Deep Drawing: Performance Surface Sound Source Localization

Lennon Seiders
lseiders3@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Alex Zhang
ziyuzha@umich.edu
University of Washington
Seattle, WA, USA

Anusha Chinthamaduka
anushac@umich.edu
University of Michigan
Ann Arbor, MI, USA

Julie Zhu
zhujulie@umich.edu
University of Michigan
Ann Arbor, MI, USA

John Granzow
jgranzow@umich.edu
University of Michigan
Ann Arbor, MI, USA

Abstract

Deep Drawing is an ongoing exploration of the sound of drawing through AI co-performance. As the sounds of a performer's drawing gestures resonate through a wooden board, they are re-created, spatially localized, and visualized in real-time using a novel deep learning approach. The system foregrounds the often-overlooked, timbrally complex sounds of drawing and frames them as a shared interpretive space between human and machine. This paper presents an accessible hardware setup and a sound source localization (SSL) model that can be trained with minimal data, enabling expressive interaction without extensive calibration or large datasets. Because SSL for performance-sized surfaces is a relatively unexplored research topic, we introduce practical techniques for this setting, including high-fidelity data capture, preprocessing techniques, and two model architectures. Deep Drawing contributes a replicable system that emphasizes performance, embodiment, and co-creative agency.

Keywords

Machine Learning, Sound-Source Localization, Digital Signal Processing, Microphone Array, Performance

1 Introduction

Whether it is a signature, a picture, or a scribble, each inscription produces a unique, expressive sound that we often take for granted. By drawing on a plywood board, these unique timbres are amplified. Our goal is to use sound source localization to visualize these gestures, enhancing the multi-modal effect of a musical performance with this drawing board. This work builds upon our ongoing project's previous efforts to visualize these sounds [18].

Deep Drawing follows a growing trend of applying artificial intelligence and machine learning to creative expressivity and performance [1, 2, 11]. Drawing has a rich history at NIME as a synthesizer input, including Sketch-to-Sound [16], Sounding Brush [13], Illusio [3], and JamSketch [9]. Rather than mapping drawn marks to synthesis parameters, we foreground the acoustic byproduct of drawing itself. Other than this project's previous work [18], no other research addresses the application of sound-source localization on a performance surface.

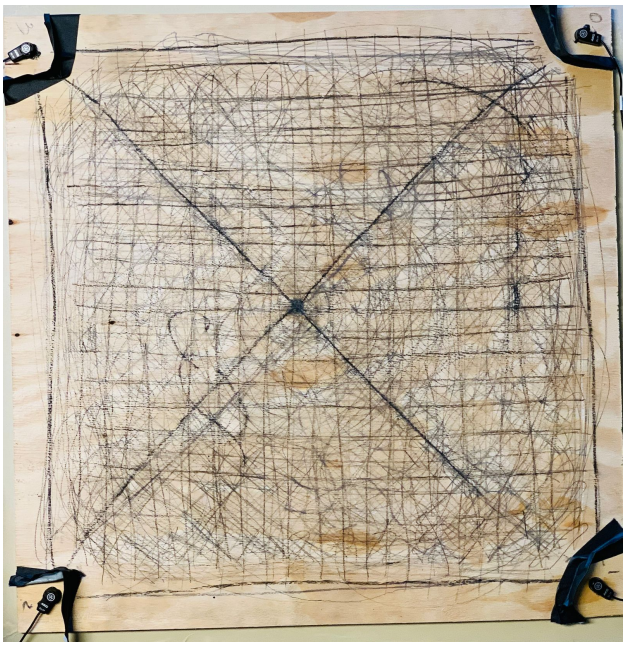

Figure 1: Plywood performance board with contact microphones.

Sound source localization (SSL) is a long-standing field in acoustics research, with many approaches relying on estimating the time difference of arrival (TDOA) between microphones [4].

Because we must localize based on sound propagating through plywood, an anisotropic material, TDOA estimation is extremely difficult in this case. Additionally, sound travels through the 3 ft × 3 ft plywood so quickly that TDOAs are challenging to measure [5]. As a result of these limitations, traditional TDOA SSL methods are not fit for this application.

More recently, deep learning–based SSL approaches that rely on frequency and phase representations have shown strong performance across a range of acoustic tasks [7, 10, 12, 14, 17]. Unlike classical methods that depend on explicit time-delay estimation, these models can learn complex propagation patterns directly from data, making them more robust to noise, reflections, and material-specific effects. Because sound transmission through plywood is highly anisotropic and difficult to model analytically, we adopt a learning-based approach for this task.

In line with our previous work, we use a ResNet-based model that operates on short spectrogram inputs from four contact microphones. Residual networks [6] are well-established in SSL [10], offering a strong balance between representational capacity and computational efficiency for real-time inference. We also

NIME '26, June 23–26, 2026, London, UK

evaluate a lightweight transformer architecture, which has shown strong performance in certain SSL tasks [12, 15].

Unlike traditional SSL applications that aim for precise geometric reconstruction, our goal is perceptual and performative. The system does not need to recover the exact physical trajectory of the pen; rather, it should produce a convincing and expressive audiovisual trace of the gesture in real time. This shifts the design priorities toward low latency, small datasets, and models that are practical in performance contexts.

This paper makes three main contributions:

(1) A simple, replicable hardware and data-capture system for sound source localization on a performance-sized drawing surface.
(2) A preprocessing and training pipeline designed for low-latency operation with relatively small datasets.
(3) A comparison of convolutional and transformer-based architectures for coordinate prediction in this setting.

2 Methodology

2.1 Hardware Setup

The Deep Drawing system is built around a simple and replicable sensing surface. The drawing surface consists of a 3 ft × 3 ft × $\frac{1}{4}$ in three-ply plywood board, chosen for its affordability and ability to clearly transmit frictional and impact sounds.

Four contact microphones (AKG C411 PP) are attached to the top of the board, one near each corner. Contact microphones are used instead of air microphones to reduce ambient noise and isolate the mechanical energy of the drawing gestures.

Audio from the microphones is recorded simultaneously at 192 kHz using a multichannel audio interface. The high sampling rate is useful because sound travels quickly through wood, resulting in very small time differences between channels.

For ground-truth labels, the performer draws with a ballpoint pen fitted with a high-contrast marker on top, tracked by an overhead GoPro Hero 13 Black at 240 fps.

Each video frame was converted into a 2D position on the board using a calibrated mapping between image space and surface coordinates. These coordinates were then time-aligned with the audio data to produce paired audio–position samples. Each video frame-length of audio was converted into a spectrogram representation, and the resulting spectrogram frames were labeled with the corresponding pen coordinates at that moment in time.

2.2 Dataset

The dataset contains approximately 590,000 labeled frames (41 minutes at 240 fps), captured as four recordings of about ten minutes each. During each recording, the performer drew freely across the surface, producing a variety of strokes, pressures, and gesture types.

Between each recording, the drawing board was rotated relative to the microphone layout. Compared to our previous work using a uniform orientation, we believe that this procedure helps the dataset better reflect the inconsistent conditions of a live performance setup.

To prevent temporal leakage between training and validation data, each recording was split sequentially rather than randomly. The first 90% of each recording was used for training, and the final 10% was reserved for validation. This ensures that the validation set contains gestures that occur later in time and are not simply near-duplicates of training samples.

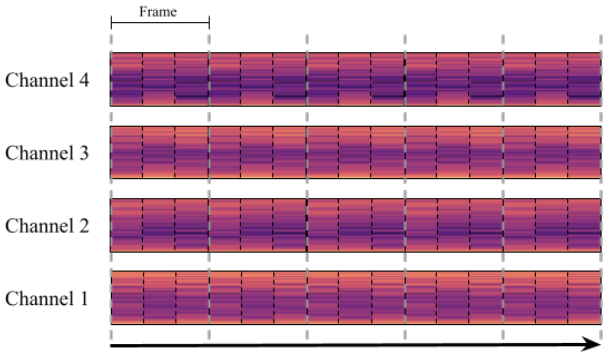

Figure 2: Dataset preprocessing format. Each video frame is labeled with a coordinate, and the corresponding audio is converted into spectra.

2.3 Model Architecture

Two architectures were evaluated for sound source localization: a convolutional model based on ResNet-50 and a transformer-based model with a CNN tokenizer front-end. Both models take as input a context window of $T = 8$ consecutive spectrogram frames, preprocessed as shown in 2 using a 256-point FFT. This context window spans about 33 ms, balancing our concern with latency and sufficient data for the model. Per-channel normalization is applied by dividing by the maximum absolute magnitude across all frames, sub-segments, and frequency bins within each channel of a given recording.

2.3.1 ResNet Model. The ResNet model formulates localization as an image-style regression problem. The input tensor is reshaped into a single-channel pseudo-image by flattening the temporal context, microphone channels, and sub-segment dimensions into one spatial axis, with frequency bins forming the other axis.

A ResNet-50 backbone [6] is trained from scratch. The first convolutional layer is modified to accept a single-channel spectrogram input, and the final classification layer is replaced with a lightweight regression head. This head consists of dropout followed by a linear projection that outputs the predicted (x, y) coordinates. (see Appendix for full details)

2.3.2 Transformer Model. The transformer model treats the multichannel spectrogram from each frame as a token in a short temporal sequence, allowing the model to capture temporal dependencies across the context window.

A lightweight convolutional tokenizer extracts spectral features from each frame and converts them into fixed-dimensional embeddings. Positional embeddings are added to encode temporal order, and the resulting sequence is processed by a stack of transformer encoder layers with multi-head self-attention.

The embedding corresponding to the most recent frame is passed through a small regression head to predict the (x, y) coordinates. (see Appendix for full details)

3 Results

To test the effectiveness of the two architectures and filtering, we test each configuration on each 10 minute recording individually in addition to the entire dataset. This way, we can evaluate the

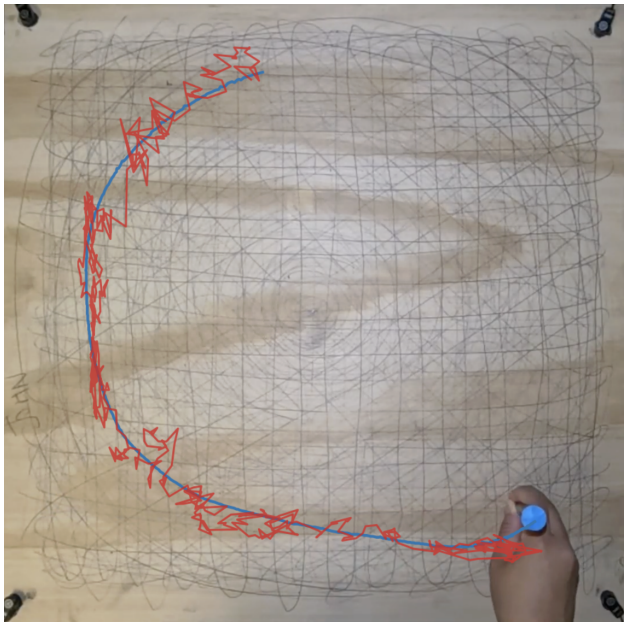

Figure 3: Example model output. The blue trace is ground truth, red is the model's prediction. For more media, see https://deepdrawing.github.io/.

ability of each model configuration to generalize across performances and gain insight into the effect of human and calibration error introduced by reconfiguring the recording setup each time.

Table 1: L1 Localization error by model and high-pass filter setting.

Model	Hi-Pass	R0	R1	R2	R3	All
ResNet50	0	0.087	0.296	0.065	0.249	0.191
ResNet50	1000	0.081	0.301	0.064	0.251	0.197
ResNet50	2000	0.082	0.295	0.068	0.251	0.195
Transformer	0	0.139	0.297	0.145	0.250	0.228
Transformer	1000	0.135	0.296	0.137	0.255	0.237
Transformer	2000	0.135	0.296	0.126	0.270	0.225

3.0.1 High-pass filtering. Across recordings, the effect of high-pass filtering was modest and inconsistent. For the ResNet model, filtering at 1000 Hz slightly improved performance on some recordings (e.g., R0), but produced little change overall. The 2000 Hz setting yielded similar aggregate performance to the unfiltered condition.

The transformer model showed small improvements at higher cutoff frequencies on certain recordings (e.g., R2), but overall differences between filtering conditions remained minor. These results suggest that while high-pass filtering can slightly stabilize performance, the models are relatively robust to low-frequency content in the input.

3.0.2 ResNet Model. The ResNet model achieved the lowest overall localization error across most configurations. Without high-pass filtering, it obtained the best aggregate performance with an average L1 error of 0.191 across all recordings.

Performance varied across individual recordings. Recordings R0 and R2 showed relatively low errors (0.087 and 0.065, respectively), while R3 produced significantly higher error (0.249). This indicates that the model's performance is sensitive to changes in board orientation and recording conditions.

Despite this variance, the ResNet architecture consistently outperformed the transformer in aggregate error and showed more stable behavior across filtering conditions.

3.0.3 Transformer Model. The transformer model achieved higher overall errors compared to the ResNet across all configurations. Its best aggregate performance was obtained with a 2000 Hz high-pass filter, yielding an average L1 error of 0.225.

Like the ResNet, the transformer exhibited large variation across recordings. For example, R1 produced significantly higher errors (around 0.296) than other recordings, suggesting sensitivity to changes in setup or performer behavior.

Although the transformer did not match the ResNet in accuracy, it achieved comparable performance on some recordings while using substantially fewer parameters. This indicates that lightweight attention-based models remain a viable option when computational constraints are critical.

4 Conclusion

This research tests two potential model architectures, and contributes preprocessing techniques and data collection methods for the task of sound source localization on a performance surface. As suggested in the Results section, the problem of reliably localizing sound with contact microphones on an anisotropic material is still far from solved; however, the proposed approach demonstrates that lightweight deep learning models can achieve meaningful spatial predictions with relatively small datasets and minimal calibration.

The ResNet model provided the most consistent performance, while the transformer's 2M parameters offer a viable lightweight alternative that shows potential. More broadly, these results indicate that learning-based approaches are viable for performance-sized surfaces where classical time-delay methods are impractical.

The variance across recordings suggests that the learned mapping is sensitive to board orientation and mounting conditions. Rather than viewing this as a limitation, we consider orientation-specific calibration to be a practical and musically acceptable solution, since performance setups are typically fixed and the training process is short.

This work contributes a replicable system that foregrounds the expressive sound of drawing and frames localization as part of a co-creative performance process. By treating drawing gestures as both sonic and spatial events, Deep Drawing opens new possibilities for instruments that translate physical inscription into interactive audiovisual experiences.

5 Discussion and Future Work

A core challenge in producing better results is the quality of recording data. While we have made significant improvements in recording fidelity, calibration and human error still makes recording and labeling a difficult task. In the future, we aim to develop a more streamlined recording procedure in order to get the most out of the drawing audio. One potential direction is the use of a pen plotter, which eliminates the need for video recording to label data and can record data for long periods of time. We have obtained a plotter, and plan to try using it in future work.

Another possible direction is a reformulation of the learning objective. Because our goal is to recreate drawing gestures from their sound, the curves and shape of a pen stroke is more important than its exact location. Taking this into account, training a model to predict the shape of the gesture rather than the exact coordinate of the pen at each point in time is worth exploring.

Acknowledgments

Deep Drawing is an ongoing research project that is part of ArtsEngine, an interdisciplinary initiative at the University of Michigan.

6 Ethical Standards

The data used in this paper was recorded and labeled by the researchers. Training of each model was done using on-campus computing resources at the University of Michigan.

References

[1] Sabina Hyoju Ahn, Ryan Millett, and Seyeon Park. 2025. Eco-Sonic Interfaces for Embodied AI Sound Exploration. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 1, 5 pages. https://doi.org/10.5281/zenodo.15698772

[2] Misagh Azimi and Mo H. Zareei. 2025. Live Improvisation with Fine-Tuned Generative AI: A Musical Metacreation Approach. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 54, 5 pages. https://doi.org/10.5281/zenodo.15698902

[3] Jerônimo Barbosa, Filipe Calegario, Veronica Teichrieb, Geber Ramalho, and Giordano Cabral. [n. d.]. Illusio: A drawing-based digital music instrument.

[4] Pierre-Amaury Grumiaux, Srđan Kitić, Laurent Girin, and Alexandre Guérin. 2022. A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America* 152, 1 (2022), 107–151.

[5] Guangping Han, Qinglin Wu, and Xiping Wang. 2006. Stress-wave velocity of wood-based panels: Effect of moisture, product type, and material direction. *Forest products journal. Vol. 56, no. 1 (Jan. 2006): Pages 28-33* (2006).

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Toni Hirvonen. 2015. Classification of spatial audio location and content using convolutional neural networks. In *Audio Engineering Society Convention 138*. Audio Engineering Society.

[8] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6980

[9] Tetsuro Kitahara, Sergio Giraldo, and Rafael Ramírez. 2017. JamSketch: a drawing-based real-time evolutionary improvisation support system. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 505–506.

[10] Adam Kujawski, Gert Herold, and Ennes Sarradj. 2019. A deep learning method for grid-free localization and quantification of sound sources. *The Journal of the Acoustical Society of America* 146, 3 (2019), EL225–EL231.

[11] Sandy Ma and Charles Patrick Martin. 2025. Touching Wires: tactility and a quilted musical interface for human-AI musical co-creation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 41, 8 pages. https://doi.org/10.5281/zenodo.15698865

[12] Christopher Schymura, Benedikt Bönninghoff, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki, and Dorothea Kolossa. 2021. PILOT: Introducing transformers for probabilistic sound event localization. *arXiv preprint arXiv:2106.03903* (2021).

[13] Sourya Sen, Koray Tahiroğlu, and Julia Lohmann. 2020. Sounding brush: A tablet based musical instrument for drawing and mark making. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 331–336.

[14] Juan Manuel Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. 2018. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors* 18, 10 (2018), 3418.

[15] Nelson Yalta, Y Sumiyoshi, and Yohei Kawaguchi. 2021. The Hitachi DCASE 2021 Task 3 system: Handling directive interference with self attention layers. *DCASE Challenge, Music Technol. Group, Universitat Pompeu Fabra, Barcelona, Spain, Tech. Rep* 29 (2021).

[16] Shuoyang Zheng, Bleiz M Del Sette, Charalampos Saitis, Anna Xambó, and Nick Bryan-Kinns. 2024. Building Sketch-to-Sound Mapping with Unsupervised Feature Extraction and Interactive Machine Learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 591–597.

[17] Jie Zhou, Binbin Mi, Jianghai Xia, Hao Zhang, Ya Liu, Xinhua Chen, Bo Guan, Yu Hong, and Yulong Ma. 2024. Noise source localization using deep learning. *Geophysical Journal International* 238, 1 (2024), 513–536.

[18] Julie Zhu, Erfun Ackley, Zhiyu Zhang, and John Granzow. 2025. Deep Drawing: Spectra and Sound-Source Localization. In *Proceedings of the International Computer Music Conference (ICMC 2025)*. International Computer Music Association, Boston, MA, USA, 446–450.

A Model Architecture

A.1 ResNet Model

The ResNet model reformulates the localization task as a single-channel image regression problem. The five-dimensional input tensor of shape $(B, 8, 4, 3, F)$, where B is the batch size and F is the number of frequency bins, is reshaped into a pseudo-image of shape $(B, 1, 96, F)$ by flattening the context-frame, channel, and sub-segment dimensions into a single spatial axis ($8 \times 4 \times 3 = 96$ rows), with frequency bins forming the column axis.

The backbone is a ResNet-50[6] initialized from scratch. All intermediate layers remain unmodified, while two modifications are made to the standard architecture:

(1) **Input layer:** The initial 7×7 convolutional layer is replaced with a single-channel variant ($\texttt{Conv2d}(1, 64, 7 \times 7, \text{stride} = 2, \text{padding} = 3)$) to accept the single-channel spectrogram input instead of three-channel RGB images.
(2) **Output head:** The 1000-class classification layer is replaced with a dropout layer ($p = 0.1$) followed by a linear projection from the 2048-dimensional feature space to two output units, corresponding to the predicted (x, y) coordinates.

A.2 Transformer Model

The transformer model treats each frame's multi-channel spectrogram as a token in a temporal sequence, explicitly modeling temporal dependencies across the context window with self-attention.

Rather than flattening each frame into a vector, a lightweight convolutional tokenizer extracts local spectral features while preserving frequency and channel structure. Each frame's spectrogram of shape $(4, 3, F)$ is passed through a three-layer CNN:

(1) $\texttt{Conv2d}(4, 16, 1 \times 7, \text{stride} = (1, 2))$ with batch normalization and GELU activation, convolving along the frequency axis;
(2) $\texttt{Conv2d}(16, 32, 3 \times 5, \text{stride} = (1, 2))$ with batch normalization and GELU, mixing across sub-segments and frequency;
(3) $\texttt{Conv2d}(32, 64, 3 \times 3, \text{stride} = (1, 2))$ with batch normalization and GELU, further refining the representation.

Adaptive average pooling reduces the output to a single vector per frame, which is projected to the model dimension $d_{\text{model}} = 128$ via a linear layer. The tokenizer is applied independently to each of the $T = 8$ frames, yielding a sequence of token embeddings.

Learned positional embeddings are added to the token sequence to encode temporal order, followed by dropout ($p = 0.1$). The sequence is then processed by a stack of $L = 6$ transformer encoder layers, each with $h = 4$ attention heads, a feedforward dimension of 1024, GELU activations, pre-layer normalization for improved training stability, and dropout of $p = 0.1$. Full self-attention is used, as all frames in the context window are available at both training and inference time.

The representation at the final temporal position (corresponding to the most recent frame) is extracted and passed through a two-layer output head: layer normalization, a linear projection

from d_{model} to $d_{\text{model}}/2 = 64$ with GELU activation, followed by a final linear projection to two output units predicting the (x, y) coordinates.

A.3 Training Objective and Optimization

Both models are trained to predict normalized two-dimensional sound source coordinates, where pixel coordinates in the range $[0, 936]$ are linearly scaled to $[0, 1]$. The loss function is the mean absolute error (L1 loss) between predicted and ground-truth coordinates.

Both models are optimized with Adam [8] using a weight decay of 1×10^{-4} and a cosine annealing learning rate schedule over 20 epochs with a batch size of 32. The ResNet model uses a learning rate of 1×10^{-3}, while the transformer model uses a lower learning rate of 1×10^{-4}.