ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

End-to-End Music Braille Transcription from Sheet Music Images

Lihua Chai[1], *Yue Fu*[1], *Zhi Yu*[2,*], *Tianyuan Huang*[3], *Zepeng Zhu*[2] *and Jiaxian He*[2]

[1] Department of Special Education, Zhejiang College of Special Education, Hangzhou, China
[2] School of Software Technology, Zhejiang University, Hangzhou, China
[3] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

SUMMARY: *An end-to-end translation paradigm was proposed to address the problems of error accumulation and low robustness in cascaded music score-to-braille translation methods. For this purpose, a large-scale staff music-braille parallel corpus was constructed, which contains 300,000 sample pairs with music-element-level alignment. An encoder-decoder model was then designed to achieve direct conversion from score images to braille symbol sequences. This was accomplished by jointly optimizing visual feature extraction, musical semantic understanding, and sequence generation. A data augmentation strategy specific to music score characteristics was also introduced to enhance the model's generalization ability towards diverse layouts and image noise. Experimental results show the model's performance is significantly superior to that of cascaded baseline methods. The generated braille exhibits high accuracy in key musical semantics, such as pitch and duration. Furthermore, a strong robustness is demonstrated against different layouts and noisy environments. A new technical solution for efficient and accurate automated braille music production is thus provided.*

KEYWORDS: *End-to-End Music Braille Transcription; Optical Music Recognition (OMR); Staff-to-Braille Parallel Corpus; Encoder-Decoder Model; Hybrid Vision Transformer (Hybrid ViT); Data Augmentation; Error Propagation; Music Accessibility*

1 Introduction

In the ongoing pursuit of information accessibility, leveraging advanced information technology to serve the cultural and educational needs of the visually impaired community has emerged as a research endeavor of profound societal impact and significant academic challenge. According to the World Health Organization, at least 2.2 billion people globally suffer from some form of vision impairment [1], with a large proportion being over the age of 50, presenting unique challenges for education and rehabilitation [2]. Within this domain, ensuring equitable access to music education stands as a cornerstone issue. Music Braille, a tactile system developed by Louis Braille, serves as the primary medium for blind and visually impaired musicians to read, write, and study musical scores. However, a stark reality persists: the global availability of Braille sheet music is alarmingly scarce. This scarcity is not a matter of chance but a direct consequence of an archaic, prohibitively expensive, and labor-intensive production pipeline. The traditional workflow requires a two-tiered expertise: first, a professional engraver must painstakingly transcribe a staff notation image into a structured digital format like MusicXML; subsequently, a Braille specialist must manually translate this symbolic representation, meticulously applying a complex, context-sensitive set of encoding rules. This

*corresponding2025@126.com
https://doi.org/10.65102/is20261078
Published: 10 June, 2026

process, heavily reliant on a dwindling pool of experts with dual proficiency, creates an insurmountable bottleneck between the supply of and demand for accessible music scores, thereby systematically limiting the educational and artistic opportunities available to the visually impaired.

For solving this key problem of accessibility shortage, automatic Braille music translating has already appeared as a hopeful technology method. In the current research, the leading paradigm is the two-step cascade model. This method first uses Optical Music Recognition (OMR) for parsing a sheet music image into a symbolic middle representation, which then is converted into a target Braille sequence through a rule-based system. Although conversion from a pure symbolic form such as MusicXML to Braille can obtain high correctness, the whole efficiency of the pipeline, therefore, is critically dependent on a completely perfect OMR stage, which is a condition that is seldom satisfied in actual practice. This cascaded structure has two basic weaknesses, which greatly reduce its effect in the real world. First, it from its own nature is easy to suffer from error propagation. OMR systems, when they face the huge difference of real world scores—including different arrangements, various topological shapes, printing noise, and picture quality decrease—will certainly bring identification mistakes in important music parts such as pitch and duration. One single wrongly recognized note or one wrongly grouped beam can cascade in a catastrophic manner through the processing pipeline. The afterwards rule-dependent translation machine, which works on this damaged input data, then honestly transmits and frequently enlarges these first mistakes, hence leading to a final Braille export that is not only not correct, but musically makes no sense and actually cannot be used.

Second, this paradigm has the problem that the rule-based systems it uses possess restricted generalization ability. The grammatical rules of music Braille are deeply dependent on the context which surrounds it. The correct coding for symbols like octave signs, chromatic alterations, or playing methods is not settled but is dynamically decided by their relational situation inside a bar, phrase, or even among many musical staves. For give an instance, an octave marking is only utilized when the octave has change and stays to have effect for all following notes till a new marking appears. A system that is built on pre-set, hand-made rules has difficulty in completely covering the nearly infinite combination complexity of music expression. Therefore, these systems show fragile working effect, hence they cannot strongly and correctly deal with musical structures that have not been seen before, complicated polyphony, or style words which are not expected by their designers. This restriction causes them to be not suitable for the abundant variety of music literal works.

In the recent years, the end-to-end (E2E) study model, which has obtained top-level effect in domains such as machine translation and speech identification, provides an attractive option. An E2E model has the goal to learn a direct mapping which goes from the raw input pixel space to the final target sequence, hence therefore it bypasses the brittle intermediate symbolic stage completely. Through carrying out combined optimization on visual feature extraction, music semantic comprehension, and sequence generation inside a unified neural framework, this method can fundamentally avoid the problem of error propagation. In the domain of OMR, recent deep learning models already have exhibited the potential of end-to-end learning for monophonic scores, and even for complex polyphonic scores, therefore highlighting that this paradigm has viability for structured recognition tasks. It possesses the theoretical ability to study the extremely non-linear, complicated conversions from data, hence promising stronger expressional ability and a higher acting upper limit. Nevertheless, the use of this progressive model is faced with a basic difficulty: as a method that relies on data, the effect of a deep learning model is crucially dependent on whether people can get a large-scale, high-quality parallel language corpus. The field of Braille music translation at present has an acute shortage of public accessible, matched data sets of "staff notation image to Braille music" sequences.

This data narrow spot has become the main hindrance that stops progress in the direction of firm, end-to-end settlement schemes.

For breaking this twofold bottleneck of data shortage and model building restrictions, this paper puts forward a whole set framework for direct, end-to-end Braille music translation that comes from sheet music pictures. Our contribution works are planned to systematically solve the above-said difficulties. First, for solving the basic problem of data shortage, we have completed the great work of building a large-scale, high-quality parallel language corpus, which includes 300,000 "staff-to-Braille" sample pairs. This data collection has fine-grained, cross-modal matching of important music elements, which gives the strong experience basis that is needed to train and strictly evaluate data-needing E2E models. Second, by the utilization of this special resource, we put forward a new type of end-to-end translation model which is based on an Encoder-Decoder structure. This model is designed by people for accepting scattered score images and directly produce Braille sequences which are correct in syntax and semantics in one single effective step, therefore thus it shows a feasible road for breaking through the restrictions of cascaded systems. At last, for the filling of the gap between laboratory working effect and real-world application possibility, we have made a multi-dimensional data increase method which is fitted to the special features of music notation. Through the simulation of many kinds of layout distortions, print quality decrease, and image noise, this method therefore greatly increases the model's firmness and generalization abilities, hence ensuring dependable performance under hard, non-ideal situations.

In summary, the primary contributions of this work are threefold:

1) We construct and will release the first large-scale, richly-annotated parallel corpus for staff-to-Braille music translation, providing an invaluable resource to catalyze future research in this under-resourced area.

2) We propose a novel end-to-end translation framework that validates the feasibility and superiority of a direct image-to-sequence conversion path, establishing a new and strong baseline for high-accuracy music Braille transcription.

3) We design a targeted data augmentation technique that yields substantial performance gains, demonstrably improving model robustness against the complexities and imperfections of real-world sheet music.

Collectively, our work provides a practical and scalable technical solution for the automated, high-fidelity production of Braille sheet music, holding significant academic value and profound societal implications for promoting information accessibility and equity in music education.

2 Related Work

Our work lies in the cross position of Optical Music Recognition (OMR) and assistive technology that serves people with visual impairment. Therefore, our review at first discusses the development of OMR from traditional working flows to end-to-end systems, and thus points out their limitations in the tasks which focus on accessibility. We then carry out examination on the domain of Braille information processing, therefore emphasizing the specific, yet unresolved problem of direct music-to-Braille transcribing that our research work faces.

2.1 Optical Music Recognition. From Cascaded Pipelines to End-to-End Systems

Optical Music Identification (OMI) has the target to automatically change sheet music pictures into a machine that can read form [6-8]. The historically leading method has been the cascade

type pipeline, which decomposes the question into a series of separated sub-tasks: staff line checking, symbol cutting apart, sorting, and meaning rebuilding to make a structured output such as MusicXML. Although this modular property provides interpretability, its main disadvantage is mistake propagation: an error that appears in any step, for example a wrongly recognized note or a cracked beam, can forever damage the last output. This therefore lets the whole pipeline inherently have brittleness and be not dependable for the applications that demand high faithfulness.

Along with the coming of deep learning, the capability of single modules has obtained very big promotion. Object detection models have been successfully applied to symbol recognition on large-scale datasets like DeepScores [10] and MUSCIMA++[11], significantly boosting accuracy [12-14]. Certain methods further incorporate Graph Neural Networks to construct the structural connections between musical marks, thus promoting semantic rebuilding. Nevertheless, these progressing achievements have not solved the basic question of error that accumulates passing between different module parts.

For solving this problem, the academic domain has more and more moved in the direction of end-to-end learning. The models which are built on CNN-RNN mixed structures, [16-19], and in newer time, Transformers, have the aim to directly make a raw image map to a symbol sequence. Take Sheet Music Transformer as example, it has proved that it has the ability to change complex polyphonic scores into a linear order of tokens through one single step, and this reduces error propagation effectively. But, a key restriction still exists in all existing end-to-end OMR systems: their target output is always a symbolic form (e.g., MEI, MusicXML, or a self-made text expression) which is made for sight-reading musicians or digital playing back. Not any among them have the design purpose to directly produce music Braille. This place has a not trivial difference, because music Braille is one highly professional language that has own grammar and context-related rules, which these systems do not have equipment to deal with. Therefore, there still exists a vacancy for a genuine entire end-to-end scheme that is customized for accessibility.

2.2 Braille Information Processing and the Music Translation Gap

The domain of Braille information handling has walked along a same technological moving path. The earlier research work depended on rule-based methods and statistic methods, to do text Braille translation and identification, and it often used traditional computer vision technologies such as HOG+SVM or Hough transformation for optical Braille identification (OBR) [25-28]. The development of deep learning has brought about more firm solutions, with models that are based on CNN being constructed for point inspection and character identification from various image sources,. The establishment of special-purpose data collections has further promoted advancement in this domain, hence giving dedicated structural designs for division and identification in multiple languages. In recent time, end-to-end models that are used for textual Braille have appeared, they directly translate images which are of Braille documents into natural language text.

But, a comprehensive scanning exposes a notable unbalance: the great majority of studies is focused on character-based Braille. The special and much more complicated field of music Braille still is a rarely researched boundary area. Although systems that have the ability to convert a neat, already-existing symbolic file such as MusicXML into Braille notation do indeed exist, they work as the last step in a traditional step-by-step work flow. Therefore, their working effect is completely reliant on a flawless, mistake-less input from an OMR system, which, as what we have discussed, is seldom attainable in actual practice. They are not able to directly carry out work on the images of sheet music, hence therefore they are susceptible to this identical problem of error propagation.

According to our knowledge, no earlier study has tried to connect this gap through making an end-to-end framework which changes sheet music pictures directly into music Braille order rows. This challenge is critical, has not been solved, and is the one our work faces. Through combining a strong visual encoder, which is common in current OMR, with a decoder that knows the special syntax of music Braille, and training this system on a large-scale, specially made parallel corpus, we put forward the first overall solution to this important cross-modal accessibility task.

3 Methodology

The objective of this present paper is to build an end-to-end model for visual translation of musical scores. According to what Figure 1 shows, the model put forward by us uses a Hybrid Vision Transformer (Hybrid ViT) structure, which has the ability to directly change staff music pictures into corresponding music Braille ordered rows. The input is made up of segmented music score pictures, while the output is a symbol sequence that conforms to the standards of musical Braille. The core design thought is to make use of the mutual supplementary merits of Convolutional Neural Networks (CNNs) for stable partial feature identification and Transformers for overall environment-based inference. This combined effect effectively solves the special, multi-level structure problems which exist in music scores, from single note parts to wide cross-measure relationships.

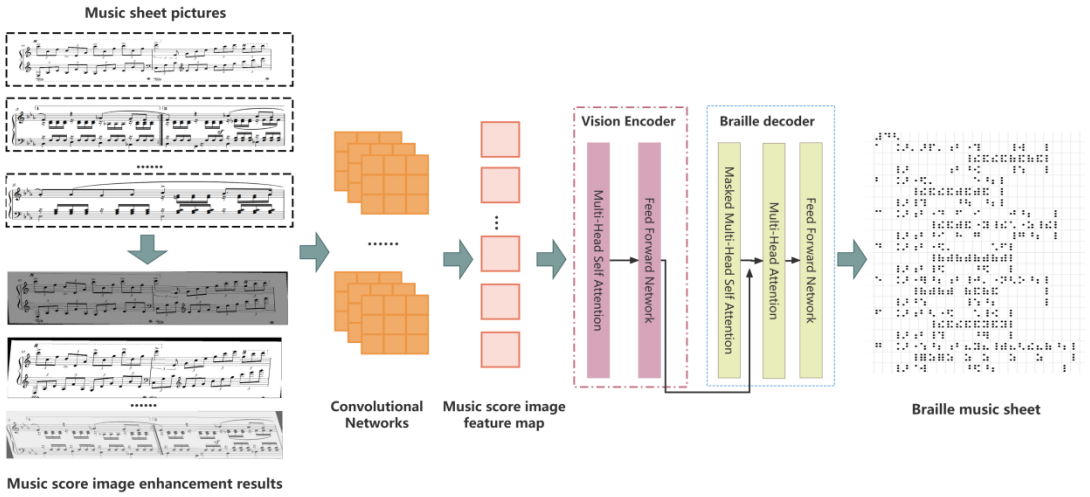

Figure 1: The integral structure of our end-to-end model which is used for translation from sheet music to Braille. This course of action starts from original or enhanced music score pictures, which are put into a CNN backbone for the extraction of local features. These characteristic points then are encoded through a Vision Transformer for the capturing of the global context of music. In the end, one Braille decoder, in an autoregressive way, produces the final Braille music sequence through paying attention to the encoded visual representations.

3.1 Input Normalization and Augmentation for Enhanced Robustness

The depth degree of input data's quality and consistency has a profound influence on the operation effect of deep learning models. Real-world music score documents, especially those that are got through scanning or photographing, display obvious variation in light condition, size, direction, and printing quality. For the guarantee that our model does not change with these surface changes and pays attention to the core music structure, we carry out a two-step input

flow: standardization and expansion.

Firstly, all image materials are transformed into a single-channel gray-scale form, and their sizes are adjusted to the fixed resolution of 512×1280. This unification of norms guarantees the size consistency for group handling. After that, we carry out the normalization processing for pixel values by utilizing the mean value $\mu = 0.7931$ and standard deviation $\sigma=0.1738$ which are calculated beforehand through the whole training dataset:

$$x_{norm}=\frac{x-\mu}{\sigma}$$

This operation carry out transformation on input data, letting them have approximately zero mean value and unit variance. This kind of normalization has the important meaning for stabilizing the training process, because it guarantees that the initial activation values inside the network have good properties, thus prevents problems such as gradient vanishing or gradient explosion, thus promotes the speed of convergence to become faster.

Second, for the enhancement of model generalization ability, we utilize a sequence of intense data augmentations in the process of running. These things contain random affine change methods such as rotation, scaling (90%-110%), and shearing. These specifically imitate common wrong products from not perfect scanning, such as page slant and non-uniform size. We also utilize photometric distortions, which include random modifications to brightness and contrast, to imitate the variations of illumination and ink wear in aged musical scores. Through letting the model contact this broad scope of variations which are produced by algorithm, we force it to study sturdy, basic characteristics of music notation, hence it does not excessively fit to accidental attributes of the training pictures. This point is of utmost importance for the practical application in real world whose input quality can not be ensured.

3.2 Music-Aware Local Feature Extraction with CNNs

The pictures of music scores are not the random placing of picture elements; They are controlled by a strong composition and space rule system. A music note, for instance, is a combination of basic parts such as a note head, a stem, and flags, whose space placement carries very much information. This built local property and translation unchangeability—the thing that a G-clef is still a G-clef no matter where it sits on the page—thus makes Convolutional Neural Networks (CNNs) a very appropriate selection for first-stage character feature taking out. A pure Vision Transformer, which lacks this inductive bias, therefore would have difficulty to learn these basic patterns from original pixels without a very huge amount of data.

Therefore our architecture first uses a deep CNN, namely ResNetV2, to act as a feature extractor which has awareness of music. The pre-activation arrangement of ResNetV2's remaining blocks guarantees no hindrance of gradient moving, thus it permits effective training work for a deep network. When the input picture goes through the convolutional layer levels, the network can learn a hierarchical expression representation. The early-layer structures perform the function of original-type detectors, thus they recognize edges, corners, and staff lines. Deeper layers put these basic elements together into more complex and semantically meaningful "visual morphemes" of music notation, for example complete noteheads, clefs, accidentals, and rests. The output result of the CNN backbone, $F_{backbone}$, is a feature map that retains spatial information, in which each vector on the channel dimension encodes the existence and category of high-level music symbols at that position.

This process essentially serves as a powerful, learned feature engineering step, abstracting the raw image into a more symbolic representation. To prepare this representation for the Transformer, we apply a final projection layer. A 1×1 convolution reduces the channel depth

of $F_{backbone}$ to the Transformer's embedding dimension, D_{model}. This is followed by Batch Normalization (BN) to stabilize the input distribution to the Transformer, and a GELU activation function for its smooth, non-monotonic properties favored in modern architectures. This can be formalized as:

$$F_{embed}=GELU\left(BN(Conv_{1\times1}(F_{backbone}))\right)$$

The resulting feature map, F_{embed}, is a semantically dense and structured representation, perfectly priming the model for the subsequent stage of global contextual analysis.

3.3 Hybrid Embedding for Hierarchical Representation

With a robust local feature map $F_{embed}\in \mathrm{R}^{H'\times W'\times D_{model}}$ from the CNN, the next challenge is to model the long-range dependencies that define musical syntax and semantics. This is achieved via a hybrid embedding strategy that bridges the local-feature space of the CNN with the sequential-processing paradigm of the Transformer.

As conceptually depicted in Figure 1, we first partition the 2D feature map F_{embed} into a sequence of non-overlapping patches. Let each patch have a size of $P\times P$. The feature map is reshaped into a sequence of $N=(H'\times W')/P^2$ flattened patches. Each patch $p_i\in \mathrm{R}^{P^2\times D_{model}}$ is a localized region of the high-level feature map. This patch sequence $Z=[p_1,p_2,\ldots,p_N]$ now represents the entire score image as a series of "visual words," where each word encapsulates the musical primitives within a small spatial area.

For letting these patches become suitable for the Transformer, we carry out flattening processing and conduct linear projection on each one to turn it into a single token embedding. In the key aspect, we add learnable one-dimensional position encodings into this token sequence. Position information is of extreme importance in music: the vertical position of one note on the staff determines its pitch height, and its horizontal position decides its time arrangement. The position encoding $\mathrm{E}_{\mathrm{pos}}$ lets the model carry out reasoning upon these absolute and relative spatial relationships. The last input sequence that the encoder uses is:

$$Z_0=[z_1,z_2,\ldots,z_N]+E_{pos}$$

This mixed method is strong: the CNN effectively deals with the low-level vision work through building a word list of music signs, while the Transformer pays attention to the high-level "language" work of studying the grammar rules that control how these signs are put together to make music.

3.4 Global Music Structure Encoding with Self-Attention

The sequence of embedded patches Z_0 is fed into a multi-layer Transformer encoder. The core of this encoder is the Multi-Head Self-Attention (MHSA) mechanism, which empowers the model to understand the global structure of the music score by dynamically relating every patch to every other patch.

The self-attention mechanism carries out the calculation of a weighted sum of values, in which the weights are decided by the compatibility (or "attention score") between one query and one group of keys. To the token which is at position i, its query $\mathrm{q_i}$ pays attention to all keys $\mathrm{k_j}$ that belong to other tokens, for the purpose of confirming how much attention should be put on their corresponding values $\mathrm{v_j}$. This flow is defined as:

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This mechanism is profoundly effective for modeling musical notation for two key reasons:

The Establishment of the Model for Vertical Co-occurrence (Harmony): In regard to one chord, many noteheads are aligned on the vertical direction. When the model carries out processing on a patch which includes one notehead of a chord, self-attention permits it to assign high attention scores to patches which are vertically aligned within the original feature map. Therefore, this permits the model to perceive the spatially separate noteheads as one single harmonic unit, hence this is essential for correct Braille transcription, in which chords are linearized in a specific order (e.g., from bottom to top).

The Establishment of Models for Horizontal Long-Distance Dependencies (Melody and Rhythm): The structures of music such as slurs, ties, repeat signs often cross over multiple measures and even complete music systems. A traditional CNN which has a limited receptive field would have difficulty to capture these long-distance connections. Self-attention, on the other hand, is able to directly link a block at the start of a slur to a block at its end, no matter what the distance between the two is. This enables the model to keep semantic consistency in the whole music section, hence making this long-range context not get lost.

Through the piling of multiple MHSA layers, the encoder constructs an ever more abstract and context-perceiving representation of the whole score image, thus capturing the complex net of connections that determine the music.

3.5 Autoregressive Braille Sequence Generation

After the encoder has produced a contextually rich representation of the music score, denoted as $\mathbf{F}_{enc}$, the decoder's task is to autoregressively generate the corresponding sequence of musical Braille symbols, $\mathbf{y}=(y_1,y_2,\ldots,y_T)$. This process is particularly challenging because musical Braille is not merely a one-to-one transliteration of musical symbols; it is a highly structured, context-dependent language with its own grammar. For example, octave markers are not repeated for every note but are stated once and remain active until a new octave marker appears. Similarly, chords are linearized into a sequence of interval prefixes followed by notes. Generating a syntactically valid Braille sequence thus requires not only recognizing visual symbols but also understanding and maintaining state over long musical passages.

To address this, we employ a standard Transformer decoder architecture, which excels at such structured sequence generation tasks. The decoder consists of a stack of identical layers, each containing three key sub-modules: masked self-attention, cross-attention, and a feed-forward network. At each decoding step t, the decoder takes as input the previously generated Braille tokens $(y_1,\ldots,y_{t-1})$ and the encoder's output $\mathbf{F}_{enc}$.

(a)Covered Self-Attention: This first sub-module, covered self-attention, permits the decoder to carry out processing on the already partially produced Braille ordered sequence. This calculation is done to get attention scores between every token $(y_1,\ldots,y_{t-1})$, therefore it lets the model hold the internal dependencies and grammatical rules of the Braille music language. The masking mechanism makes it so that the prediction of token y_t is only able to rely on the earlier tokens, hence it maintains the autoregressive property. This point is of vital importance for the correct processing of stateful Braille structural units. For example, when we have produced an octave mark (for example, "5th octave"), the self-attention mechanism lets the model "memorize" this condition when it produces later notes, hence until a new octave mark is needed and produced.

(b) Cross-Attention: Let Predictions Get Root in Visual Proof. The second, and the most critical sub-module, is the cross attention mechanism. It is exactly this place that the textual (Braille) and visual (score image) modalities undergo a process of fusion. The partially produced Braille order (expressed as queries from the self-attention output) pays attention to the output keys and values of the encoder (F_{enc}). In the formal perspective, the query vectors Q_{dec} are obtained from the decoder, meanwhile, the key vectors K_{enc} and the value vectors V_{enc} are obtained from the encoder:

(c)This mechanism enables the decoder to dynamically place its next prediction on the most related visual proof from the score. For instance, when the decoder makes decision on whether produce a note or a rest, it can give attention to the specific image region which corresponds to the current time position. When we do transcription work for a complex chord, the cross-attention mechanism is able to focus on those noteheads which are stacked in the vertical direction, thus it sequentially extracts the information which is needed for generating the linearized Braille representation. This direct connection with vision input reduces the danger of illusion and guarantees the produced sequence keeps loyalty to the original music score.

At the last, the output which comes from the cross-attention layer is passed by a position-wise feed-forward network, hence the whole of this process is repeated through the stack that consists of decoder layers. The output of the final layer is projected onto the dimension of the Braille vocabulary through a linear transformation, which is followed by a softmax function for generating a probability distribution on the next possible token $p(y_t|y_{<t},F_{enc})$.

3.6 Optimization with Masked Cross-Entropy Loss

This model is trained from end to end through the minimization of the discrepancy that exists between the predicted Braille sequence and the ground-truth sequence. Because the target sequences have different lengths, a standard calculation of cross-entropy loss is not enough, because it will be biased by the random quantity of filling tokens that are used to group sequences of different lengths. These filling tokens do not carry any semantic information, and the putting of them into the loss calculation therefore would bring noise into the gradients, hence making training not stable.

For solving this problem, we utilize a Masked Cross-Entropy Loss function. The core thought is to carry out calculation of the loss only on the real, non-padding tokens that are in the target sequence. The loss function L which belongs to one single training example is defined as the negative logarithmic likelihood of the real correct sequence, averaged on its all effective tokens:

$$L=-\frac{1}{\sum_{t=1}^{T} I(y_t \neq y_{pad})} \sum_{t=1}^{T} I(y_t \neq y_{pad}) \cdot \log p(y_t|y_{<t},\mathbf{F}_{enc})$$

where:

y_t is the ground-truth token at timestep \$t\$.

y_{pad} is the special padding token.

$p(y_t|y_{<t},F_{enc})$ is the probability which is predicted by the model for the token y_t.

$I(\cdot)$ is the function that marks, which is equal to 1 when the input parameter of it holds true, and 0 in other cases. This function acts in the role of the mask, therefore it effectively makes zero the loss contribution which comes from padding tokens.

The denominator is one normalization factor, which equals the real length of the sequence, therefore it makes sure the loss for every sequence is on a comparable scale, no matter what its

original length is.

Through the adoption of this masking strategy, we can ensure that the parameters of the model are updated only on the basis of its capability to predict meaningful Braille symbols. This therefore brings more stable training motion, quicker convergence, and hence finally, a model that is more accurately adjusted to the work of correct Braille music character conversion writing. The overall loss which belongs to one batch is the average value of the losses that are calculated for every sequence inside this batch.

4 Experiment

4.1 Experimental Dataset

Because there is no public paired data set which is made of staff notation pictures and corresponding music Braille sequences, this study designs and builds a new, high-quality data set for supporting the OMR to Braille transcription work. The construction flow of data set is shown in Figure 2, it includes the following key steps:

Intermediate Expression and Source Obtaining: We have adopted MusicXML to be the middle annotation form for data producing. MusicXML is an opening standard for the expression of musical notation, it can structurally and accurately code every element in a music score, including pitch, duration, rests, key signatures, time signatures, dynamic marks, slurs, and other expressive symbols. A diverse collection of MusicXML files is downloaded from the open-source MuseScore community to ensure musical variety and representativeness in the dataset.

Staff Notation Image Rendering: Using the command-line batch processing functionality of MuseScore 4, the collected MusicXML files are automatically rendered into high-resolution staff notation images in PNG format. Rendering parameters—like page side empty space, proportion parameters, and staff dimension—are made fixed in all files to guarantee look consistency and cut down unimportant changes that are brought by rendering differences.

Picture and MusicXML Segmenting with Alignment: For realizing fine-grained structure matching between staff pictures and MusicXML notes, we use an automatic cutting and matching method that is based on projection analysis. In specific words, the high-resolution images that have been produced are at the first step converted into grayscale and then into binarized form. Then we carry out the computation for a horizontal projection outline of pixel intensity. Through detecting periodic low points in the projection (which correspond to positions of staff lines) and carrying out statistical clustering upon the spaces between lines, the boundary limits of every music system (that is, one horizontal block of staves) can be robustly obtained by identification. People afterwards cut this image into many small pieces, every piece includes one or more whole measures. At the same time, the original MusicXML document is divided into parts on the basis of its internal <measure> structure, and the breaks of system are decided by the <print new-system="yes"> label. This therefore guarantees severe time and structure corresponding relation among image pieces and XML segments. The consistence that the number of image segments and XML segments have further gives verification to the correctness of their pairing.

Generation of Braille Label Sequences: We have realized a rule-based conversion engine from MusicXML to Braille. The said engine is composed of two core part compositions:

(1) Symbol Mapping Table: A all-sided, one-to-many lexicon which stipulates the mapping from every MusicXML symbol (for example, one quarter note C4, sharp mark, 4/4 time signature) to its corresponding Braille Unicode character(s).

(2) Sequential and Structure Rules: One group of syntactic rules which is based on the

International Standards for Braille Music Notation, these rules handle connection relations among music elements, and organize the mapped Braille symbols into a linear order that accords with Braille reading and writing customs. This contains processing chords (vertical piled notes changed into horizontal series), beat grouping, octave markers, and the conversion of special symbols (e.g., dynamics and articulations).

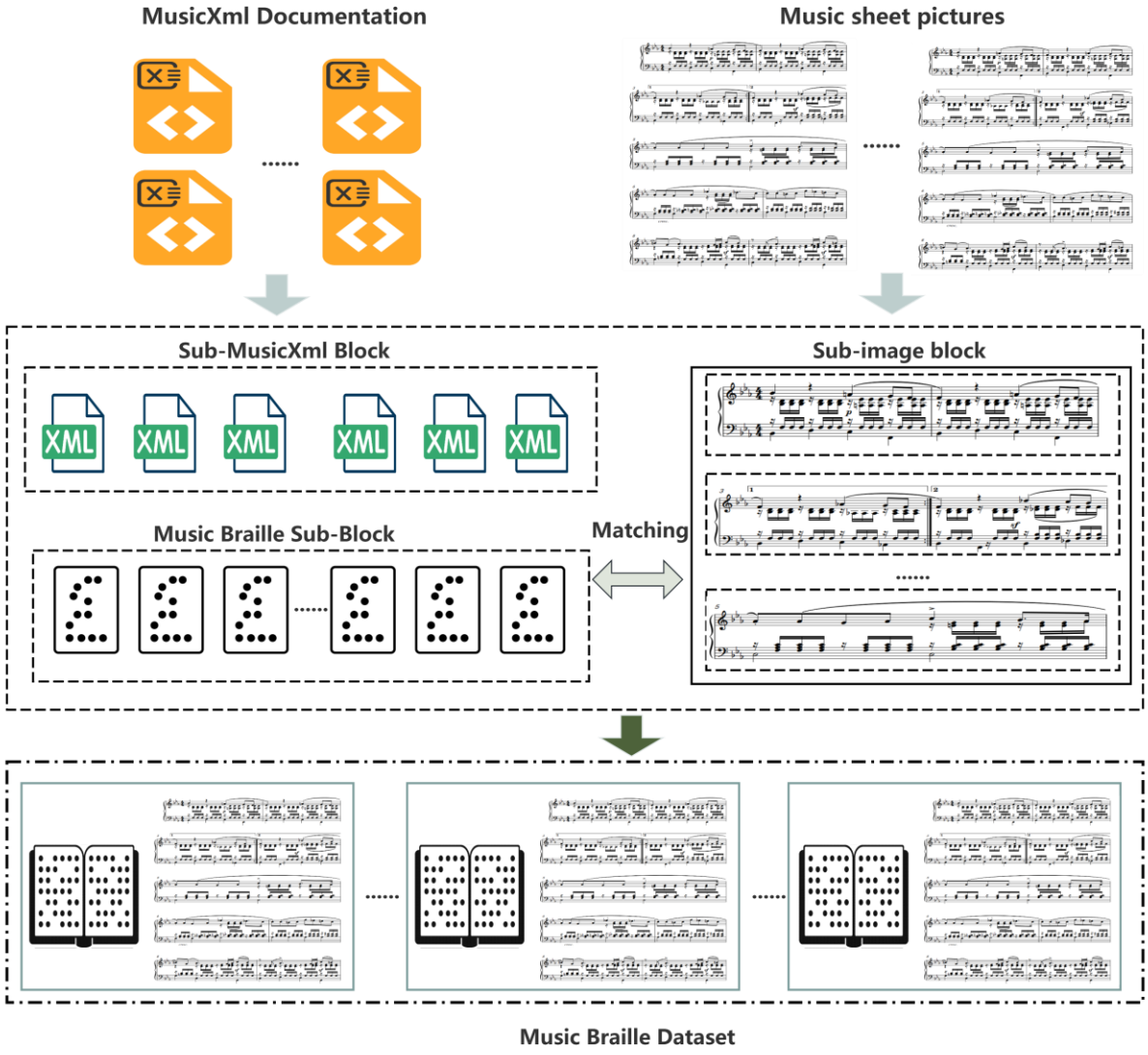

Figure 2: Music Braille Dataset Production Flowchart

Man-made Checking and Quality Control: The automatically produced paired data can have mistakes or unclear meanings. For the purpose of guaranteeing high annotation accuracy, we have introduced one manual verification stage. Experts who are skilled at music Braille carry out inspections based on samples, with focus on amending transcribing mistakes in complicated music structures—such as multiple voices, decorative notes, and rare note marks. This step greatly promotes the quality of the data set, making it not only fit for data-driven model training but also a dependable reference standard for performance assessment.

4.2 Dataset Statistics and Analysis

For the purpose of depicting the vocabulary and structure constitution of the data set, we carry

out classification for the Braille transcribing sequences into semantic token types and thus analyze their distribution situation. This analysis on data level discloses the frequency and diversity of music elements that are covered, hence it reflects that the dataset has capability to support the learning of both basic notation patterns and complex expressive structures.

Table 1 give out the statistical data and representative instances for main token classification groups. According to what has been displayed, Note_Root_Tokens occupy the main position of vocabulary with more than 12.4 million occurrences, which correspond to note time values such as quarter, eighth, and half notes—this forms the core rhythm and melody content. Clef_Tokens (430,218), Key_Tokens (308,928), and Time_Signature_Tokens (217,456) offer the necessary context-related information which is used for pitch and meter explanation.

Table 1: Statistics and Examples of Music Token Categories in Our Dataset

Token Category	Count	Examples
KEY TOKENS	308.928	C major, G minor, F# major
TIME_SIGNATURE_TOKENS	217,456	4/4,3/4,6/8
CLEF_TOKENS	430,218	Treble clef, Bass clef, Alto clef
NOTE_ROOT_TOKENS	12,488,438	Treble clef, Bass clef, Alto clef
DYNAMICS_TOKENS ROOT	602.617	Half note, Quarter note, Eighth note
REPEAT_TOKENS	3,422	Start repeat, End repeat, D.C. al Fine

This dataset also contains a very large quantity of expressive components, such as Dynamics_Tokens_Root (602,617), which cover dynamic marks including p (piano), f (forte), and mf (mezzo-forte), these are of great importance for the transmission of interpretation intention. Structural guiding marks—for example, "Begin repeat", "Finish repeat", and "D.C. al Fine"—are included in the Repeat_Tokens classification (3,422 cases), therefore allowing precise modeling of complicated music structures even when their occurrence frequency is not high.

This token-level decomposition verifies that our data collection includes a big scope of music constructions, from basic written symbols to high-level expressive and structure meanings. This balanced and moreover realistic distribution can support the training of robust end-to-end models which have the ability to handle diverse inputs that are rich in music, therefore hence making it a valuable resource for the research work in the domain of music accessibility and multimodal translation.

4.3 Experimental Setup

For the measurement of the consistency that exists between the generated braille music sequences and the reference sequences, this study utilizes two metrics that have wide use: BLEU and ROUGE.

The BLEU (Bilingual Evaluation Understudy) metric carries out the evaluation of model performance through the computation of the n-gram overlapping between the sequences that are generated and the reference sequences. Its core thought is that higher n-gram accuracy shows better local pattern matching with the reference text. The calculation of BLEU score is carried out as:

$$BLEU=BP\cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where p_n denotes the precision of matching n-grams between the generated and reference sequences, w_n represents the weight assigned to each n-gram order (typically set to uniform

weights, i.e., $w_n=\frac{1}{N}$), BP is the brevity penalty that penalizes overly short outputs. Specifically, BP=1 if the generated sequence length exceeds or equals the reference length; otherwise, $BP=\exp(1-\frac{r}{c})$, where r is the reference length and c is the candidate length.

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) family of metrics focuses on recall, measuring how well the generated sequence covers the content of the reference. Common variants include ROUGE-1, ROUGE-2, and ROUGE-L, defined as:

$$ROUGE\text{-}n=\frac{\sum_{ref\in R}\sum_{gram_n\in ref}Count_{match}(gram_n)}{\sum_{ref\in R}\sum_{gram_n\in ref}Count_{ref}(gram_n)},$$

where $gram_n$ denotes an n-gram in the reference sequence, $Count_{match}(gram_n)$ is the number of n-grams co-occurring in both the generated and reference sequences (clipped by the reference count), and the denominator represents the total number of n-grams in the reference. ROUGE-L refers to the longest common subsequence-based F1 score.

4.4 Experimental Environment

We carry out experiments upon a large-scale self-built data set which consists of paired staff notation pictures and corresponding braille music sequences. All models have been evaluated upon the identical test collection utilizing standard sequence making metrics from natural language process: BLEU, ROUGE-1, ROUGE-2, and ROUGE-L. These measurement indexes together carry out assessment on the similarity that is between generated braille sequences and reference braille sequences. To speak specifically, BLEU puts emphasis on n-gram precision, while ROUGE metrics put emphasis on recall, hence they provide a balanced evaluation of accuracy and completeness.

The training of models is conducted by using the Adam optimizer, which has an initial learning rate of 1×10^{-4}, and is combined with the strategies of learning rate warmup and cosine annealing. The input images are resized into a fixed resolution, and the training work is carried out with a batch size of 16 through 20 epochs on one single NVIDIA A40 GPU. In the decoding stage, the beam search that has a beam size of 5 is utilized to produce the ultimate braille symbol sequence.

4.5 Ablation Study

For researching the contribution that different components give to model performance, we have carried out a systemic ablation study, and it focuses on the influence of image augmentation strategies upon model robustness and generalization. Concretely speaking, we make a comparison between two kinds of training set collocations on one data collection which has 100,000 samples: one kind has the image increase transformation and the other one does not have it. The evaluating normatives include BLEU, ROUGE-1, ROUGE-2, and ROUGE-L. The obtained results are summarized within Table 2.

The experiment's outcome data indicate that the bringing in of image augmentation brings about obvious enhancements in every evaluation measurement index. Concretely speaking, the BLEU score has an increment of 1.7%, ROUGE-1 has an increment of 2.2%, therefore ROUGE-2 and ROUGE-L obtain improvements of 2.1% and 2.2%, respectively. This promotion shows that through simulating real world degradations like noise, illumination changes, staff line blur, and geometry distortions, image increasing effectively strengthens the model's adaptive ability to various input conditions. The augmentative training lets the model attain higher stability and fault-bearing capability. Therefore, data expanding not only promotes total performance but also thus greatly promotes the model’s robustness in actual application

scenarios.

Table 2: Image Enhancement Contrast Chart

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Translation Model	0.954	0.966	0.961	0.962
Sheet Music Enhancement	0.971	0.988	0.982	0.984

4.6 Analysis of Comparative Experimental Results

For carrying out systematic research on the influence that training data size has upon model convergence and generalization, we have done training of the end-to-end music-to-Braille transcription model under three kinds of configurations: 10k, 100k, and 300k synthesized samples. All model types had their evaluation done on the same kept-aside test set through utilization of BLEU and ROUGE score metrics.

According to what is displayed in figure 3, the model's performance obtains dramatic enhancement when the data size becomes larger. When there are only 10k samples, this model obtains a BLEU score of 0.322, hence it shows not enough study of the complicated structure mapping that exists between music notation and Braille coding. When the data quantity grows to 100k, the BLEU score rises to 0.954, therefore it shows powerful learning ability and exact translation action. Further enlarging the scale to 300k samples gives a small promotion to 0.972, therefore it shows that performance has already reached saturation. This tendency shows that large-scope, top-grade artificial data is necessary for training deep nerve models in this field, especially for capturing complex rhythm structures, multi-voice organization forms, and high-level symbol systems.

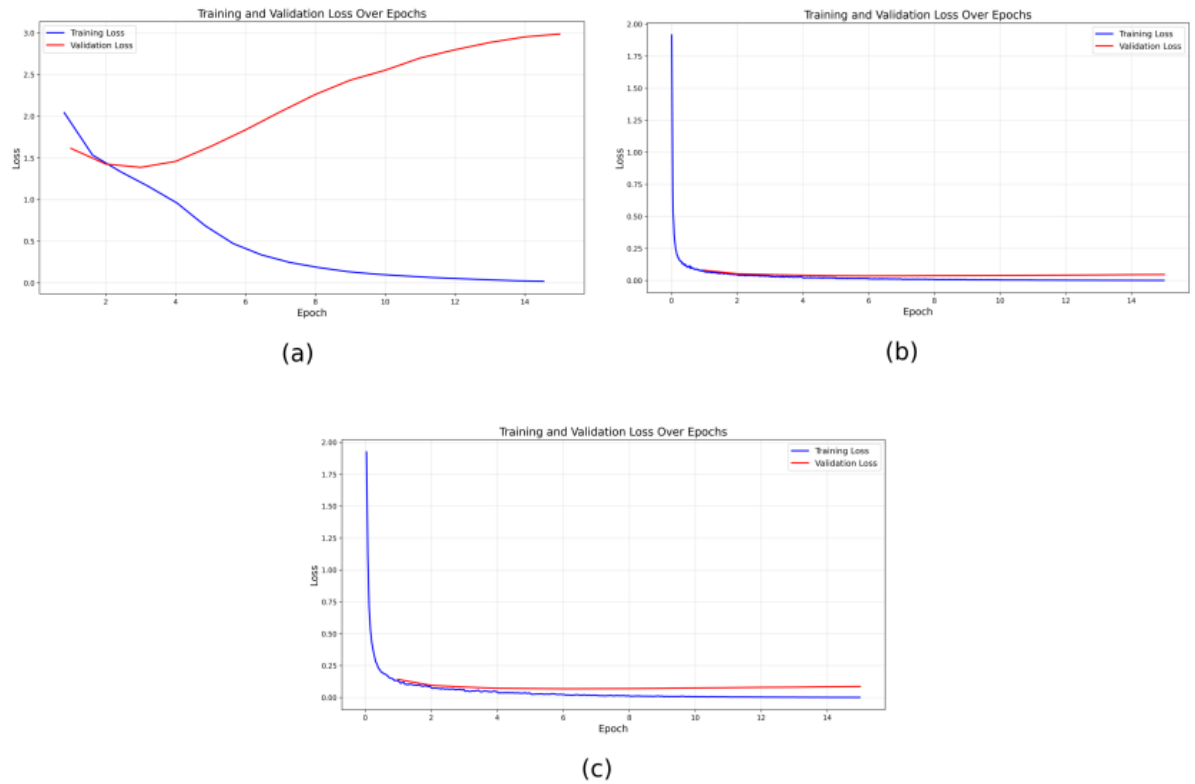

Figure 3: centering Model performance on loss with different training data scales.

A deeper inspection of the training moving processes (Figure 4) therefore discovers key

understandings regarding convergence and generalization. Under the 10k environment, the training loss continuously drops from the starting value of 2.05 in the direction of zero, this shows that the complete memorization of the training set is achieved. But the verification loss at first reduces, after that it increases, this is an obvious sign of overfitting. This gives the indication that when data diversity is not large enough, the model is not able to study patterns that can be generalized, and hence it instead overfits to examples which have unique characteristics.

By comparison, the 100k and 300k setups all show the tendency of continuous decrease in both training loss and validation loss, and the difference between these two curves is very small. We do not see any upward inflection in the validation loss, this confirms that optimization is stable and generalization has improvement. The large number of training samples lets the model study robust feature expressions that hence can perform good generalization on inputs which have not been seen before.

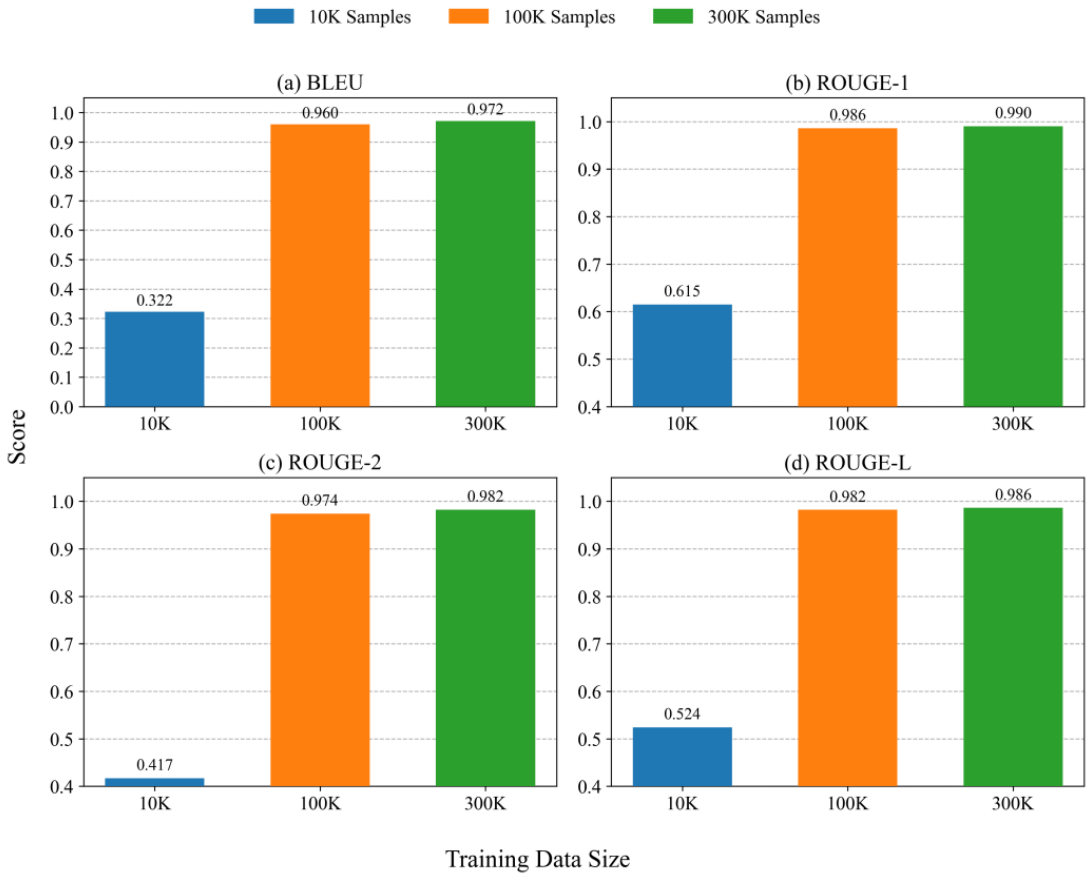

Figure 4: The performance of the model on BLEU and ROUGE metrics under different training data size situations. The outcome demonstrates a big promotion when one scales from 10K to 100K samples on all evaluation norms. When we further increase the data size to 300K, the promotion of performance becomes very small, which shows that we have already arrived at the position of decreasing repayment. This point hence emphasizes the very important function of a big-size data group for the training of the end-to-end model in an effective way.

To make the summary, the results have proven that data scale not only promotes the final translation accuracy but also in the fundamental way makes the training process stable and thus enhances the generalization ability of the model. The overfitting that has been observed in the small-data domain has put high light on the frangibility of end-to-end learning which is without

enough supervision, while the unchanging promotions that come with bigger data collections hence highlight the significance of extensible data generation for constructing dependable assistive music technologies.

4.7 End-to-End vs. Two-Stage Pipeline

For the purpose of evaluating the merits of the end-to-end transcription framework which we have put forward, when compared with traditional methods, we carry out a comparison between it and a typical two-stage working flow. The tradition method is constituted by two steps that follow order: first, an OMR system makes analysis on the input staff notation image to produce one middle symbol expression—usually it uses the MusicXML format; Second, a rule-based change engine changes this symbolic mark into a braille music row according to the standard braille music notation rules which are already set. Although this kind of module-type design provides the possibility of component explanation and repeated utilization, it naturally has the issue of error transmission: any wrong situation that occurs in the OMR step—like wrongly recognized note pitch values, wrong beam grouping, or missed dynamic marking symbols—will be brought to the following steps, and often get amplified in the follow-up rule-based conversion work, therefore it causes braille output that has wrong semantic meaning or invalid structure.

By comparison, our end-to-end frame structure skips the middle symbol expression completely, and through one united training nerve model, maps the input picture directly to a braille symbol row sequence. This method removes the reliance on possibly noisy and unfinished symbol outputs from OMR systems, therefore it reduces the danger of continuous mistakes. Through carrying out optimization on the whole transcription process in an end-to-end way, the model can learn to pay attention to visual features which have the most relation to correct braille producing, instead of being limited by the not-good enough discretization and parsing choices that come from an independent OMR module.

According to what Table 3 shows, the results of experiment prove that the end-to-end method we put forward has obviously better performance than the two-stage baseline method on all evaluation metrics. The end-to-end model has obtained a BLEU score which is 0.972, hence it represents a 15.4% relative enhancement compared with the two-stage model (0.842). In the same way, it obtains improvements of 7.0% on ROUGE-1 (0.990 compares to 0.925), 10.3% on ROUGE-2 (0.982 compares to 0.890), and 7.3% on ROUGE-L (0.986 compares to 0.919). These continuous promotions on many measurement indexes therefore prove that the braille sequences produced have more good fluency, accuracy, and structure loyalty.

To make a summary, the end-to-end framework has obvious advantages when compared with the two-stage method. Through the method of directly producing braille sequences from pictures, it is therefore able to avoid error spreading that comes from middle OMR output results, and hence allows the whole transcription process to be jointly optimized. The model's whole-domain attention method can effectively seize long-distance music structures, for example cross-measure links and many-voice textures, hence it can bring more accurate and consistent braille output. In addition, the simplified work flow promotes robustness and deployable ability. These advantages are always manifested in higher-level achievements through all measurement standards of assessment.

Table 3: End-to-End vs. Two-Stage Model Performance Comparison

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Two-Stage Model	0.842	0.925	0.890	0.919
End-to-End Model	0.972	0.990	0.982	0.986

4.8 Qualitative Analysis and Error Case Study

Although the quantitative measurement methods such as BLEU and ROUGE can give a high-level outline of working effect, the qualitative analysis about the error patterns can provide the deeper understanding toward the actual merits of our end-to-end frame. Figure 5 gives a contrast case research on a complicated music section that includes polyphony, it shows the usual mistake modes of every model and thus exposes the basic weaknesses of the cascaded method.

This example sufficiently manifests that the two-stage pipeline easily has catastrophic breakdowns, thus it generates big chunks of "hallucinated" or \textbf{false content} (marked in green) which do not have any matching relation with the original music. This kind of mistake is the direct result that comes from the error propagation problem. In this special situation, the first OMR module was not able to correctly analyze the polyphonic structure of the music. This system had no ability to tell apart and split the upper and lower sounds inside the five-line staff, therefore it led to an intermediate MusicXML expression which had structure problems, in which notes coming from the two sounds were wrongly put together or wrongly understood. The following rule-based converter, though it is logically correct, then accurately translated this damaged MusicXML, hence leading to a semantically senseless and syntactically incorrect Braille sequence. This model is in essence carrying out a "correct" translation work upon "rubbish", therefore it causes a thorough break in the faithfulness of transcript recording.

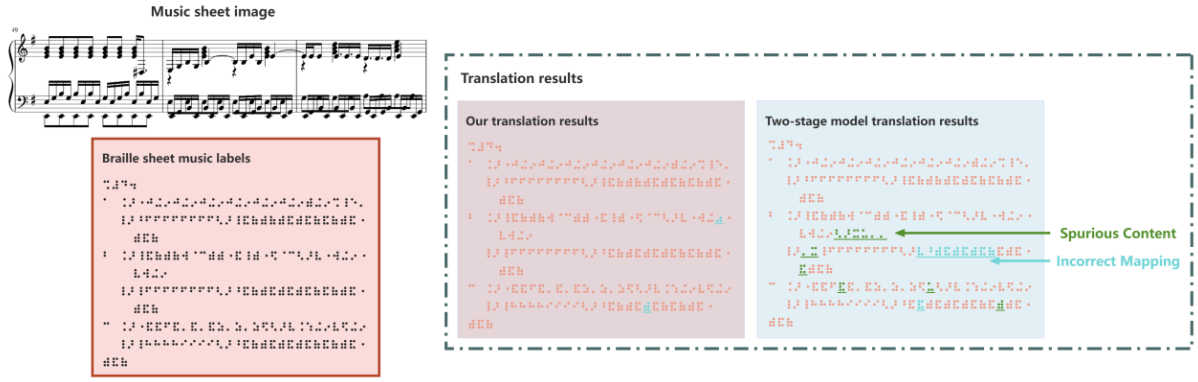

Figure 5: The qualitative comparison that shows the end-to-end model of us has superiority compared with the two-stage pipeline. The two-stage model (top right) undergoes serious mistake expansion, thus causing big problems like the production of big sections of \textbf{false content} (green) and many cases of \textbf{wrong matching} (blue). Although the translation result of our model (which is at the bottom right) is not completely without mistakes, as what one small matching error has shown, hence it has obtained a very big decrease in translation mistakes. Most significantly, it reduces the very serious illusion problem, thus showing its improved firmness and usefulness in lessening the continuous breaking problems that often appear in multi-step systems.

By the strong opposite comparison, our end-to-end model shows astonishing anti-interference ability toward these complicated situations. According to what its output shows, the serious hallucination problem is completely got rid of. This model can correctly recognize that two mutually independent voices exist, and can accurately produce the corresponding linearized Braille sequence, this sequence obeys the standard rule for expressing polyphony (for instance, first write down one voice then write another, with suitable indicators). This achievement can be hence attributed to the architectural design that the model has. The global self-attention mechanism that the vision encoder has is not subjected to constraints by local parsing decisions. It is able to process the whole visual environment of the measure at the same

time, therefore permitting it to study the hidden regulations of vertical arrangement and time overlapping which determine polyphony. Through directly throwing these learned picture patterns onto the right Braille output, our model walks around the demand for an explicit, complete symbolic expression, hence it proves firm in facing the type of structure analysis mistakes which bring trouble to the two-stage working flow.

This architecture superiority goes beyond multi-voice to other complicated music structures that include long-distance dependencies. As an example, a tie or a slur which stretches over a measure or system break is able to bring a big difficulty to a two-stage model when its OMR part wrongly divides the picture at the boundary, thus breaking the meaning connection. The self-attention mechanism of our end-to-end model may with ease build connections among faraway visual patches—for example the beginning and the ending of a slur—no matter what there are measure lines between them, thus it preserves the important musical context and hence guarantees a correct translation.

To give a summarization, this case research offers solid proof that the quantitative advantage of our model (which is shown in Table~\ref{table:end2end_vs_cascade}) therefore comes from a basic structural superiority. Through the overall study of the mapping relationship between pixel points and Braille, our end-to-end framework can effectively reduce the possibility of cascaded mistakes, therefore obtaining much more stable and correct transcription results, especially when we encounter the complicated, multi-layer structures which are existing in the actual music score of real environment.

5 Conclusion

This current work puts forward an end-to-end model for Braille music translation, which through the joint optimization of visual understanding and sequence generation, realizes high-accuracy direct conversion from sheet music images into Braille sequences. Our research model proves that one-step mapping can basically avoid the error gathering that exists inside series structures, hence greatly getting better results than traditional two-step methods. The put-forward framework gives a new, expandable method for automatic, high-precision Braille music making, thus having quite big theoretical and practical meanings. Even so, difficulties still exist in dealing with complicated real situation examples like handwritten music scores or extremely damaged pictures—future research will put emphasis on increasing training data and improving model firmness to solve these problems.

Acknowledgment

This work was supported by the China Disabled Persons' Federation under the Special Research Program on Assistive Devices for Persons with Disabilities (Project No. 2024CDPFAT-42).

References

[1] World Health Organization, “World report on vision,” World Health Organization, Geneva, 2019.

[2] A. Mariotti and D. Pascolini, “Global estimates of visual impairment,” British Journal of Ophthalmology, vol. 96, no. 5, pp. 614–618, 2012.

[3] D. Goto, T. Gotoh, R. Minamikawa-Tachino, and N. Tamura, “A transcription system

from MusicXML format to Braille music notation," EURASIP Journal on Advances in Signal Processing, vol. 2007, no. 1, p. 042498, 2006.

[4] A. Liu, L. Zhang, Y. Mei, B. Han, Z. Cai, Z. Zhu, and J. Xiao, "Residual recurrent CRNN for end-to-end optical music recognition on monophonic scores," in Proceedings of the 2021 Workshop on Multi-modal Pre-training for Multimedia Understanding, 2021, pp. 23–27.

[5] A. Rios-Vila, J. Calvo-Zaragoza, and T. Paquet, "Sheet music transformer: End-to-end optical music recognition beyond monophonic transcription," in International Conference on Document Analysis and Recognition, Springer, 2024, pp. 20–37.

[6] J. Calvo-Zaragoza, J. Hajič Jr., and A. Pacha, "Understanding optical music recognition," ACM Computing Surveys (CSUR), vol. 53, no. 4, pp. 1–35, 2020.

[7] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. Marcal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," International Journal of Multimedia Information Retrieval, vol. 1, no. 3, pp. 173–190, 2012.

[8] E. Shatri and G. Fazekas, "Optical music recognition: State of the art and major challenges," arXiv preprint arXiv:2006.07885, 2020.

[9] F. J. Castellanos, C. Garrido-Munoz, A. Rios-Vila, and J. Calvo-Zaragoza, "Region-based layout analysis of music score images," Expert Systems with Applications, vol. 209, p. 118211, 2022.

[10] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, and T. Stadelmann, "DeepScores—a dataset for segmentation, detection and classification of tiny objects," in 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 3704–3709.

[11] J. Hajič and P. Pecina, "The MUSCIMA++ dataset for handwritten optical music recognition," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, IEEE, 2017, pp. 39–46.

[12] E. Shatri and G. Fazekas, "DoReMi: First glance at a universal OMR dataset," arXiv preprint arXiv:2107.07786, 2021.

[13] Z. Huang, X. Jia, and Y. Guo, "State-of-the-art model for music object recognition with deep learning," Applied Sciences, vol. 9, no. 13, p. 2645, 2019.

[14] E. Shatri and G. Fazekas, "Knowledge discovery in optical music recognition: Enhancing information retrieval with instance segmentation," arXiv preprint arXiv:2408.15002, 2024.

[15] A. Baró, P. Riba, and A. Fornés, "Musigraph: Optical music recognition through object detection and graph neural network," in International Conference on Frontiers in Handwriting Recognition, Springer, 2022, pp. 171–184.

[16] J. Calvo-Zaragoza and D. Rizo, "End-to-end neural optical music recognition of monophonic scores," Applied Sciences, vol. 8, no. 4, p. 606, 2018.

[17] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, “Handwritten music recognition for mensural notation with convolutional recurrent neural networks,” Pattern Recognition Letters, vol. 128, pp. 115–121, 2019.

[18] F. J. Castellanos, J. Calvo-Zaragoza, and J. M. Iñesta, “A neural approach for full-page optical music recognition of mensural documents,” in Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 2020, pp. 558–565.

[19] M. Alfaro-Contreras, J. M. Iñesta, and J. Calvo-Zaragoza, “Optical music recognition for homophonic scores with neural networks and synthetic music generation,” International Journal of Multimedia Information Retrieval, vol. 12, no. 1, p. 12, 2023.

[20] A. Rios-Vila, J. M. Iñesta, and J. Calvo-Zaragoza, “On the use of transformers for end-to-end optical music recognition,” in Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2022, pp. 470–481.

[21] A. Rios-Vila, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, “End-to-end optical music recognition for pianoform sheet music,” International Journal on Document Analysis and Recognition (IJDAR), vol. 26, no. 3, pp. 347–362, 2023.

[22] M. Jiang, X. Zhu, Y. Xia, G. Tan, B. Yuan, and X. Tang, “Segmentation of Mandarin Braille word and Braille translation based on multi-knowledge,” in WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings, vol. 3, IEEE, 2000, pp. 2070–2073.

[23] X. Wang, Y. Yang, J. Zhang, W. Jiang, H. Liu, and Y. Qian, “Chinese to Braille translation based on Braille word segmentation using statistical model,” Journal of Shanghai Jiaotong University (Science), vol. 22, no. 1, pp. 82–86, 2017.

[24] X. Wang, Y. Yang, H. Liu, and Y. Qian, “Chinese-Braille translation based on Braille corpus,” International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC), vol. 8, no. 2, pp. 56–63, 2016.

[25] A. Antonacopoulos and D. Bridson, “A robust Braille recognition system,” in International Workshop on Document Analysis Systems, Springer, 2004, pp. 533–545.

[26] T. D. S. H. Perera and W. K. I. L. Wanniarachchi, “Optical Braille recognition based on histogram of oriented gradient features and support-vector machine,” in 2018 International Conference on Computing and Communication Engineering, 2018.

[27] Z. Khanam and A. Usmani, “Optical Braille recognition using Circular Hough Transform,” arXiv preprint arXiv:2107.00993, 2021.

[28] G. A. Venugopal-Wairagade, “Braille recognition using a camera-enabled smartphone,” International Journal of Engineering and Manufacturing, vol. 4, pp. 2–3, 2016.

[29] I. G. Ovodov, “Optical Braille recognition using object detection neural network,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1741–1748.

[30] A. Al-Salman and A. AlSalman, “Fly-LeNet: A deep learning-based framework for

converting multilingual Braille images," Heliyon, vol. 10, no. 4, 2024.

[31] R. Li, H. Liu, X. Wang, and Y. Qian, "DSBI: double-sided Braille image dataset and algorithm evaluation for Braille dots detection," in Proceedings of the 2018 2nd International Conference on Video and Image Processing, 2018, pp. 65–69.

[32] A. Shyna, J. Raju, R. M. George et al., "BrailleSegNet: A novel methodology for Braille dataset generation and character segmentation," Displays, p. 103145, 2025.

[33] N. A. Asfaw, B. H. Belay, and K. M. Alemu, "A deep learning approach for line-level Amharic Braille image recognition," Scientific Reports, vol. 14, no. 1, p. 24172, 2024.

[34] Y. Zheng, H. Xing, Z. Yu, X. Zhang, F. Gao, and J. Bu, "End-to-end solution for Braille-to-Chinese," in 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE), IEEE, 2023, pp. 1304–1308.

[35] A. Wu, Y. Yuan, and M. Zhang, "Vision-Braille: An end-to-end tool for Chinese Braille image-to-text translation," arXiv preprint arXiv:2407.06048, 2024.