

# Audio Matters Too! Enhancing Markerless Motion Capture with Audio Signals for String Performance Capture

YITONG JIN\*, Central Conservatory of Music, China and Tsinghua University, China  
 ZHIPING QIU\*, Central Conservatory of Music, China and Tsinghua University, China  
 YI SHI, Central Conservatory of Music, China and Tsinghua University, China  
 SHUANGPENG SUN, Tsinghua University, China  
 CHONGWU WANG, Central Conservatory of Music, China  
 DONGHAO PAN, Central Conservatory of Music, China  
 JIACHEN ZHAO, Tsinghua University, China  
 ZHENGHAO LIANG, Weilan Tech, China  
 YUAN WANG, Central Conservatory of Music, China  
 XIAOBING LI, Central Conservatory of Music, China  
 FENG YU, Central Conservatory of Music, China  
 TAO YU†, Tsinghua University, China  
 QIONGHAI DAI†, Tsinghua University, China

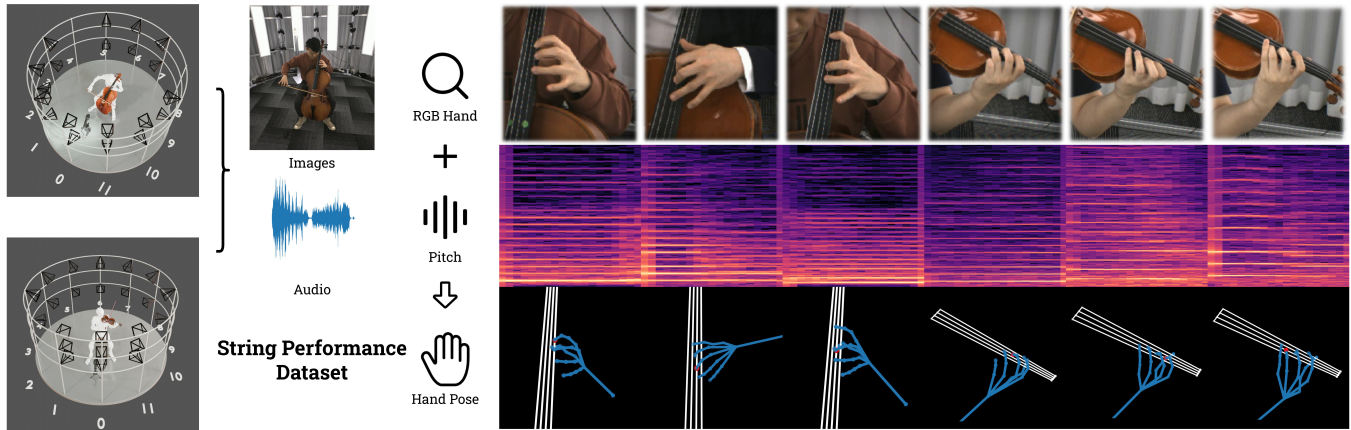


Fig. 1. We present the String Performance Dataset with an audio-guided multi-modal framework enhancing markerless motion capture for string performance.

\*Equal Contribution

†Corresponding Author

Authors' addresses: Yitong Jin, Central Conservatory of Music, China and Tsinghua University, China, jinyitong@mail.com.edu.cn; Zhiping Qiu, Central Conservatory of Music, China and Tsinghua University, China, zhiping\_qiu@mail.com.edu.cn; Yi Shi, Central Conservatory of Music, China and Tsinghua University, China, shiyi@mail.com.edu.cn; Shuangpeng Sun, Tsinghua University, China, pengcheng786@gmail.com; Chongwu Wang, Central Conservatory of Music, China, 1225@ccom.edu.cn; Donghao Pan, Central Conservatory of Music, China, pdh0227@sina.com; Jiachen Zhao, Tsinghua University, China, zhao\_jiachen@163.com; Zhenghao Liang, Weilan Tech, China, liangzhenghaothu18@163.com; Yuan Wang, Central Conservatory of Music, China, 22a056@mail.com.edu.cn; Xiaobing Li, Central Conservatory of Music, China, lxiaobing@ccom.edu.cn; Feng Yu, Central Conservatory of Music, China, yufengAI@ccom.edu.cn; Tao Yu, Tsinghua University, China, ytrock@mail.tsinghua.edu.cn; Qionghai Dai, Tsinghua University, China, qhdai@mail.tsinghua.edu.cn.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

In this paper, we touch on the problem of markerless multi-modal human motion capture especially for string performance capture which involves inherently subtle hand-string contacts and intricate movements. To fulfill this goal, we first collect a dataset, named String Performance Dataset (SPD), featuring cello and violin performances. The dataset includes videos captured from up to 23 different views, audio signals, and detailed 3D motion annotations of the body, hands, instrument, and bow. Moreover, to acquire the detailed motion annotations, we propose an audio-guided multi-modal motion capture framework that explicitly incorporates hand-string contacts detected from the audio signals for solving detailed hand poses. This framework serves as a baseline for string performance capture in a completely markerless manner without imposing any external devices on performers, eliminating the potential of introducing distortion in such delicate movements. We argue that the movements of performers, particularly the sound-producing gestures, contain subtle information often elusive to visual methods but can be inferred and retrieved from audio cues. Consequently, we refine the

© 2024 Copyright held by the owner/author(s).  
 0730-0301/2024/7-ART90  
<https://doi.org/10.1145/3658235>

vision-based motion capture results through our innovative audio-guided approach, simultaneously clarifying the contact relationship between the performer and the instrument, as deduced from the audio. We validate the proposed framework and conduct ablation studies to demonstrate its efficacy. Our results outperform current state-of-the-art vision-based algorithms, underscoring the feasibility of augmenting visual motion capture with audio modality. To the best of our knowledge, SPD is the first dataset for musical instrument performance, covering fine-grained hand motion details in a multi-modal, large-scale collection. It holds significant implications and guidance for string instrument pedagogy, animation, and virtual concerts, as well as for both musical performance analysis and generation. Our code and SPD dataset are available at [https://github.com/Yitongishere/string\\_performance](https://github.com/Yitongishere/string_performance).

CCS Concepts: • **Computing methodologies** → **Motion capture**.

Additional Key Words and Phrases: Marker-less Motion Capture, String Performance, Multi-modality

#### ACM Reference Format:

Yitong Jin, Zhiping Qiu, Yi Shi, Shuangpeng Sun, Chongwu Wang, Donghao Pan, Jiachen Zhao, Zhenghao Liang, Yuan Wang, Xiaobing Li, Feng Yu, Tao Yu, and Qionghai Dai. 2024. Audio Matters Too! Enhancing Markerless Motion Capture with Audio Signals for String Performance Capture. *ACM Trans. Graph.* 43, 4, Article 90 (July 2024), 10 pages. <https://doi.org/10.1145/3658235>

## 1 INTRODUCTION

In the broad spectrum of human movement, the playing of instruments stands as an intricate demonstration of human fine motor skills. Thus, traditional video recording often fails to reveal such details. The challenge lies in the delicate dance of a musician’s fingers, especially within the realm of string instruments, where the absence of fixed keys—characteristic of pianos or brass instruments—imparts a remarkable degree of freedom to the performer. While current motion capture (MoCap) methodologies excel in seizing common movements, they often stumble in accurately replicating these domain-specific actions, primarily due to a lack of specialized motion data. Despite these formidable obstacles, the capturing and subsequent recreation of string instrument performance holds vast potential in applications such as string instrument pedagogy, virtual concerts, and animation industries [Wheatland et al. 2015]. These fields are currently underserved by existing technologies, implying significant development opportunities. It is worth noting that string performance is an art form that intertwines both visual and auditory modalities, engaging both senses to culminate in a comprehensive musical expression. The movements of performers serve as the physical foundation for the music creation, embodying an ambiguous but inevitable connection between action and sound. As a result, our key observation involves establishing and further leveraging the subtle correlations between the music conveyed through audio and movement depicted visually during string performance. This approach may facilitate breakthroughs in this domain-specific motion capture, transcending related solutions grounded solely on a single modality.

The capture and analysis of musical instrument performance movements have received increasing attention in recent years [Papiotis et al. 2016; Perez-Carrillo 2019; Wanderley 2022; Zhang et al. 2022]. Currently, a common challenge in this field is the lack of available data, primarily due to the prohibitively high cost associated

with manual annotation. Table 1 lists some relative data sources, however, some of them are limited in scale, resolution, and number of viewpoints or only provide raw video without extracting motion information. Naturally, MoCap becomes indispensable, enabling in-depth analysis and understanding beyond the scope of standard video footage. MoCap methods are typically classified into marker-based and markerless (or marker-free) approaches depending on whether markers are involved. At present, the majority of existing studies utilize marker-based approaches with devices such as infrared sensors, electromagnetic systems and inertial systems to capture motion for various musical instrument performances, including clarinets [Teixeira et al. 2015], guitars [Perez-Carrillo 2019], drums [Gonzalez-Sanchez et al. 2019], violins [Volpe et al. 2017; Young and Deshmane 2007], cellos [Gonzalez-Sanchez et al. 2019], pianos [Payeur et al. 2014; Tits et al. 2015], instrument duos [Jakubowski et al. 2017; Thompson et al. 2017] and string quartets [Papiotis et al. 2016]. The primary advantage of marker-based methods lies in their precise positioning without the need for additional visual algorithms. However, they require special environmental setups such as infrared or magnetic fields, imposing certain demands on the experimental environment. Moreover, it is inevitable that markers deployed on the performer’s torso, limbs, and hands introduce varying degrees of hindrances to the musical performance. This obstacle makes it impractical to place markers on every finger joint, thereby preventing performers from executing their performance smoothly [Perez-Carrillo 2019], resulting in the loss of capturing the most critical sound-producing finger movements. On the other end of the spectrum, the markerless approach does not impede performers and enables the investigation of as many points as desired in standard imagery, laying the necessary groundwork for exploring detailed finger movements. However, the markerless approach lacks guaranteed accuracy, and currently, advanced pose estimators are trained on generic data without validation for performance in specific scenarios like musical instrument playing, where complex human-object occlusion and interaction relationships are involved, indicating a need for improvement. As of now, there has been no previous work focused exclusively on exploring fine-grained instrument-playing motions using the markerless MoCap approach, suggesting potential opportunities ahead. Another noteworthy aspect is that regardless of the involvement of markers, both MoCap approaches struggle to capture the contact information between the performer’s note-playing fingers and the instrument. However, this contact is crucial as it serves as a key signal for motion capture or motion generation during musical performance.

To address these challenges, we first captured a multi-view multi-modal dataset of string instrument performance on a totally marker-free basis, named String Performance Dataset (SPD). We publish the dataset and hold the belief that this non-intrusive setup is pivotal in enabling musicians to perform with freedom and naturalness, devoid of any movement distortion. This ensures that every subtle gesture they display holds inherent value, thereby enhancing the significance of our precise annotation method. In practice, we found that even with 23 cameras and the state-of-the-art MoCap networks, we still cannot obtain accurate hand poses due to the complex finger movements and the absence of constraints on hand-string interaction. Thus, we also present a baseline framework that

Table 1. A summary of commonly used music performance datasets, detailing their focused instruments, number of excerpts, duration, and forms of content.

| Dataset                              | Instrument              | Pieces     | Duration     | Camera Views      | Mocap Annotation                    |
|--------------------------------------|-------------------------|------------|--------------|-------------------|-------------------------------------|
| <i>Marker / Sensor Based Dataset</i> |                         |            |              |                   |                                     |
| TELM1 [Volpe et al. 2017]            | Violin                  | 41         | 2.4 h        | 3 + 13 (infrared) | Body, Instrument, Bow               |
| QUARTET [Papiotis et al. 2016]       | String quartet          | 30         | 0.5 h        | 1 + 26 (infrared) | Body, Instrument, Bow               |
| MMG [Perez-Carrillo et al. 2016]     | Guitar                  | 10         | 0.17 h       | N/A               | Body, Hands, Instrument             |
| EEP [Marchini et al. 2014]           | String quartet          | 23         | N/A          | 0 (wired EMF)     | Bow                                 |
| Bowstroke [Young and Deshmane 2007]  | Violin                  | N/A        | N/A          | 1                 | Bow                                 |
| <i>Markerless Dataset</i>            |                         |            |              |                   |                                     |
| CCOM-HuQin [Zhang et al. 2022]       | HuQin                   | N/A        | 1.29 h       | 3                 | N/A                                 |
| URMP [Li et al. 2018]                | Multi-instrument        | 44         | 1.3 h        | 1                 | N/A                                 |
| C4S [Bazzica et al. 2017]            | Clarinet                | 54         | 4.5 h        | 1                 | N/A                                 |
| ENST-Drums [Gillet and Richard 2006] | Drum kit                | N/A        | 3.75 h       | 2                 | N/A                                 |
| <b>SPD (ours)</b>                    | <b>Cello and Violin</b> | <b>120</b> | <b>3.0 h</b> | <b>23</b>         | <b>Body, Hands, Instrument, Bow</b> |

models both the performer and instrument with a combination of audio signals to enhance the vision-based MoCap. The proposed framework finally yields pose annotations aligned with hand-string interactions, resulting in more reasonable and accurate MoCap outcomes, especially for subtle finger movements. To conclude, our main contributions are: 1) The first large-scale multi-modal MoCap dataset for string instrument performance with fine-grained instrument-sound-aligned hand poses. 2) An audio-guided multi-modal framework for enhancing markerless motion capture for string performance.

## 2 RELATED WORK

**Instrument performance dataset:** Motion capture in musical instrument performance requires data of visual modality. However, visual resources are scarce compared to audio data. Table 1 lists several datasets featuring visual modalities or MoCap data of musical instrument performance, along with their detailed information. The datasets on a marker-free basis like URMP [Li et al. 2018], C4S [Bazzica et al. 2017], ENST-Drums [Gillet and Richard 2006] and CCOM-HuQin [Zhang et al. 2022] merely contain raw videos from limited views, without capturing performance movements or providing related motion data. Most marker-based datasets only contain coarse-grained MoCap annotation without hand details [Marchini et al. 2014; Papiotis et al. 2016; Volpe et al. 2017; Young and Deshmane 2007], consequently missing the most essential movements in instrument performance scenarios. Although Multi-modal Guitar [Perez-Carrillo et al. 2016] includes finger movements for guitar playing, the dataset is quite limited, containing only 10 data entries totaling 10 minutes. Additionally, [Simon et al. 2017] collects cello performance videos to validate their hand keypoint detection algorithm. However, this work does not specifically focus on instrument performance, and as a result, the amount of data is quite limited and fails to model the instrument.

**Marker-based instrument performance MoCap:** This type of approach utilizes sensors or markers attached to performers and instruments to capture their spatial location. Specific technologies include infrared markers [Rozé et al. 2018], electromagnetic field (EMF) sensing [Papiotis et al. 2016], and inertial measurement unit (IMU) [Young and Deshmane 2007]. Some scholars also integrate

multiple sensors to improve MoCap accuracy. [Schoonderwaldt and Demoucron 2009] combine the infrared markers and IMU to capture motion for violin performance. [Gonzalez-Sanchez et al. 2019; Volpe et al. 2017] introduce the electromyogram (EMG) as an auxiliary tool to assist in instrumental performance MoCap. Among these technologies, the infrared method is most popular owing to its high capturing accuracy with relatively small marker sizes (3 mm around), which has also been extended to multiple musicians like violin duets [Thompson et al. 2017], string quartets [Maestre et al. 2017], and musical ensembles [Hilt et al. 2019]. However, the attached marker inevitably disrupts the hand or finger movements of the performer. [Perez-Carrillo 2019; Perez-Carrillo et al. 2016] we mentioned above, investigate the finger-string interaction in guitar playing, placing markers on the dorsal side of each finger articulation. However, the contact between fingers and strings primarily occurs on the palm side. This deviation, compounded by errors stemming from marker size, becomes substantial when analyzing intricate finger movements, rendering this approach both intrusive to the playing and unreliable regarding accuracy. Therefore, most of the above works only focus on larger body parts like arms and torso to explore expressive movements, failing to effectively capture the sound-producing actions of the performer’s hands.

**Vision-based instrument performance MoCap:** Vision-based motion capture technology facilitates the analysis of myriads of points from videos, which is not constrained by the number of markers. This provides a necessary foundation for exploring the intricate finger movements of musicians. [Hadjakos et al. 2013] apply a purely vision-based approach to analyze violin performances with Kinect. However, they focus mainly on the performer’s head movements and do not delve into finger movements. Beyond research specifically targeting instruments, some general human pose estimators can also capture these scenarios with vision-based algorithms [Cao et al. 2017; Lugaresi et al. 2019; Yang et al. 2023; Zhang et al. 2020]. Yet, they fall short in training with domain-specific data, underscoring the necessity to improve their accuracy in detecting the unique postures typical in musical performances.

**Multi-modal approach in musical performance:** Recent works integrating audio modality focus on generating movements for violin playing [Hirata et al. 2022, 2021; Kao and Su 2020; Shrestha

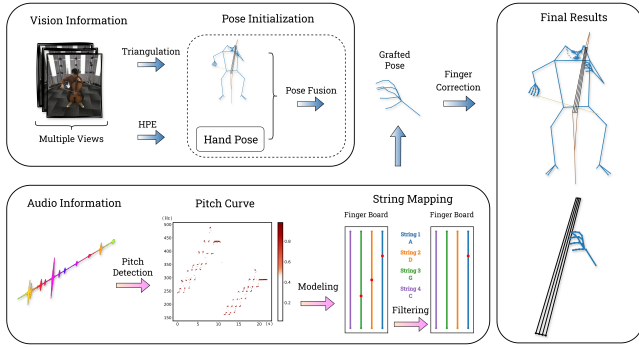


Fig. 2. The pipeline combines the information extracted from the visual and auditory inputs. The following sections elaborate on detailed explanations of each part.

et al. 2022] and singing [Pan et al. 2022]. Earlier studies on hand and finger modeling and animating explore mapping sheet music (or symbolic music notation) to finger movements using rule-based approaches, involving animation generation for guitar [ElKoura and Singh 2003], violin [Kim et al. 2000], and piano [Kugimoto et al. 2009; Zhu et al. 2013]. However, these vague mappings from audio or sheet music to performance are unable to precisely replicate the real-world actions, resulting in relatively rough playing motions that fail to capture the nuanced original movements.

### 3 METHODOLOGY

In this section, we present the process of constructing the String Performance Dataset by providing an in-depth look at the complete pipeline of our MoCap framework, which is specially designed to capture string instrument performance. This includes detailed coverage of all stages from data acquisition to final output. As illustrated in Fig.2, under the premise of not involving markers or sensors, we integrate domain knowledge of the specific string instruments (cello or violin) and features extracted from the music produced by the instrument playing to constrain and further optimize the reconstruction results achieved solely through vision-based methods. We concentrate on the cello and violin, two of the most mainstream classical string instruments, which exhibit significant differences in playing postures, instrument sizes, and musical ranges.

#### 3.1 Data Acquisition

MoCap for the string performance suffers from severe occlusion caused by the complex interactions between the performer and the instrument. To minimize this challenge, we establish a multi-view MoCap system, as shown in Fig.3. The system consists of up to 23 cameras distributed across 12 fixed poles positioned at the "hour marks" around the performer, covering a cylindrical space with a diameter of approximately 4 meters. While our system can function effectively in both cello and violin scenarios with identical camera settings, we adopt slightly different camera distributions to achieve a more comprehensive Line-of-Sight coverage by positioning additional cameras in front of the violin performers. This arrangement

addresses challenges inherent in capturing violin performances, including variations in fingerboard orientations and the smaller size of the target compared to the cello. Each camera is focused on the performer at the center of the venue, and solid-colored curtains are placed around the venue to eliminate any potential distractions from the background. All used cameras are FLIR's ORX-10G-245S8C, featuring a resolution of 2656x2300, a frame rate of 30 fps, and input/output synchronization. For audio recording, we use a Sony ICD-PX470 microphone fixed in front of the performer. In addition, a manual clap is applied to align the video and audio signals before each performance recording.

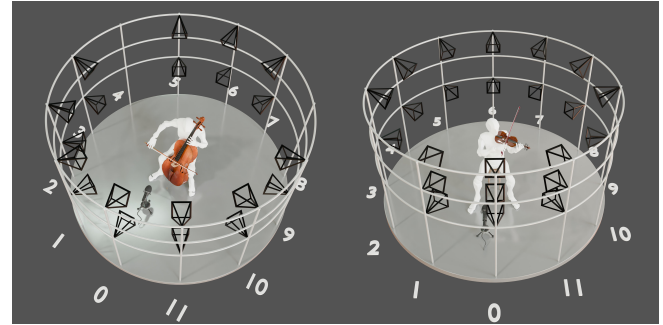


Fig. 3. Our recording setup, with slight differences between the cello and violin scenarios in the camera numbers and positions. 20 cameras for the cello and 23 cameras for the violin.

The data we collected includes performances from a total of 9 cellists and violinists, ranging from amateurs and conservatory students to acclaimed professionals. We apply the standard of 440 Hz for the A4 note during string tuning. The repertoire spans a diverse range, including scales, études, classical compositions, and modern pop music, encompassing 120 pieces with a cumulative duration exceeding 3 hours. This breadth ensures a robust and versatile dataset, suitable for a wide range of applications in musical research and practice. Most importantly, all performances take place in a completely marker-free setting, allowing musicians to fully express their movements. This greatly enhances the value of our subsequent detailed and nuanced capture of these actions.

#### 3.2 Pose Initialization

In this subsection, we describe how we initialize the spatial information of the performer and instrument. Spatial information requires reconstruction from multiple views, hence the necessity to first obtain 2D-keypoint locations. Given that our points of interest are distributed across three parts: the human, the instrument, and the bow, each having distinct motion patterns, we employ different algorithms to extract them separately based on 2D images.

**Performer pose initialization:** Identifying body or hand movements is invariably complex, yet thankfully, a wealth of advanced methods, devices, and datasets [Li et al. 2019; Wheatland et al. 2015; Zheng et al. 2023] exist to support our endeavors. We apply the open-source *DWPose* model [Yang et al. 2023], a state-of-the-art whole-body pose estimator that surpasses two widely used models *OpenPose* [Cao et al. 2017] and *MediaPipe* [Lugaresi et al. 2019;

Zhang et al. 2020], to identify keypoints of the performer from 2D images, including fingers, limbs, torso, and facial expressions, totaling 133 keypoints. DWPose model achieves precise localization in human keypoint detection, however, it may cause hand distortions that do not follow anthropometrics. Some extreme relationships of finger positions are unusual in daily life but frequently occur on the left hand of string instrument performers when conducting complicated playing techniques. To tackle this issue, we specifically train a hand pose estimator (HPE) using the combination of several commonly used datasets for hand pose estimation, including InterHand2.6M [Moon et al. 2020], DARTset [Gao et al. 2022], BlurHand [Oh et al. 2023], and HanCo [Zimmermann et al. 2021]. In the HPE implementation, we follow the backbone architecture and implementation details of InterNet which is outlined in [Moon et al. 2020]. However, we make a notable modification to its output layer. Instead of outputting hand joint positions, the final layer is transitioned to predict a 6D representation for 3D hand joint rotation [Zhou et al. 2019], constrained by the MANO model [Romero et al. 2022]. This modification yields three pivotal effects. First, integrating the MANO model introduces anthropometrics priors of the human hand, thereby guiding the HPE to emphasize the correctness and reasonableness of hand poses. Secondly, the prediction of 3D hand joint rotation provides a complete kinematic chain from the wrist joint to each fingertip, bringing forth the feasibility of further hand pose correction based on inverse kinematics methods, as elaborated later. Lastly, supported by comprehensive empirical results in [Zhou et al. 2019], the continuous 6D representation has been proven to be beneficial for training deep neural networks, leading to faster convergence and lower loss when compared to commonly used but discontinuous ones, such as quaternions and Euler angles. As for the inference process of performers' hand poses, we conduct predictions of the HPE model on various views of imagery. Subsequently, in order to achieve the integration of the results from multiple perspectives, the inferred 6D representations are converted into quaternions, following which they will undergo interpolation based on the assigned weights of each camera. We assign weights according to the relative orientation between the camera and the performer, prioritizing perspectives that offer clearer views of the left hand. Up to this point, we obtain the performer's left-hand pose prediction by the HPE model across multiple viewpoints. However, the question arises: how do we "graft" such hand pose prediction onto the predictions of the rest of the body parts from DWPose to fuse the results? The key lies in determining the angle at which the left wrist joint connects. First, we align the wrist positions from both results. Then we adopt the approach of inverse kinematics, taking the hand pose prediction from DWPose as the target and continuously rotating and translating the whole hand pose prediction from the HPE model around the wrist joint until getting the minimal Euclidean distance of all corresponding hand joints between the two sets of results. This optimization process can be expressed as:

$$(\theta_w^*, T^*) = \underset{\theta_w, T}{\operatorname{argmin}} \left( \sum \|J_{\text{HPE}}(\theta_w, T) - J_{\text{DWPose}}\| \right), \quad (1)$$

where  $J_{\text{HPE}}$  and  $J_{\text{DWPose}}$  represent the positions of all hand keypoints predicted by the HPE and DWPose respectively.  $\theta_w$  and  $T$  denote the rotation matrix at the wrist joint and the translation



Fig. 4. Illustration of cello keypoints.

vector to be optimized from the HPE prediction. As for the rest of the body parts beyond the left hand, we use the results predicted by the DWPose model directly.

**Instrument pose initialization:** Modeling the instrument is essential, particularly the strings, as they are the most direct interactive elements between the performer and the instrument, encompassing key sound-producing finger movements. Currently, there is no established ready-to-use method for localizing and modeling string instruments. However, due to the rigid structure of the instrument, we can infer its overall state by determining a few keypoints after taking its geometric structure as prior knowledge. We leverage Google's TAPNet [Doersch et al. 2023] for keypoints tracking as it exhibits exceptional performance in tracking such gentle and slight rotation or displacement of the instrument body during the performance. We specify the points of our interest in the first frame of the video being analyzed, and the corresponding positions of these points will be automatically retrieved in the subsequent frames. In Fig.4, we illustrate how we select keypoints to be tracked on the cello. A similar setup is used for the violin, except for removing the end pin and the tail gut. Once the nut and bridge are located, the positions of the strings can be deduced based on geometric relationships. Direct positioning is avoided because the strings are less visible in the imagery.

**The bow:** The pose of the bowing hand (right hand) does not vary much when performing, but the speed, angle, and relative position between the bow and strings are viewed as key elements of the playing action. The mentioned TAPNet is not robust for bow tracking due to the distinct characteristics of bow movements compared to those of the instrument body. Bow movements involve extensive and high-speed movements, especially when playing intense notes. We trained our bow detection model based on YOLOv8 architecture [Jocher et al. 2023] to identify the frog and tip plate of the bow.

Next, we integrate keypoints from synchronized frames across various views through triangulation, obtaining the three-dimensional spatial information of the scenario. In this process, no single camera can independently capture all keypoints of interest due to occlusions caused by the instrument and the performer, self-occlusion by the performer, and even the obstructive relationship between the bow and the instrument. To address this issue, we apply the Random Sample Consensus algorithm (RANSAC) to exclude outliers in the reconstruction of each spatial point.

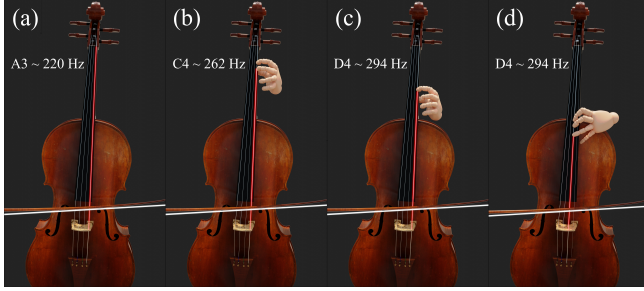


Fig. 5. Using the cello as a reference, we present examples of various note-playing finger positions alongside their corresponding vibrating lengths (highlighted in red) and pitch values. (a) represents an open string note, producing the lowest pitch achievable on the excited string. (b) and (c) demonstrate different finger positions on the same string, resulting in different pitches. (c) and (d) illustrate instances in which varying finger positions on different strings produce the same pitch.

All the motion capture processes described above are completed on a purely visual and marker-free basis, achieving the reconstruction of keypoints in three-dimensional space for the performer, the instrument, and the bow in string instrument performance scenes. In the following subsection, we elaborate on how we use the audio-guided approach to further optimize the results.

### 3.3 Audio-guided Multi-Modal Motion Capture

In string instrument performance, playing movements serve as the physical basis of music production. Building on this idea, we believe that music itself contains key information to infer the interaction between the fingers and the instrument. By utilizing our designed Pitch-Finger model and incorporating domain knowledge from string instruments, we map the musical pitch information extracted from audio onto specific locations on the fingerboard of the instrument. This provides the foundation for correcting our initial motion results obtained from the visual modality.

**3.3.1 Pitch Detection.** Unlike noise or human speech, string instruments produce sound with a regular and repetitive waveform, resulting in a distinct sound frequency, i.e., the pitch. The performers alter the pitch by adjusting their left-hand position on the fingerboard and the strings they touch. Meanwhile, the mapping between the playing movements and pitch is much clearer than that of the other sound elements such as timbre, duration, and intensity. We utilize the CREPE pitch tracker [Kim et al. 2018], which achieves up to 99.9% accuracy with a threshold of 25 cents in monophonic pitch detection, to determine the pitch curve of our recorded audio. In this process, we maintain consistency between the frequency of applying pitch estimation in music and the frame rate of our video recording (30 fps), ensuring a one-to-one correspondence between each frame and its associated pitch value.

**3.3.2 Pitch-Finger Model.** The vibrating length refers to the portion of the string that actually vibrates to produce sound, highlighted in red in Fig.5, indicating the pitch. The relationship between pitch  $P$

and vibrating length  $L_{vib}$  can be presented as:

$$P = \frac{F \cdot L_{fund}}{L_{vib}}, \quad (2)$$

where  $L_{fund}$  is the full length of the string, known as fundamental length.  $F$  represents the frequency of vibration along the fundamental length (Fig. 5-a). When a string is excited by the bow, the lowest sound it can produce is its fundamental frequency, while theoretically, there is no upper limit for the highest pitch. Therefore, given a note whose pitch is above the fundamental frequency of a certain string on a string instrument, the vibrating length of the target string can be uniquely determined by Equation (2).

Performers use their left hands to adjust the vibrating length by changing the note-playing finger position, at which the string is pressed (Fig. 5-b and c). With the detected pitch value, we aim to determine the expected note-playing finger position deduced from the audio, referred to as the audio-guided position in the following discussion. The audio-guided position serves as the target for correcting finger movements. However, string instruments typically have four strings, with some overlap in pitch range. This overlap can lead to ambiguous finger placement, as the same pitch might be produced by multiple finger positions on different strings (demonstrated in examples c and d of Fig. 5). Therefore, accurately identifying the audio-guided position solely based on pitch poses a significant challenge. To determine it precisely, we use our initial results from the vision-based algorithm for the left hand as guidance. Initially, we identify multiple potential finger positions based on the detected pitch, all of which are reasonable for producing the target note frequency. Then, we ascertain the actual finger position of the performers by referring to their wrist position from the results obtained in section 3.2. The feasibility of this step is twofold. On the one hand, as illustrated in Fig. 5 -c and d, the apparent difference in finger position of the same note on different strings greatly reduces the threshold of filtering the final audio-guided position from potential ones. On the other hand, benefiting from the well-adjusted multi-view setup, our previous results based solely on visual detection are quite reliable for further optimization according to the audio signals. Finally, once the audio-guided position is determined, we bind it to its nearest fingertip, forming a pair between the target and the fingertip to be corrected. At each time step, we focus on adjusting the note-playing finger to the corresponding time.

In addition, we incorporate domain knowledge of string instruments to guide the algorithm in binding the audio-guided position to the note-playing finger. When performing the vibrato effect, the pitch exhibits rapid and slight fluctuations, and the note-playing finger for these consecutive frames remains unchanged. The physical premise of this vibrato effect is the fast overall trembling of the left hand. However, this trembling can change the distance between the audio-guided position and each fingertip, affecting the binding relationship. To avoid such circumstances, we assess the pitch curve by detecting the pattern of pitch variation. If the pitch variation across consecutive frames is within  $\pm 30$  cents (100 cents per semitone), vibrato is confirmed, and the binding relationship from the onset to the end of the vibrato remains constant.

Identifying the binding relationship is crucial for further detailed optimization of the note-playing finger. The external force exerted

by the fingerboard on the left hand during string instrument performance can cause varying degrees of joint deformation. Such deformation induces deviation in the prediction as our used hand-pose estimator (HPE) is trained on data where the hand is not subjected to external forces. Hence, we further employ inverse kinematics to guide the note-playing fingertip towards the audio-guided position, which can be expressed as:

$$(\theta_d^*, \theta_p^*) = \underset{\theta_d, \theta_p}{\operatorname{argmin}} \|J_{\text{tip}}(\theta_d, \theta_p) - J_{\text{audio-guided}}\|. \quad (3)$$

We take the distance between the audio-guided position ( $J_{\text{audio-guided}}$ ) and the tip of the note-playing finger ( $J_{\text{tip}}$ ) as the cost function. We then fine-tune the rotation angles ( $\theta_d, \theta_p$ ) of the Proximal Interphalangeal joint and the Distal Interphalangeal joint of the note-playing finger using the L-BFGS-B algorithm [Zhu et al. 1997], a method previously employed in [Tsang et al. 2005], affirming its efficiency in rectifying hand poses. This ensures optimal alignment between the note-playing fingertip and the audio-guided position. Even though there might theoretically be multiple solutions for the target finger angles, the strong constraints of audio-guided position and the proximity of our initial results to these optima enable us to consistently achieve desirable finger postures within a limited solution space. Additionally, employing shorter optimization iteration step size helps avoid unreasonable solutions that adhere only to mathematics but not to anthropometrics.

#### 4 RESULTS AND COMPARISON

With our proposed multi-modal MoCap framework, 3D movement information can be automatically extracted from the original multi-view videos. This 3D representation not only includes motion capture data of the performer and instrument but also accurately replicates the performer's hand contact and relative position with the instrument. In Fig.6, we present the global 3D results observed from multiple views and validate the reprojection results by overlaying the original images. Our 3D performance motion reconstruction demonstrates reasonable results for both cello and violin performance.

Moreover, we conduct an ablation study to gain a more comprehensive understanding of how our audio-guided approach enhances vision-based MoCap results. For quantitative analysis, we utilize mean per joint position error (MPJPE) as our metric to assess both the entire left hand and specifically the note-playing finger. The MPJPE metric measures Euclidean distance (in millimeters) between estimated coordinates and the ground-truth coordinates after aligning their root joints (i.e., wrist joints) through translation. While acquiring ground-truth 3D annotations is not straightforward since complex hand poses and occlusions are ubiquitous in string performance. To obtain ground-truth 3D annotations of each frame, we manually annotate the performer's left hand on 2D images from four different perspectives (simultaneously captured, with a clear presentation of the left hand), before triangulating these 2D annotations to 3D skeletons for metric calculation. All the annotators possess certain professional experience in playing string instruments to guarantee a reliable level of accuracy in annotating.

A total of 200 frames are involved in the evaluation, covering comprehensive holds (wrist positions or finger placements with

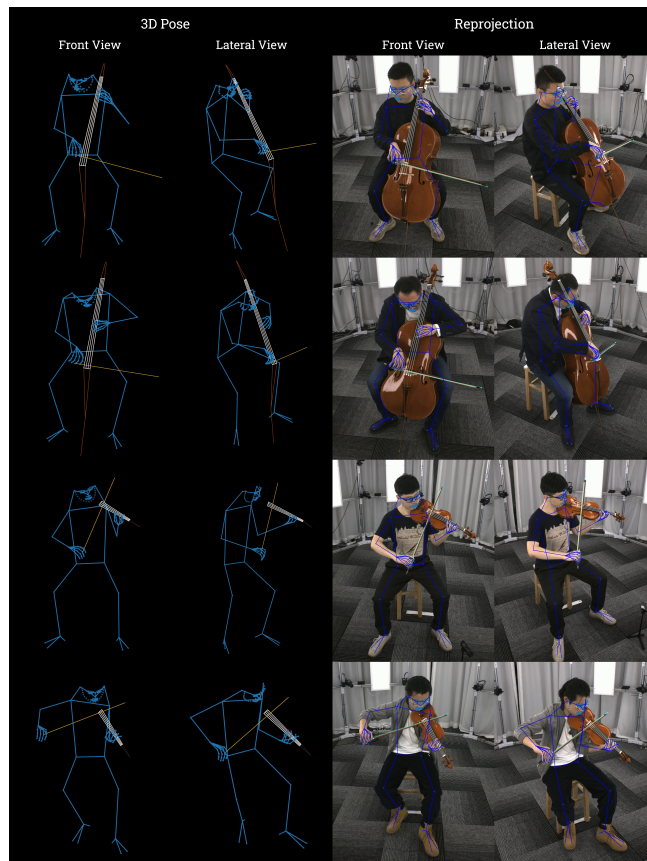


Fig. 6. Demonstration of our results through reprojection and 3D visualization from different views.



Fig. 7. Some examples from the test set are presented, with a frontal zoomed-in view of the left hand, covering various holds. The test set contains 200 frames, each of which corresponds to 4 images of different viewpoints. We manually annotate these images before the triangulation to obtain the ground-truth 3D annotations of corresponding frames.

various hand poses on the fingerboard to produce specific notes) for both cello and violin performances. Some examples from our test set are listed in Fig.7. Table 2 summarizes the MPJPE results obtained from the DWPose model (the state-of-the-art vision-based model), our HPE model but without the audio-guided module, and our full approach respectively. In both the evaluation in terms of the entire hand and specifically the note-playing finger, our HPE model outperforms DWPose, and moreover, the integration of the audio-guided approach further enhances performance. Given that the note-playing finger is pivotal in sound production for string instrument

Table 2. Ablation study: Evaluation on MPJPE and contact deviation, along with the improvements over the baseline (DWPose) results.

| Method                 | MPJPE (whole hand) |          | MPJPE (note-playing finger) |          | Contact Deviation |          |
|------------------------|--------------------|----------|-----------------------------|----------|-------------------|----------|
| DWPose                 | 17.00              | —        | 16.14                       | —        | 22.40             | —        |
| HPE (w/o Audio-guided) | 15.27              | ↓ 10.2 % | 14.95                       | ↓ 7.4 %  | 16.06             | ↓ 28.3 % |
| HPE + Audio-guided     | <b>14.93</b>       | ↓ 12.2 % | <b>13.27</b>                | ↓ 17.8 % | <b>5.19</b>       | ↓ 76.8 % |

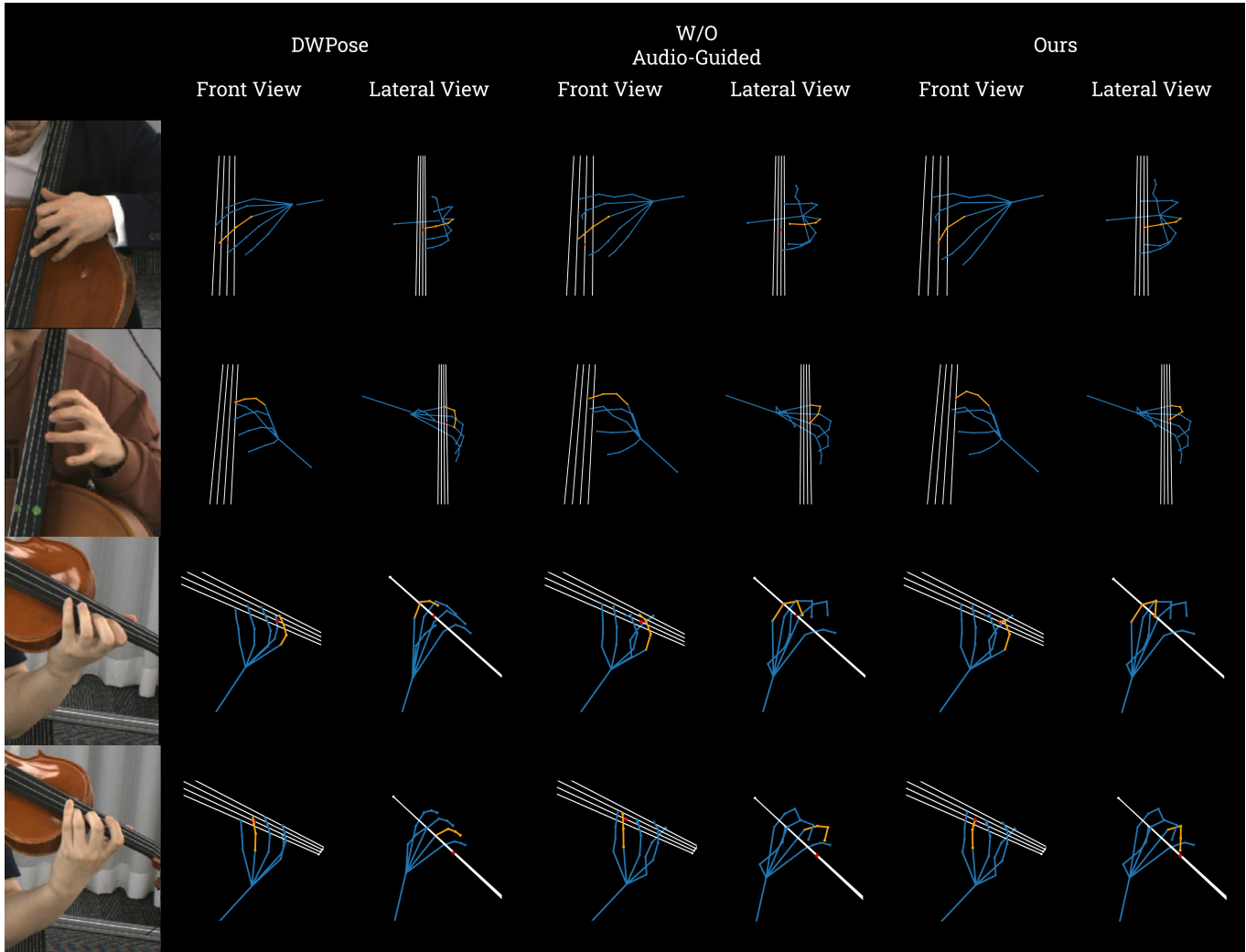


Fig. 8. Close-up details show a comparison of various hand positions during performance with different degrees of optimization. Column "DWPose" presents the results from the state-of-the-art vision-based model without applying either our HPE model or the audio-guided approach, while the middle column shows the results from the HPE model only. The final column, labeled "Ours", incorporates our full approach.

performance, we conduct an analysis focusing exclusively on this crucial element, where the audio-guided approach exhibits a more pronounced effect.

Apart from MPJPE evaluation, we calculate the contact deviation to determine whether the contact relationship between the performer and the instrument is accurately reconstructed. This deviation is the Euclidean distance between the estimated position

of the note-playing fingertip and the theoretical finger position inferred from the pitch. Before evaluation, these theoretical positions representing the expected locations of the note-playing finger's tip were manually verified. This measurement is presented in the last column of Table 2. The evaluation from this perspective indicates that integrating the audio-guided approach greatly improves the details of finger placement for pressing strings, thereby clarifying



the contact relationship between the note-playing fingers and the instrument during the performance.

To further provide qualitative insights, we present detailed close-up views of the performer's left hand from various perspectives in Fig.8, which corroborate the findings of the quantitative evaluation.

## 5 DISCUSSION

The proposed motion capture framework and the String Performance Dataset have limitations that suggest future work. Unlike harmonic instruments like pianos, strings, as melodic instruments, mainly employ a monophonic playing style. However, polyphonic performance, where multiple strings are triggered simultaneously, may occur in certain pieces. The accuracy of polyphonic pitch detection poses a limitation of our currently employed approach, therefore, we do not include such a scenario as unreliable detection may mislabel note-playing fingers. In addition, we primarily focus on using the audio-guided approach to assist in capturing the performer's left hand. The potential for expanding the application of the audio-guided motion capture to include other body parts is on the horizon, necessitating the acquisition of additional constraints from the musical aspect. Apart from pitch information, other elements inherent in the audio, such as volume and note duration, remain unexplored. This exploration holds the promise of providing us with further insights for optimizing the results of the bow-holding hand. Furthermore, the recording frame rate is set at 30 frames per second, potentially posing a limitation for the upper bound of the fast motion.

Our contribution is capable of supporting research in string performance analysis and string instrument pedagogy. Future efforts aiming at achieving comprehensive coverage of string instruments could include the exploration of the viola and double bass, both of which share significant similarities with the strings family. Besides, with the availability of SPD, there is potential to create mappings between music and playing movements of string instruments. This advancement could facilitate the achievement of high-fidelity and professional-grade motion generation in animations or virtual avatars. On the other hand, in pursuit of flexibility in utility and a lightweight data collection system, the research could extend to monocular MoCap (from online videos or real-life performances recorded on mobile devices), which presents super challenges, especially with subtle instrumental performances. Nevertheless, we believe our audio-guided method demonstrates a promising and practical solution. It can also be combined with other newly proposed MoCap technologies via VR hardware [Han et al. 2022], RF-vision [Zhang et al. 2023b], MEMS-ultrasonic sensors [Zhang et al. 2023a], and even special patterns attached to the human body or hand [Chen et al. 2021]. We leave this for our forthcoming research.

A broader perspective suggests that the application of the audio-guided approach can be extended to fields like sports (where the sound of ball impacts is linked with striking actions), dance performance (with movement transitions synchronized with musical cues), and other domains where human movements are closely associated with sounds. This could break through the inherent Line-of-Sight subjection in conventional camera-based MoCap to a certain extent, effectively improving results in situations with occlusion or contact.

## 6 CONCLUSION

In summary, this paper presents the String Performance Dataset (SPD), which contains 3 hours of multi-view videos, corresponding 3D MoCap annotations, and synchronized audio signals. We also propose an audio-guided multi-modal framework to capture the fine-grained finger movements and instrument poses on a completely marker-free basis. It leverages the definitive correlations between audio and hand position, resulting in precise hand pose estimation that matches the exact instrument's sound. Such a novel audio-guided approach can serve as an inspiration for using audio information to enhance visual restoration in other scenarios. To this end, SPD is the first large-scale multi-modal MoCap dataset for string instrument performance with precise instrument-sound-aligned hand poses. The SPD surpasses existing similar datasets in terms of data volume, shooting angle variety, and the granularity of captured movement, effectively mitigating the data scarcity in the field of instrument performance. Based on the current progress, research on more generalized MoCap settings and the task of string performance generation will be carried out in the near future.

## ACKNOWLEDGMENTS

We thank Haotian Zhou, Xinghong Wang, Ziyi Huang, Shihao Yao, Yutong Ding, and Yuetonghui Xu for their performances in data acquisition. This work was supported in part by the National Key R&D Program of China (No.2022YFF0902204), in part by the National Natural Science Foundation of China under Grant No.62171255, in part by the Tsinghua University-Joint research and development project under Grant R24119F0 JCLFT-Phase 1, in part by the "Light Field Generic Technology Platform" (Z23111000290000) of Beijing Municipal Science and Technology Commission, in part by the Guoqiang Institute of Tsinghua University under Grant No.2021GQG0001, in part by the Special Program of National Natural Science Foundation of China under Grant No.T2341003, in part by the Major Program of the National Social Science Fund of China under Grant No.21ZD19, in part by the Advanced Discipline Construction Project of Beijing Universities, and in part by the Nation Culture and Tourism Technological Innovation Engineering Project (Research and Application of 3D Music).

## REFERENCES

- Alessio Bazzica, JC Van Gemert, Cynthia CS Liem, and Alan Hanjalic. 2017. Vision-based detection of acoustic timed events: a case study on clarinet note onsets. *arXiv preprint arXiv:1706.09556* (2017).
- Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.
- He Chen, Hoyoon Park, Kutay Macit, and Ladislav Kavan. 2021. Capturing detailed deformations of moving human bodies. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–18.
- Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. 2023. TAPIR: Tracking Any Point with per-frame Initialization and temporal Refinement. *ICCV* (2023).
- George Elkoura and Karan Singh. 2003. Handrix: animating the human hand. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 110–119.
- Daiheng Gao, Yulian Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. 2022. DART: Articulated hand model with diverse accessories and rich textures. *Advances in Neural Information Processing Systems* 35 (2022), 37055–37067.
- Olivier Gillet and Gaël Richard. 2006. Enst-drums: an extensive audio-visual database for drum signals processing. In *International Society for Music Information Retrieval*

- Conference (ISMIR).
- Victor Gonzalez-Sanchez, Sofia Dahl, Johannes Lunde Hatfield, and Rolf Inge Godøy. 2019. Characterizing movement fluency in musical performance: Toward a generic measure for technology enhanced learning. *Frontiers in psychology* 10 (2019), 84.
- Aristotelis Hadjakos, Tobias Großhauser, and Werner Goebel. 2013. Motion analysis of music ensembles with the Kinect. In *Conference on New Interfaces for Musical Expression*. 106–110.
- Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. 2022. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.
- Pauline M Hilt, Leonardo Badino, Alessandro D’Ausilio, Gualtiero Volpe, Ser a Tokay, Luciano Fadiga, and Antonio Camurri. 2019. Multi-layer adaptation of group coordination in musical ensembles. *Scientific reports* 9, 1 (2019), 5854.
- Asuka Hirata, Keitaro Tanaka, Masatoshi Hamanaka, and Shigeo Morishima. 2022. Audio-Driven Violin Performance Animation with Clear Fingering and Bowing. In *ACM SIGGRAPH 2022 Posters*. 1–2.
- Asuka Hirata, Keitaro Tanaka, Ryo Shimamura, and Shigeo Morishima. 2021. Bowing-Net: Motion Generation for String Instruments Based on Bowing Information. In *ACM SIGGRAPH 2021 Posters*. 1–2.
- Kelly Jakubowski, Tuomas Eerola, Paolo Alborno, Gualtiero Volpe, Antonio Camurri, and Martin Clayton. 2017. Extracting coarse body movements from video in music performance: A comparison of automated computer vision techniques with motion capture data. *Frontiers in Digital Humanities* 4 (2017), 9.
- Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. YOLO by Ultralytics. <https://github.com/ultralytics/ultralytics>
- Hsuan-Kai Kao and Li Su. 2020. Temporally guided music-to-body-movement generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 147–155.
- Junhwan Kim, Frederic Cordier, and Nadia Magnenat-Thalmann. 2000. Neural network-based violinist’s hand animation. In *Proceedings Computer Graphics International 2000*. IEEE, 37–41.
- Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 161–165.
- Nozomi Kugimoto, Rui Miyazono, Kosuke Omori, Takeshi Fujimura, Shinichi Furuya, Haruhiro Katayose, Hiroyoshi Miwa, and Noriko Nagata. 2009. CG animation for piano performance. In *SIGGRAPH’09: Posters*. 1–1.
- Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. 2018. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia* 21, 2 (2018), 522–535.
- Rui Li, Zhenyu Liu, and Jianrong Tan. 2019. A survey on 3D hand pose estimation: Cameras, methods, and datasets. *Pattern Recognition* 93 (2019), 251–272.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
- Esteban Maestre, Panagiotis Papiotis, Marco Marchini, Quim Llimona, Oscar Mayor, Alfonso P erez, and Marcelo M Wanderley. 2017. Enriched multimodal representations of music performances: Online access and visualization. *Ieee Multimedia* 24, 1 (2017), 24–34.
- Marco Marchini, Rafael Ramirez, Panos Papiotis, and Esteban Maestre. 2014. The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research* 43, 3 (2014), 303–317.
- Gyeongsik Moon, Shooou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX* 16. Springer, 548–564.
- Yeonguk Oh, Joonkyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee. 2023. Recovering 3D hand mesh sequence from a single blurry image: A new dataset and temporal unfolding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 554–563.
- Yifang Pan, Chris Landreth, Eugene Fiume, and Karan Singh. 2022. VOCAL: Vowel and Consonant Layering for Expressive Animator-Centric Singing Animation. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.
- Panagiotis Papiotis et al. 2016. *A computational approach to studying interdependence in string quartet performance*. Ph. D. Dissertation. Universitat Pompeu Fabra.
- Pierre Payeur, Gabriel Martins Gomes Nascimento, Jillian Beacon, Gilles Comeau, Ana-Maria Cretu, Vincent D’Aoust, and Marc-Antoine Charpentier. 2014. Human gesture quantification: An evaluation tool for somatic training and piano performance. In *2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) Proceedings*. IEEE, 100–105.
- Alfonso Perez-Carrillo. 2019. Finger-string interaction analysis in guitar playing with optical motion capture. *Frontiers in Computer Science* 1 (2019), 8.
- Alfonso Perez-Carrillo, Josep-Llu s Arcos, and Marcelo Wanderley. 2016. Estimation of guitar fingering and plucking controls based on multimodal analysis of motion, audio and musical score. In *Music, Mind, and Embodiment: 11th International Symposium, CMMR 2015, Plymouth, UK, June 16-19, 2015, Revised Selected Papers 11*. Springer, 71–87.
- Javier Romero, Dimitrios Tzionas, and Michael J Black. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610* (2022).
- Jocelyn Roz e, Mitsuko Aramaki, Richard Kronland-Martinet, and Solvi Ystad. 2018. Assessing the effects of a primary control impairment on the cellists’ bowing gesture inducing harsh sounds. *IEEE Access* 6 (2018), 43683–43695.
- Erwin Schoonderwaldt and Matthias Demoucron. 2009. Extraction of bowing parameters from violin performance combining motion capture and sensors. *The Journal of the Acoustical Society of America* 126, 5 (2009), 2695–2708.
- Snehesht Shrestha, Cornelia Ferm uller, Tianyu Huang, Pyone Thant Win, Adam Zukerman, Chethan M Parameshwara, and Yiannis Aloimonos. 2022. AIMusicGuru: Music Assisted Human Pose Correction. *arXiv preprint arXiv:2203.12829* (2022).
- Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1145–1153.
- Euler CF Teixeira, Mauricio A Loureiro, Marcelo M Wanderley, and Hani C Yehia. 2015. Motion analysis of clarinet performers. *Journal of New Music Research* 44, 2 (2015), 97–111.
- Marc R Thompson, Georgios Diapoulis, Tommi Himberg, and Petri Toivainen. 2017. Interpersonal Coordination in Dyadic Performance. In *The Routledge Companion to Embodied Music Interaction*. Routledge, 186–194.
- Micka el Tits, Jo elle Tilmann, Nicolas D’Alessandro, and Marcelo M Wanderley. 2015. Feature extraction and expertise analysis of pianists’ Motion-Captured Finger Gestures. In *ICMC*.
- Winnie Tsang, Karan Singh, and Eugene Fiume. 2005. Helping hand: an anatomically accurate inverse dynamics solution for unconstrained hand motion. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 319–328.
- Gualtiero Volpe, Ksenia Kolykhalova, Erica Volta, Simone Ghisio, George Waddell, Paolo Alborno, Stefano Piana, Corrado Canepa, and Rafael Ramirez-Melendez. 2017. A multimodal corpus for technology-enhanced learning of violin playing. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*. 1–5.
- Marcelo M Wanderley. 2022. *The Oxford Handbook of Music Performance*. Vol. 2. Oxford University Press. 465–494 pages.
- Nkenge Wheatland, Yingying Wang, Huaguang Song, Michael Neff, Victor Zordan, and Sophie J org. 2015. State of the art in hand and finger modeling and animation. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 735–760.
- Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4210–4220.
- Diana Young and Anagha Deshmane. 2007. Bowstroke database: a web-accessible archive of violin bowing data. In *Proceedings of the 7th international conference on New interfaces for musical expression*. 352–357.
- Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).
- Qiang Zhang, Yuanqiao Lin, Yubin Lin, and Szymon Rusinkiewicz. 2023a. Hand Pose Estimation with Mems-Ultrasonic Sensors. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.
- Shujie Zhang, Tianyue Zheng, Zhe Chen, Jingzhi Hu, Abdelwahed Khamis, Jiajun Liu, and Jun Luo. 2023b. OCHID-Fi: Occlusion-Robust Hand Pose Estimation in 3D via RF-Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15112–15121.
- Yu Zhang, Ziya Zhou, Xiaobing Li, Feng Yu, and Maosong Sun. 2022. CCOM-HuQin: an Annotated Multimodal Chinese Fiddle Performance Dataset. *arXiv preprint arXiv:2209.06496* (2022).
- Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *Comput. Surveys* 56, 1 (2023), 1–37.
- Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5745–5753.
- Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)* 23, 4 (1997), 550–560.
- Yuanfeng Zhu, Ajay Sundar Ramakrishnan, Bernd Hamann, and Michael Neff. 2013. A system for automatic animation of piano performances. *Computer Animation and Virtual Worlds* 24, 5 (2013), 445–457.
- Christian Zimmermann, Max Argus, and Thomas Brox. 2021. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*. Springer, 250–264.