ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

# Using Image Recognition Technology to Assist in Music Score Recognition and Instruction in Distance Music Education

*Dan Shen*[1,*] *and Xuandong Sun*[2]

[1] School of Art, South China University of Technology, Guangzhou, 510006, China
[2] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, 510006, China

**SUMMARY:** *Addressing the challenges in remote music education-such as the complex sources of sheet music images, significant variations in image quality, the heavy burden of manual identification on teachers, and the difficulty of directly applying recognition results to teaching-this paper focuses on the structural restoration of low-quality classroom sheet music images and the generation of instructional prompts. This paper at first builds a music score data arrangement plan which is specially made for remote teaching situations, it unites publicly obtainable OMR data together with screenshots, mobile telephone pictures, classroom video frames, and teacher-marked manuscripts under one single training and evaluation frame. This hereby constructs the RemoteScore-Teach data object, which has the integration of symbol-level annotations, MusicXML alignment outcomes, and bar-level teaching-oriented labels. Building on this foundation, we propose the STG-OMR model, which integrates visual encoding, positional embedding, scale embedding, symbol relationship modeling, sequence decoding, and instructional hint generation into a single recognition pipeline. This enables the system to simultaneously output structured musical score results and bar-level instructional hints. Experimental results demonstrate that the proposed method achieves superior performance on both public benchmarks and remote teaching test sets, with Symbol F1, SeqAcc, and Hint-P may attain 95.4%, 91.7%, and 89.3% upon RemoteScore-Teach, respectively, and it displays higher stability in the situations which include photographed scores, reflective screen captures and annotated scores. The cutting experiments further prove that the score position prior information, the relation graph restriction conditions and the teaching hint branch structure are the main sources that bring performance promotion. This research provides practical technical support for pre-class preparation, in-class identification of key measures, and post-class assignment screening, while also offering a new implementation path for intelligent score analysis in remote music education.*

***KEYWORDS:*** *remote music education; music score recognition; optical music score recognition; instructional prompt generation; structured score analysis*

## 1 Introduction

The teaching effect of distant music education in large part depends on that whether teachers can quickly understand the music scores which are submitted by students. In the network classes, achievement results are commonly handed in through mobile telephone pictures, screen catches, flat panel uploads, or live flowing screen shots. These pictures frequently encounter the

*shendan2025@163.com
https://doi.org/10.65102/is2026755
Published: 30 April, 2026

problems of skewing, glare, compression traces, cut-off edges, and artificial hand-written notes. Before giving formal teaching, teachers always must firstly recognize clefs, time signatures, accidentals, slurs, and relations between staves, therefore then decide which measures need emphases and which parts are appropriate for practice homework tasks. If this procedure occupies an excessively long time, the progression of the class is broken, hence students' holding of the score's framework is weakened. For long-distance music teaching, the readable degree of music score is no longer simply a problem about arranging materials; it gives direct influence to the order of explanations, the effect degree of demonstrations, and the level of after-class feedback [1-4].

This problem is especially important because reading music notation is in itself a high-burden activity in the process of music learning. The students are required to finish pitch distinguishing, rhythm type identification, wrong pitch adjustment, and structure connection across measures in a restricted time period. This is especially the case in piano teaching, where the matching between left-hand and right-hand staff lines, the keeping of chord time lengths, and partial rhythm changes greatly raise the pressure of score reading. For solving these difficult problems, current studies have given assistance in fields such as the visual cognition of note identification, network piano learning intervening measures, mixed-reality-aided exercising, and reflection-based interface design [5-7]. These research works point out that for a teaching system which can be effectively carried out in actual classrooms, it is not enough to only supply recorded outcomes; it must also arrange key information from the music score into prompting contents that can be conveniently got by teachers, therefore helping them rapidly find difficult measures and sections with much knowledge.

The technical method that most closely matches this demand is optical music recognition (OMR). The early stage OMR mainly depended on symbol dividing, template matching, and rule-based combinations, hence it is suitable for music scores that have relatively stable layout and simple structure. Along with the deep learning's progress, the research emphasis has gradually moved toward object detection, sequence modeling, end-to-end transcription, and layout-aware Transformers, and the application scope has been extended from single-part music scores to piano double-staff scores, harmonic scores, and complex full-page layouts. [8-14]. On public standard test sets, current methods have obtained comparatively steady restoration of note kinds, order connections, and some page structures. The bringing forward of unified expression forms and evaluation frames has also made comparisons among different models more possible to carry out. This development shows that the research focus of music score recognition has moved towards more fine-grained directions: whether structural information can be kept under the condition of degraded images; whether the obtained results can still keep their reliability when the output objects become more complex; and whether the recognition outcomes still can keep being explainable and usable when people put them into use in actual real situations.

The problem is located in the great differences which exist between sheet music pictures in distance classrooms and the standard OMR reference standards. Classroom teaching materials are frequently obtained from camera photographs or screen text records, in which visual angle deformation, partial bright spots, shadow blocking, and compression false traces often happen at the same time. The correlative investigation on OMR which uses cameras has already proven that such domain changes can directly destroy the stability of recognition [15]. Furthermore, one individual music work may correspond to many uploaded editions in the teaching course, hence further expanding the difference between the training distribution and the deployment distribution. Although public datasets give an important technical base, they are not able to completely include the real input forms that are met in distance teaching rooms.

The output goals of current methods also do not align with the requirements of teaching.

The great majority of existing OMR research works take note identification, sequence writing-back, or MusicXML recovery as the main research endpoints. Although these output results achieve the digitization of music scores and the preservation of documents, they therefore have difficulty in directly completing the explanation tasks that are needed in the classroom. What the teachers truly deeply care about is which kinds of measures have more complex rhythmic structures, which kinds of positions are easy to have wrong reading of accidentals, and which cross-stave relationships need that people give advance warning in advance. If the system only is able to give out transcription results, teachers therefore still must by themselves carry out secondary screening work; The identification module stays in the document handling stage and has not yet entered the teaching arrangement stage.

Moreover, a number of studies which have close relation to digital music education put more focus on immersive interaction and learning experiences, yet it has not sufficiently discussed the standard staff notation structure [16, 17]. Although this kind of work is helpful for promoting participation, it possibly cannot solve the most pressing problem that teachers meet in remote classrooms: the identification of music notation. To teaching activities which need explanations, demonstrations, and giving out exercises that center on specific measures, the system’s supporting function will have restriction if it cannot steadily fix attention on the notation structure itself. At the same time, the output expression forms and assessment indicators of OMR are still being step by step made into standards. The comparative outcomes that exist among diverse methods on the symbol, sequence, and page levels cannot at present be directly transformed into application choices that are inside teaching situations [18]. The expenses of deployment also cannot be neglected. Long-distance teaching normally needs quick reaction speeds and accurate location placement; if a model depends on too much calculation resources, therefore its actual use value will become lower.

Under this background, this article puts its focus on the work of classroom help in distance music education. This research puts its focus on three mutually connected problems: the way to stably obtain score structures from low-grade classroom images; how to carry out the organization of recognition results into bar-level prompting materials that teachers can directly make use of; and how to verify the effect and deployable property of this method in the real remote situations which exist in actual environment. For solving these difficult problems, the present paper builds a data arrangement frame which is specially made for long-distance teaching, it combines publicly accessible OMR data and actual classroom pictures into one whole data set. On the method level, we have designed a united music notation recognition model which integrates teaching direction, therefore making both score rebuilding and the production of teaching hints to be completed in one single framework. With regard to verification, we assess this method on many different aspects-containing identification correctness, scene stability, component contributions, and calculation speed-by making use of both public standard datasets and our own constructed remote teaching dataset. This thesis attempts to push music score identification from a plain transcription work to a classroom help function, hence offering a steadier base for score analysis and more straight help for finding key points in long-distance music teaching.

# 2 Research on Music Score Recognition and Teaching Assistance in Remote Music Education Settings

## 2.1 Research on Sample Organization and Data Processing

This section puts emphasis on data origin, arrangement, and pretreatment methods, with the aim to turn the dispersed, varied, and visually non-uniform sheet music pictures existing in

distant classrooms into calculable, comparable, and repeatable experiment objects. The data processed in this study are not standard sheet music images in the conventional sense, but rather the actual sheet music materials exchanged and used by teachers and students on remote teaching platforms.These materials include both high-resolution electronic score screenshots and versions captured via mobile phone oblique shots, webcam frames, and images annotated with classroom markings. Images from different sources exhibit significant variations in layout, lighting distribution, character clarity, and background noise; therefore, a method design relying solely on a single standard data distribution would struggle to support subsequent teaching assistance outputs.To balance comparability against public benchmarks with the authenticity of remote teaching scenarios, this paper organizes the data into two layers: public pre-training and scenario fine-tuning. The pre-training phase incorporates MUSCIMA++, DeepScoresV2, CVC-MUSCIMA, and Camera-PrIMuS to cover the fundamental structural distributions found in handwritten scores, printed scores, real-world camera-captured scores, and symbol detection scenarios [19-22]. The fine-tuning phase constructs the RemoteScore-Teach dataset to address the actual image noise and instructional labeling requirements in remote classrooms. The organization of remote teaching samples is illustrated in Figure 1.

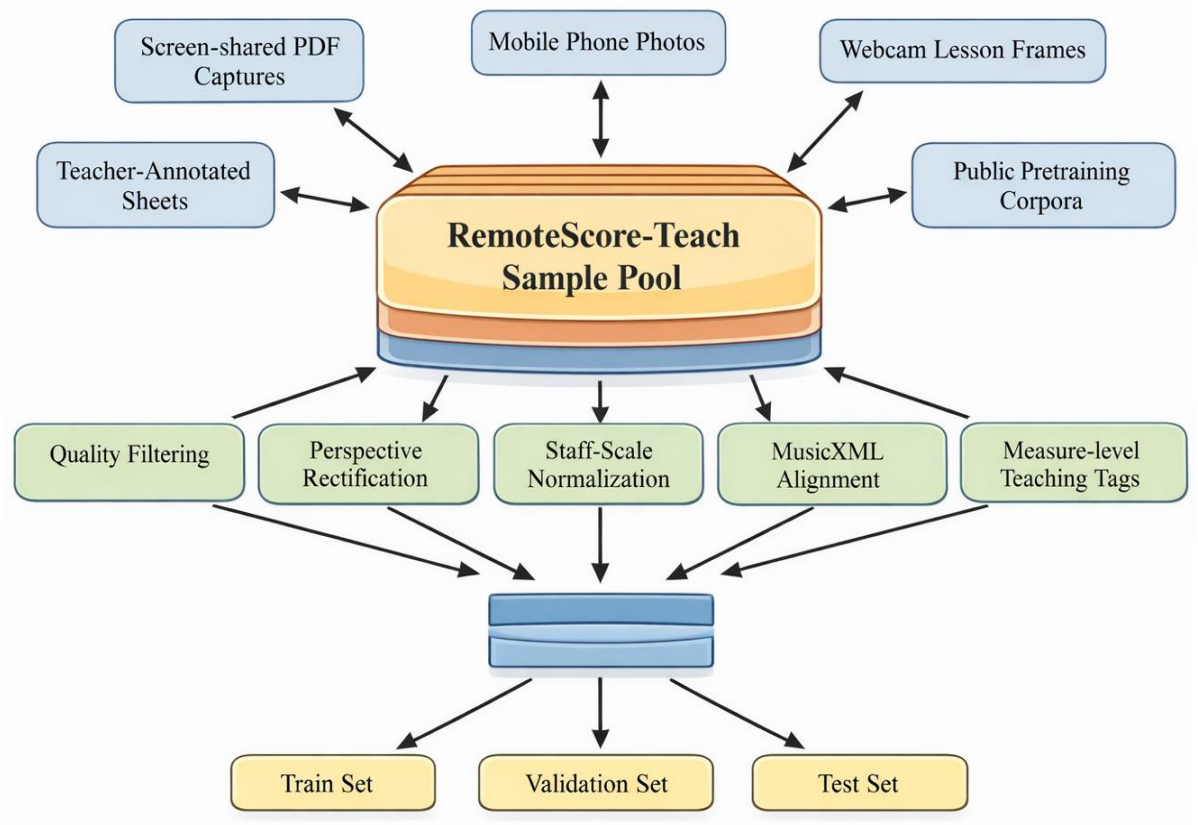


*Figure 1: Diagram of remote teaching sample construction and data organization mechanisms.*

Figure 1 has put together data sources, quality examination, layout adjustment, MusicXML matching, and teaching label corresponding into one work flow, thus it guarantees that training, verification, and test data sets all use consistent data standard before they are input into the model. Table 1 has given out the component condition of the RemoteScore-Teach data set.

*Table 1: Composition and Annotation Statistics of the RemoteScore-Teach Dataset*

| Subset | Page Count | Section Count | Symbol Instances | Teaching Tags | Main Features |
|---|---|---|---|---|---|
| Screen Share PDF Screenshot | 1020 | 9184 | 45672 | 2816 | High clarity but noticeable compression artifacts |
| Mobile Photo Scores | 988 | 8742 | 43915 | 2704 | Tilt, reflections, and shadows present |
| Camera Classroom Frames | 816 | 7388 | 36124 | 2195 | Low resolution, noticeable edge blurriness |
| Teacher Annotations | 416 | 3402 | 16987 | 1226 | Handwritten circles, arrows, and text interference |
| Total | 3240 | 28716 | 142698 | 8941 | Work categorized into 2268/486/486 pages |

In Table 1, the data collection totally has 3,240 sheet music pictures, 28,716 measures, 142,698 symbol individual cases, and 8,941 teaching labels, among which are 1,020 pages of screen-shared PDF screen shots, 988 pages of sheet music taken by mobile telephone, 816 pages of classroom pictures recorded by network camera, and 416 pages of teacher-marked handwritten manuscripts. In contrast with conventional OMR data collections, the present research not only keeps symbol-level position notes and sequence-level MusicXML matching but also gathers teaching focus information into bar-level labels. Every page sample writes down three kinds of message: firstly, the coordinate positions and sorts of basic symbols, for example, note symbols, rest symbols, slur symbols, time signature symbols, key signature symbols, and strength symbols; secondly, the structured order that is in alignment with MusicXML; third, bar-level prompting labels which show elements that teachers ought to give first priority to in the process of instruction. This research carries out categorization of instructional labels into six kinds: concentrated pitch jumps, thick accidentals, complicated rhythmic divisions, obvious syncopation or dotted note structures, clear cross-staff connections, and disturbance from fingerings and expression marks. This method's goal is to discuss both "whether the recognition outcomes are correct" and "whether the recognition outcomes can directly serve for teaching" inside the same research frame.

Because the same work can show many times in distance classrooms through different recording ways, this paper uses work-level cut instead of random page-level cut when it splits the data. To speak concretely, different recorded editions of the same work, screen captures from different phases of one course, and versions that have annotations and that do not have annotations therefore will not appear at the same time in the training, verification, and test sets. The ultimate training set, validation set, and test set are respectively divided into 2268 pages, 486 pages, and 486 pages. This division can lower the overestimation that is brought by content leakage, therefore it makes the test results more able to reflect the actual situation of putting into use in real-world classrooms. In the meanwhile, the proportion values of the four picture origins keep basically same in the three sub-collections, hence decreasing the deviation which is brought by the too high gathering of one single origin in one certain sub-collection.

In the preprocessing step, this thesis firstly carries out direction adjustment and view angle adjustment for the primitive pictures to solve the most frequent problems in mobile phone photographing and camera frame collecting, such as rotation deviation, edge shrinking, and partial extending. After that, we carry out illumination equalization, contrast standardization

and spectrum scale standardization, thus making samples from different sources enter a comparable range on the aspects of line interval and stroke breadth. Because teacher markings, hand-drawn arrows, and class notes themselves are the important components of the remote teaching environment, this thesis does not use strong noise reduction methods to directly eliminate non-spectral areas. On the contrary, it is through the combination of saliency suppression and local enhancement that it decreases their obstructive action upon the extraction of core features. This method retains the original properties of real classroom pictures, meanwhile it prevents the losing of structural information that is brought by overmuch cleaning.

For the increment of the adaptability of the model to distortion modes in distance teaching, extra augmentations-which include Gaussian blur, partial exposure, perspective deformation, edge blocking, and JPEG compression-were added in the training stage. The strength of these improvements was determined according to the statistical distribution of real classroom pictures and modified together with alterations of error on the verification set, hence it can prevent obvious differences between training improvements and testing noise. With respect to annotation quality control, this research firstly aligned symbol-level annotations to MusicXML, and thereafter carried out a two-round examination of bar-level teaching labels. In the first turn, annotators who have the ability to read music notations have given initial labels; In the second round, inconsistencies were manually corrected one bar by one bar, hence only highly consistent labels are finally kept in the training set. The dataset that we have got supports both the work of symbol recognition and the following research on the mapping of instructional cues.

## 2.2 Research on the Structural-Instructional Joint Recognition Model

After the data goes into the model, it must at the same time solve two problems: how to rebuild the music content and how to find out the teaching key points. For this purpose, this paper has built the Structure-Teaching Combined Identification Model, STG-OMR. The model is composed by image correction and local enhancement, hierarchical vision coding, a symbol relation graph, a sequence decoder, and an instruction cue head. Figure 2 has shown the mutual connection relations among these modules and the dual-output interface.

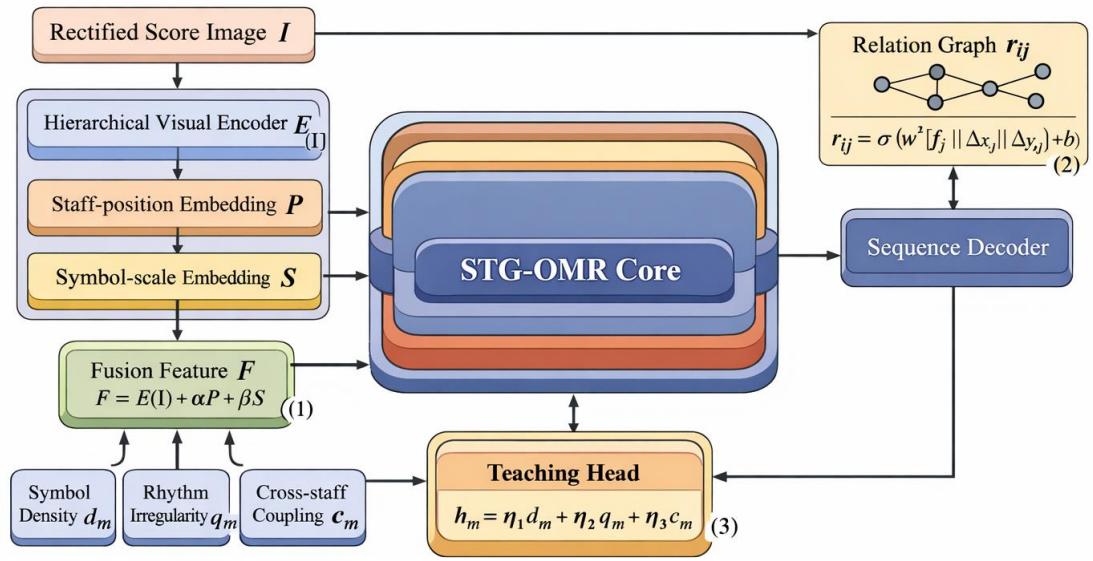


*Figure 2: Schematic diagram of the mechanism of the structure-based music notation recognition model.*

The corrected image first enters the visual encoding stage, forming a feature representation that integrates texture, position, and scale information, as shown in Equation (1).

$$F=E(\tilde{I}) + \alpha P + \beta S \tag{1}$$

where $\tilde{I}$ represents the corrected image, $E(\tilde{I})$ is the output of the backbone encoder, $P$ is the staff position embedding, $S$ is the symbol scale embedding, and $\alpha$ and $\beta$ are learnable weights. By directly incorporating position and scale information into the main features, the model can simultaneously leverage staff geometric priors and local dimensional differences to maintain semantic distinctions when processing dots, rests, accidentals, and small ties.

After generating the fused features, the model extracts candidate symbol nodes and restores structural relationships at the node level. The assignment of noteheads to stems, the attachment of slurs, the corresponding positions of temporary accidentals, and cross-staff chord connections all require explicit node relationship constraints. To this end, the model estimates the association strength between candidate nodes as shown in Equation (2).

$$r_{ij} = \sigma\left(w^{\top}\left[f_i \| f_j \| \Delta x_{ij} \| \Delta y_{ij}\right] + b\right) \tag{2}$$

In the equation, $r_{ij}$ represents the relationship strength between node $i$ and node $j$, where $f_i$ and $f_j$ are node features, $\Delta x_{ij}$ and $\Delta y_{ij}$ are relative displacements, $w$ and $b$ are learnable parameters, and $\sigma(\cdot)$ is the Sigmoid function. This relationship graph provides structural constraints prior to sequence decoding, enabling the model to maintain more stable durations and connection relationships in scenarios involving long measures, high-density chords, and dual-staff notations.

After the structural relationships have been gotten determination, the sequence decoder, which is a module of the model, outputs a linear token sequence that can be compatible with MusicXML. The sequence of decoding is arranged as "measure-stave-symbol" for the purpose of reducing semantic shift which is brought by reordering among measures and staves. In order to let recognition outcomes can be directly put into teaching work flow, the model, after the decoder, adds a teaching prompt head, and defines a teaching attention score on measure level, which is displayed in Equation (3).

$$h_m = \eta_1 d_m + \eta_2 q_m + \eta_3 c_m \tag{3}$$

In the equation, $h_m$ represents the instructional focus score for the $m$ th measure; $d_m$ denotes symbol density; $q_m$ denotes rhythmic irregularity; $c_m$ denotes cross-staff coupling strength; and $\eta_1$, $\eta_2$, and $\eta_3$ are weights learned from the validation set. Measures with scores exceeding the threshold enter the prompt ranking module and are mapped to corresponding instructional labels. Thus, the system's output simultaneously covers symbol-level results, structured sequences, and measure-level focal points that teachers can directly utilize.

The training work adopts a stage-by-stage united optimization method. In the first 40 training cycles, the model makes the visual coding, node checking and sequence decoding get stable. After that, relation consistency restriction conditions and instruction label losses are slowly put forward to reduce the amplification of early identification noise on instruction outputs. When the inference step is carried out, the model at the same time gives out symbol-level outcomes, MusicXML orderings, and bar-level guiding signals. The key innovation of this model just is that it puts positional prior knowledge, structural restriction and teaching mapping together into one whole identification flow, and this therefore guarantees that the structural reconstruction got from classroom pictures keeps synchronous with the teaching output results.

## 2.3 Experimental Setup and Evaluation Protocol

The design of experiment must at the same time assess recognition correctness, the usable degree of instruction, and the expenses of deployment. For the guarantee of consistent comparison standards, the whole working procedure uses a two-step training process and multi-aspect assessment. The workflow of training and evaluation is displayed in Figure 3.

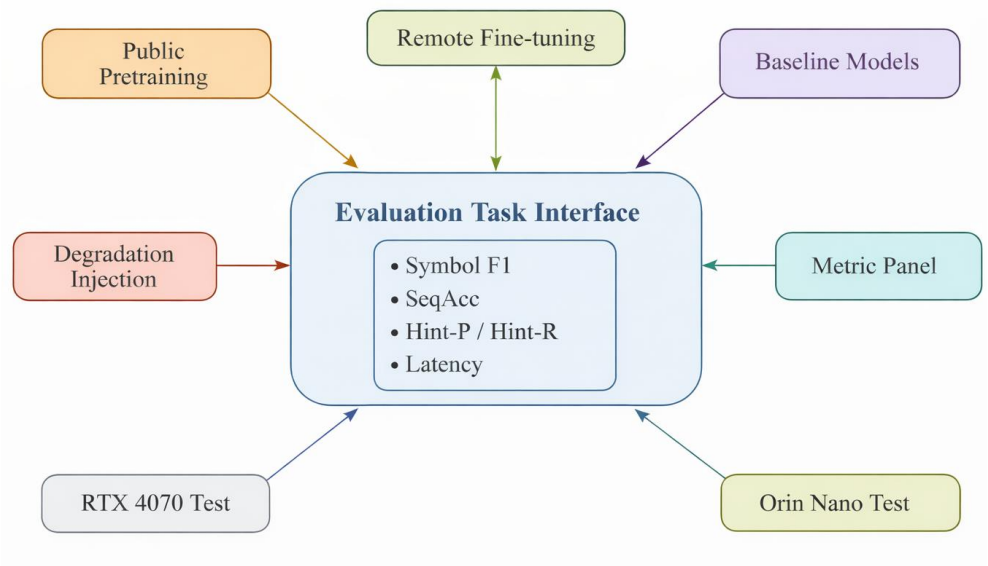


*Figure 3: Schematic diagram of the experimental protocol, comparison strategy, and evaluation interface.*

The first stage involves pre-training on public datasets to enable the model to fully learn the fundamental structural distributions in printed, handwritten, and camera-captured sheet music; the second stage involves fine-tuning on the RemoteScore-Teach dataset to adapt to the challenges of remote classrooms, such as glare, compression, low resolution, and annotation interference.The comparison methods cover four representative approaches, including the single-part end-to-end recognition model CRNN-E2E, the piano-specific Pianoform-E2E, Sheet Music Transformer, and the implicit layout-aware Transformer.These methods represent four major technical approaches-convolutional sequence modeling, end-to-end transcription for piano scores, global Transformer decoding, and layout-aware modeling-and effectively cover the current mainstream benchmarks for OMR.

We carry out the assessment of evaluation metrics at the same time in three aspects: the performance of recognition, the usability in teaching, and the efficiency of deployment. The section that is about recognition performance utilizes symbol-level F1 score and sequence accuracy for measuring local recognition accuracy and overall structural recovery quality; the teaching usability part uses prompt accuracy and recall to examine the model's capacity for finding high-risk measures; the deployment efficiency part computes average inference delay time and parameter number to evaluate the model's work feasibility on desktop and edge equipment. For guaranteeing consistent measurement indexes among different models, the output results of all compared methods are transformed into a unified symbol aggregate and sequence expression form before the corresponding measurement indexes are calculated.

From the perspective of carrying out the work, the training input is carried out standardization processing to have a 1536 px long side. We utilize the AdamW optimizer, the initial learning rate is $1\times10^{-4}$, the batch size is 6, the total number of training epochs is 120. Both pre-training work and fine-tuning work use the identical data increasing methods;

Nevertheless, in the fine-tuning stage, the sampling ratios for blurring, non-uniform exposure, and perspective deformation are raised in order to better conform to the distribution of images in the environment of distance classrooms. The verifying collection is mainly utilized to confirm model weight parameters, prompt threshold values, and the weight coefficients of each loss item, while the testing collection is only employed after the model has been solidified. This arrangement assists in preventing overfitting judgments, when we consider the limited size of remote scene data.

For the evaluation of what extent the model can adapt to distortions that exist in real world classrooms, this our paper carries out three extra groups of gradual degradation experiments. The first group changes the Gaussian blur radius between 0 and 3 px for observing the influences of camera out-of-focus and compression blur; the second group changes the perspective slope angle from 0 degree to 12 degree to imitate geometry distortions which are brought by sloped mobile phone photographs and desk surface photographing; The third group changed the non-uniform illumination coefficient from 0 to 0.5, for considering classroom conditions that include reflections, shadows and local overexposure. We have carried out deployment experiments on RTX 4070 and Jetson Orin Nano at the same time; The former one was utilized for standard training and high-efficiency inference, hence the latter one simulated a portable teaching assistant equipment. By carrying out combined experiments on these two kinds of equipment, this paper can at the same time carry out evaluation on both the effect of the algorithm and the expenses for its deployment. Under the above-mentioned framework, the Methods part constitutes a comparatively full research cycle: data objects are arranged uniformly and input into the combined structure-instruction model, and the outputs of the model are hence verified via multi-dimension indicators and multi-scenario pressure tests.

# 3 Recognition Performance, Error Characteristics, and Educational Adaptability Analysis

## 3.1 Overall Recognition Performance and Cross-Scenario Adaptation Analysis

The first question which we need to deal with is that whether the model we put forward can at the same time promote both speech recognition ability and the level of teaching guiding words in long-distance classrooms. The whole performance of each method on public benchmark datasets and remote teaching test sets is displayed in Table 2.

*Table 2: Comparison of Key Results on Public Benchmarks and Remote Teaching Scenarios*

| Method | Camera-PrIMuS Acc/% | MUSCIMA++ F1/% | RemoteScore-Teach Symbol F1/% | RemoteScore-Teach SeqAcc/% | Hint-P/% | Latency/ms·page$^{-1}$ |
|---|---|---|---|---|---|---|
| CRNN-E2E [9] | 96.3 | 85.1 | 88.7 | 79.4 | 71.8 | 146 |
| Pianoform-E2E [10] | 97.1 | 87.3 | 90.9 | 83.6 | 75.2 | 171 |
| SMT [11] | 97.8 | 91.6 | 93.5 | 87.2 | 82.6 | 228 |
| LA-Transformer [13] | 97.9 | 92.1 | 93.7 | 88.1 | 83.4 | 246 |
| STG-OMR (Ours) | 98.1 | 94.2 | 95.4 | 91.7 | 89.3 | 184 |

In the Table 2, STG-OMR has obtained the most excellent results on all three test sets: Camera-PrIMuS, MUSCIMA++, and RemoteScore-Teach. Concretely speaking, on Camera-PrIMuS, the precision has achieved 98.1%, which is an enhancement of 1.8 percentage points

compared with CRNN-E2E; On the method of MUSCIMA++, the Symbol F1 score attained 94.2 percent, which is an improvement that has 2.1 percentage points more than LA-Transformer; On the RemoteScore-Teach method, Symbol F1, SeqAcc, and Hint-P have achieved 95.4%, 91.7%, and 89.3% respectively, therefore these results represent 1.7, 3.6, and 5.9 percentage points of promotion compared to the existing advanced baseline at present. The main outcome points out that the merits of structural modeling are obvious in the symbol level, sequence level, and phrase grade teaching inspiration level. The whole merits do not get even distribution among various picture origins. In order to carry out the comparison of the models' adaptability towards different remote image sources, the changes of sequence accuracy for every method are displayed in Figure 4.

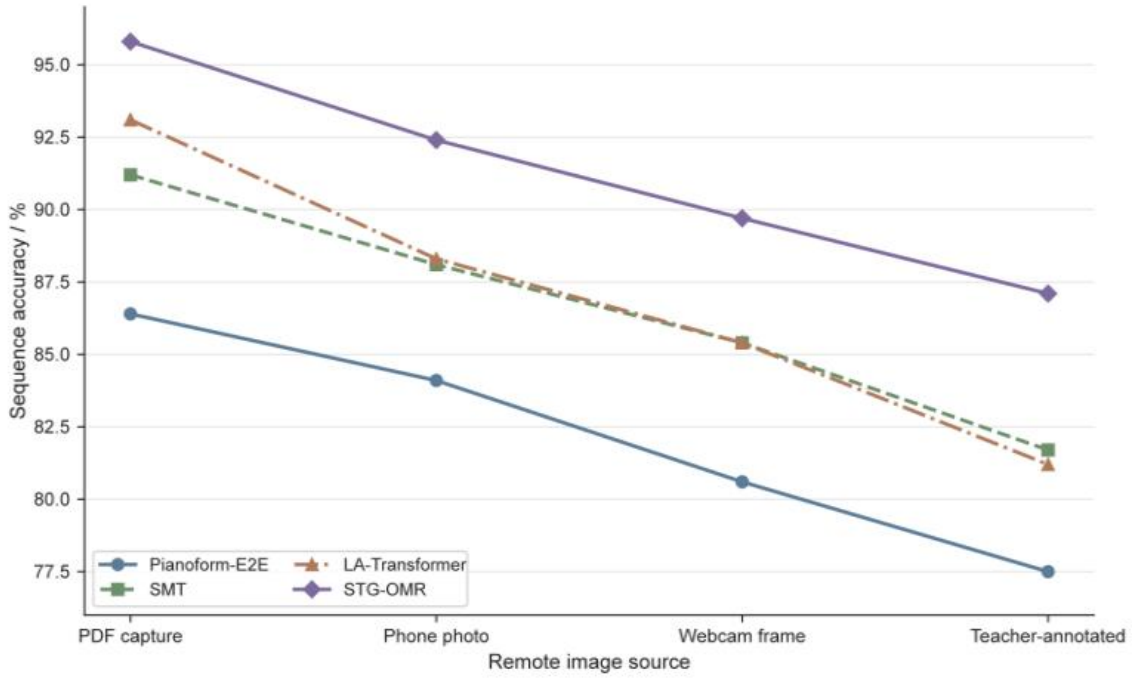


*Figure 4: Comparison of sequence accuracy curves across different remote image sources.*

In Figure 4, the STG-OMR obtains SeqAcc that is 95.8%, 92.4%, 89.7%, and 87.1% in four kinds of scenes: PDF capture, mobile photograph, web camera frame and teacher marking, respectively. When we compare with LA-Transformer, these data represent the promotion of 2.7, 4.1, 4.3, and 5.9 percentage points, respectively. The performance difference slowly becomes larger from electronic spectrum screen photographs to handwritten pictures, camera pictures, and teacher-marked examples, therefore showing that position prior knowledge and relation graphs are more sensitive to view changes, part covering, and handwork marks, hence the improvements mainly happen on the most difficult samples which appear in actual classroom environments. Besides separated test spots, it is also needed that we observe the model's tolerant ability toward continuous distortion changes. Figure 5 shows us the alterations of recognition stability among different models when continuous degradation conditions exist.

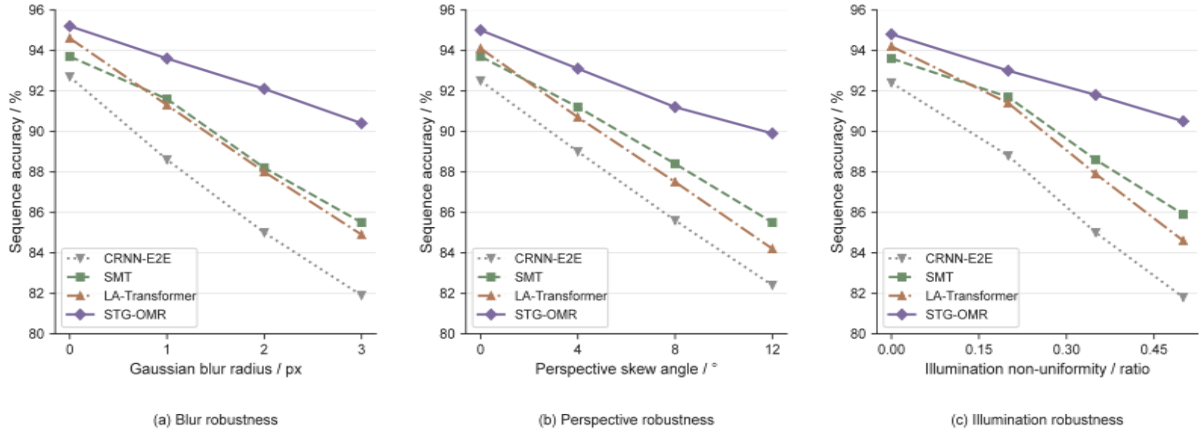


*Figure 5: Coupling curve between degradation intensity and recognition stability.*

In Figure 5(a), as the Gaussian blur radius increases from 0 to 3 px, STG-OMR's SeqAcc drops from 95.2% to 90.4%, a decrease of 4.8 percentage points; LA-Transformer drops from 94.6% to 84.9%, a decrease of 9.7 percentage points.In Figure 5(b), as the perspective skew increases from 0° to 12°, the SeqAcc of STG-OMR decreases from 95.0% to 89.9%, while that of SMT decreases from 93.7% to 85.5%.In Figure 5(c), as the illumination unevenness coefficient increased from 0 to 0.5, STG-OMR decreased from 94.8% to 90.5%, and CRNN-E2E decreased from 92.4% to 81.8%.The three sets of curves collectively indicate that the models exhibit a more gradual performance decline under conditions of blurring, geometric distortion, and local exposure variations. These results are consistent with the design of the method.The spatial and scale embeddings in the fusion features enhance the ability to distinguish accidentals, sharps and flats, and vertical notes within the staff, while the relationship graph reduces the accumulation of stem misclassifications, slur breaks, and cross-staff misconnections during the sequence stage. Although the instructional hint header does not directly contribute to the calculation of recognition metrics in the main results table, it exposes structural differences in high-risk measures to the training process in advance, thereby improving the restoration quality of complex measures.

## 3.2 Module Contribution, Robustness, and Efficiency Analysis

The overall results validate the effectiveness of the method. The next step is to determine which modules contribute to the performance gains and whether the trade-offs associated with these modifications are within an acceptable range. The module ablation results and efficiency statistics for STG-OMR are shown in Table 3.

*Table 3: Module Ablation and Efficiency Analysis*

| Variant | Symbol F1/% | SeqAcc/% | Hint-P/% | Params/M | Latency/ms·page$^{-1}$ |
|---|---|---|---|---|---|
| w/o rectification | 93.1 | 88.5 | 83.4 | 42.8 | 169 |
| w/o staff-position embedding | 94.2 | 89.6 | 85.7 | 43.1 | 181 |
| w/o relation graph | 93.9 | 88.9 | 84.8 | 38.2 | 175 |
| w/o teaching head | 94.9 | 90.8 | 78.5 | 41.7 | 178 |
| Full STG-OMR | 95.4 | 91.7 | 89.3 | 43.6 | 184 |

In Table 3, after we have carried out the removal of the image correction module, the Symbol F1 has a drop from 95.4% to 93.1%, the SeqAcc drops from 91.7% to 88.5%, therefore, the Hint-P drops from 89.3% to 83.4%. This alteration indicates that perspective changes and non-uniform illumination in long-distance classrooms are accumulated in the whole recognition chain, therefore making front-end rectification a precondition for stable follow-up output. After we have taken away the staff position embedding, the values of Symbol F1, SeqAcc, and Hint-P have fallen to 94.2%, 89.6%, and 85.7%, each one separately. Position prior information is especially important for telling apart connected notes, short detached notes, temporary pitch modifiers, and pitch modifiers on extra ledger lines. After we take away the relationship graph, SeqAcc fell to 88.9%, and cross-stave chords, connections between note values inside one measure, and long-distance correspondences are the things that have the earliest degradation. After we take away the teaching head part, the main identification indicators only have a small drop, but Hint-P falls from 89.3% to 78.5%, therefore this shows that only depending on the post-processing matching of identification results brings difficulty to steadily get back the bar-level difficulty information which teachers really care about. The trade-off of accuracy and inference delay for various methods is displayed in Figure 6.

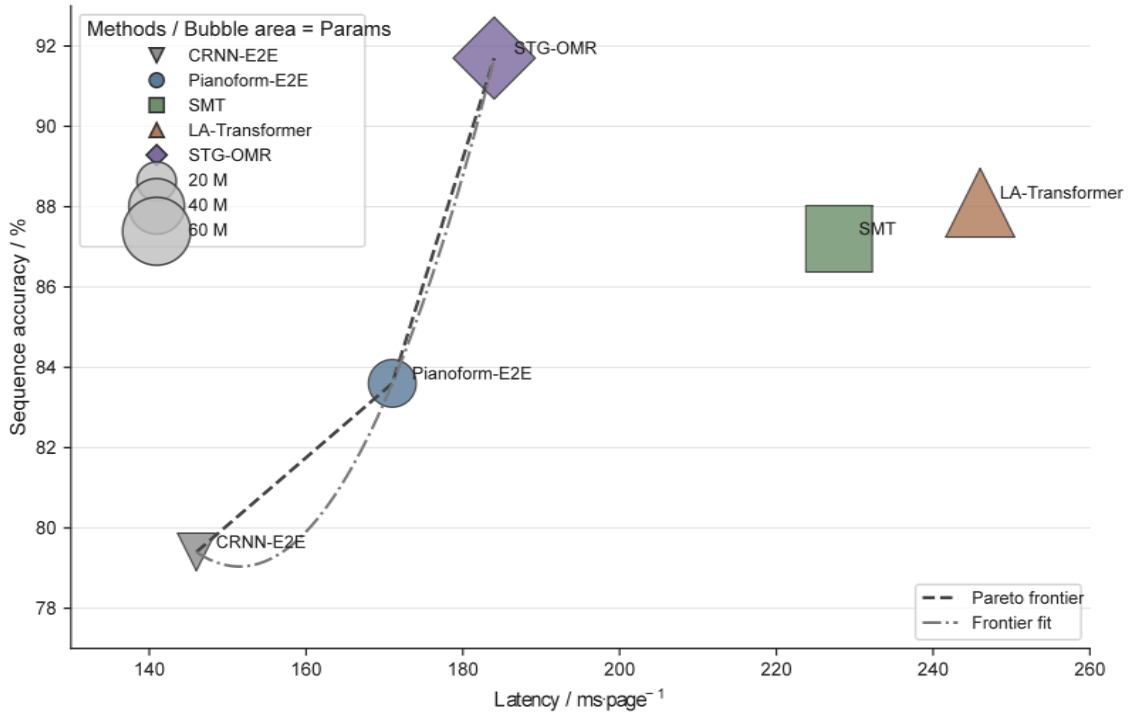


*Figure 6: Heatmap of key symbol confusion and density map of bar-level teaching focus.*

In Figure 6, STG-OMR obtains 91.7% SeqAcc, its average inference time consumption per page is 184 ms, and it has 43.6 M parameters, therefore it is placed on the Pareto front of accuracy against latency. When we make a comparison with SMT, our method can decrease the time delay by 44 ms·page$^{-1}$ and elevate the accuracy by 4.5 percentage points; When we make comparison with LA-Transformer, it is able to cut down the waiting delay by 62 ms·page$^{-1}$ and raise the accurate degree by 3.6 percentage points. At the aspect of device terminal, the model's average inference time delay each single page is 184 ms when use RTX 4070, and 0.91 s when use Jetson Orin Nano. After we combine Table 3 and Figure 6 and carry out analysis, we can find image correction, spectral table embedding and relationship graph are the main factors that give contribution to recognition effect, therefore the teaching head plays a very key function on

linking recognition outcomes to educational usage; the entirety of calculation expense still stays inside a scope that can be put to use.

## 3.3 Analysis of Error Sources, Case Performance, and Educational Deployment

While average metrics indicate the model's overall effectiveness, the key issue at the deployment level remains: where does the model primarily make errors, and are these errors concentrated in pedagogically sensitive areas? Figure 7 illustrates the model's primary error types and their distribution in high-risk instructional sections.

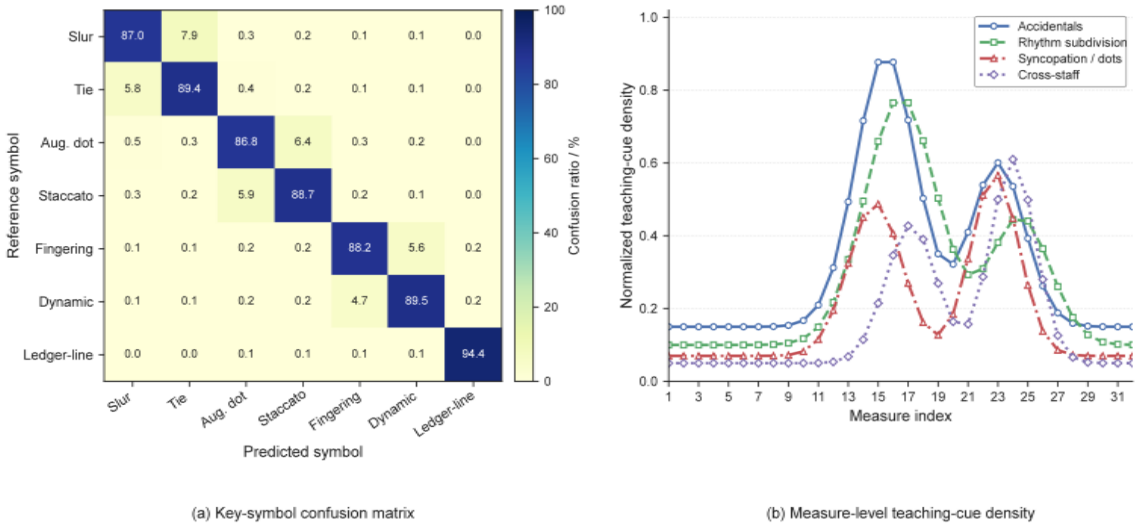


*Figure 7: Heatmap of key symbol confusion and density map of instructional focus at the section level.*

In Figure 7(a), high-value non-diagonal errors are primarily concentrated in several categories of confusion: Slur→Tie, Aug. dot→Staccato, Fingering→Dynamic, and Ledger-line pitch shift, with corresponding error rates of 7.9%, 6.4%, 5.6%, and 5.2%, respectively.These errors collectively point to a common characteristic: the local visual patterns are similar, and distinguishing them requires stronger contextual cues or clearer notational relationships. Figure 7(b) shows the distribution of instructional focus density at the bar level. High-density areas primarily appear in bars 14-18 and 22-25, with the main labels being accidentals, rhythm subdivision, syncopation/dots, and cross-staff.This distribution aligns closely with the areas identified as challenging in classroom instruction. When accidentals appear consecutively, students are prone to misreading pitch placement; when dotted notes and syncopated rhythms occur densely, maintaining a sense of meter becomes significantly more difficult; and passages with concentrated cross-staff chords often involve hand positioning and voice part comprehension.The correlation revealed in Figure 7 indicates that the areas where the model currently makes the most errors are also the areas that teachers need to prioritize in their instruction. To further examine the model's output characteristics in real classroom images, Figure 8 shows a comparison of recognition results and instructional prompts for a typical case.

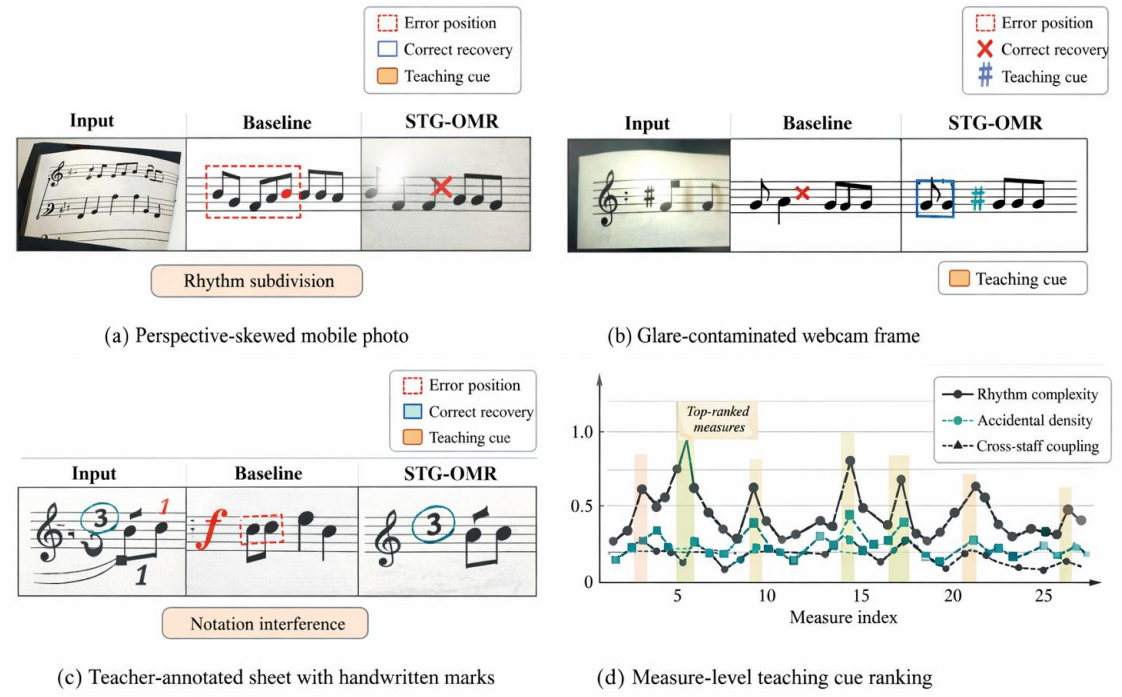


*Figure 8: Comparison of recognition results and instructional prompts in a typical remote classroom case.*

In Figure 8(a), the image which is obliquely taken by a mobile phone has obvious perspective distortion which is very great. The baseline model makes wrong reading on the dotted eighth-note-sixteenth-note group as a quarter-note-eighth-note combination, therefore STG-OMR keeps the original rhythm grouping and thus marks the corresponding measure as "complex rhythmic subdivision." In the Figure 8(b), two temporary sharp points have been lost from the camera frame on account of light glare; the baseline output has lead the whole change of the following pitch order. STG-OMR, by using position embedding and relation restriction, got tonal relation back, hence marked this section as "dense temporary accidentals". In Figure 8(c), on the score, the teacher has by hand written circles and numbers. Although the model has correctly found out the main notes and bar boundary lines, it has wrong identified a fingering number to be a dynamic marking, this shows that dense handwritten annotations are still a weak place for current models.

From an application perspective, STG-OMR is best suited for three types of tasks: pre-class preparation assistance, in-class local explanation support, and post-class assignment preliminary screening. It can help teachers identify high-risk measures in advance, provide structured prompts when explaining rhythm, accidentals, or cross-stave relationships, and perform preliminary analysis in assignment scenarios with a high volume of low-quality images.At the same time, conditions involving strong glare, severe occlusion, and overlapping handwritten annotations continue to cause local confusion; for advanced teaching tasks that require integration with performance audio, note durations, or expression markings to reach a conclusion, the current system still needs to be used in conjunction with the teacher's expertise.

# 4 Conclusion

This paper carries out discussion on difficulties of music score image identification and teaching help in long-distance music education through building a overall research frame that

covers data arrangement, structure recovery and teaching hint production. Unlike ordinary OMR tasks which only pay attention to document writing transcription, this research clearly limits its research range to real picture situations in classrooms, putting the first place on the processing of low-quality music materials including mobile phone photographs, screen catching pictures, camera frame catching pictures, and teacher-marked manuscripts, and further connects recognition outcomes to the finding of teaching difficulties and the producing of bar-layer guiding suggestions. Therefore, this paper not merely focuses on whether the sheet music can have transcription done, but also focuses on whether the transcription outcomes can be put into the processes of the teacher's explanation, assessment, and feedback.

(1) This present thesis has finished the data arrangement work which faces to long-distance teaching. Through bringing in publicly open OMR data and building the RemoteScore-Teach data object, this study puts together symbol-level annotations, MusicXML alignment, and bar-level teaching labels into one task framework, hence guaranteeing consistency of the training, validation, and test sets on the object level. This organizing method enhances the matching degree between experiment outcomes and actual classroom situations, hence it also provides a steady data base for later combined identification and teaching model construction.

(2) This treatise puts forward the STG-OMR model, which is a united structured teaching and music score identification model. Beyond the visual encoding work, this model explicitly brings in staff position embedded expressions, scale embedded expressions, and node relation restriction terms, and furthermore adds guiding prompt headers after the sequence decoding work, hence it enables the system to at the same time output structured music score achievement results and teaching-related information contents. Experiment consequences indicate that the method put forward by us attains 95.4% Symbol F1, 91.7% SeqAcc, and 89.3% Hint-P on the remote teaching test set, while it keeps relatively steady performance under the conditions of camera angle, glare, and annotation interference. These findings show that position prior knowledge, connection construction, and guide prompt branches together promote the model's adaptive ability to complicated classroom pictures.

(3) This article also points out the shortcomings that existing methods have. Sheet music that has strong light reflection, serious covering, or thick handwritten notes still causes partial wrong classification, and mixing between finger-count numbers, strength marks, and curve-shaped symbols has not been completely solved. In the future, research work can additionally add more detailed teacher label data, cross-modal combined modeling of music score and performance voice frequency, and more light-weight edge arrangement methods, therefore to promote the reaction speed and stability of the system in the real-time classroom environment.

# About the Author

Dan Shen was born in 1980 in Hengyang, Hunan Province, P.R. China. She studied at the Belarusian State Academy of Music and received her master's degrees in both piano performance and chamber music performance in 2004. She is a dedicated lecturer at the School of Art, South China University of Technology. Her research interests include music education and piano performance.

Xuandong Sun was born in 1972 in Jining, Shandong Province, P.R. China. He obtained his doctoral degree from the Chengdu Institute of Computer Applications, Chinese Academy of Sciences, in 2012. He is currently working at the School of Computer Science, Guangdong University of Technology. His research focuses on computer information technology and graphics processing.

# References

[1] Yu, X., Ma, N., Zheng, L., et al. (2023). Developments and applications of artificial intelligence in music education. Technologies, 11(2), 42.

[2] Merchán Sánchez-Jara, J. F., González Gutiérrez, S., Cruz Rodríguez, J., et al. (2024). Artificial intelligence-assisted music education: A critical synthesis of challenges and opportunities. Education Sciences, 14(11), 1171.

[3] Pan, H., & Wu, W. (2024). Online learning to play the piano: Perspectives and achievements / Aprender a tocar el piano en línea: perspectivas y logros. Culture and Education, 36(2), 370-393.

[4] Carvalho, A., Vieira, C., Santos, I. G., et al. (2024). The online teaching of the performing arts in higher education in a context of confinement as a creativity challenge. Media Practice and Education, 25(4), 326-342.

[5] Wong, Y. K., & Fang, J. F. (2025). Learning and teaching of fluent musical note recognition: The visual perceptual perspective. Frontiers in Cognition, 4, 1439-439.

[6] Amm, V., Chandran, K., Engeln, L., et al. (2024). Mixed reality strategies for piano education. Frontiers in Virtual Reality, 5, 1397-154.

[7] Suzuki, A., Ginsborg, J., Phillips, M., et al. (2024). Developing an online intervention to equip tertiary piano students with skills and strategies for effective practice. Music & Science, 7, 2059-2043241262612.

[8] Calvo-Zaragoza, J., Hajič, J. Jr., & Pacha, A. (2020). Understanding optical music recognition. ACM Computing Surveys, 53(4), Article 77.

[9] Calvo-Zaragoza, J., & Rizo, D. (2018). End-to-end neural optical music recognition of monophonic scores. Applied Sciences, 8(4), 606.

[10] Ríos-Vila, A., Rizo, D., Iñesta, J. M., et al. (2023). End-to-end optical music recognition for piano-form sheet music. International Journal on Document Analysis and Recognition, 26(3), 347-362.

[11] Ríos-Vila, A., Calvo-Zaragoza, J., & Paquet, T. (2024). Sheet Music Transformer: End-to-end optical music recognition beyond monophonic transcription. In Document Analysis and Recognition - ICDAR 2024 (pp. 20-37). Cham: Springer.

[12] Mayer, J., Straka, M., Hajič, J. Jr., et al. (2024). Practical end-to-end optical music recognition for piano-form music. In Document Analysis and Recognition - ICDAR 2024 (pp. 55-73). Cham: Springer.

[13] Ríos-Vila, A., Fuentes-Martínez, E., & Castellanos, F. J. (2025). An implicit layout-aware transformer for full-page end-to-end optical music recognition. International Journal of Multimedia Information Retrieval, 14, 34.

[14] Alfaro-Contreras, M., Iñesta, J. M., & Calvo-Zaragoza, J. (2023). Optical music recognition for homophonic scores with neural networks and synthetic music generation.

International Journal of Multimedia Information Retrieval, 12, 12.

[15] Liu, Y., Wu, R., Wu, Y., et al. (2023). A stave-aware optical music recognition on monophonic scores for camera-based scenarios. Applied Sciences, 13(16), 9360.

[16] Banquiero, M., Valdeolivas, G., Trincado, S., et al. (2023). Passthrough mixed reality with Oculus Quest 2: A case study on learning piano. IEEE MultiMedia, 30(2), 60-69.

[17] Karolus, J., Sylupp, J., Schmidt, A., et al. (2023). EyePiano: Leveraging gaze for reflective piano learning. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (pp. 1209-1223).

[18] Torras, P., Biswas, S., & Fornés, A. (2024). A unified representation framework for the evaluation of optical music recognition systems. International Journal on Document Analysis and Recognition, 27(3), 379-393.

[19] Hajič, J. Jr., & Pecina, P. (2017). The MUSCIMA++ dataset for handwritten optical music recognition. In 2017 14th IAPR International Conference on Document Analysis and Recognition (pp. 39-46).

[20] Tuggener, L., Satyawan, Y. P., Pacha, A., et al. (2021). The DeepScoresV2 dataset and benchmark for music object detection. In 2020 25th International Conference on Pattern Recognition (pp. 9188-9195).

[21] Fornés, A., Dutta, A., Gordo, A., et al. (2012). CVC-MUSCIMA: A ground truth of handwritten music score images for writer identification and staff removal. International Journal on Document Analysis and Recognition, 15(3), 243-251.

[22] Calvo-Zaragoza, J., & Rizo, D. (2018). Camera-PrIMuS: Neural end-to-end optical music recognition on realistic monophonic scores. In Proceedings of the 19th International Society for Music Information Retrieval Conference (pp. 248-255).

[23] Baró, A., Riba, P., & Fornés, A. (2022). Musigraph: Optical music recognition through object detection and graph neural networks. In Frontiers in Handwriting Recognition (pp. 171-184).

[24] Ayllon, E., Castellanos, F. J., & Calvo-Zaragoza, J. (2023). A weakly-supervised approach for layout analysis in music score images. In IbPRIA 2023 (pp. 170-181). Cham: Springer.

[25] Gallego, A. J., & Calvo-Zaragoza, J. (2017). Staff-line removal with selectional auto-encoders. Expert Systems with Applications, 89, 138-148.