Research Article

Applying Object Detection to Automatic Drum Transcription

Dylan Li

Tsinghua International School, Zhongguancun North St. Haidian District, Beijing, 100084, China

ABSTRACT

Automatic music transcription (AMT) is a fundamental problem in music information retrieval (MIR), involving the conversion of audio recordings into symbolic representations such as MIDI. This study presents a novel approach to automatic drum transcription (ADT) by reframing it as a computer vision object detection problem. Using the YOLO11 model, drum notes were predicted and transcribed with bounding boxes in spectrograms generated from the Expanded Groove MIDI Dataset (E-GMD). Two-second audio segments were extracted via a sliding window, converted into 640×640 grayscale spectrograms, and annotated with bounding boxes corresponding to onset times and instrument classes. The model achieves strong detection performance, with results mAP@0.5 of 0.943, precision of 0.892, and recall of 0.846. Results demonstrate YOLO11's ability to handle polyphonic, temporally dense drum passages without explicit onset separation. This work highlights the potential of adapting computer vision techniques to audio-based event detection, paving the way for broader MIR applications beyond percussion, such as multi-instrument transcription and real-time performance analysis.

Keywords: Automatic Music Transcription; Object Detection; YOLO; Spectrogram; Drum Transcription; Audio Segmentation; Music Information Retrieval

INTRODUCTION

Automatic music transcription (AMT) remains a longstanding challenge in the field of music information retrieval (MIR), an interdisciplinary science combining computer science and music. Applications of MIR range from music genre classification and recommendation systems to track separation and transcription. Within AMT, automatic drum transcription (ADT) focuses specifically on transcribing the rhythmic content of drum set performances.

Traditional drum transcription methods often rely on onset detection from hand-crafted audio features or spectrograms, followed by classification. Recent work, such as OaF-Drums by Callender, Hawthorne, and Engel (1), has improved transcription quality using deep learning approaches—employing recurrent neural networks (RNNs) for onset timing detection and convolutional neural networks (CNNs) for instrument classification. However, other advances in deep learning and computer vision open promising new directions.

One such advancement is the You Only Look Once (YOLO) (2) family of object detection algorithms,

Corresponding author: Dylan Li, E-mail: dylan.li.2026@this.edu.cn.

Accepted September 16, 2025
https://doi.org/10.70251/HYJR2348.35282287

widely recognized for real-time detection accuracy in visual tasks. In this work, I apply YOLO11 (3) to drum transcription by reframing the task as a visual object detection problem, mapping drum hits to temporal "objects" within spectrogram images. This allows simultaneous detection of overlapping events across multiple drum classes—closely mirroring the polyphonic nature of real drum performances. Moreover, by demonstrating that object detection frameworks can be applied to spectrogram representations, this study lays the groundwork for adapting similar methods to other MIR problems, including multi-instrument AMT, onset tracking in complex ensembles, and more.

METHODS AND MATERIALS

Dataset

This study uses the Expanded Groove MIDI Dataset (E-GMD) (1), which contains high-quality, human-performed drum recordings paired with MIDI annotations. The dataset was filtered to exclude recordings shorter than two seconds and those using drum kits with significantly non-acoustic timbres ("808 Simple," "909 Simple," "Nu RNB," "Ele-Drum," "Custom1") as they harmed model performance. Additionally, due to class imbalance and a low amount of tom drum samples, the three separate tom classes in E-GMD were combined into one.

The filtered dataset contained 38 distinct drum kit sounds. Audio data augmentation was applied using the audiomentations library (4) to improve generalization, including time stretch, pitch shift, Gaussian noise, air absorption simulation, and shelf filtering.

Audio Processing and Annotation

Recordings were sampled at 44.1 kHz and segmented with a 2-second sliding window (1-second hop). Mel spectrograms were generated with 256 mel bins and a hop length of 441 samples, which is equivalent to 10 ms of audio. Spectrograms were converted to grayscale, resized to 640 × 640 pixels, and annotated with bounding boxes spanning the full vertical range at each onset time. Bounding box width corresponded to 50 ms. Each was labeled with the appropriate instrument class (bass drum, snare, tom, closed hi-hat, open hi-hat, crash cymbal, or ride cymbal).

Using this technique, 300 random E-GMD audio samples were selected and converted to 13,292 spectrogram images split 80% train, 20% validation. Table 1 summarizes the number of annotated onsets per class in the filtered dataset. As shown, snares and closed hi-hats are the most frequent classes, while crash cymbals and open hi-hats occur less often, reflecting the natural class imbalance in E-GMD.

Model Training

A YOLO11-medium (3) model was trained using the Ultralytics Python library on the generated dataset. Training ran for 100 epochs with batch size 32, stochastic gradient descent optimizer, and linear learning rate decay from 0.01 to 0.0001. All built-in Ultralytics data augmentation was disabled except hsv_v = 0.4, scale = 0.5, mosaic = 0.5 with close_mosaic = 40, and erasing = 0.3. Finally, the loss weights were adjusted to 7.5 for box loss, 2.0 for classification loss, and 0.5 for distribution focal loss to emphasize classification accuracy and deemphasize distribution focal loss as it would converge within less than 10 epochs of training due to the uniform size and shape of the bounding boxes.

Inference and Post-processing

Inference mirrored training segmentation. Predictions from the central 1-second region of each window were retained, except for start and end windows, which were adjusted to capture edge events. Bounding boxes were mapped back to onset times and instrument labels for MIDI reconstruction and visualization.

Table 1. Number of annotated drum onset events per class in the training and validation sets of the filtered E-GMD dataset

Split	Kick	Snare	Toms	Open Hi-Hat	Closed Hi-Hat	Crash Cymbal	Ride Cymbal
Training	37960	61490	16510	3791	44516	2390	18235
Validation	9779	15599	4635	847	11233	580	4613
Total	47739	77089	21145	4638	55749	2970	22848

Counts illustrate dataset imbalance (e.g., snare dominates) that affects evaluation and model learning.

RESULTS

Quantitative Evaluation

Evaluation metrics included precision, recall, and mean average precision (mAP) at IoU thresholds of 0.5 and 0.5–0.95 (Table 2). A confusion matrix is shown in Figure 1 (normalized in Figure 2), from which it is evident that the largest issue the model faces currently is bounding box prediction, with classification being more accurate in comparison.

In addition to object detection metrics (precision, recall, mAP), transcription accuracy was evaluated using standard onset-based metrics. A prediction was considered correct if its onset fell within ±50 ms of

Table 2. Per-class and overall precision, recall, and mean average precision (mAP) scores for YOLO11 on the filtered E-GMD dataset

Class	Precision	Recall	mAP@0.5	mAP@0.5-0.95
All	0.892	0.846	0.943	0.636
Kick	0.903	0.858	0.952	0.662
Snare	0.880	0.885	0.955	0.685
Toms	0.860	0.855	0.948	0.654
Open Hi-Hat	0.895	0.857	0.934	0.644
Closed Hi-Hat	0.899	0.863	0.951	0.628
Crash Cymbal	0.922	0.731	0.901	0.544
Ride Cymbal	0.885	0.871	0.957	0.640

Values are shown at IoU=0.5 and averaged over 0.5–0.95, highlighting strong performance on kick, snare, and hi-hat, with lower recall on crash cymbal.

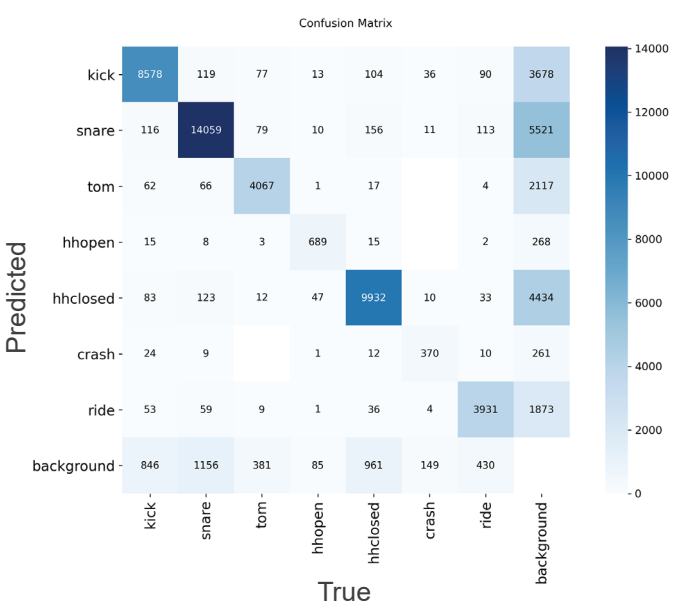

Figure 1. Confusion matrix comparing YOLO11 predictions with ground-truth labels across all drum classes and background. High values along the diagonal indicate accurate classification, while off-diagonal entries show misclassifications (e.g., among cymbals).

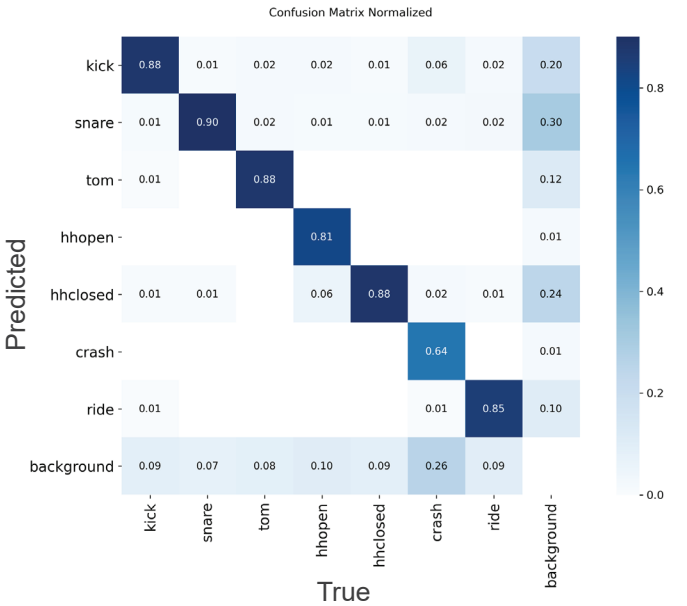

Figure 2. Normalized matrix comparing YOLO11 predictions with ground-truth labels across all drum classes and background, with each column summing to 1. High values along the diagonal indicate accurate classification, while off-diagonal entries show misclassifications (e.g., among cymbals).

a ground-truth onset of the same class, using one-to-one nearest-neighbor matching. Precision, recall, and F1 per class and overall metrics are reported in Table 3. Precision–recall (PR) and F1 curves were plotted by sweeping detection confidence thresholds and are shown in Figure 3 and Figure 4, respectively.

Table 3. Onset-based precision, recall, and F1 scores with ±50 ms tolerance and 0.5 confidence threshold on YOLO predictions, reported per class and overall

Class	Precision	Recall	F1
All	0.881	0.893	0.887
Kick	0.875	0.886	0.880
Snare	0.864	0.910	0.886
Toms	0.874	0.888	0.881
Open Hi-Hat	0.870	0.876	0.873
Closed Hi-Hat	0.908	0.887	0.897
Crash Cymbal	0.917	0.766	0.835
Ride Cymbal	0.897	0.892	0.895

A one-to-one nearest-neighbor matching procedure was applied. These metrics align with evaluation practices in prior ADT work and enable direct comparison with OaF-Drums and other baselines.

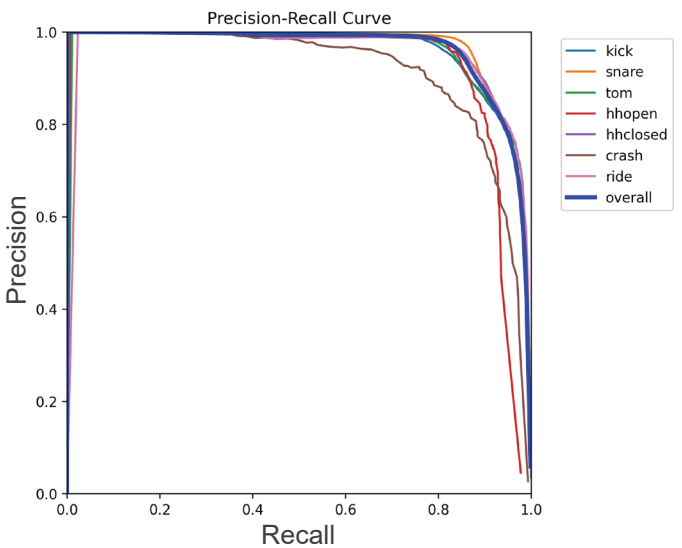

Figure 3. Precision–recall (PR) curves by drum class. Curves are computed by sweeping confidence thresholds on YOLO predictions and evaluating onset-based matches within ±50 ms. This shows the tradeoff between detection precision and recall across instruments.

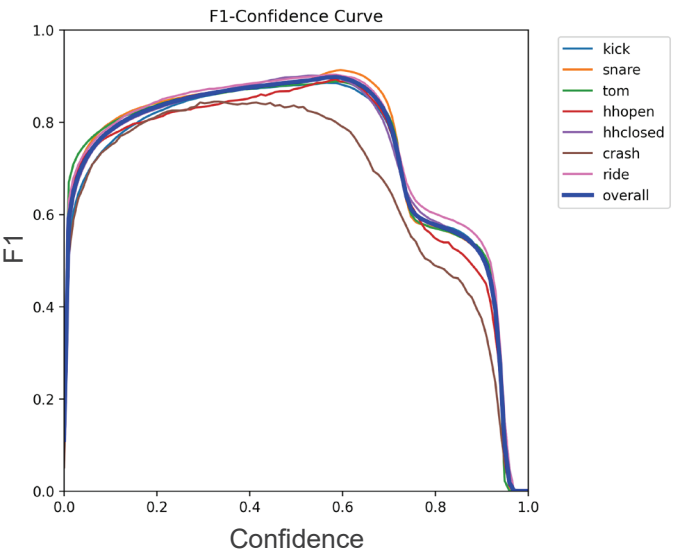

Figure 4. F1 curves by drum class. Curves are computed by sweeping confidence thresholds on YOLO predictions and evaluating onset-based matches within ±50 ms. These curves indicate the operating point where each instrument achieves peak performance.

Qualitative Examples

Figure 5 shows ground truth and predicted bounding boxes for 2-second spectrogram windows, demonstrating successful predictions and also common errors such as occluded notes and predicting multiple boxes for one note. Figure 6 shows about 6 seconds of output from the model when run on a funk drum beat recording and visualized in a symbolic notation, demonstrating a possible real-world application.

DISCUSSION

The results support the viability of reframing drum transcription as an object detection task. By leveraging spectrogram time–frequency structure, YOLO11 detected multiple drum classes concurrently, even in dense rhythmic passages.

A key strength of the work presented by this paper is the ability to handle polyphonic events without explicit onset separation, allowing for more flexibility and easier scaling to more instruments. Additionally, this work is also able to robustly detect prominent drum hits, such as the snare and bass drums, with high precision and recall.

In contrast, some limitations are that low-energy, soft, or occluded hits can often be missed, multiple boxes are occasionally predicted for one note, which

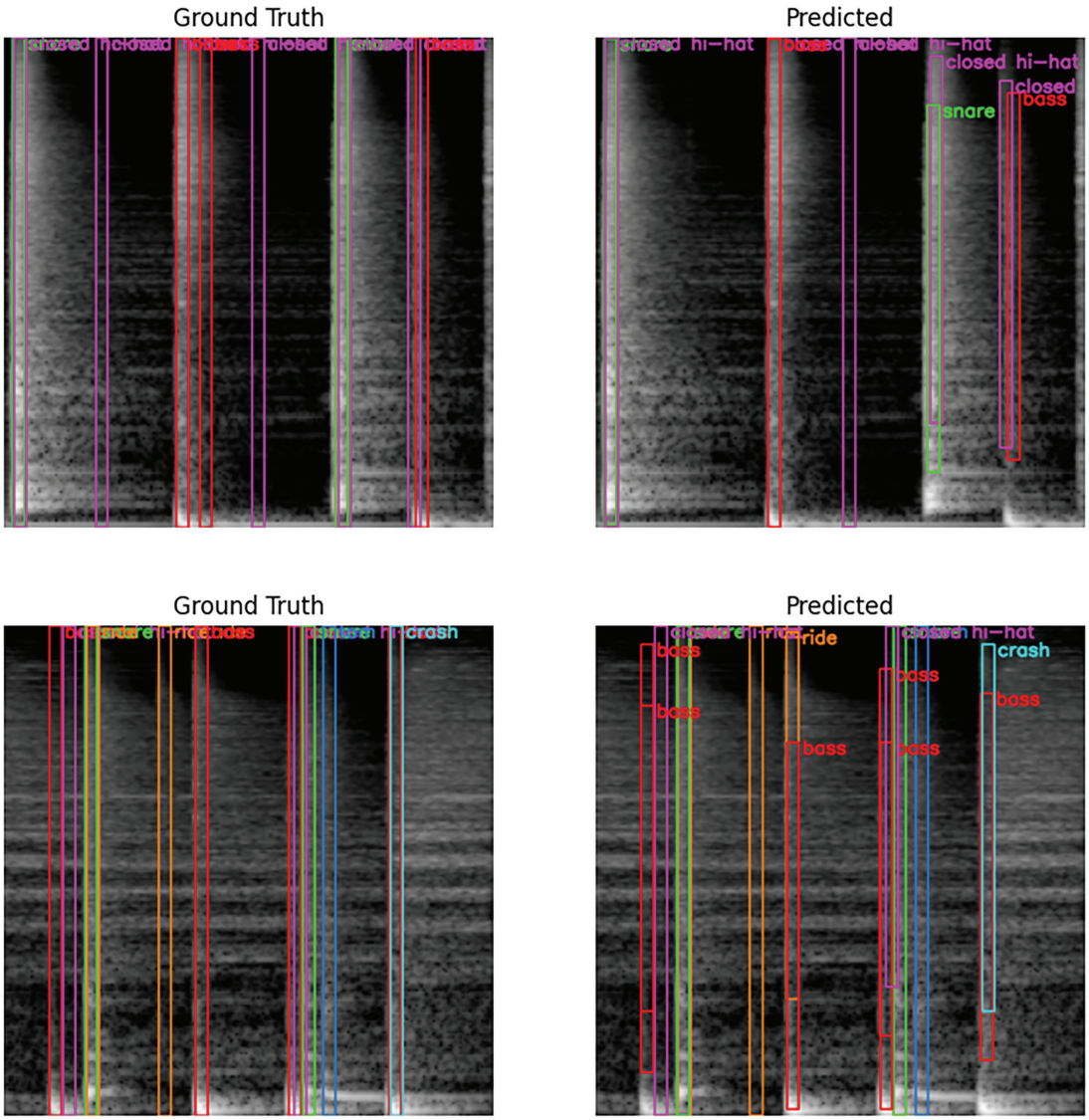

Figure 5. Example spectrogram excerpts (2-second windows) with ground-truth onset annotations on the left and YOLO11 predictions on the right. Successful predications and common errors such as occluded notes and predicting multiple boxes for one note can be seen on the right.

Figure 6. Six seconds of output from the model on a funk drum beat recording and visualized in a symbolic notation, demonstrating a possible real-world application. Time is represented along the width of the image, and different notes are along the height.

is perhaps due to YOLO's anchor settings, and rapid note sequences lead to reduced accuracy due to a fixed bounding box width.

This paper only provides a new baseline approach to ADT, with much growth still left. As the dataset excluded synthetic kits such as 808/909 to focus on acoustic realism, generalization to purely electronic or hybrid drum kits may be limited. Future work will test on mixed acoustic–electronic data.

The dataset exhibits noticeable class imbalance, particularly with fewer tom events compared to kick, snare, or hi-hat. In this work we addressed this by merging tom classes into a single 'toms' category. Alternative strategies, such as using focal loss or weighted sampling, could help further mitigate imbalance by emphasizing underrepresented classes. If finer tom granularity (high, mid, low) is important, a hierarchical classification approach (e.g., first predicting tom-family vs. non-tom, then subclassifying tom type) may improve performance. This classification could potentially use traditional audio techniques for pitch detection. These directions are left for future work.

Finally, future work on this subject can also be potentially focused on experimenting with adaptive bounding box widths to capture varied note durations, incorporating temporal smoothing or context-aware post-processing (e.g., transformer-based sequence modeling), exploring transfer learning from larger audio-version datasets to improve generalization, extending to multi-instrument AMT, and a possible real-time application of this technique.

While the current work focuses on percussion, its core methodology—treating time–frequency audio features as images for object detection—can be generalized to other musical instruments and even non-musical event detection in audio. In MIR, such cross-domain techniques open the door to unified models capable of detecting diverse sound events from a single visual representation. Integrating temporal modeling, adaptive bounding box strategies, and cross-modal learning could further enhance performance, making real-time, multi-instrument transcription systems a practical reality.

CONCLUSION

This study introduces and validates a novel framework for automatic drum transcription by applying the YOLO11 object detection model to spectrogram images, treating drum hits as temporal "objects." The approach achieved high performance (mAP@0.5 = 0.943, precision = 0.892, recall = 0.846) across multiple drum classes, demonstrating the effectiveness of visual object detection techniques in handling polyphonic, densely timed musical events. Beyond percussion, this framework offers a foundation for expanding AMT systems to multi-instrument contexts, hybrid real-time transcription tools, and other MIR tasks where precise event localization is critical. By bridging computer vision and audio signal processing, this work contributes to the growing toolkit for music analysis and opens new directions for cross-domain model design.

ACKNOWLEDGEMENTS

I would like to thank the creators of the Expanded Groove MIDI Dataset for their valuable contribution to the research community, and Gerry Chen for guidance and support throughout this project.

FUNDING SOURCES

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Lee Callender, Curtis Hawthorne, and Jesse Engel. "Improving Perceptual Quality of Drum Transcription with the Expanded Groove MIDI Dataset." 2020. arXiv:2004.00188.
2. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; pp. 779-788. https://doi.org/10.1109/CVPR.2016.91
3. Jocher G & Qiu J. (2024). Ultralytics YOLO11. Available from: https://github.com/ultralytics/ultralytics (accessed on 2025-8-13)
4. Jordal I. "audiomentations: A Python library for audio data augmentation," version 0.42.0. Available from: https://github.com/iver56/audiomentations (accessed on 2025-08-14)