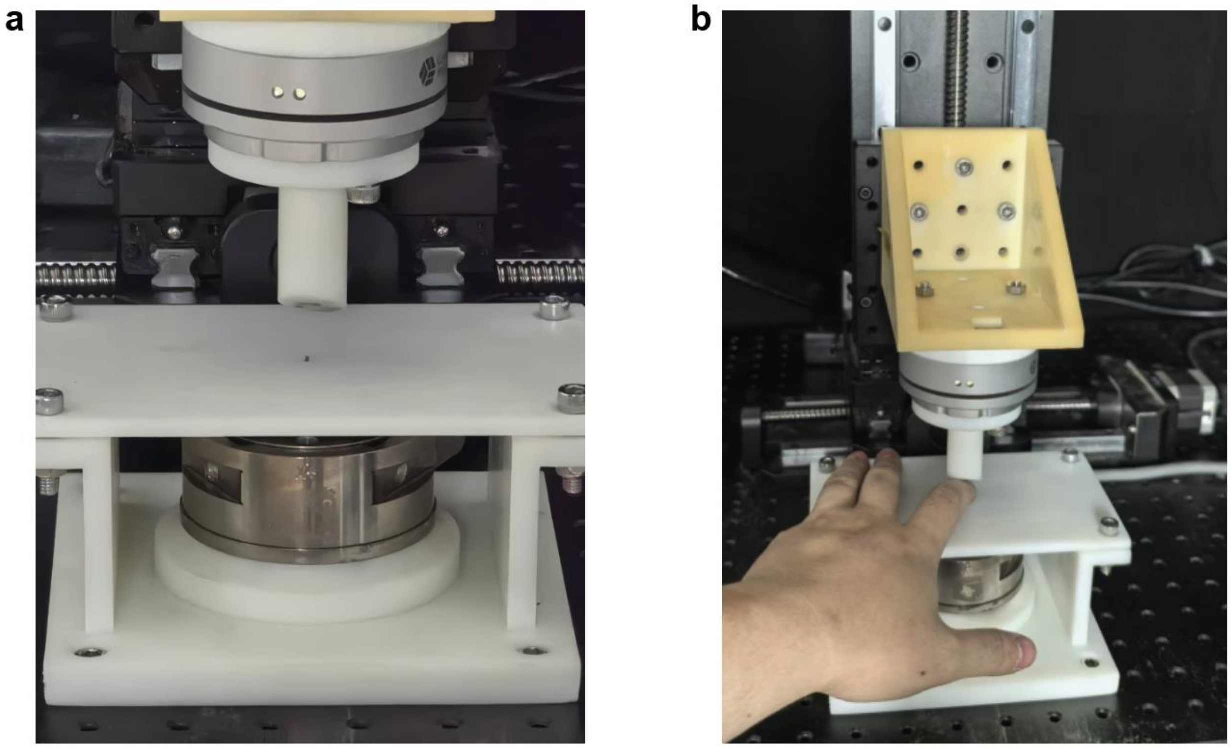


Fig. S4. Human finger force data testing and collection device. The figure shows the custom-built apparatus used to acquire ground-truth force data from a human finger for comparative analysis. **a**, The testing platform in its neutral state, comprising a high-precision multi-axis force/torque sensor mounted beneath a flat contact plate. **b**, A human participant applies force to the contact plate. The underlying force/torque sensor records the complete three-axis force vector (Fx, Fy, Fz) generated by the finger during dynamic interaction, providing a quantitative benchmark for the biomimetic performance of the HydroPalm.

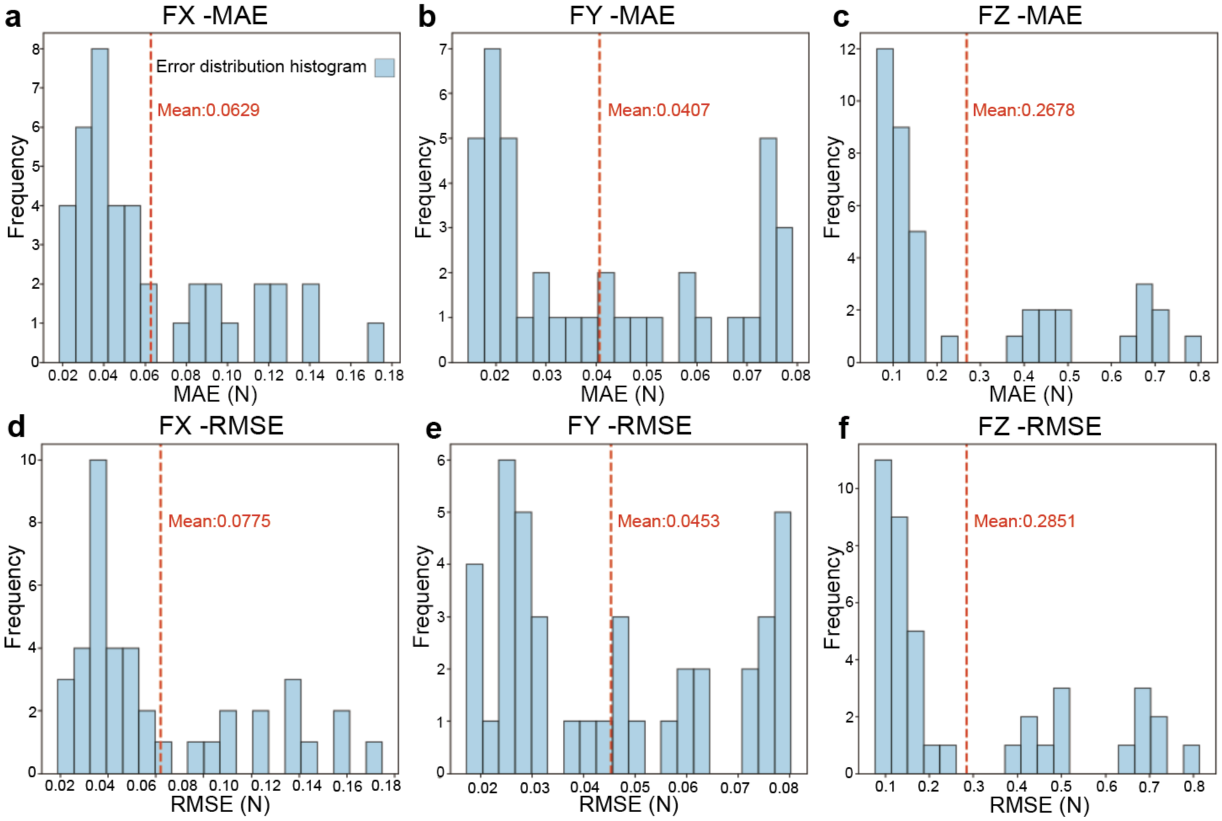


Fig. S5. The prediction error of all test sequences. a-c, Histograms showing the distribution of mean absolute error (MAE) for force predictions in x-direction **a**, y-direction **b**, and z-direction **c**. The y-axis represents the frequency of errors occurring within each range in all test sequences. Sky blue bars represent the error distribution across test sequences (including types of main text), and red dashed lines indicate the mean MAE values (Fx: 0.0629 N, Fy: 0.0407 N, Fz: 0.2678 N). **d-f**, Histograms showing the distribution of Root Mean Square Error (RMSE) for force predictions in x-direction **d**, y-direction **e**, and z-direction **f**. Sky blue bars represent the error distribution across 41 test sequences, and red dashed lines indicate the mean RMSE values (Fx: 0.0775 N, Fy: 0.0453 N, Fz: 0.2851 N). These performance indicators demonstrate the reliability of our algorithm.

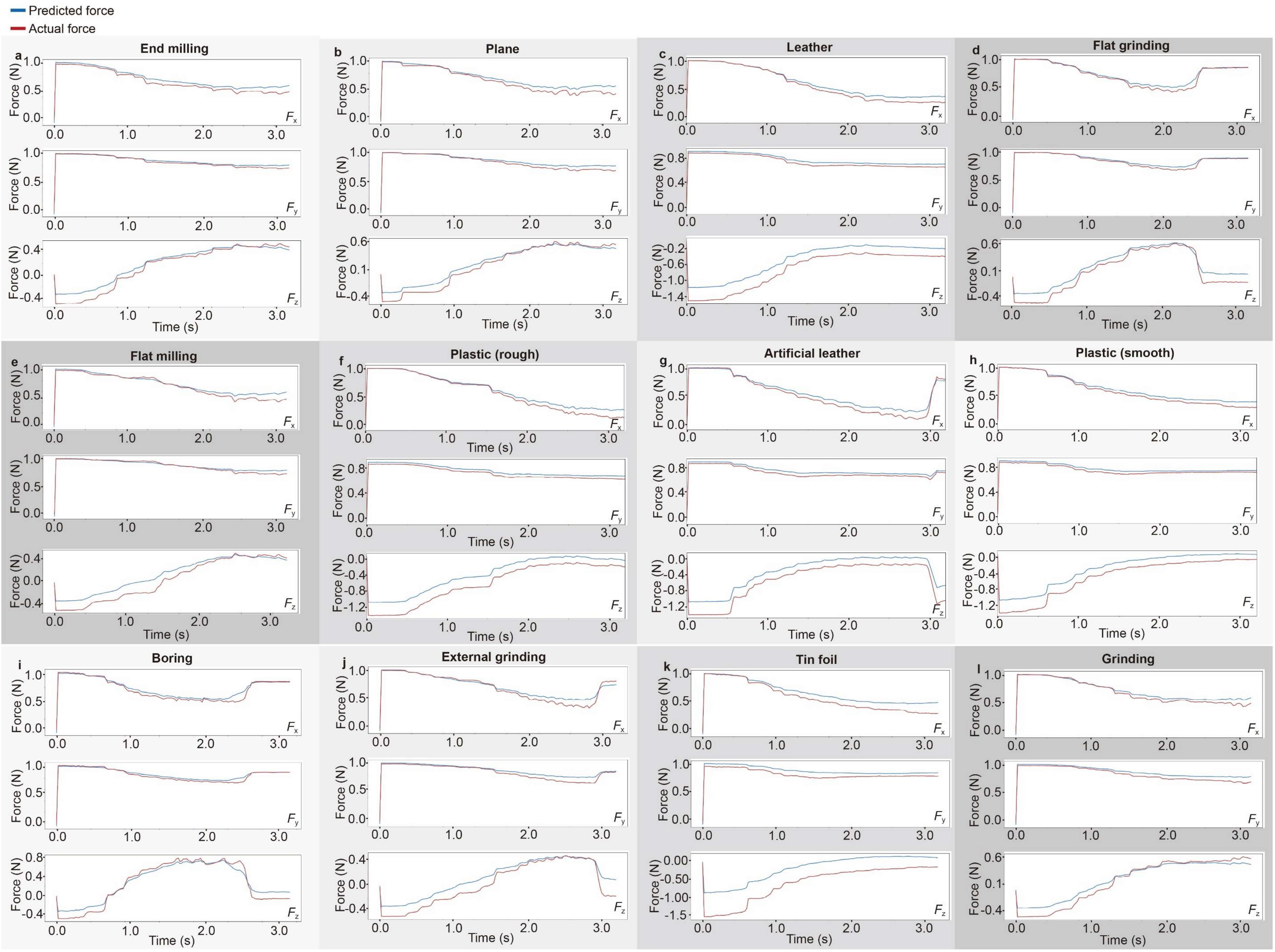


Fig. S6. Comparison between actual and predicted Fx, Fy, and Fz force components of various materials. Panels show three stacked traces per condition (top: Fx, middle: Fy, bottom: Fz). blue lines denote predicted forces and red lines denote actual forces after post-processing. The x-axis represents time (0-3.2 s) and the y-axis represents force (N). **a**. End milling; **b**, Plane; **c**, Leather; **d**, Flat grinding; **e**, Flat milling; **f**, Plastic (rough); **g**, Artificial leather; **h**, Plastic (smooth); **i**, Boring; **j**, External grinding; **k**, Tin foil; **l**, Grinding.

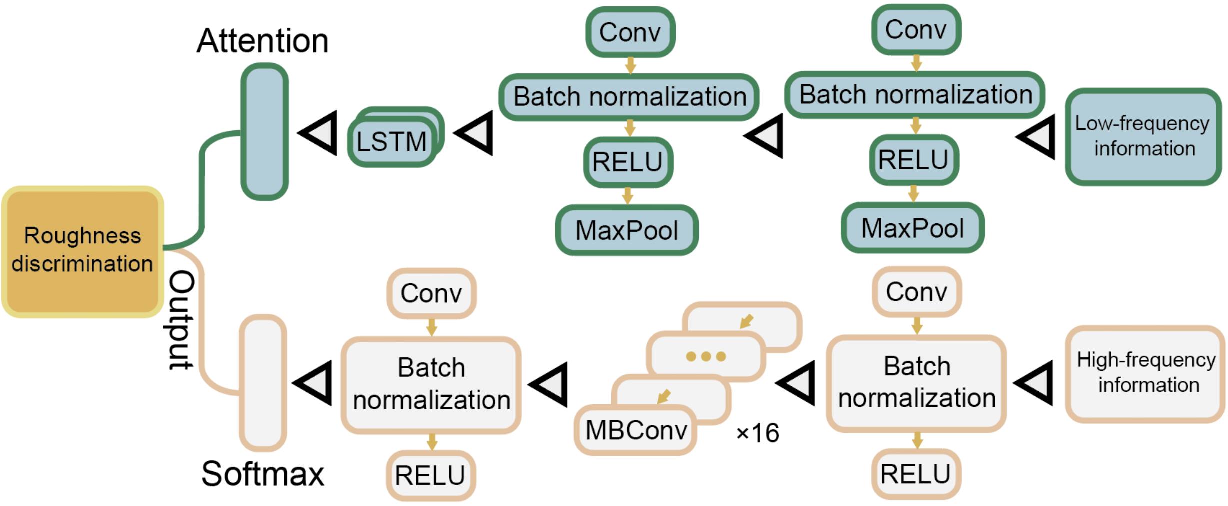


Fig. S7. Network architecture for tactile roughness discrimination. The network employs a dual-stream architecture for multi-modal information processing. The upper stream, composed of a Convolutional Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) layer and an attention mechanism, processes low-frequency force information. This stream extracts sequential features from force data, with the attention module dynamically weighing relevant temporal segments. The lower stream, based on an EfficientNet-B5 architecture, extracts the features from high-frequency tactile (vibration) information, specifically designed for robust texture representation. Outputs from both streams are then fused through a strategic multi-modal fusion method (details provided in Supplementary Text 3), which is subsequently fed into a softmax layer for roughness discrimination and output. Each convolutional block within both streams typically includes a convolution layer, batch normalization, and a ReLU activation function. Max pooling layers are utilized for downsampling in the low-frequency stream, while the EfficientNet-B5 block comprises multiple MBConv layers.

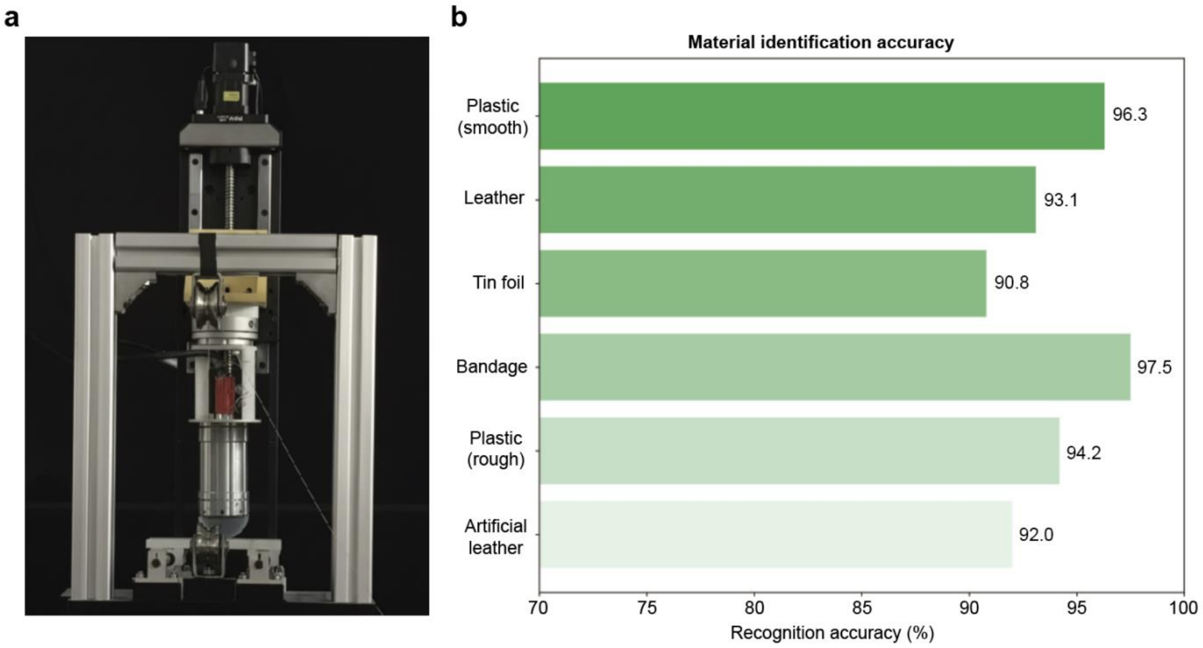


Fig. S8. Material identification accuracy. a, Experimental setup for texture recognition testing in air environment. **b**, The sensor's recognition accuracy for six different materials, including the following materials: artificial leather, plastic (rough), bandage, tin foil, leather, and plastic (smooth)

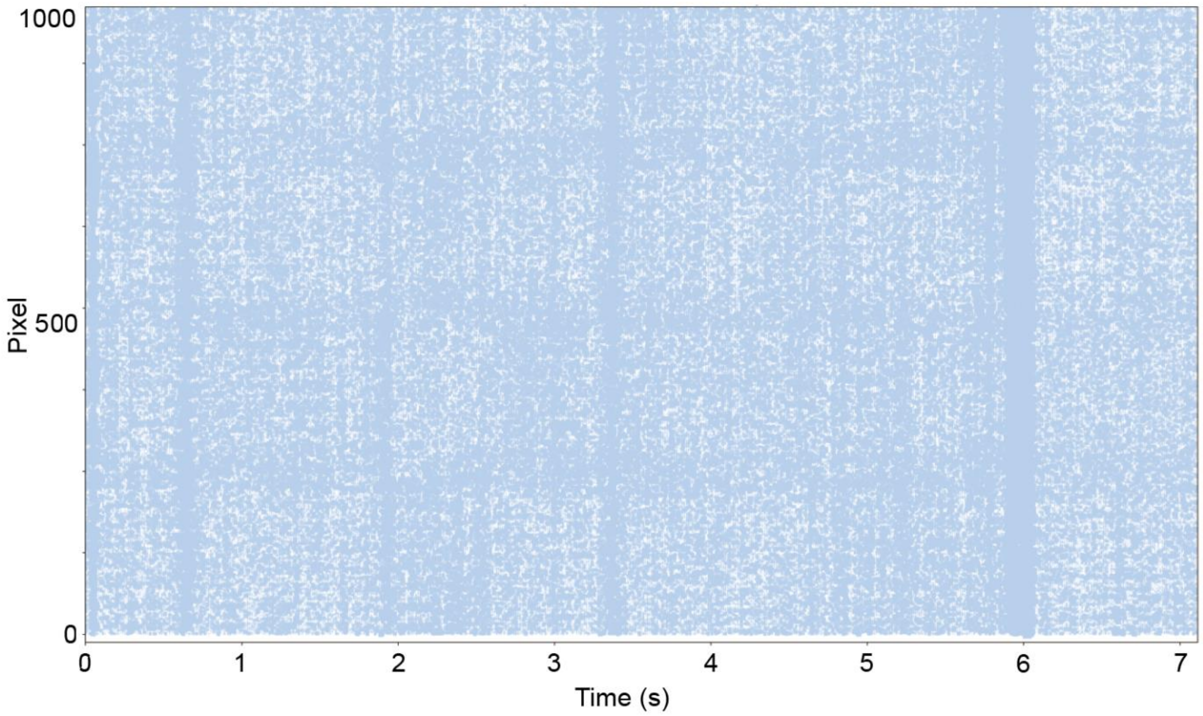
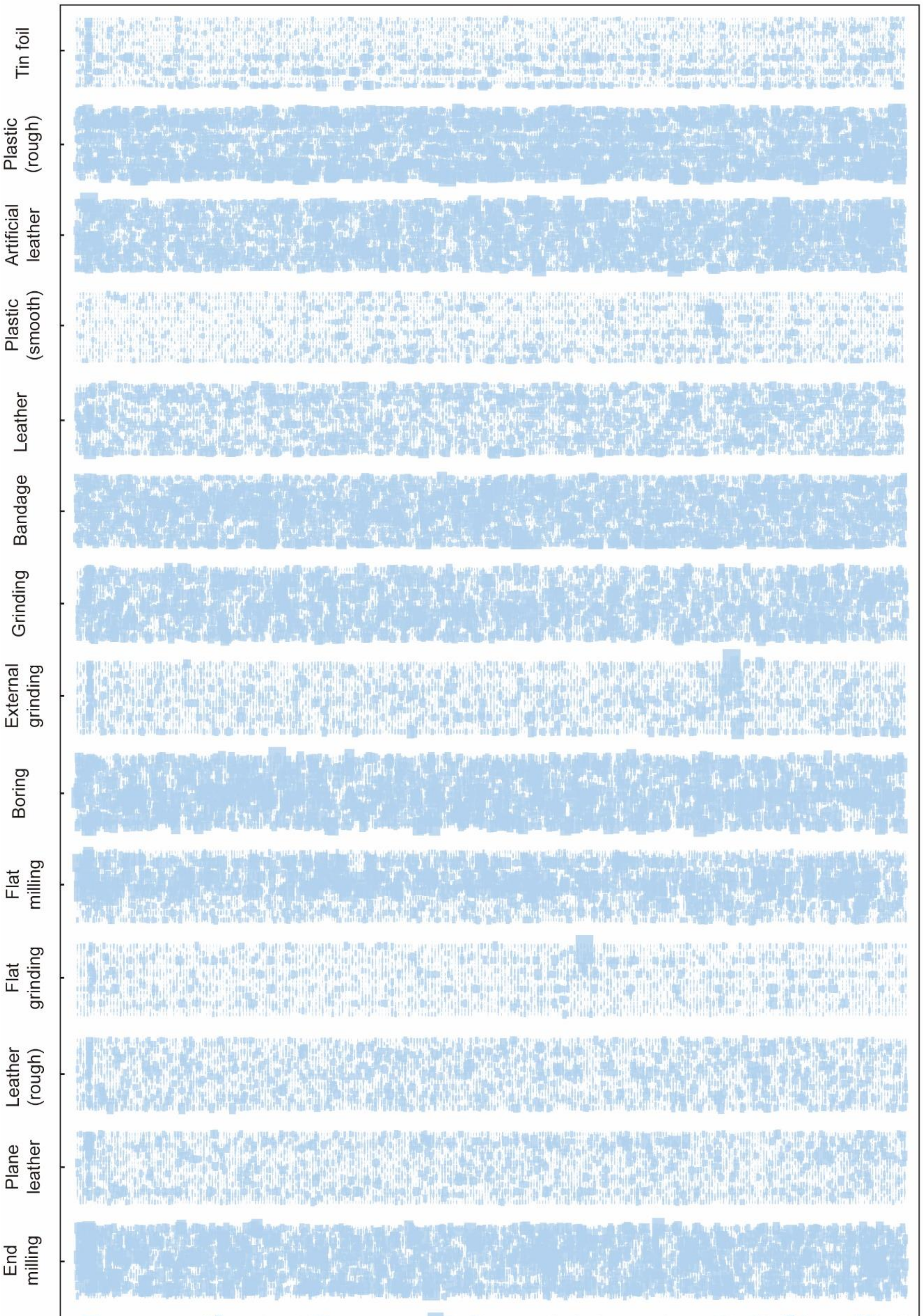


Fig. S9. Temporal texture-based pulse representation for material surface characterization. High-resolution pulse diagram showing temporal texture variations across the surface of a bandage material during tactile interaction. The visualization is based on 961 sampling points extracted from the central region (31×31 grid) of the tactile sensor video frame.



Tin foil
Plastic (rough)
Artificial leather
Plastic (smooth)
Leather
Bandage
Grinding
External grinding
Boring
Flat milling
Flat grinding
Leather (rough)
Plane leather
End milling
Turn
0
0.2
0.4
0.6
0.8
1.0
Time (s)

Fig. S10. Temporal high-frequency pulse information derived from a 20-pixel texture patch. Materials include: tin foil, plastic(rough), artificial leather, plastic(smooth), leather, bandage, grinding, external grinding, boring, flat milling, flat grinding, leather(rough), plane leather, end milling, turn.

Legends for Supplementary Videos:

Supplementary Video 1. Measurement of the 3D contact forces for a human finger and HydroPalm.

Supplementary Video 2. Dynamic texture recognition on various material surfaces.

Supplementary Video 3. Slip prediction and compliant motion control experiments conducted on HydroPalm.

Supplementary Video 4. Underwater pipe-assembly task conducted on HydroPalm.