

Virtual Instrument Performances (VIP): A Comprehensive Review

T. Kyriakou^{1,2} , M. Álvarez de la Campa Crespo² , A. Panayiotou^{1,2} ,
Y. Chrysanthou^{1,2} , P. Charalambous²  and A. Aristidou^{1,2} 

¹Department of Computer Science, University of Cyprus, Cyprus

²CYENS - Centre of Excellence, Cyprus

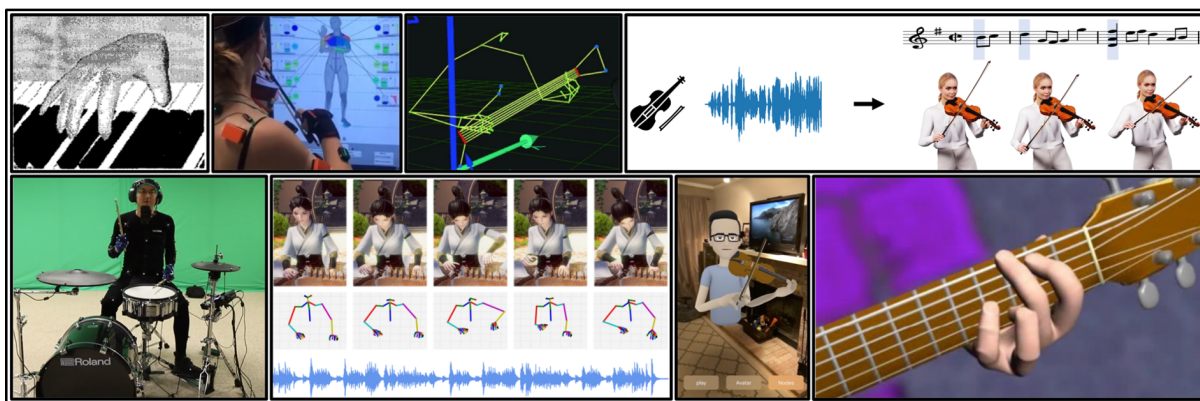


Figure 1: Modern multi-modal performance capture techniques enable us to record movement (body, hands, face, instruments), audio, and other data such as EEG signals from performers while they play musical instruments. This data can be used to enable more advanced analysis of performances, synthesis of novel performances, and the creation of new experiences within XR environments. Images extracted from [SE00, Cye23, PCAW16, HTHM22, Cin21, CFZ⁺21, SDSKS18, ES03].

Abstract

Driven by recent advancements in Extended Reality (XR), the hype around the Metaverse, and real-time computer graphics, the transformation of the performing arts, particularly in digitizing and visualizing musical experiences, is an ever-evolving landscape. This transformation offers significant potential in promoting inclusivity, fostering creativity, and enabling live performances in diverse settings. However, despite its immense potential, the field of Virtual Instrument Performances (VIP) has remained relatively unexplored due to numerous challenges. These challenges arise from the complex and multi-modal nature of musical instrument performances, the need for high precision motion capture under occlusions including the intricate interactions between a musician's body and fingers with instruments, the precise synchronization and seamless integration of various sensory modalities, accommodating variations in musicians' playing styles, facial expressions, and addressing instrument-specific nuances. This comprehensive survey delves into the intersection of technology, innovation, and artistic expression in the domain of virtual instrument performances. It explores musical performance multi-modal databases and investigates a wide range of data acquisition methods, encompassing diverse motion capture techniques, facial expression recording, and various approaches for capturing audio and MIDI data (Musical Instrument Digital Interface). The survey also explores Music Information Retrieval (MIR) tasks, with a particular emphasis on the Musical Performance Analysis (MPA) field, and offers an overview of various works in the realm of Musical Instrument Performance Synthesis (MIPS), encompassing recent advancements in generative models. The ultimate aim of this survey is to unveil the technological limitations, initiate a dialogue about the current challenges, and propose promising avenues for future research at the intersection of technology and the arts.

CCS Concepts

• **Computing methodologies** → **Animation; Motion capture; Motion processing; Machine learning;**

1. Introduction

The digital evolution of performing arts, including musical experiences, in virtual settings, stands at the forefront of a transformative era driven by Extended Reality (XR), the Metaverse, the widespread adoption of Artificial Intelligence (AI), and recent advances in real-time computer graphics. This shift has significantly altered the performing arts landscape, unlocking unparalleled possibilities for inclusivity, creativity, and live performances in diverse locations. Beyond the challenges accentuated by the recent pandemic, which served as a catalyst for these possibilities, our motivation is firmly grounded in the inherent potential for innovation and growth within virtual and mixed-reality spaces.

The digitization and visualization of the performing arts play a pivotal role in enhancing accessibility to art, preserving cultural heritage in an intangible format, and reaching a diverse global audience. This transformation not only ensures the long-term preservation, documentation, and analysis of these art forms for future generations but also serves educational purposes. It enables students to delve into the intricacies of various art forms, exploring their historical, cultural, and technical aspects. In this revolutionary era where the digital realm seamlessly connects with artistic expression, the performance of musical instruments in extended reality (XR) represents a dynamic evolution. We define this as *Virtual Instrument Performance* (VIP), a multimedia presentation that encompasses the comprehensive execution of a musical instrument within a virtual environment. This multidisciplinary art form combines musical skill with advanced audiovisual technologies for high-quality audio production, animations, and interactive elements, creating a holistic and captivating experience for both performers and audiences. Within this digital realm, the performer plays their instrument in a computer-generated world, where visual effects and animations synchronize with the music, enhancing sensory engagement. This blurs the boundaries between reality and virtual worlds, expanding the possibilities of musical expression and entertainment in the digital age. VIP represents a boundary-pushing fusion of music, visuals, and interactivity, transcending traditional barriers tied to physical presence and allowing artists to connect with worldwide audiences. The Metaverse, a shared virtual space, and digital twins, enhanced virtual replicas of physical entities, act as catalysts, driving performing arts into new dimensions. In this expansive digital canvas, artists craft immersive experiences that transcend geographical boundaries, leading to an era where live performances are not restricted to a specific stage but resonate across borders. Moreover, the flexibility introduced by recording and broadcasting liberates audiences from the constraints of time zones and schedules, allowing them to enjoy concerts at their convenience. Performing in a virtual environment unleashes new horizons for creativity, freeing artists from the constraints of the physical, tangible world while offering innovative means to engage and interact with their creations. Virtual spaces unlock a realm of boundless potential, from altering appearances and introducing virtual entities to defying gravity, unlocking new dimensions of scalability and creativity. Viewers can now access exclusive vantage points and unique perspectives on performances, and even participate in ways that were previously unimaginable.

However, creating virtual characters that convincingly play mu-

sical instruments presents significant challenges. Firstly, there is the issue of the loss of the live experience, as watching a performance on a screen or in a virtual environment lacks the energy and connection between the audience and performers. Additionally, there are several other hurdles to overcome, including maintaining quality and authenticity (for example, the camera angles, sound recording, and post-production editing can affect the viewer's perception of the performance), technical obstacles (for example capturing the essence of a live performance and presenting it in a visually appealing way requires skill and equipment), and financial constraints since the cost of digitization and visualizations can be prohibitive. Like other performing arts, playing musical instruments encompasses intricate, multi-modal performances with complex and fine detail subtle movements, making their acquisition, analysis, comprehension, and synthesis inherently demanding. In particular, data acquisition involves integrating and synchronizing various types of data, capturing precise motion with high fidelity, accommodating variations in musicians' playing styles, addressing occlusion challenges, and dealing with instrument-specific nuances. On the other hand, generating convincing animations of musicians playing musical instruments requires replicating instrument sounds accurately, synthesizing complex and multi-modal animations (covering pose, wrist, facial expressions, and instrument animations), infusing emotional expression, ensuring real-time interaction, and efficiently managing computational resources. Balancing these aspects necessitates advanced technology, including cutting-edge motion capture systems, sound modeling techniques, and advanced AI algorithms, all crucial for achieving the realism and expressiveness required for convincing virtual musical performances.

The importance of this domain, along with an acknowledgment of its challenges, has been emphasized by the attention it has received from various global organizations. These organizations provide financial support to numerous projects with the aim of shaping the future of performing arts digitization, visualization, and the advancement of their virtual enhancements. Among several others, European projects like PREMIERE [PRE23], SHARESPACE [SHA23a], CAROUSEL+ [CAR21], Apollo project [APO23], the PHENICX project [LGS15] are key players in this dynamic landscape. For instance, PREMIERE is dedicated to developing a comprehensive ecosystem of digital applications powered by advanced AI, XR, and 3D technologies to cater to the diverse needs of individuals involved in performing arts productions. Simultaneously, SHARESPACE paves the way for inclusive hybrid societies by facilitating remote interactions within a shared sensorimotor space. The CAROUSEL project allows online users to participate on online performing art creations, such as dance, despite physical separation, addressing issues of isolation and loneliness. These developments also lay the foundation for novel forms of online communication and expression. On the other hand, the Apollo project adds a physical dimension to this digital landscape, establishing a permanent exhibition in the foyer of the Konzerthaus, providing visitors with insights into the Berlin's musical heritage, and the opportunity to experience virtual reality. The PHENICX project utilizes new digital methods to make classical music performances more accessible and engaging through innovative multi-modal enhancements. These projects serve as exemplary illustrations of the significant contributions that funding and collaborative efforts can

make in shaping the future of performing arts. However, it's worth noting that none of these projects primarily focuses on VIPs, underscoring the untapped potential for future research in this domain.

In addition to global funding organizations, the past few years have witnessed a surge in interest from the industry within the realm of VIP. This transformative landscape has magnetized prominent artists who discern the immense potential of virtual concerts. Collaborative ventures with platforms like Roblox [Rob23], Meta's Horizon Venues [Met23a], WaveXR [Wav], and Epic Games' Fortnite [For] have given birth to immersive experiences that transcend conventional musical performances. Renowned figures such as John Legend [Leg20], who seamlessly combined vocals and piano, and acclaimed bands like Foo Fighters [Met22] and 21 Pilots [Mov23a], have boldly ventured into this digital frontier. Their efforts have not only gained a huge audience and attention but also made virtual concerts a profitable business, changing how we see art and redefining the landscape of artistic expression and entertainment.

Despite the transformative potential in virtual instrument performances, this dynamic and ever-evolving field has not received the attention in research it deserves, mostly due to the formidable challenges it presents that often act as barriers to further exploration. This survey serves as a groundbreaking state-of-the-art report, offering a comprehensive exploration of the intricate fusion of technology, innovation, and artistic expression in this domain. It goes beyond being a mere response to global challenges and instead positions itself as an enlightening guide to the boundless possibilities that the virtual world opens up for musical experiences. While a comprehensive musical performance encompasses a multitude of elements, our survey specifically emphasizes the instrumental dimension, focusing on the delicate nuances of musicians' movements and the audio quality of the music.

In particular, this survey explores the recent advancements in data acquisition, with a specific focus on the multi-modal aspects within this field. Our study extends to existing multi-modal repositories, particularly those centered around musical instruments and musicians, which may serve as valuable resources for training AI networks and models. We have carefully assessed data acquisition methods and systems, which encompass a wide array of techniques, including motion capture, facial expression recording, and the capture of audio and MIDI data. Our evaluation highlights the strengths of these methods while also addressing the limitations and challenges they present. Furthermore, our study delves into recent techniques for Music Information Retrieval (MIR) tasks, with a particular emphasis on the Musical Performance Analysis (MPA) field, and offers an overview of various works in the realm of Musical Instrument Performance Synthesis (MIPS), encompassing recent advancements in generative models (e.g., methods that take MIDI information as its sole input and generate realistic animations featuring individuals playing musical instruments). Our analysis covers both the progress made in this area and the limitations that these innovative techniques currently face. The primary objective of our survey is to shed light on the current technological constraints, discuss ongoing challenges, and propose future research pathways in this continually evolving intersection of technology and the arts. Figure 1 displays representative examples of the VIPs, highlight-

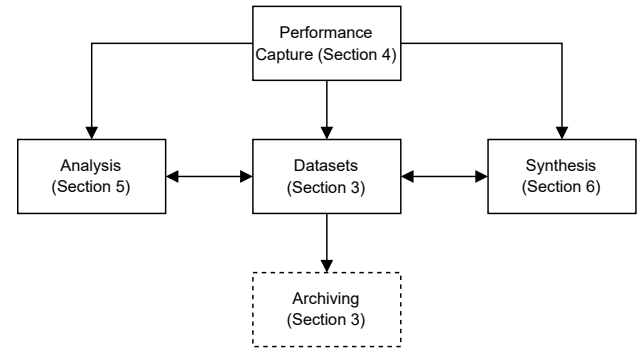


Figure 2: Structural interconnection between the different sections of this survey. To analyze and synthesize new VIPs, performance capture technologies are used to record multi-modal data of performers and their instruments. The data is then either directly used by analysis and synthesis systems or stored in databases using appropriate formats and representations. In some cases, these datasets are used for archiving purposes and are therefore enriched with metadata, analysis, synthesized data, and semantic annotations by experts in the respective domains.

ing their diversity and multi-modality, while Figure 2 provides a visual representation of the structural interconnection among the various sections within this survey.

Our survey is structured as follows: in Section 2, we begin by presenting the various representations and formats employed in repositories that store virtual instrument performances, including pose, facial, and audio files. Moving on to Section 3, we provide a comprehensive exploration of the existing datasets related to performing music. These repositories are categorized based on their modality, e.g., audio-modality or multi-modality, as well as their scope and the range of instruments they encompass. In Section 4, we delve into the technologies utilized for data acquisition. Here, we explore a multitude of methodologies and technologies that capture human movements, spanning from pose and facial expressions to finger dexterity and audio aspects. Section 5 offers an in-depth view on methods to analyse musical performances. This section serves as a canvas where we extract diverse musical properties and explore the nuances of artistic expression, considering inputs such as posture and finger extensions. Section 6 unfolds examples of Musical Instrument Performance Synthesis, presenting various methodologies and recent machine learning models employed to generate musical performances with different instruments. In Section 7, we engage in a thoughtful discussion, addressing the challenges and limitations encountered throughout our exploration of the virtual musical performances pipeline, and conclude our survey with closing remarks that encapsulate the essence of our journey across the vast realm of virtual instrument performance. This section not only provides a reflective analysis of the insights we have accumulated, but also gives practical recommendations and outlines future research directions in this multi-disciplinary domain.

2. Background Knowledge

This section explores the vital concept of data representation in the context of VIP. From capturing the gestures of instrumentalists to the audio itself, data representation serves as the bridge connecting the world of art performance to the digital realm, enabling new possibilities for artistic expression and analysis. We start by mentioning various audio representations, and then we explore motion representation of performers.

2.1. Audio Representation and Storage

This section delves into the diverse methods used for storing, describing, and documenting sound in the realms of music and technology. It encompasses various protocols that facilitate communication between audio hardware devices, a collection of music annotations that provide detailed descriptions of sounds, and a range of audio file formats optimized for music storage. Firstly, let's explore two popular *Communication Protocols*: MIDI and OSC. Musical Instrument Digital Interface (MIDI), is a standardized protocol and set of specifications used for the digital communication and control of electronic musical instruments and computer systems. MIDI enables the exchange of musical information and instructions between different devices. A MIDI message starts with a status byte indicating its type and channel, followed by pitch and velocity data bytes. For example, to play a note in MIDI, a "Note On" message is transmitted, with an assigned "velocity" setting that influences the note's volume [MIDa, Epi]. OpenSoundControl (OSC), similar to MIDI, serves as a protocol for the real-time exchange of messages between software and hardware in various applications [MIDb]. OSC is a newer protocol that can transmit a wider range of data types than MIDI, such as numerical values, strings, arrays, and even user-defined data structures, but it is also more complex and less widely supported. OSC is more suitable for a wider range of creative applications beyond traditional music, including interactive installations, multimedia performances, and communication between various types of software and hardware devices.

Secondly, we list the *Basic Music Elements* [Sar16], the fundamental concepts that define and give structure to a piece of music. They contribute to the mood, harmony, and rhythm of what we hear. Pitch: the frequency of the note's vibration (how high or low the sound is); Duration: How long or short the sound is; Dynamics: the volume (how loud/quiet the sound is); Timbre: the unique sound of an instrument, for example an electric guitar sounds different from an acoustic guitar (tone color of a sound); Melody: a succession of musical notes; Harmony: the simultaneous, vertical combination of notes, usually forming chords (multiple pitches played at the same time); Tempo: beats per minute (how fast or slow a piece of music is played); Texture: the density (thickness or thinness) of layers of sounds, melodies, and rhythms in a piece (a complex orchestral composition will have more possibilities for dense textures than a song accompanied only by guitar or piano).

Thirdly, we proceed with *Audio Formats* which they encapsulate the diverse ways in which digital sound is stored and represented. While some formats might prioritize minimizing file size for easier sharing and storage, others might focus on retaining the utmost audio fidelity for professional applications. Choosing the right audio format depends on the need for quality and usage. WAV, AIFF,

FLAC, and PCM provide high-quality, uncompressed or lossless audio, ideal for editing and archiving, though with larger file sizes. For online distribution, compressed formats like MP3, AAC, and OGG offer smaller files at the cost of potential quality loss. M4A and WebM are versatile, supporting various codecs and are suited for web use and Apple devices. Ultimately, the choice should balance audio quality and file size, considering the end-user's platform and needs.

Lastly, we provide a reference list of acronyms and terminology associated with musical performances that will be employed in subsequent sections of this survey. MFCCs (Mel-frequency Cepstral Coefficients): These are features used to simplify audio signals, making them more amenable to analysis and pattern recognition. MFCCs are particularly valuable in speech and audio processing applications; Onset/Offset: The identification of the starting and ending points of musical notes. This process is crucial for the precise analysis of various musical elements, including tempo, pitch, and more; String quintets: This term refers to a musical composition designed for five string players, often involving combinations of violins, violas, cellos, and double bass; Vibrato: A musical technique where the pitch of a note is subtly varied, typically through small, rapid oscillations in pitch, to add expressiveness and depth to the sound. Vibrato is commonly used by string players and singers to enhance the emotional quality of their performance.

2.2. Motion Representation and Storage

Typically, character animation is represented using joint/bone hierarchies; each bone's transformation is relative to its parent and bones are used to drive parts vertices of a mesh with specific influence (weights). These hierarchies allow for efficient manipulation and animation of the entire skeleton through techniques such as keyframe animation, forward and inverse kinematics and motion capture. Rotations in these representations are typically represented using Euler Angles, Quaternions [PGA18], Rotation Matrices (or 6D representations) [ZBL*19] or variations such as Dual Quaternions [AAC22]. The storage and retrieval of this type of data is usually achieved using suitable motion capture file formats and protocols. One of the most common used motion capture formats is BVH (Biovision Hierarchical Data). It is divided into two sections: the first delineates the skeleton's hierarchical structure and initial pose, while the second captures the motion, providing channel data for each frame [MM*01]. Another format that the last years is gaining popularity is SMPL [LMR*15]. The Skinned Multi-Person Linear Model (SMPL) is a data-driven model that accurately captures a wide range of human body shapes and poses using a vertex-based approach. It utilizes parameters derived from the rest pose template, blend weights, pose-dependent blend shapes, identity-dependent blend shapes, and a regressor from vertices to joint locations.

Furthermore, in combination with motion capture data, that are able to realistically animate the body of a virtual avatar, *Facial capture* has surged in prominence with the advancing horizons of technology, and offers accurately translations of the subtle movements of our faces into digital form for realistic representation (more details in Section 4.4). Central to this is the concept of "blendshapes". This technique involves a set of predefined facial expressions that can be blended in various combinations to represent a spectrum

of human emotions. When these blendshapes are integrated into 3D models, they allow these models to emulate real-world facial expressions with incredible precision. To store and transfer these complex datasets, formats like FBX [AUT], Alembic [SL], and COLLADA [AB06] are utilized. These formats not only encapsulate the blendshape data but also ensure compatibility across different software and platforms.

2.2.1. Future Research

We argue that future research should incorporate facial expression data when capturing musical performances data, as they offer significant insights into the emotions and intentions behind the music. The interplay of facial expressions with musical elements provides a richer context, allowing for a deeper understanding and appreciation of the performance.

3. Multi-modal Datasets of Performing Music

Creating multi-modal repositories of musical performances data is a complex task that requires careful organization and systematic presentation. It also involves addressing significant challenges in data acquisition, including the capture of high-fidelity data, curation, and synchronization across various modalities (see Section 4). The intricate nature of music-related performance capture data adds an additional layer of complexity, with challenges like data occlusion, the capture of nuanced dexterous movements of the performer, and the need for standardized metadata to ensure the repository's quality, usability, and comprehensiveness. In this section, we provide an overview of various databases and repositories, each offering a unique perspective on musical content. These repositories encompass a wide range of data types, ranging from sheet music, audio recordings, video, and MoCap, to a diverse spectrum of musical instruments, genres, and styles. Exploring those databases is a valuable step in the research process, enabling researchers to access, evaluate, and leverage existing resources to advance their work, validate algorithms, promote interdisciplinary collaboration, facilitate data integration, train machine learning models, and inspire innovative research directions, thereby contributing to the growth of knowledge in the field. In the following subsections, we briefly discuss about music data archiving and then, we categorize various databases firstly based on their data modality(audio and multi) and secondly based on their primary intended usage; it is worth noting that certain datasets may be well-suited for multiple tasks, but we group them according to their predominant use cases. Our organization partially relies on the approach presented by Li et al. [LLD*19], offering a structured exploration of this rich landscape. Our analysis additionally enlists recent repositories not covered in the original paper, along with datasets that exhibit greater variability and are not closely associated with URMP [LLD*19], ensuring a more comprehensive review of the available resources. While our primary focus lies on multi-modal datasets, we have also chosen to include repositories centered around audio and MIDI, recognizing their potential utility for the research community. Finally, this section includes a concise discussion on music composition.

3.1. Archiving Musical Performances

The organization and accessibility of any database play a pivotal role, with metadata serving as the hub, providing the essential descriptions of the underlying data. In essence, metadata functions as a documentation system for the data at hand. These metadata can be categorized into five primary types, each shedding light on different facets of the resources [DIHB08]:

1. *Descriptive* metadata aids in the discovery and identification of resources. It captures elements such as the pitch contour of a vocal line, the genre, or specific instrument types used in a musical composition.
2. *Structural* metadata delves into the organization of data, elucidating details like the sequencing of note annotations in a musical score or the hierarchy of layers in a multi-track recording.
3. *Administrative* metadata comes into play when documenting aspects like the date of a song's annotation, the file type of a performer's video, or rights related to the usage of motion capture data of a dancer interpreting the music.
4. *Reference* metadata might describe standard classifications, like predefined categories of emotion or sentiment, or reference points for motion capture data related to standard body movements.
5. *Statistical* metadata within these datasets could reveal patterns, such as the frequency of a particular emotion across several songs or common movements found in motion capture data across multiple performances.

To systematically and cohesively organize data, it is imperative to implement and delineate metadata schemas. These schemas illustrate the interconnections among various metadata components [Sic14]. The primary role of metadata is to assist users in locating information, exploring resources, and conducting in-depth examinations of the content and structure of the data. This is particularly vital for managing electronic resources and ensuring the digital preservation of information and assets. Similarly to the work of Aristidou et al. [ASC19] that deals with the acquisition of dance data and proposed a schema for comprehensive archiving of dance performances, a similar schema for musical instrument performances should be established. While numerous schemas are focused on music data, they often overlook the multi-modal aspects of musical instrument performance data. To our knowledge, the closest resemblance to a schema describing multi-modal musical performance data is RepoVizz [MLMG11], a data repository and visualization tool that offers structured storage browsing of multi-modal recordings. This tool stores data as DataPacks, which are essentially tree documents with nodes categorizing data, providing descriptions, or pointing to different data files, but it does not rely on a specific structural schema. Hence, we assert that the creation of a suitable metadata schema or protocol, designed to facilitate the organization and maintenance of a substantial volume of multi-modal musical performance data, is of paramount importance and will significantly benefit future research endeavors. In this survey, we will not be delving deeply into the details of music metadata and archiving. However, for those seeking a preliminary exploration of this subject, we recommend referring to the work of Serra et al. [SMB*13], that provides detailed insights and discussions on various aspects of music data, its organization, and preservation.

It serves as an excellent initial reference for anyone interested in delving into the specifics of music archiving.

3.2. Audio/MIDI-modal Datasets

This subsection briefly reviews several performance datasets which are mainly focused on audio and MIDI modality. These datasets are categorized into four groups according to their predominant applications. The first category, “Pitch Estimation, Transcription, and Analysis”, focuses on the foundational process of understanding music by extracting notations and interpreting individual notes, forming the basis for further analysis. The second category, “Music Information Retrieval and Instrument Recognition”, centers on extracting metadata and differentiating between various musical sources. Moving on to the third category, “Music Generation and Composition”, it delves into the creative aspect of music. It emphasizes tools and algorithms designed to create novel sounds and automate musical composition. The final category, “Source Separation, Mixing, and Signal Processing”, delves into the technical aspects, focusing on refining audio quality, isolating vocals or instruments, and ensuring an optimal listening experience. Table 1 lists the audio/MIDI-modality repositories, categorized based on their primary scope. This table provides details regarding the instruments featured, duration, and data content and formats for each repository.


Pitch Estimation, Transcription, and Analysis: The MAPS database [EBD10] was designed as a robust resource for the music information retrieval community. Comprising MIDI-annotated piano recordings, its intent is to further the evolution of pitch estimation and automatic transcription techniques. It boasts an array of sounds captured under varied conditions. Furthering the discourse on transcription, the LabROSA dataset [PE07] is a collection of 130 pieces of audio and MIDI, recorded on a Yamaha Disklavier grand piano, mainly aids research into classification-based transcription methods. Drawing attention to stringed instruments, the GuitarSet [XBP*18] stands out with its use of a hexaphonic pickup. This comprehensive dataset includes numerous acoustic guitar excerpts accompanied by time-aligned annotations, pivotal for transcription and performance analytics. For ensemble works, the TRIOS dataset [FP13] is a valuable resource, offering separated tracks from five chamber music trio recordings, along with their corresponding MIDI scores. The dataset by Su et al. [SY16] involves an innovative approach where a musician recreates nine musical excerpts, where they are later checked for accuracy with annotated MIDI, with the aim of resolving any possible mismatches. Delving into classical realms, the Bach10 dataset [DPZ10] is tailor-made for polyphonic music research. Featuring ten J.S. Bach chorales, it provides a blend of audio recordings and accurate ground-truth data for each part played by distinct instruments. Shifting focus to orchestral compositions, the PHENICX-Anechoic dataset [MCOB*16] offers denoised recordings for four symphonic pieces, accompanied with note annotations sourced from the Anechoic Dataset [PPL08]. Concluding this category, the MusicNet dataset [THK17] contains classical music tracks from 10 composers and 11 instruments, spanning 34 hours, each annotated with precise, time-specific labels from 513 classes.

Music Information Retrieval and Instrument Recognition: The Wood Wind Quintet (WWQ) dataset [BED09] provides insights from a single classical quintet, releasing a 54-second snippet for public use, which has served as a benchmark for the MIREX Multi-F0 Estimation And Tracking task [MIR23]. Moving to a broader spectrum, the RWC Music Database [GHNO02, GHNO03, G*04] contains six unique collections, featuring everything from popular music to classical tunes, totaling 315 musical pieces. A standout aspect of this database is its exhaustive compilation of 50 instruments, capturing diverse playing styles and dynamics. Likewise, provides original audio signals, corresponding standard MIDI files, and, for song entries, supplementary text files containing lyrics. Furthermore, Nlakh [KPJ*23], combines the NSynth [ERR*17] and Lakh [Raf16] datasets, offered in two distinct versions focusing on solo and mixed tracks. It caters to a wide instrument range and is notable for its large size. The SSMD dataset [HKS12] offers individual ground-truth annotated audio tracks from cover versions of popular western songs, majorly spotlighting vocals, with its library of 104 songs. Venturing into regional tunes, the iKala dataset [CYF*15] contains high-quality Chinese pop songs, each paired with human-annotated pitch contours and time-marked lyrics, challenging separation algorithms with its inclusion of non-vocal segments. Last but certainly not least, the Free Music Archive (FMA) [DBVB17] sets itself apart as a vast repository, covering 343 days of audio from over 100K tracks, neatly categorized into 161 genres, while also offering a plethora of metadata.

Music Generation and Composition: The MAESTRO dataset [HSR*18] presents around 200 hours of audio and MIDI recordings from ten years of the International Piano-e-Competition. The recordings have been synchronized to maintain an accuracy close to 3 ms, and each piece is thoroughly annotated, offering insights into composers, titles and performance years. Another contribution comes from the ADL (Augmented Design Lab) Piano MIDI dataset [FLW20], which showcases a collection of piano compositions, spanning various genres. Extracted and refined from the larger Lakh MIDI dataset [Raf16], this dataset emphasizes compositions associated with “Piano Family” instruments. Adding to this category, the dataset developed by Benetos et al. [BKD12] emerges as an instrumental tool for automatic piano tutoring. The dataset consists of seven real-world recordings, intentionally captured with a moderately detuned Yamaha U3 Disklavier. Each recording is a true reflection of human performances, complete with occasional mistakes, which are precisely documented in the MIDI ground-truth. The NSynth dataset [ERR*17] offers a collection of 306K musical notes from 1,006 instruments, each categorized by its distinct pitch, timbre and envelope. Notably, each note is a monophonic audio snippet, covering every pitch on a standard MIDI piano and five distinct velocities. Notes are further annotated with details like their sound production source, their instrument family and various sonic qualities. The Nintendo Entertainment System Music Database (NES-MDB) [DMM18] features tracks synthesized by the iconic NES and spans approximately 46 hours of chiptunes. Each track in the dataset provides a score for four instrument voices, accompanied by details on dynamics and timbre. The POP909 dataset [WCJ*20] includes piano arrangements for 909 songs, totaling 60 hours, produced by expert musicians. These songs, available in MIDI format, feature

Table 1: Audio/MIDI-modal Music Performance Datasets

Name	Content	Quantity		Annotation							
		Samples	Duration	N	M	PC	G	I	L	E	m
Pitch Estimation, Transcription, and Analysis											
MAPS [EBD10]	Piano	270	18.6 h	✓	✓						
LabROSA [PE07]	Piano	130	2.7 h	✓	✓						
GuitarSet [XBP*18]	Guitar	360	3 h	✓	✓	✓					
TRIOS [FP13]	Multi	5	3.2 m	✓	✓						
Su et al. [SY16]	Multi	10	5 m	✓							
Bach10 [DPZ10]	Multi	10	5.5 m	✓	✓	✓					
PHENICX-A [MCOB*16,PPL08]	Multi	4	10.6 m	✓	✓	✓		✓			
MusicNet [THK17]	Multi	330	34 h	✓	✓			✓			✓
Music Information Retrieval and Instrument Recognition											
WWQ [BED09]	Multi	1	1 m	✓	✓						
RWC [GHNO02]	Multi	215	N/A	✓	✓				✓		✓
Nlakh-multi [KPJ*23]	Multi	110K	153 h	✓	✓		✓	✓			✓
SSMD [HKS12]	Songs	104	6.8 h								✓
FMA [DBVB17]	Songs	106.5K	343 days				✓				✓
iKala [CYF*15]	Songs	252	2.1 h			✓			✓		
Music Generation and Composition											
MAESTRO [HSR*18]	Piano	1.18K	172.3 h	✓	✓						
ADL Piano MIDI [FLW20]	Piano	11K	N/A	✓	✓						
SiPT [BKD12]	Piano	7	6.4 m	✓	✓						✓
NSynth [ERR*17]	Multi	306K notes	340.1 h	✓	✓			✓			✓
NES-MDB [DMM18]	Songs	5.2K	46.1 h	✓	✓			✓			
POP909 [WCJ*20]	Vocal, Lead instrument	909	60 h	✓	✓						✓
Groove [GRE*19]	Drums	1.15K	13.6 h	✓	✓						✓
Bach Doodle [HHR*19]	MIDI	21.6M	N/A	✓	✓						✓
MusicCaps [ADB*23]	Music-Text Pairs	5.5K	N/A								✓
Source Separation, Mixing, and Signal Processing											
MASS [VIN08]	Multi	6	4.8 m						✓		✓
Mixploration [CPR14]	Multi	12	4.9 m								✓
MedleyDB [BST*14]	Multi	122	7.3 h			✓	✓	✓			✓
MUSDB18 [RLS*17]	Multi	150	10 h			✓	✓	✓			✓
MTG-Jamendo [BWT*19]	Multi	55.5K	3.7K h				✓	✓			✓
Slakh2100 [MWSLR19]	Multi	2.1K	145 h	✓	✓		✓	✓			✓
Emotion and Sentiment Analysis/Generation in Music											
VGMIDI [FW19]	Piano soundtracks	823	N/A	✓	✓					✓	✓
EMOPIA [HCD*21]	Songs	1.1K	N/A	✓	✓					✓	✓
DEAM [AYS16]	Songs	1.8K	N/A							✓	✓

N: Note, M: MIDI, PC: Pitch contour, G: Genre, I: Instrument, L: Lyrics, E: Emotion, m: metadata, : Unavailable/Non-working Link

vocal and instrument melodies alongside piano accompaniments, all aligned with the original audio; annotations include tempo, beat, key, and chords. The Groove MIDI Dataset [GRE*19] offers 13.6 hours of electronic drum performances from 10 professional drummers, paired with relevant metadata like style annotations and tempo, all in MIDI format. The Bach Doodle dataset [HHR*19] stems from an interactive tool [Bac], allowing users to craft melodies harmonized in Bach's style by the Coconet [HCR*19] model. This resulted in over 21.6 million compositions across 8.5 million sessions, detailing user melodies, harmonizations, and various metadata attributes. Finally, MusicCaps [ADB*23], which focuses on text to music generation, contains musical snippets sourced from AudioSet [GEF*17]. Each of these clips is matched

with its English text description. For every 10-second snippet, there are a descriptive caption and a list of music aspects.

Source Separation, Mixing, and Signal Processing: The MASS (Music Audio Signal Separation) dataset [VIN08] provides short song excerpts, lasting between 10 to 40 seconds. Each of these excerpts offers Stereo Microsoft PCM WAV files at 44.1Khz and 24 bits, capturing every instrumental track, where based on production settings, may or may not have effects. On a parallel note, the "MIXPLORATION" Dataset [CPR14] is designed to provide an analysis of audio mixing and includes four root components: the raw source audio files, the specific mixing parameters, survey data capturing the listener feedback on these mixes and a time-series log of the mixing adjustments. Furthermore, regarding melody extraction, MedleyDB [BST*14] is a collection of melody annotated,

royalty-free multi-track recordings designed mainly for melody extraction research. While 14 tracks don't have a defined melody, they still play a significant role in other musical research areas. The dataset also delivers instrument activation annotations and extensive metadata. The MUSDB18 dataset [RLS*17] offers 150 full-length tracks, from a spectrum of genres. For each track, its original stems, isolating elements like vocals, drums and bass are provided. Meanwhile, the MTG-Jamendo Dataset [BWT*19] emerges as an open framework for music auto-tagging. Sourcing its music from Jamendo, it comes with over 55K tracks tagged across multiple categories such as mood, genre and instruments. Finally, the Slakh dataset [MWSLR19] integrates multi-track audio files and aligned MIDI. Stemming from the Lakh MIDI dataset [Raf16], it employs high-quality virtual instruments to render individual MIDI tracks, which are then combined to form complete musical compositions. Its current version, Slakh2100 offers 2.1K tracks, generated from a diverse range of 187 patches across 34 categories.

Emotion and Sentiment Analysis/Generation in Music: The VGMIDI dataset [FW19] is a collection of 823 pieces extracted from video game soundtracks in MIDI format. These tracks, converted to piano arrangements, are of varying lengths, with some as short as 26 seconds and others extending up to 3 minutes. The selection criteria focus on the pieces' emotional intensity, with 95 pieces annotated based on valence, indicating the emotion's positivity or negativity, and arousal, denoting the emotion's intensity. The EMOPIA dataset [HCD*21] centers around the perceived emotion in pop piano music, combining both audio and MIDI formats. The emotion detected in each clip is verified through labels provided by a team of four annotators, ensuring a comprehensive understanding of the emotional content. Furthermore, the DEAM dataset [AYS16] offers a more expansive perspective on Western popular music genres, including but not limited to rock, pop, electronic, and jazz. It includes 58 full-length tracks and 1,744 45-second excerpts.

3.3. Multi-modal Datasets

This subsection is dedicated to the examination of performance datasets featuring multi-modal data. These databases incorporate a range of modalities, extending beyond audio and note annotations and may encompass visual data, motion capture (MoCap) data, and information related to style and emotion. To facilitate efficient organization, we categorize them into four groups: The first category "Music Information Retrieval and Analysis" offers a structured approach for extracting and analyzing essential musical elements and patterns. The second category, "Music Generation and Composition" delves into the tools and techniques used in crafting and automating music generation. The third category, "Emotion and Sentiment Analysis/Generation in Music" specifically contain emotional information. Notably, datasets including emotion labels often employ the Circumplex model of emotion [Rus80]. This model utilizes a two-dimensional circular space, featuring arousal and valence dimensions, where valence represents positive versus negative emotion, and arousal indicates emotional intensity. Finally, the last category, "Musical Motion, Interaction and Learning", underscores the convergence of music and motion, encompassing areas like movement analysis, pose estimation, and the study of how music influences or interacts with physical movements. Details about

these datasets are presented in Table 2, and regarding motion data, we specify which elements were captured, between upper-body, lower-body, fingers and instrument.

Music Information Retrieval and Analysis: The dataset presented by Perez et al. [PCAW16] is a collection showcasing guitar performances that encapsulates ten musical segments, each played by two distinct guitarists. It integrates audio, 3D motion data, and details from the musical score, including note onset/offset, pitch, and precise data on the plucked string, plucking finger, fret, and left-hand fingering. Next, the C4S dataset [BVGLH17] focuses on clarinet performances and contains 54 videos spanning 4.5 hours from nine clarinetists. It includes ground-truth onsets and specific coordinates for facial landmarks with four regions of interest (ROIs): the mouth, left hand, right hand, and clarinet tip. The EEP Dataset [MRPM14] focuses on string quartet performances offering 23 multi-modal recordings, with tracking motion and bowing descriptors of each musician and a score alignment. ENST-Drums [GR06] is a research dataset for automated drum transcription and processing. It provides recordings from three professional drummers, spanning various genres, and using different drum kits, capturing audio on 8 separate channels and additionally offers three stereophonic files. Performances were video captured from two angles providing both a frontal and a right-side perspective. The Abesser dataset [ALD*11] focuses on ensemble performances across blues, funk, and swing genres. It not only provides multi-track audio but also delves into the rhythmic quality, onset detection, and other intricate musical annotations. URMP [LLD*19] includes 44 classical chamber music pieces, varying from duets to quintets, accompanied with visual information. Each piece comes with musical scores, individual audio tracks and detailed ground-truth annotations with both frame-level and note-level transcriptions. The Lakh MIDI dataset [Raf16], with its vast collection of MIDI files, offers ground truth data for audio content-based music information retrieval, transcription, meter, lyrics, and advanced musicological characteristics. Lastly, the YouTube-100M dataset [HCE*17], while not exclusively a music dataset, has been used for soundtrack classification. The dataset contains around 100 million YouTube videos, which have been auto-labeled with multiple labels out of a set of 30K topic labels, averaging 5 labels per video, based on information, context, and visuals.

Musical Motion, Interaction, and Learning: The QUARTET dataset [PMPCM14] captures both intricate audio details and bowing motion data from string quartet exercises conducted under two experimental scenarios: solo and ensemble. MAPdat (Music Assisted Pose dataset) [SFH*22] provides ground truth motion capture, audio, and video recordings for four master musicians with a total duration of 33.5 hours. The TELMI Dataset [VKV*17] concentrates on violin performers, documenting motion capture, audio, video, depth, and physiological data, ensuring an all-round perspective. Sarasua et al. [SCTO17] provide two datasets capturing instrumental gestures from five violinists and two pianists with expressive variations. They employ a diverse array of tools, ranging from EMG devices to gyroscopes, to ensure comprehensive data collection. Furthermore, the Solos dataset [MSH20] stands in alignment with the URMP dataset [LLD*19], offering detailed recordings across a plethora of instruments. These are further sup-

Table 2: Multi-modal Music Performance Datasets

Name	Content	Quantity		Annotation				
		Samples	Duration	N	M	m	V	Mo
Music Information Retrieval and Analysis								
Multi-modal Guitar [PCAW16]	Guitar	10	10 m	✓		✓	✓	U f I
C4S [BVGLH17]	Clarinet	54	4.5 h	✓		✓	✓	
EEP [MRPM14]	String quartet	23	N/A	✓	✓	✓		f I
ENST-Drums [GR06]	Drum kit	N/A	3.75 h	✓	✓	✓	✓	
Abeßer et al. [ALD*11]	Multi	N/A	1.12 h	✓		✓	✓	
URMP [LLD*19]	Multi	44	1.3 h	✓	✓	✓	✓	
Lakh [Raf16]	Multi	176.5K	N/A	✓	✓		✓	
YouTube-100M [HCE*17]	Multi	100M	5.4M h			✓	✓	
Musical Motion, Interaction, and Learning								
QUARTET [PMPCM14]	String Quartet	96	50 m	✓			✓	U f I
MAPdat [SFH*22]	Violin	40.2K	33.5 h	✓			✓	U L
TELMi [VKV*17]	Violin	41	2.4 h	✓			✓	U L f I
Gesture Datasets [SCTO17]	Piano, Violin	N/A	50 m	✓	✓		✓	U
Solos [MSH20]	Multi	755	66.2 h	✓	✓		✓	U f
RepoVizz [MLMG11]	Multi	N/A	N/A	✓		✓	✓	U L f I

N: Note, M: MIDI, m: metadata, V: Video, Mo: MoCap (U: Upper body, L: Lower Body, f: fingers, I: Instrument),
 □: Unavailable/Non-working Link

plemented with audio, MIDI, and skeletons and video resources with clearly visible hands. Finally, RepoVizz [MLMG11] emerges as a tool tailored for the needs of the scientific community studying music performance, while is not only a data repository but also an effective visualization tool. It provides structured storage and user-friendly access to multi-modal recordings, spanning audio, video, motion capture, and much more. The goal of RepoVizz is to enable seamless online access to a shared music performance database, enabling collaboration and innovation among researchers.

3.3.1. Challenges and Limitations

As we conclude this section, it is essential to highlight the open challenges and limitations in the realm of multi-modal music databases. While these repositories are invaluable for various research areas, there are ongoing challenges related to data acquisition, documentation, and organization, such as the quality of the data, the synchronization of the multiple modalities, dealing with data occlusions, interoperability, stylistic variations, and metadata standards. Moreover, as evident from Table 2, most of the datasets that feature motion capture data are predominantly centered on stringed instruments. This concentration on a specific subset of instruments represents a limitation within the domain of musical instruments. In light of this observation, it is imperative that future research initiatives direct their efforts toward the establishment of repositories that encompass a more diverse array of musical instruments. The anticipated result of such endeavors is the creation of resources that are not only more expansive in their coverage but also more readily accessible. This expansion and increased accessibility would be of great benefit to both the music and virtual instrument research communities, as it would provide a richer and more representative datasets for exploring and advancing the field.

3.4. Music Composition

In recent years, advancements in music composition technologies have provided new opportunities for research. These tools enable digital music composition, enabling researchers to investigate and analyze musical constructs with greater precision and depth. Notably, these technologies present the opportunity to create customized datasets for further research initiatives, such as training machine learning models or augmenting existing datasets. While not the main focus of this survey, this subsection offers a brief overview of recent studies that employ diverse methods to compose computer-generated music. Over the last few years, *Text-to-Music Generation* has gained significant popularity. Agostinelli et al. [ADB*23] proposed MusicLM, a generative model that delivers high-quality music, maintaining consistency over extended durations and accurately adhering to text-based conditioning cues. Similarly, the MusicGen by Copet et al. [CKG*23], a single Language Model that operates over music tokens, can generate high-quality samples, influenced by textual descriptions or melodic attributes, offering enhanced control over the resulting output. Schneider et al. introduced Moûsai [SKJS23], a novel text-to-music generation model using latent diffusion, capable of real-time producing several minutes of music, ensuring both high musical quality and effective text-audio integration. MuseCoco [LXK*23], presented by Lu et al., is a data-efficient system for generating symbolic music from text descriptions by leveraging musical attributes. Moving towards to *Multi-Track Generation*, Ren et al. developed PopMAG [RHT*20], a pop music accompaniment generation framework, based on a novel multi-track MIDI representation which encodes multi-track MIDI events into a single sequence, and utilize a sequence-to-sequence model. Lv et al. presented GETMusic [LTL*23], a framework for generating music with any arbitrary source-target track combinations, which relies on a novel music representation combined with a diffusion model. Furthermore,

some works focused on *Music Form/Structure Generation*. Lu et al. developed MeloForm [LTY*22], an expert system to construct melodies from motifs to phrases using a predefined musical form, while they employed a transformer-based refinement model to enhance the richness. Museformer, proposed by Yu et al. [YLW*22] also use a transformer that incorporates novel fine- and coarse-grained attention mechanisms for music generation, capturing both music structure-related correlations and additional contextual information, leading to high-quality, well-structured long music sequences. An interesting work by Wang et al. [WLL*20] introduces an algorithm for synthesizing interactive background music based on visual content. Using neural networks for scene sentiment analysis and a cost function for music synthesis, it ensures emotional consistency between visual and auditory elements, as well as music continuity. Moreover, it is imperative to acknowledge the existence of several studies in the domain of *Song Writing*. Specifically, the works by Sheng et al. [SST*20], Xue et al. [XSW*21], and Ju et al. [JLT*22] have made noteworthy contributions to this field. We encourage readers, who are interested in exploring this topic in greater detail, to refer to the comprehensive analysis conducted by Ji et al. [JYL23] which extensively explore the current popular music generation tasks using deep learning techniques. Likewise, Siphocly et al. [SEHS21], describe and analyze various AI algorithms and techniques available for composing computer music.

3.4.1. Conclusions

It is crucial to underscore that the results generated in this section are not flawless. They stem from various challenges, such as limitations in creativity and musical structure, difficulties in conveying emotion, restricted user interactivity, and inconsistencies in music evaluation criteria. Nonetheless, given the rapid progress in this field, we expect that more advanced tools for automated music generation will soon emerge, which will be well-suited for research purposes.

4. Musical Performance Capture

In the context of musical instrument performance, the interplay between a musician's bodily movements, finger dexterity, and facial expressions, combined with the characteristics of the musical instrument and the resulting auditory experience, collectively shape the expressive and artistic delivery of the music. When it comes to digitizing a musical performance for archiving, documentation, streaming, analysis, and synthesis, it is essential to capture all the elements that are integral to the overall experience. This holistic approach to digitization and documentation is crucial for faithfully preserving the essence of the performance.

These elements encompass auditory aspects, such as capturing the unique timbre of the instruments and obtaining a high-quality audio recording that faithfully reproduces the voices of the performers. This audio component plays a pivotal role in retaining the emotional depth of the performance. Beyond the auditory aspects, digitization also extends to the visual components of the performance, including the appearance of the performers, their attire, the stage, lighting, shapes, colors, instruments, and any objects used. Moreover, the digitization process encompasses dynamic and

kinesthetic components, including the performers' postures, the nuanced movements of their fingers during instrumental play, and the emotions conveyed through their facial expressions. The extent to which each of these elements is captured in detail can vary based on the objectives of replicating the performance in a virtual environment. Moreover, apart from capturing the movements of the performers and the sounds of the instruments, in some cases, it is also necessary to capture the kinematics of props. This includes the movement of instruments on the stage and the dynamics of instrument accessories, such as the drumsticks of a drum kit.

In the scope of this survey, our primary focus centers around the interaction between artists and their instruments, their ability to convey emotion, and the quality of the sound they produce and transmit within a virtual context. Therefore, our concentration is mainly directed towards capturing the dynamic movements of performers, which include their postures, finger actions, and facial expressions, as well as achieving a faithful reproduction of the instruments' sound as played by the performers. Within this section, we will explore various systems and technologies designed to capture each of these critical modalities. We will assess their suitability within the context of VIP, highlighting their advantages and drawbacks, and addressing the challenges they present. The aim is to provide a well-informed basis for selecting the most suitable capture technology that aligns with the specific needs of users in the realm of virtual instrument performance.

4.1. Music

Recording instruments can be accomplished through various techniques, with the resulting sound being saved as audio files, often annotated with the corresponding instrument(s) that produced the sound. Some instruments, particularly electronic ones (e.g., electronic keyboards, electronic drums, and certain wind instruments), support the automatic retrieval of MIDI data in addition to capturing the sound. This MIDI data is valuable for further processing and analysis.

One of the initial and most common methods for capturing instrument sounds is to use microphones. When recording acoustic instruments, a common practice is to position a microphone in front of the instrument, as depicted in the left image of Figure 3. Conversely, when capturing the sound of electric instruments, the microphone is frequently situated in front of the amplifier to record the amplified sound, as demonstrated in the right image of Figure 3.

Alternatively, audio interfaces offer an effective means of audio acquisition. Among the available audio interfaces, the "Scarlett" [Foc23], produced by Focusrite, stands out as one of the most renowned and widely used options. To use these interfaces, instruments are connected to the audio interface using a jack cable, and the interface is then linked to a computer. Another notable audio interface is the "iRig" [Mul23], known for its portability. It allows for direct connections to smart devices like iPhones, iPads, or personal computers. An alternative method involves directly connecting an electronic musical instrument to a computer, provided that the instrument is electronic and compatible. This setup enables the automatic acquisition of both the instrument's sound and corresponding MIDI data. As mentioned earlier, this approach is particularly



Figure 3: Recording instruments with microphones in front of the instrument: a cello on the left [Zie23], and an amp on the right [Bra23].

useful for electronic instruments that have built-in MIDI recording capabilities.

MIDI files are often used as the ground-truth transcription for music, enabling precise representation of the musical data. For instruments that lack integrated MIDI recording capabilities, manual annotation remains the most accurate method for creating ground-truth transcriptions. However, this manual approach is labor-intensive and time-consuming. To address this challenge, several software solutions for converting audio to MIDI are available. Some popular options include Basic Pitch by Spotify [Spo23], Piano Scribe by Google [Goo23], and Logic Pro [App23b], each offering various features for MIDI conversion.

Finally, in the domain of audio and music editing, as well as notation, a wide range of software applications is available for recording, post-production and musical composition. These applications cater to different user needs and preferences, ranging from industry-standard commercial tools [App23b, Abl23, Stu23, Ste23, Ado23] to software designed for small businesses or home users [App23a]. Additionally, there are research-based solutions for specialized applications [Aud23, CLS10, MRL*15, LL21]. However, the discussion of these methods and tools is beyond the scope of this survey. For a comprehensive review of audio editing methods and tools, readers are encouraged to refer to the following works [Col13, Mat23a], which provides an in-depth exploration of this topic.

4.2. Body Movement Capture

Motion capture technology has played a pivotal role in digitizing, preserving, and disseminating intangible creations, such as dance performances [ASC19], or sport performances [vdKMR18]. Recent years have witnessed a growing demand for realistic 3D animation in various sectors, including media, entertainment, research, and training, prompting industries to seek effective 3D motion capture solutions. The advantages and disadvantages of these technologies have been extensively reviewed in surveys, such as [WF02, MHK06]. This technology has found widespread application in the entertainment industry, notably in the production of animated films, video games, and virtual reality experiences. Recent advancements in hardware and software, including high-speed cameras, inertial measurement units, and depth sensors, have sig-



Figure 4: The left image shows musicians playing violin, motion captured using an optical MoCap system with reflective markers, tracking their body and instrument motion (image extracted from [Fut]). The right image shows a musician playing piano, while the movements of the finger joints are tracked with 5mm reflective markers (image extracted from [Ger]).

nificantly enhanced the sophistication and accuracy of motion capture.

Motion capture can be broadly categorized into marker-based and marker-less systems. The choice of the most suitable system depends on the required quality and purpose of the application (e.g., mobility, interaction), as well as the desired level of accuracy and precision, allowing for the capture of even the most subtle movements and expressions, such as finger gestures, facial expressions, and even eye movements. In the following subsections, we offer a concise review of the prevalent technologies and systems utilized for capturing human motion. This will encompass a comparative analysis of different methods and an examination of more intricate movements, including fingers and facial expressions.

4.2.1. Marker-based Systems

Marker-based systems necessitate the attachment of sensors, markers or stickers to the bodies of performers. These systems can be categorized into two main types: optical and inertial-based systems.

4.2.1.1. Optical Systems Optical motion capture systems use fiducial markers near joints for real-time data acquisition. Popular in studios, these markers enable 3D positioning via high-speed cameras using triangulation. Passive systems like Vicon [Sys23] and NaturalPoint's OptiTrack [Opt23b] use retroreflective balls, offering high accuracy but are sensitive to lighting and marker swapping issues. Active systems like PhaseSpace [Pha23b] and Qualisys [Qua23] use LEDs for cleaner and labelled data but require wires and power sources. While precise, optical systems are costly, intrusive, lack portability, and require extensive setup. Data cleaning, especially for occlusions, remains a challenge [AL13, PHLW15, LC10, SDB*12], with recent attempts using Deep Learning (DL) for denoising and restoring missing markers [Hol18, CWZ*21]. Wheatland et al. [WWS*15] survey several systems and technologies, highlighting their advantages and limitations. An example of full-body and finger tracking using an optical motion capture system with reflective markers is shown in Figure 4.

4.2.1.2. Inertial-based Systems Inertial systems, including XSens [Mov23c] and Rokoko [Rok23b], use micro-inertial measurement units (IMUs), biomechanical models, and sensor fusion algorithms for motion capture. These systems measure rotational rates using gyroscopes, magnetometers, and accelerometers, translating them into a skeleton model. Tesla Suit [Sui23b] has introduced a suit that additionally encompasses a full-body haptic feedback system that uses electro muscle stimulation and transcutaneous electrical nerve stimulation. While inertial-based systems offer advantages such as cost-effectiveness, portability, and suitability for outdoor use, they are not without challenges. They can be complex, lack precise orientation measurement, and suffer from positional accuracy issues and drift over time. Despite these challenges, they are gaining popularity among independent game developers due to their quick setup. More recently, there is a trend towards reducing the equipment and body attachments for motion tracking using only six or even less inertial sensors, such as Sony's MoCap [Den22]; several machine learning techniques using sparse sensors show promise, especially in applications like virtual reality and sports training [YZH*22, PYA*23, DKP*23]. However, these methods are still in research development and face challenges in capturing highly dynamic and heterogeneous movements.

4.2.2. Markerless Systems

The markerless family of methods and systems is less intrusive than the previous two families of methods as it eliminates the need for subjects to wear specialized tracking equipment. Typically, the subject's outline silhouettes is captured from various angles using single or multiple vision or RGB-depth-sensitive cameras along with specialized software. A voxel-based representation of the subject's body evolves over time, and animation is achieved by fitting a skeleton into the 3D model [DAST*08, GSdA*09, VBMP08, LSG*11, LGS*13]. Over the last decade, numerous methods have been proposed; in this work, we draw insights from two key surveys, the work of Desmarais et al. [DMSM21] and the work of Xia's et al. [XGL*17]. Additional insights can be found in related studies [HXZ*19, XCZ*18].

Voxel-based representation encompasses three primary methodologies: *generative*, *discriminative*, and *hybrid* approaches. Generative motion capture methods (model-based) determine a person's pose and body shape by fitting a template model to data extracted from images. By inputting a set of model parameters, such as body shape, bone lengths, and joint angles, a representation of the model is generated, capturing the pose and shape of the body [GPKT10, WZC12, YLH*12, HYXC15, YSD*16]. On the other hand, discriminative approaches (model-free) map image features directly to pose descriptions or search a database of poses to find the closest match to the current image, as seen in studies such as [SSK*13, TSSF12, PMTS*15, YGTW15]. A blend of the previously mentioned strategies is utilized in hybrid methods [BMB*11].

Many researchers prefer using single-camera setups for markerless motion capture because of their cost-effectiveness, simplicity, and speed. Monocular systems, being generally less expensive than multi-camera configurations, with quicker setup times and fast data processing. Single-camera setups have been seen

in various recent studies [PCG*19, ZPT*19, YZZ*20]. To address the challenges of occlusions in monocular systems, and to achieve greater accuracy and precision, along with a full 360-degree coverage, the use of multiple cameras has become more prevalent [HAF*16, OERF*16, DDF*17], at the cost of a more complex configuration, and increased processing demands. In the recent era of DL, there has been a significant increase in efforts dedicated to pose reconstruction, utilizing both single and multi-camera setups, to enhance accuracy, adaptability, and automation. DL models have the capacity to automatically extract and learn complex features from raw data, perform end-to-end learning, adapt to various poses and conditions, and deliver high accuracy. It benefits from extensive data availability, parallel processing capabilities, and continuous advancements in model architectures, making it a versatile and powerful approach for accurately estimating poses in diverse applications [MSM*20, SAA*20, HZZ*21]. There are several commercial motion capture systems and companies in use today, that fuse markerless technology with DL, e.g., Microsoft's Kinect [Mic23], Move.AI [AI23], DeepMotion [Dee23], Plask [Pla23], Mediapipe [LTN*19], FreeMoCap [Mat23b], etc. Nonetheless, they have not yet achieved the level of accuracy and fidelity seen in optical MoCap systems.

The main advantage of these methods lie in their affordability, portability, the absence of body-attached sensors, and ease of setup. However, they encounter challenges when the articulated body is obscured from cameras due to self-occlusions or occlusions by other objects, subject clipping, or when the subject wears extensive clothing like bulky costumes. Furthermore, localizing subjects in a global coordinate system becomes extremely challenging without multiple synchronized video sources. Proper lighting conditions are essential, given that performances may vary from low-light conditions to illumination from several light sources. The clothing of performers and the complexity of the environment add to the challenges of obtaining desired outcomes. To address these challenges, controlled lighting and controlled background environments are typically employed. Despite these efforts, capturing multiple characters becomes problematic when other elements in the scene obstruct the subject's view, especially in scenarios like performances on a stage crowded with multiple people and objects. In comparison to optical or IMU-based systems, these methods have not yet achieved the same level of fidelity and versatility.

Recently, volumetric capturing has gained prominence among various vision-based methods. This approach constructs 3D models from multiple 2D images or videos, as demonstrated by companies like 4Dviews Studios [4DV23] and Evercoast [Eve23]. While useful for virtual reality and animation, it requires numerous cameras and view-angles in order to produce a detailed and accurate 3D model, while it is sensitive to lighting and moving objects. Heavy clothing poses or other elements in the scene pose challenges, obscuring body shape and creating shadows. Moreover, characters are usually represented as one combined mesh with their clothes, which makes it challenging to separate different costume layers, or rigging and skinning.

4.2.3. Discussion on the MoCap Categories

Optical motion capture technology stands as the industry standard for capturing the movements of intangible entities, including mu-

Table 3: The 3 primary Mocap categories with their advantages and disadvantages.

	Optical	Inertial	Markerless
Methodology	Cameras track reflective or light-emitting markers.	Utilizes accelerometers, gyroscopes, magnetometers.	Computer vision algorithms without physical markers.
Advantage	High accuracy and reliability.	Portable, feedback, versatile.	More flexible, convenient, often less expensive.
Disadvantage	Expensive, occlusions, less portable.	Prone to drift, magnetic interference, may cause discomfort to the subject.	Less accurate, sensitive to lighting, occlusions.

sical instruments. It finds widespread use in various fields, such as film production, video game development, biomechanics, and medical research. By utilizing cameras to track reflective or light-emitting markers, it is known for its remarkable precision and reliability, even when capturing subtle or intricate movements, and high-frame acquisition. However, it is generally costly, requires camera calibration, and can be hindered by occlusions. Conversely, inertial motion capture systems are prized for their portability and versatility. However, these methods are susceptible to issues like positional drift and magnetic interference, and the precision of data acquisition is somewhat diminished. Finally, markerless (or vision-based) motion capture systems leverage advanced computer vision algorithms, providing flexibility and convenience, albeit at the cost of reduced pose accuracy. They are notably sensitive to lighting conditions and occlusions. For a summarized overview of the advantages and limitations of each of these methods, please refer to Table 3.

4.3. Fingers Capture

Finger motion capture differs from full-body acquisition due to the intricate nature of hand movements, demanding advanced precision in capturing the highly articulated motions of the fingers. This precision is particularly crucial for applications such as surgery, sign language interpretation, or playing musical instruments. These systems encounter unique challenges that set them apart from whole-body tracking, including self-occlusion and precise contact modeling. They often require specialized hardware, such as gloves or infrared cameras.

The level of detail and precision needed to capture hand and finger movements can vary greatly based on the project's specific requirements. Some applications require high-fidelity tracking of hand and finger motions, while others prioritize capturing broader body movements. In contexts like musical performances, finger and hand movements play a critical role in instrument playing. To achieve accurate tracking of these intricate movements, a reliable motion capture system is essential. Therefore, assessing a system's ability to capture hands and fingers is vital for tailoring it to a project's specific needs.

Numerous commercial motion capture systems offer specialized gloves designed for finger motion capture. These systems include optical-based products such as Vicon, Optitrack, and PhaseSpace gloves [Sys23, Opt23a, Pha23a], as well as inertial-based products like Rokoko Smartgloves [Rok23c], Xsens Metagloves [Mov23b], Tesla Glove [Sui23a], and Perception Neuron Studio Gloves [Neu23]. Moreover, the MANUS Quantum MoCap

Metagloves can be integrated in most of the industry standard motion capture systems [Met23b]. These products inherit both the advantages and limitations of their respective family systems, as described in the previous section. It's important to note that using gloves may not be suitable for musicians as they can interfere with their ability to play instruments effectively. One potential solution is to employ an optical motion capture system that uses small markers directly applied to the fingers and hands or thin gloves with markers, as demonstrated in relevant research papers [Ari18, KMO*09, PPHB18].

Another approach is to use camera-based systems that do not require physical markers attached to the body, such as the Leap Motion Controller [Ult23]. There is a wide range of specialized hand tracking methods that rely on silhouette extraction principles and achieve animation by fitting a skeleton into a 3D model. Recent advancements in this field, exemplified by methods like DeepMotion [Dee23], Move.AI [AI23], Google's MediaPipe [LTN*19], and the Free Motion Capture Project (FreeMoCap) [Mat23b], have expanded their capabilities to encompass tracking of complete body parts, including the face, hands, and fingers, even for multiple individuals, using only monocular video input.

4.3.1. Markerless Systems Accuracy

The accuracy of markerless systems in hand tracking and reconstruction surpasses that of full-body tracking, primarily due to the more constrained articulation of hand movements. However, while these markerless systems may not be the primary choice for hand tracking in musical instrument applications due to sensitivity to lighting and environmental conditions and susceptibility to occlusions, they are more commonly used in hand tracking compared to full-body tracking solutions.

4.4. Face Capture

Facial capture is specialized and distinct from full-body motion capture due to the unique challenges associated with capturing the complexity and subtlety of facial expressions, as well as its critical role in conveying emotions and character in various applications. The technical challenges involved in capturing minute facial details, and the priority on realism over efficiency in facial capture setups. These distinct requirements contribute to the specialization of facial capture as a field within motion capture technology.

Facial expression capture and motion transfer to virtual characters have been subjects of research for over three decades. Central to this has been the Facial Action Coding System (FACS), developed by Paul Ekman and Wallace V. Friesen in the 1970s [EF78],



Figure 5: Use of a motion capture system to holistically track a musician (full-body, face, fingers), drum sticks, and drums. Image extracted from [Cin21].

which categorizes facial expressions into distinct ‘action units’ corresponding to muscle movements. This system has been essential in both psychological studies and animation. Parallel to FACS, blendshapes in animation adjust a character’s neutral facial mesh to various predefined expressions. By combining these shapes, diverse facial expressions are achieved, making blendshapes a prevalent technique in film and gaming [LAR*14]. In a recent work by Choi et al. [CEM*22] is introduced the advanced facial animation system “Anatomy” which deviates from FACS-blendshape systems, offering detailed, anatomically accurate control and easy animation transfer from actor to virtual character. In this survey, we will focus on contemporary systems widely utilized today, rather than an extensive discussion of facial expression capture research. For an in-depth review of such research, readers are directed to comprehensive works by Vilchis et al. [VPMRGM23] and Wen et al. [WZHC22].

Industry standards in facial motion capture have converged towards the utilization of Head-Mounted Cameras (HMCs), exemplified by systems such as Vicon’s Cara [Car]. These specialized helmets can accommodate both cameras and smartphones, ensuring a stable and consistent perspective of the actor’s face, even during head movements. This prevents any blurriness in the expressions captured and maintains the quality of the data. The lightweight and comfortable design of HMCs ensures that the artist’s freedom of movement is preserved, providing a seamless experience during performances. Moreover, the capabilities of HMCs extend beyond facial motion capture, encompassing comprehensive motion capture setups that include the entire body and musical instruments, as illustrated in Figure 5. This exemplifies the versatility and wide-ranging applications of HMCs in motion capture.

There are two principal categories within the domain of HMCs: marker-based and markerless systems. Marker-based systems use physical markers tracked by cameras. Examples include reflective marker systems like Vicon and OptiTrack, which use small reflective spheres and infrared cameras; painted or sticker markers applied directly to the actor’s skin. Despite their accuracy and reliability, these systems can be intrusive and time-consuming to set up. On the other hand, markerless systems eliminate the need for physi-

cal markers, relying on computer vision and machine learning algorithms to directly track facial movements, for example, by drawing dots on the face to extract facial expressions. This category includes depth-sensing cameras (e.g., Apple’s Face ID using TrueDepth Camera technology [App23c]) that generate a three-dimensional map of the face, RGB cameras combined with software algorithms (e.g., Faceware [Fac23b]), and smartphone applications capable of markerless motion capture. A trend towards markerless systems is evident, marked by a transition from compact cameras such as Go-Pro to devices featuring Apple’s TrueDepth Camera technology. This technology captures facial data by projecting and analyzing numerous invisible dots, generating a depth map and concurrently recording an infrared image of the face at high resolution (up to 4K) and frame rate (up to 240 fps).

Both marker-based and markerless systems have their distinct advantages and are suited to different applications. Marker-based systems, while potentially intrusive, offer unparalleled accuracy, especially for subtle facial expressions. Markerless systems, in contrast, provide rapid setup and are less obtrusive, but may not achieve the same level of precision. The selection between these two types of systems should be informed by the specific requirements of the project and the resources available.

In terms of software, there are numerous specialized applications designed to animate characters based on the facial data captured by the camera. Among the many options available, such as Maya [Aut23] and Blender [Ble23], MetaHuman Animator [Gam23] from Epic Games stands out as a leading solution. This software enables the rapid and precise translation of real-world performances into high-fidelity facial animations, compatible with both iPhone and stereo HMCs. Other applications such as Live Link Face [Fac23a], Rokoko’s Face Capture [Rok23a], and iClone [Rea23] also offer real-time facial motion capture and are compatible with Apple smart devices.

4.5. Challenges in Multi-modal Synchronization

Synchronizing multi-modal data captured during an instrumental music performance is of critical importance. The research conducted by Li et al. [LLD*19] addresses the intricate challenge of synchronizing concurrent sound sources when generating multi-track datasets. Furthermore, when integrating data from diverse modalities using varying capture devices, achieving synchronization among these devices is vital to attain the desired output. These devices may possess distinct processing speeds, capture frequencies, and data transfer rates, which can introduce inconsistencies. While manual synchronization of all devices is possible, this approach is labor-intensive and susceptible to errors. An efficient alternative is to employ a global clock. Timecode generators are commonly used for this purpose, maintaining local synchronization across devices by assigning a unique code to each frame or data packet. This ensures a consistent timeline across multiple devices, thereby facilitating precise data alignment. Another method that can aid in synchronization, though it may not completely solve the issue, is Genlock, which ensures that all devices operate at the same capture frequency. This is particularly crucial in scenarios where even minor differences in data capture rates can result in significant inconsistencies in the final output. Most MoCap sys-

tems support both Genlock and Timecode [Opt, XSe]. Typically, a central control PC or synchronization unit is utilized to initiate and conclude recordings on the various devices within the same network.

5. Musical Performance Analysis

The analysis of musical performances involves evaluating a range of modalities that stem from playing a musical instrument. This assessment encompasses not only the music itself but also the performer's posture, including body language, finger movements, and facial expressions. Both methods provide distinct perspectives, offering valuable insights into the nuances of the performance and contributing to a comprehensive understanding of the artistic expression.

5.1. Audio Analysis

Music Information Retrieval (MIR) and Musical Performance Analysis (MPA) are two closely related research fields, both centered on aspects of music. MIR concentrates on developing algorithms and techniques to extract information from music audio signals, which can serve various purposes, including music genre classification [TC02], instrument classification [HBKD06], beat detection [PBDL23], music recommendation [ZSQJ12] and music transcription and melody extraction [SG12]. The work presented in [W*03], which revolves around audio identification, has been successfully incorporated into applications like Shazam [Sha23b]. Shazam stands as an exemplary MIR application, capable of identifying songs by analyzing short audio samples and matching them against an extensive audio database. Within the realm of MIR, there are several noteworthy surveys that provide valuable insights, including [Dow03, TWV05, Ori06, CVG*08, SGU*14, SNA19, KR12].

On the other hand, MPA focuses on the evaluation of live musical performances, examining how musicians interpret a piece and highlighting nuances in variations and expressiveness that go beyond the original score. For instance, one application of MPA is tutoring musical instrument learning, where students perform and receive feedback [EMNS20]. Another MPA example is illustrated by the PHENICX project [LGS15], which focuses on visualizing information within orchestral music, incorporating elements from the musical score and performance-related aspects. It is important to recognize that the interpretation of a musical piece during a performance can profoundly influence listeners' perceptions. Even when working with the same musical score, different renditions can lead to distinct preferences and interpretations among listeners. The parameters of music audio performance can be categorized along the same fundamental dimensions as audio: tempo and timing (musicians adjust tempo and timing during performance for expressive effect), dynamics (performers make decisions about volume variations based on their musical judgement), pitch (musicians enhance musical expression by employing techniques like vibrato, adding nuances to the prescribed pitches in the score), and timbre (performers shape the timbre of a musical piece through their playing techniques and instrument configurations) [Ler12, LAPG19, LAPG21]. Within the domain of

MPA, there are relevant surveys that offer valuable insights, including [Gab99, Gab03, GDDP*08, Ler12, LAPG19, LAPG21].

5.2. Pose Analysis

Similar to the research conducted in the field of sports analysis and physiology [CECS18, BNWY23, HW23], this subsection provides an overview of methods that evaluate performers' posture and musculoskeletal systems. These methods are designed to promote both their physical and mental well-being, prevent injuries, and enhance the quality of their performance.

Pose Analysis using Motion Capture: The field of instrumental performance analysis has been significantly enriched through the integration of motion capture technology, providing intricate insights into musicians' motor skills for improved training and injury prevention. The Tone project [Cye23] introduces a virtual mirror, offering musicians real-time musculoskeletal feedback and the ability to analyze their posture and muscle activity from various perspectives. Additionally, Ancillao et al. [ASGA17] examines upper limb and bow positioning in violin players, emphasizing the criticality of quantitative assessments in skill evaluation and motor disorder diagnosis. Investigating finger movement coordination during piano playing, the study conducted by Wings et al. [WF15] discerns the nuanced differences in technique between professional and amateur musicians. Wolf et al. [WMB*19], introduced a marker-based method to explore upper body movements, with a particular focus on addressing musculoskeletal disorders of high string players (violin and viola) – see Figure 6. In the pursuit of injury prevention and performance optimization, Shan et al. [SV03] delve into Overuse Syndrome in violinists, advocating for training strategies that emphasize physical economy.

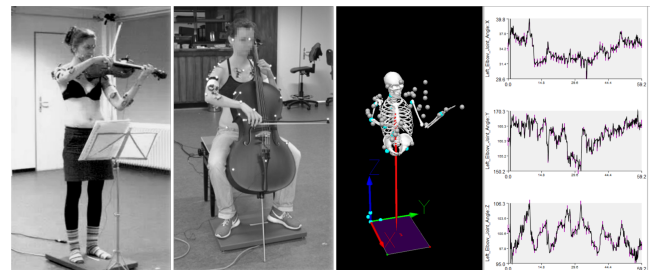


Figure 6: Analysis of 3D upper body kinematics of high string players during performance. Image extracted from [WMB*19].

Spahn et al. [SWEN14] investigate how different playing positions affect body movements and weight distribution in violinists, highlighting the implications for health and performance. Furthermore, the ergonomic risk factors associated with hand movements in pianists are explored by Sakai et al. [SLS*06]. Hopper et al. [HCW*17] provide a comprehensive analysis of movement patterns in cellists, contributing valuable insights for teaching methods and injury management. Rabuffetti et al. [RCBF07] investigated how different violin shoulder rest setups impact players' performances, using an optoelectronic motion capture system to analyze

fifteen violinists playing a G scale under three shoulder rest conditions. The study found that a higher rest led to less rotations of head, left shoulder, and left acromion elevation, but increased left shoulder flexion and left forearm pronation, emphasizing the musicians' ability to adapt their body rather than their bowing technique. These results suggest that tailored assessments and improved shoulder rest designs could enhance player comfort and adaptability without compromising sound quality. Together, these studies underscore the transformative impact of motion capture technology in the realm of musical performance, fostering a data-driven approach to skill development, injury prevention, and the enhancement of training methodologies.

Pose Analysis using Computer Vision: Blanco-Pineiro et al. [BPDP15] investigated the postures of 100 music students, utilizing video and photo analyses performed by expert evaluators. The methodology included recording musicians in both seated and standing positions, as well as capturing still images in “ready-to-play” static poses. They examined 11 variables related to overall and specific body part postural quality, identifying common postural flaws and the contexts in which they occur. The aim was to highlight these issues and promote corrective measures for better postural habits during musical performances. Araujo et al. [ACML09] investigated postural flaws in four student violinists from an orchestra, using 20-minute frontal video recordings and anatomic markers. The study aimed to categorize and evaluate the frequency of these postural flaws. The findings revealed that all the violinists displayed postural flaws during their performance, highlighting that these flaws were unnecessary and could be avoided as they are not intrinsic to standard instrumental techniques. Chan et al. [CDA13] took a different approach by evaluating the effectiveness of a 10-week intervention programs on the posture of 57 professional orchestral musicians. Utilizing photographs for pre-and-post intervention assessments, they found improvements in Exercise Therapy showcasing the potential of visual assessment tools in evaluating postural changes. In a different study, Bejjani et al. [BH89] examined how body measurements influence the postures of 16 professional trumpeters while performing standing up. Through detailed photographs and anthropometric data collection, providing data on the physical limitations that can impact a musician's performance. Longo et al. [LDSR*20] contributed by investigating the impact of body posture on voice performance during simultaneous singing and instrument playing. The study, which included 17 musicians, involving guitarists and pianists, utilized the Multi-Dimensional Voice Program (MDVP) for voice analysis and visual assessments for evaluating posture. Results underscored the complex relationship between a musician's physicality and their auditory output. Shifting the focus to ergonomics, Valenzuela-Gomez et al. [VGRGAG20] investigated the postural implications of different guitar supports (guitar cushion, rigid lap support and footstool) on classical guitarists. By integrating REBA and 3DSSPP software with subjective questionnaires, their work highlighted the ergonomic challenges and the need for improved support designs to enhance comfort and performance. Finally, Islan et al. [IBP*18] provided a comprehensive analysis of the glenohumeral joint dynamics in violinists, employing a multifaceted approach involving the RULA (Rapid Upper Limb Assessment) method, CATIA software for geometric modeling, and ANSYS software for FEM



Figure 7: Analysis of the upper body posture of a musician playing high-stringed bow instrument using 3D back scans. Image extracted from [OMB*18].

analysis. This study enriched the understanding of musculoskeletal strains in musicians, particularly the impact of repetitive movements and prolonged training, offering crucial insights for future ergonomic interventions.

Other approaches: Utilizing ultrasound topometry, Piatek et al. [PHG*18] assessed the spinal kinematics of fourteen alto saxophonists, exploring the back strain associated with different saxophone-carrying systems (neck-strap, shoulder-strap, and Sax-holder). Additional tests with various saxophone weights indicated that the instrument's weight had a more substantial impact on body balance than the carrying system used. In another vein, Park et al. [PKH*12] employed Electromyography (EMG) and 3D motion analysis to assess the relationship between neck pain and playing posture by comparing muscle activity and neck motion between nine students with neck pain and nine without. This study found that students with neck pain had more neck strain and muscle activity, which shows how dangerous it is to play an instrument in an asymmetrical posture. Complementing this, Yagisan et al. [YKGG09] utilized digital photogrammetry to examine upper right limb positions in nine violinists, aiming to refine teaching methodologies and avert musculoskeletal issues. Chung et al. [CRO*92] analyzed wrist movements of nine pianists through biaxial electrogoniometers. In different approaches, Ohlendorf et al. [OMC*18, OMB*18] employed 3D back scans, videorasterstereography, and pressure mapping to explore the upper body posture of musicians playing high-stringed bow instruments and to assess the effects of varied ergonomic chairs on posture and seating pressure (see Figure 7). The studies unveiled substantial postural alterations and pressure variations, accentuating the influence of chair design and instrument use on musician health. Coker et al. [CBHC04] conducted a study with 14 percussionist to investigate how percussive exercise complexity affects postural sway and to explore the impact of a 5-week flexibility program on the participants' postural stability. The participants, divided into a flexibility-training group and a control group, underwent pre and posttests involving eight varying complexity exercises

while standing on a Center Of Pressure (COP) platform. In a unique approach, Clemente et al. [CLC*14] studied head and cervical postures in piano players during performances through accelerometry, providing valuable data on postural tendencies. Additionally, goniometry and electromyography were employed by Baadjou et al. [BvEBV*17] and Cattarello et al. [CVD*18] to analyze the connection between body posture, muscle activity, and sound quality in clarinetists, as well as the impact of different chairs on the postures of violin and viola players. These studies collectively underscore the significance of ergonomic considerations and the potential of postural exercise therapy in enhancing musicians' performance and well-being.

5.2.1. Conclusions

In this section, we emphasize the importance of evaluating both the auditory output and physical movements in a musical performance. We delve into various methods and applications for analyzing musical performances, which play a pivotal role in assessing quality and enhancing artistic development. These analyses have the potential to enrich the overall experience for both artists and their audience. For example, the audience can enjoy a more immersive experience by gaining additional insights into the performance, such as detailed note transcriptions or experiencing dynamic lighting adjustments that align with the mood and artistic intent. Musicians, on the other hand, can derive multiple benefits from such analyses. They can use the insights to prevent potential injuries resulting from repetitive movements or improper posture during performances and to refine their techniques. They can serve as a valuable tool in music education, helping musicians refine their skills. Furthermore, these insights can assist instrument manufacturers in creating more ergonomic instruments and supportive accessories, ultimately contributing to the prevention of musculoskeletal issues among performers. Future research direction could be benefited by the recent developments in volumetric capturing that can enhance the analysis. When combined with other data modalities such as ECG, EEGs, dynamic 3D scans (i.e., 4D scans), and muscle deformations, the analysis can further improve our understanding of performance quality.

6. Musical Performance Synthesis

Musical performance synthesis refers to the intricate process of replicating the nuances of a physical musical performance using technology [DZBKM22]. This interdisciplinary field brings together elements of music theory, sound science, and computer methodologies to capture more than just the fundamental notes of a piece. It aims to encapsulate the true essence of a performance, encompassing elements such as motion, unique expressions, dynamics, and the variations introduced by an artist. Within this section, our primary emphasis is on techniques designed to produce human motion in direct response to audio or MIDI input. An essential aspect in the faithful replication of a musician's performance on an instrument lies in our ability to capture the subtleties of their gestures, posture, fingers, and emotions.

A related area of research in human animation synthesis involving musical instrument performance is audio-driven dance motion

synthesis [YWJ*20], sign language generation [RKES21], and gesture generation [GFH*23] to audio. In the context of dance motions synthesis, numerous studies have leveraged machine learning methods to create realistic human animations. However, when it comes to generating motion based on audio input, a significant challenge lies in achieving temporal consistency and synchronization between the audio and the motion. Various techniques have been explored, such as recurrent neural networks (RNNs) [GMK*19] but are susceptible to temporal error accumulation issues and may result in static poses, particularly when dealing with inputs not present in the training data or when noise is introduced. To address this, temporal convolution was introduced [GBK*19] to generate simple gestures. Capturing complex and varied dance movements or nuanced musician motions presents a great challenge due to their intricate long-term spatial-temporal and kinematic characteristics. Recent studies have explored techniques like variational autoencoders (VAEs) [LYL*19], generative adversarial networks (GANs) [SWC*21], auto-regressive models [ZWC*22], acLSTM for simulating dance and music-related motions with global structure consistency [AYA*23], transformers [LYRK21], and choreography-oriented graph-based frameworks [CTL*21]. Also, the work of Zhou et al. [ZLZ*23] addresses synchronization problems by dynamically adapting animations to match the tempo of an audio file.

It is important to acknowledge that dancing and musical performances share certain similarities, such as their reliance on rhythm and timing to create a sense of movement and flow [Bri]. Both are considered as creative forms of expression capable of evoking emotions, telling stories, and conveying messages. However, they also exhibit notably different characteristics, particularly in the context of playing instruments. Musicians frequently engage with a diverse array of instruments, making the capturing of nuanced motions, particularly subtle finger movements, a challenging task. Moreover, the demand for high precision at contact points, especially in intricate finger positioning, further increases the difficulty of motion synthesis. Musicians typically exhibit more static and delicate movements in contrast to the expansive stage spaces commonly used by dancers. Moreover, in musical performances, the motion itself generates the sound, in contrast to dancing where the audio complements the movement, necessitating precise synchronization between the captured motions and the resulting audio. For all these reasons, the field of musical instrument motion synthesis has experienced limited progress. Another factor contributing to the underdevelopment of this field is the scarcity of accessible motion repositories. Previous works tend to overlook the multi-modality inherent in musical instrument movements, often focusing solely on either upper body actions or finger movements, thus failing to comprehensively encompass the entirety of the motion. As a result, most available methods generate partial body animations. In this section, we review the most prominent methods for synthesizing musical instrument performances, organized according to the specific musical instrument being emulated.

Piano: The piano is a well-known musical instrument that is conventionally played with the performer seated close to the instrument. When playing the piano, the upper body is primarily engaged in striking the piano keys, while the feet are responsible for

manipulating the pedals. Achieving the desired musical notes on a piano necessitates an extremely precise placement of the fingers on the keys. Most approaches to piano playing concentrate on the movements of the hands and fingers, overlooking the broader physical engagement required for a nuanced performance. Early methods for generating piano animations, like those by Sekiguchi and Eiho [SE00], used a virtual space simulator and hand movement generator. They assigned fingers, positioned hands using spline functions, and calculated finger angles based on note difficulty. Nagata et al. [KMO*09] used motion capture to acquire piano fingering movements, rendering visualization of the fingering, and automatic fingering generation utilizing optimized algorithms. Zhu et al. [ZRH13] used motion planning and optimization methods based on graph theory for 3D piano animations, where, similarly to Yamamoto et al. [YUS*10], they used Inverse and Forward Kinematics for hand modeling and animation, from MIDI files. More recently, there has been a shift towards the use of machine learning and generative approaches. The first category of methods utilized Long Short-Term Memory networks (LSTMs), mainly due to their capabilities in modeling sequential data and capturing the temporal dependencies. For instance, Li et al. [LMD18] developed a deep neural network system that translates MIDI note data and metric structures into a real-time skeleton sequence of a pianist playing a keyboard instrument. Their approach combined Convolutional Neural Networks (CNNs) and LSTMs to generate human-like piano performances. Similarly, Shlizerman et al. [SDSKS18] transformed audio recordings of piano (and violin) performances into animations. They trained an LSTM network on internet-sourced videos and applied the predicted points to rigged avatars (see Figure 8). Bogaers et al. [BYV21] introduced a music-driven method that generates expressive musical gestures for virtual humans using 3D motion capture data and LSTM networks. In contrast, Guo et al. [GCZ*21] introduced an augmented reality training system for piano, using MIDI data to generate 3D hand animations based on pre-trained Hidden Markov Models. The Viterbi algorithm determined the optimal finger path, and optimization methods modeled different fingerings and skills. Xu et al. [XLW*22] used Reinforcement learning (RL) to create piano finger animations. They employed an end-to-end RL approach to train an agent for piano playing using touch-augmented hands on a simulated piano. They designed touch- and audio-based reward functions and utilized the Soft Actor Critic (SAC) method for training the RL agent. The results showed that tactile sensor feedback enhanced learning efficiency, leading to proficient piano playing in a fixed number of training iterations. In a recent study, Zakka et al. [ZWS*23] introduced a system that builds upon the work of Xu et al., by utilizing deep reinforcement learning techniques to train anthropomorphic robotic hands in piano playing, resulting in the synthesis of dexterous robotic hand performance.

Violin: The violin is a renowned musical instrument traditionally played by a musician holding it close to their body. When playing the violin, the performer uses the bow to draw across the strings, while their fingers press on the strings to produce specific musical notes. Achieving the desired tones on a violin requires precise finger placement and control of the bow's speed and pressure to produce accurate and expressive music. Several studies have explored the use of neural networks and deep learning in generating

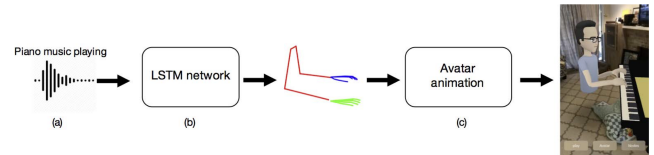


Figure 8: Synthesizing piano playing movements: (a) input an audio signal (b) fed into LSTM network to predict body movement points, (c) animate an avatar and show it playing the input music on a virtual piano. Image extracted from [SDSKS18].

animations for violin performances. Kim et al. [KCMT00] developed a system using a neural network to control hand movements and optimize training examples, including automatic finger placement via best-first search. Liu et al. [LLH*20] adopted a divide-and-rule approach, employing separate models for bowing, hand position, and upper body expression with various network architectures, including CNNs and CRNNs. Kao and Su [KS20] built on Shlizerman's work [SDSKS18] with LSTM by enhancing neural network models for 3D violinist skeleton generation, integrating an encoder-decoder architecture, self-attention, beat tracking, and bowing attack inference. Lin et al. [LKT*20] enabled real-time interaction between humans and virtual musicians using Dynamic Time Warping (DTW) for music tracking and an RNN with LSTM units for 3D body movement. Hirata et al. [HTSM21, HTHM22] estimated bowing dynamics and body motion from audio for more detailed violin performance animations (Figure 9). More recently, Shrestha et al. [SFH*22] introduced a method using transformers to generate fine-grained corrections and visual information for 3D violin animations, additionally offering a multi-modal dataset called MAPdat for this purpose.



Figure 9: Synthesizing violin playing movements: (a) Input: violin audio (b) Output: performance animation. Image extracted from [HTHM22].

Other Instruments: ElKoura and Singh [ES03] investigate realistic hand movements for guitar playing, employing a 27-degrees model, specifically addressing complex finger positioning. They introduce a data-driven algorithm utilizing k-Nearest Neighbor search to map hand configurations to more realistic ones and offer a procedural algorithm for animating the fretting hand during guitar performance, intended for music education and analysis. In a similar fashion to piano and violin synthesis, Shirai and Sako [SS21] extend the work of Liu et al. [LLH*20] by developing a 3D double bass player movement generation method. They leverage a 2-layer LSTM network and a novel motion dataset based on authentic performances, incorporating both bowing and fingering data. Their approach involves the assessment of different model structures and employs the mean absolute error as a loss function, op-

timized with Adam. On the other hand, Chen et al. [CFZ*21] introduce a deep learning-based system for synthesizing upper body animations driven by Guzheng music. They harness the power of a generative adversarial network (GAN) to capture the dynamic relationship between music and human motion. GANs are gaining traction in the field of motion synthesis due to their ability to learn from data, generate diverse and coherent motion sequences, capture temporal dependencies, and provide a feedback mechanism for producing realistic and creative animations (Figure 10). Furthermore, Zhu et al. [ZLZ*21] present a multi-stage framework for generating performance videos from audio clips. It derives global appearance and localized spatial details, converting audio to body and hand keypoints and coarse video representations. The final stage employs a Structured Temporal UNet (STU) to extract structured information and temporal consistency, showcasing its superior performance on the Sub-URMP dataset and offering promising prospects for the future of audio-visual computation research. Finally, Li [Li22] explores the use of Augmented Virtuality for musical performances by recording musicians' performances with any camera. These recordings are transformed for use in virtual environments, utilizing deep learning for real-time video matting with MODNet and realistic video shading rendering.

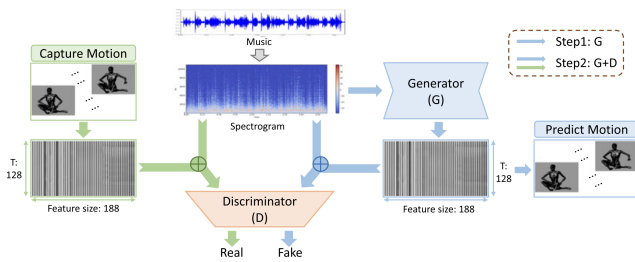


Figure 10: Overview of the music-to-motion framework proposed by [CFZ*21]. The framework consists of a generator and a discriminator.

6.1. Conclusions

In summary, it is clear that the field of musical performance synthesis has undergone a substantial transformation in recent years. It has shifted from optimizing fingering based on predefined rules and heuristics to utilizing machine learning techniques, particularly deep learning, to create natural and expressive musical performances. These advancements have not only facilitated the automatic generation of lifelike 3D animations and human-like performances but have also raised the prospect of exciting applications in virtual performances, interactive entertainment, music education, and humanoid robotics. However, the development of this domain faces challenges due to the limited availability of multi-modal musical instrument repositories and the difficulties in capturing and synchronizing them. Most existing research has primarily concentrated on partial body reconstruction, with a strong emphasis on fine details in hand and finger animations. To push this field forward, future research should broaden its focus from merely hitting the right notes to generating expressive musical gestures and harnessing tactile feedback. A recent trend in this field is the adoption

of generative models, including progressive GANs like GANimator [LAZ*22] and diffusion models like MDM [TRG*23], Motion Diffusion in Latent Space [CJL*23], and TEDi [ZLAH23]. These models, although not directly related to audio or music inputs, excel in handling other multi-modal inputs like text and emotions. They generate temporally consistent, high-fidelity, and natural long motions with the potential for sentimental control. These advancements, with the incorporation of additional constraints for precise motion synthesis, such as physics-based constraints [YSI*23], could provide a robust framework for use in musical instrument performance synthesis. This promising direction opens doors to even more sophisticated and expressive musical performances in the future.

7. Conclusions and Discussion

In this report, we provide an overview of the current state of Virtual Instrument Performances which can be a powerful tool for digitizing and visualizing the performing arts. This process plays a pivotal role in preserving cultural heritage, expanding access to a global audience, fostering creativity, and enriching educational resources. We have explored various aspects, including methods for storing motion and audio data (Section 2) and a comprehensive list of significant multi-modal datasets (Section 3). However, a universal schema for capturing and storing musical performances is still lacking. The imperative need for a common format to represent multi-modal data is evident. By defining appropriate encodings for each data type, we can effectively capture the nuances of motion and audio, benefiting the broader community and advancing future research.

Next, we present methods for capturing instrument-based performances (Section 4). These methods encompass capturing audio directly (MIDI) or indirectly from the instruments (raw audio) and capturing the motion of performers, including their body, fingers, and face. Each aspect presents unique challenges, and we summarize the pros and cons of each technology. We emphasize the importance of synchronized data from various sources, as high-quality data is indispensable for subsequent tasks, such as training models.

High-quality multi-modal data enables the development of innovative solutions to analyze performances (Section 5). These solutions help us understand how performers interact with their instruments for both performance and health reasons. We cover approaches related to audio and pose analysis separately, presenting several solutions based on technologies such as motion capture, vision, and photogrammetry. We identify the potential of high-quality motion capture systems and newer approaches like volumetric capture, which allow for non-intrusive analysis of body interactions with instruments, especially when combined with various data modalities including, ECG, EEGs, dynamic 3D scans, and muscle deformations; these methods are expected to further enhance our comprehension of performance quality.

Importantly, high-quality data opens the door to disruptive approaches that can enhance artists' creativity and change the possibilities in virtual performances (Section 6). Generative deep learning systems are at the heart of these possibilities, enabling the generation of new motion that respects the properties of source data,

based on different control signals such as audio (e.g., MIDI, raw audio), emotions/style, and text. Recent trends in this area highlight the importance of diffusion models and physics-based constraints to generate physically plausible and expressive instrument performances.

7.1. Recommendations

Based on the insights we have gained on our journey, we recommend that, within today's technological landscape, a comprehensive pipeline for capturing and storing musical instrument performances should consider the following factors. Indeed, capturing musical performances requires a delicate equilibrium between recording auditory and visual intricacies. When it comes to audio recording, sound techniques like employing microphones and audio interfaces provide precise representation, particularly for electronic instruments that can directly interface with computers. While MIDI files offer precision in musical representation, they necessitate manual transcription for instruments lacking MIDI outputs. On the other hand, the choice of motion capture technology for pose performance acquisition is a critical decision. Marker-based systems offer precision but may be cost-prohibitive, they are not as portable as other solutions, and impose movement restrictions due to sensors attached to the body or the instruments themselves. In contrast, markerless systems offer greater versatility but may compromise on fidelity. Capturing the subtleties of finger movements presents its own set of challenges, including self-occlusion and the need for specialized equipment like gloves, which can be restrictive for musicians. Recent advancements in deep learning and computer vision offer reliable methods in controlled environments, primarily due to the highly constrained articulation of the hand, aiding in pose prediction. Facial capture, crucial for conveying emotion, relies on systems like FACS for expression categorization. While HMCs ensure consistent facial capture, the choice between marker-based and markerless techniques introduces its own challenges, ranging from intrusiveness to potential precision limitations. Lastly, the synchronization of multi-modal data is of utmost importance, especially when different devices can introduce inconsistencies. Techniques such as timecode generators and Genlock are essential, with many motion capture systems supporting both methods for seamless synchronization. Ultimately, the selection of technology hinges on striking the right balance between achieving precision and practicality in capturing musical performances.

7.2. Challenges and Future Work

The field of virtual instrument performances continues to face a range of emerging challenges, which we aim to address and outline future directions for innovation and advancement in this domain. An interesting avenue for exploration lies in the collaborative synthesis of musical instrument performances by various entities, including human artists, robots, and AI agents. While there are numerous works in the literature on virtual performances, there is notably limited discourse on the interaction between performers, whether they are virtual or real. A notable example of work shedding light on this aspect is the research conducted by Chakraborty et al. [CDT21]. Such collaborations raise intriguing questions about the division of roles, creative decision-making, and the integration

of AI into artistic expression. Future research can delve into the possibilities and challenges of this multi-agent collaboration, exploring how it might redefine the boundaries of virtual instrument performances.

Within the domain of creative and artistic performances, the visualization of virtual instrument performances is of paramount importance. Future research should focus on the development of innovative methods to visualize these performances, encompassing advanced techniques for rendering lifelike avatars, creating immersive virtual concert halls, and generating interactive visual representations of the performer's emotional state. Effective visualization not only enhances the audience's experience but also offers valuable insights for performers and researchers. Moreover, the advent of XR technologies introduces unique challenges when dealing with multiple performers. Future research should explore the intricacies of XR settings, investigating how VR devices and sensors might offer new ways for performers to interact with virtual instruments. Understanding the dynamics of multiple performers in these environments, such as collaborative improvisation or synchronized actions, is essential for pushing the boundaries of virtual instrument performances.

Furthermore, future research should place a stronger emphasis on collecting data that describes the style and conditions of performances. This includes monitoring the performer's heart rate, tracking improvisational moments, and observing emotional fluctuations during the performance. This data can provide a deeper understanding of the performer's state and creative choices, ultimately leading to more immersive and emotionally resonant virtual instrument performances. Investigating the integration of biofeedback data and improvisational tracking can pave the way for groundbreaking developments in this field.

Another avenue for future research involves exploring the possibilities of simulating audio and interactions in virtual environments that differ from the settings in which the performance was originally captured. It is evident that the spatial and environmental context significantly impacts the audio and visual experience of a performance. For instance, capturing a performance in a studio but simulating it in a large auditorium or a confined space can result in distinct audio characteristics and altered expressions in motion. Similarly, investigating how the presence or absence of an audience influences a performer's emotion and expression is also a promising avenue. Future research should delve into audio simulation techniques, environmental modeling, and their effects on the overall virtual instrument performance.

We consider the thoughtful and careful study of *privacy* and *ethics* around the application of these methods to be an important issue. The digitization of an artist and their performance includes sensitive private data, such as motion, playing style, and biofeedback data. Additionally, there is the issue of unauthorized use of a performer's data to generate novel performances, raising issues about ownership of synthesized content, similar to the discussion around generative models for images and audio [Bar23] and Large Language Models [WMR*21, Har23]. Since DL systems are trained on data, it is essential to select data in such a way that racial, sex, or body-related biases are minimized or ideally removed completely. Ethics is a crucial topic in several domains surround-

ing XR environments and DL methods; we recommend reading the interesting study by Slater et al. [SGLH*20] for more information.

Addressing the challenges and opportunities presented in these areas will undoubtedly lead to groundbreaking developments in this dynamic field. Researchers and practitioners are strongly encouraged to explore these themes, pushing the boundaries of what is achievable in the world of virtual instrument performances.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy. This work has received funding from the European Union under grant agreement No 101061303. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

- [4DV23] 4DViews: Studios. <https://www.4dviews.com>, 2023. [Accessed: October 2023]. 12
- [AAC22] ANDREOU N., ARISTIDOU A., CHRYSANTHOU Y.: Pose representations for deep skeletal animation. *Computer Graphics Forum* 41, 6 (dec 2022), 155–167. doi:10.1111/cgf.14632. 4
- [AB06] ARNAUD R., BARNES M. C.: *Collada: Sailing the Gulf of 3d Digital Content Creation*. AK Peters Ltd, 2006. 5
- [Abl23] ABLETON: Ableton live, 2023. Accessed: October 2023. URL: <https://www.ableton.com/en/live/>. 11
- [ACML09] ARAÚJO N. C. K. D., CÁRDIA M. C. G., MÁSCULO F. S., LUCENA N. M. G.: Analysis of the frequency of postural flaws during violin performance. *Medical problems of performing artists* 24, 3 (2009), 108–112. doi:10.21091/mpa.2009.3024. 16
- [ADB*23] AGOSTINELLI A., DENK T. I., BORSOS Z., ENGEL J., VERZETTI M., CAILLON A., HUANG Q., JANSEN A., ROBERTS A., TAGLIASACCHI M., SHARIFI M., ZEGHIDOUR N., FRANK C.: Musiclm: Generating music from text, 2023. [arXiv:2301.11325](https://arxiv.org/abs/2301.11325). 7, 9
- [Ado23] ADOBE: Adobe audition, 2023. Accessed: October 2023. URL: <https://www.adobe.com/audition>. 11
- [AI23] AI M.: MOVE.AI. <https://www.move.ai>, 2023. Accessed: August 2023. 12, 13
- [AL13] ARISTIDOU A., LASENBY J.: Real-time marker prediction and CoR estimation in optical motion capture. *The Visual Computer* 29, 1 (Jan 2013), 7–26. doi:10.1007/s00371-011-0671-y. 11
- [ALD*11] ABESEER J., LARTILLOT O., DITTMAR C., EEROLA T., SCHULLER G.: Modeling musical attributes to characterize ensemble recordings using rhythmic audio features. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 189–192. doi:10.1109/ICASSP.2011.5946372. 8, 9
- [APO23] APOLLO: Preserving berlin's musical treasures with augmented and virtual reality, 2023. Accessed: October 2023. URL: https://ec.europa.eu/regional_policy/en/projects/Germany/preserving-berlins-musical-treasures-with-augmented-and-virtual-reality. 2
- [App23a] APPLE: Garageband, 2023. Accessed: October 2023. URL: <https://www.apple.com/mac/garageband/>. 11
- [App23b] APPLE: Logic pro, 2023. Accessed: October 2023. URL: <https://www.apple.com/logic-pro/>. 11
- [App23c] APPLE: Truedepth camera, 2023. Accessed: August 2023. URL: <https://support.apple.com/en-in/102381>. 14
- [Ari18] ARISTIDOU A.: Hand tracking with physiological constraints. *The Visual Computer* 34, 2 (2018), 213–228. doi:10.1007/s00371-016-1327-8. 13
- [ASC19] ARISTIDOU A., SHAMIR A., CHRYSANTHOU Y.: Digital dance ethnography: Organizing large dance collections. *J. Comput. Cult. Herit.* 12, 4 (Nov. 2019). doi:10.1145/3344383. 5, 11
- [ASGA17] ANCILLAO A., SAVASTANO B., GALLI M., ALBERTINI G.: Three dimensional motion capture applied to violin playing: A study on feasibility and characterization of the motor strategy. *Computer Methods and Programs in Biomedicine* 149 (2017), 19–27. doi:10.1016/j.cmpb.2017.07.005. 15
- [Aud23] AUDACITY: Audio tool, 2023. Accessed: October 2023. URL: <https://www.audacityteam.org>. 11
- [AUT] AUTODESK: FBX file format. <https://www.autodesk.com/products/fbx/overview>. [Accessed: October 2023]. 5
- [Aut23] AUTODESK: Maya, 2023. Accessed: November 2023. URL: <https://www.autodesk.es/products/maya>. 14
- [AYA*23] ARISTIDOU A., YIANNAKIDIS A., ABERMAN K., COHEN-OR D., SHAMIR A., CHRYSANTHOU Y.: Rhythm is a dancer: Music-driven motion synthesis with global structure. *IEEE Transactions on Visualization and Computer Graphics* 29, 8 (aug 2023), 3519–3534. doi:10.1109/TVCG.2022.3163676. 17
- [AYS16] ALAJANKI A., YANG Y.-H., SOLEYMANI M.: Benchmarking music emotion recognition systems. *PloS one* (2016), 835–838. doi:10.1371/journal.pone.0173392. 7, 8
- [Bac] BACHDOODLE: Celebrating johann sebastian bach. Accessed: August 2023. URL: <https://www.google.com/doodles/celebrating-johann-sebastian-bach>. 7
- [Bar23] BARNETT J.: The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023), pp. 146–161. 20
- [BED09] BAY M., EHMANN A. F., DOWNIE J. S.: Evaluation of multiple-f0 estimation and tracking systems. In *Proceedings of the 10th International Society for Music Information Retrieval Conference* (2009), ISMIR, pp. 315–320. doi:10.5281/zenodo.1418241. 6, 7
- [BH89] BEJJANI F. J., HALPERN N.: Postural kinematics of trumpet playing. *Journal of Biomechanics* 22, 5 (1989), 439–446. doi:10.1016/0021-9290(89)90204-2. 16
- [BKD12] BENETOS E., KLAURI A., DIXON S.: Score-informed transcription for automatic piano tutoring. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (2012), IEEE, pp. 2153–2157. 6, 7
- [Bl23] BLENDER: Blender, 2023. Accessed: August 2023. URL: <https://www.blender.org/>. 14
- [BMB*11] BAAK A., MÜLLER M., BHARAJ G., SEIDEL H.-P., THEOBALT C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In *2011 International Conference on Computer Vision* (2011), pp. 1092–1099. doi:10.1109/ICCV.2011.6126356. 12
- [BNWY23] BAO W., NIU T., WANG N., YANG X.: Pose estimation and motion analysis of ski jumpers based on eca-hrnet. *Scientific Reports* 13, 1 (2023), 6132. URL: <https://doi.org/10.1038/s41598-023-32893-x>, doi:10.1038/s41598-023-32893-x. 15
- [BPDPM15] BLANCO-PIÑEIRO P., DÍAZ-PEREIRA M. P., MARTÍNEZ A.: Common postural defects among music students. *Journal of bodywork and movement therapies* 19, 3 (2015), 565–572. doi:10.1016/j.jbmt.2015.04.005. 16

- [Bra23] BRAREN N.: Recording techniques and mics for guitar, bass and piano, 2023. Accessed: August 2023. URL: <https://upayasound.com/record-guitar-bass-piano/>. 11
- [Bri] BRITANNICA E.: Dance. Accessed: November 2023. URL: <https://www.britannica.com/art/dance/Music>. 17
- [BST*14] BITTNER R. M., SALAMON J., TIERNEY M., MAUCH M., CANNAM C., BELLO J. P.: Medleydb: A multitrack dataset for annotation-intensive mir research. In *15th International Society for Music Information Retrieval Conference* (2014), vol. 14, pp. 155–160. 7
- [BvEBV*17] BAADJOU V., VAN EIJSDEN-BESSELENG M., VERBUNT J., DE BIE R., GEERS R., SMEETS R., SEELEN H.: Playing the clarinet: Influence of body posture on muscle activity and sound quality. *Medical problems of performing artists* 32, 3 (2017), 125–131. doi:10.21091/mppa.2017.3021. 17
- [BVGLH17] BAZZICA A., VAN GEMERT J., LIEM C. C., HANJALIC A.: Vision-based detection of acoustic timed events: a case study on clarinet note onsets. *arXiv preprint arXiv:1706.09556* (2017). 8, 9
- [BWT*19] BOGDANOV D., WON M., TOVSTOGAN P., PORTER A., SERRA X.: The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)* (Long Beach, CA, United States, 2019). URL: <http://hdl.handle.net/10230/42015>. 7, 8
- [BYV21] BOGAERS A., YUMAK Z., VOLK A.: Music-driven animation generation of expressive musical gestures. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (New York, NY, USA, 2021). ICMI '20 Companion, Association for Computing Machinery, p. 22–26. doi:10.1145/3395035.3425244. 18
- [Car] CARA V.: Vicon cara documentation. Accessed: November 2023. URL: <https://docs.vicon.com/display/CD/Cara+Documentation>. 14
- [CAR21] CAROUSEL+: Embodied online dancing and partying with digital characters, 2021. Accessed: October 2023. URL: <https://www.carouseldancing.org/>. 2
- [CBHC04] COKER C. A., BUGBEE F., HUBER A., COOK M.: Postural sway of percussionists: a preliminary investigation. *Medical Problems of Performing Artists* 19, 1 (2004), 34–38. doi:10.21091/mppa.2004.1006. 16
- [CDA13] CHAN C., DRISCOLL T., ACKERMANN B.: Can experienced observers detect postural changes in professional musicians after interventions? In *4th International Symposium on Performance Science 2013* (2013), European Association of Conservatoires (AEC), pp. 181–186. 16
- [CDT21] CHAKRABORTY S., DUTTA S., TIMONEY J.: The cyborg philharmonic: Synchronizing interactive musical performances between humans and machines. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–9. doi:10.1057/s41599-021-00751-8. 20
- [CECS18] COLYER S. L., EVANS M., COSKER D. P., SALO A. I. T.: A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Medicine - Open* 4, 1 (2018), 24. Article 24. doi:10.1186/s40798-018-0139-y. 15
- [CEM*22] CHOI B., EOM H., MOUSCADET B., CULLINGFORD S., MA K., GASSEL S., KIM S., MOFFAT A., MAIER M., REVELANT M., LETTERI J., SINGH K.: Anatomy: An animator-centric, anatomically inspired system for 3d facial modeling, animation and transfer. In *SIGGRAPH Asia 2022 Conference Papers* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. doi:10.1145/3550469.3555398. 14
- [CFZ*21] CHEN J., FAN C., ZHANG Z., LI G., ZHAO Z., DENG Z., DING Y.: A music-driven deep generative adversarial model for guzheng playing animation. *IEEE Transactions on Visualization and Computer Graphics* 29 (2021), 1400–1414. doi:10.1109/TVCG.2021.3115902. 1, 19
- [Cin21] CINEMATOGRAPHYDATABASE: Tracking a Drummer LIVE in Unreal Engine — youtube.com. <https://www.youtube.com/watch?v=2MhmfewlFuE>, 2021. [Accessed October 2023]. 1, 14
- [CJL*23] CHEN X., JIANG B., LIU W., HUANG Z., FU B., CHEN T., YU G.: Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), CVPR, pp. 18000–18010. 19
- [CKG*23] COPET J., KREUK F., GAT I., REMEZ T., KANT D., SYNNAEVE G., ADI Y., DÉFOSSÉZ A.: Simple and controllable music generation, 2023. [arXiv:2306.05284](https://arxiv.org/abs/2306.05284). 9
- [CLC*14] CLEMENTE M., LOURENÇO S., COIMBRA D., SILVA A., GABRIEL J., PINHO J.: Three-dimensional analysis of the cranio-cervico-mandibular complex during piano performance. *Medical problems of performing artists* 29, 3 (2014), 150–154. doi:10.21091/mppa.2014.3031. 17
- [CLS10] CANNAM C., LANDONE C., SANDLER M.: Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference* (Firenze, Italy, October 2010), pp. 1467–1468. 11
- [Col13] COLLINS M.: *Pro Tools 11: Music Production, Recording, Editing, and Mixing*. Taylor & Francis, 2013. doi:10.4324/9780203066416. 11
- [CPR14] CARTWRIGHT M., PARDO B., REISS J.: Mixploration: Rethinking the audio mixer interface. In *Proceedings of the 19th international conference on Intelligent User Interfaces* (2014), pp. 365–370. doi:10.1145/2557500.2557530. 7
- [CRO*92] CHUNG I.-S., RYU J., OHNISHI N., ROWEN B., HEADRICH J.: Wrist motion analysis in pianists. *Medical Problems of Performing Artists* 7, 1 (1992), 1–5. URL: <https://www.jstor.org/stable/45440445>. 16
- [CTL*21] CHEN K., TAN Z., LEI J., ZHANG S.-H., GUO Y.-C., ZHANG W., HU S.-M.: Choreomaster: Choreography-oriented music-driven dance synthesis. *ACM Trans. Graph.* 40, 4 (July 2021). doi:10.1145/3450626.3459932. 17
- [CVD*18] CATTARELLO P., VINELLI S., D'EMANUELE S., GAZZONI M., MERLETTI R.: Comparison of chairs based on hdsemg of back muscles, biomechanical and comfort indices, for violin and viola players: A short-term study. *Journal of Electromyography and Kinesiology* 42 (2018), 92–103. doi:10.1016/j.jelekin.2018.06.013. 17
- [CVG*08] CASEY M. A., VELTKAMP R., GOTO M., LEMAN M., RHODES C., SLANEY M.: Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE* 96, 4 (2008), 668–696. doi:10.1109/JPROC.2008.916370. 15
- [CWZ*21] CHEN K., WANG Y., ZHANG S.-H., XU S.-Z., ZHANG W., HU S.-M.: Mocap-solver: A neural solver for optical motion capture data. *ACM Trans. Graph.* 40, 4 (jul 2021). doi:10.1145/3450626.3459681. 11
- [Cye23] CYENS: Tone project, 2023. Accessed: October 2023. URL: <https://neomove.cyens.org.cy/research/tone-project/>. 1, 15
- [CYF*15] CHAN T.-S., YEH T.-C., FAN Z.-C., CHEN H.-W., SU L., YANG Y.-H., JANG R.: Vocal activity informed singing voice separation with the ikala dataset. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), IEEE, pp. 718–722. doi:10.1109/ICASSP.2015.7178063. 6, 7
- [DAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. *ACM Trans. Graph.* 27, 3 (aug 2008), 1–10. doi:10.1145/1360612.1360697. 12
- [DBVB17] DEFFERRARD M., BENZI K., VANDERGHEYNST P., BRESSON X.: FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)* (2017). [arXiv:1612.01840](https://arxiv.org/abs/1612.01840). 6, 7

- [DDF*17] DOU M., DAVIDSON P., FANELLO S. R., KHAMIS S., KOWDLE A., RHEMANN C., TANKOVICH V., IZADI S.: Motion2fusion: Real-time volumetric performance capture. *ACM Trans. Graph.* 36, 6 (nov 2017). doi:10.1145/3130800.3130801. 12
- [Dee23] DEEPMOTION: Deepmotion. <https://www.deepmotion.com/>, 2023. Accessed: August 2023. 12, 13
- [Den22] DENT S.: Sony steps into the metaverse with the 'mocopi' motion tracking system, 2022. URL: <https://www.engadget.com/sony-mocopi-movement-tracker-metaverse-avatars-131721036.html>. 12
- [DIHB08] DIANE I. HILLMANN R. M., BRADY C.: Metadata standards and applications. *The Serials Librarian* 54, 1-2 (2008), 7–21. doi:10.1080/03615260801973364. 5
- [DKP*23] DU Y., KIPS R., PUMAROLA A., STARKE S., THABET A., SANAKOYEU A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, jun 2023), IEEE Computer Society, pp. 481–490. doi:10.1109/CVPR52729.2023.00054. 12
- [DMM18] DONAHUE C., MAO H. H., MCAULEY J.: The nes music database: A multi-instrumental dataset with expressive performance attributes. In *19th International Society for Music Information Retrieval Conference (ISMIR)* (2018). 6, 7
- [DMSM21] DESMARAIS Y., MOTTET D., SLAGEN P., MONTESINOS P.: A review of 3d human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding* 212 (2021), 103275. doi:10.1016/j.cviu.2021.103275. 12
- [Dow03] DOWNIE J. S.: Music information retrieval. *Annual review of information science and technology* 37, 1 (2003), 295–340. doi:10.1002/aris.1440370108. 15
- [DPZ10] DUAN Z., PARDO B., ZHANG C.: Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 8 (2010), 2121–2133. doi:10.1109/TASL.2010.2042119. 6, 7
- [DZBKM22] DONG H.-W., ZHOU C., BERG-KIRKPATRICK T., MCAULEY J.: Deep performer: Score-to-audio music performance synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022), IEEE, pp. 951–955. 17
- [EBD10] EMIYA V., BADEAU R., DAVID B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 6 (2010), 1643–1654. doi:10.1109/TASL.2009.2038819. 6, 7
- [EF78] EKMAN P., FRIESEN W. V.: Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978). doi:10.1037/t27734-000. 13
- [EMNS20] EREMENKO V., MORSI A., NARANG J., SERRA X.: Performance assessment technologies for the support of musical instrument learning. *CSEDU 2020 The 12th International Conference on Computer Supported Education* (2020), 629–640. doi:10.5220/0009817006290640. 15
- [Epi] EPIC GAMES: Midi in unreal engine. Accessed: October 2023. URL: <https://docs.unrealengine.com/4.27/en-US/WorkingWithAudio/MIDI/>. 4
- [ERR*17] ENGEL J., RESNICK C., ROBERTS A., DIELEMAN S., NOROUZI M., ECK D., SIMONYAN K.: Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (2017), ICML'17, JMLR.org, p. 1068–1077. 6, 7
- [ES03] ELKOURA G., SINGH K.: Handrix: Animating the human hand. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Goslar, DEU, 2003), SCA '03, Eurographics Association, p. 110–119. 1, 18
- [Eve23] EVERCOAST: Studio. <https://evercoast.com>, 2023. [Accessed: October 2023]. 12
- [Fac23a] FACE L. L.: Live link face - recording facial animation from an ios device, 2023. Accessed: August 2023. URL: <https://docs.unrealengine.com/4.27/en-US/AnimatingObjects/SkeletalMeshAnimation/FacialRecordingiPhone/>. 14
- [Fac23b] FACEWARE: Faceware, 2023. Accessed: August 2023. URL: <https://facewaretech.com>. 14
- [FLW20] FERREIRA L. N., LELIS L. H. S., WHITEHEAD J.: Computer-generated music for tabletop role-playing games. In *Proceedings of the Sixteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (2020), AIIDE'20, AAAI Press. doi:10.48550/arXiv.2008.07009. 6, 7
- [Foc23] FOCUSRITE: Scarlett, 2023. Accessed: August 2023. URL: <https://focusrite.com/scarlett>. 10
- [For] FORTNITE: Fortnite. URL: <https://www.fortnite.com>. 3
- [FP13] FRITSCH J., PLUMBLEY M. D.: Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), IEEE, pp. 888–891. doi:10.1109/ICASSP.2013.6637776. 6, 7
- [Fut] FUTURELEARN: Motion Capture - Online Course — futurelearn.com. <https://www.futurelearn.com/courses/motion-capture-course>. [Accessed: October 2023]. 11
- [FW19] FERREIRA L. N., WHITEHEAD J.: Learning to generate music with sentiment. *Proceedings of the Conference of the International Society for Music Information Retrieval* (2019). doi:10.48550/arXiv.2103.06125. 7, 8
- [G*04] GOTO M., ET AL.: Development of the rwc music database. In *Proceedings of the 18th international congress on acoustics (ICA 2004)* (2004), vol. 1, Citeseer, pp. 553–556. 6
- [Gab99] GABRIELSSON A.: 14 - the performance of music. In *The Psychology of Music (Second Edition)*, Deutsch D., (Ed.), second edition ed., Cognition and Perception. Academic Press, San Diego, 1999, pp. 501–602. doi:10.1016/B978-01213564-4/50015-9. 15
- [Gab03] GABRIELSSON A.: Music performance research at the millennium. *Psychology of Music* 31, 3 (2003), 221–272. doi:10.1177/03057356030313002. 15
- [Gam23] GAMES E.: Metahuman animator, 2023. URL: <https://www.unrealengine.com/en-US/blog/delivering-high-quality-facial-animation-in-minutes-metahuman-animator-is-now-available>. 14
- [GBK*19] GINOSAR S., BAR A., KOHAVI G., CHAN C., OWENS A., MALIK J.: Learning individual styles of conversational gesture, 2019. arXiv:1906.04160. 17
- [GCZ*21] GUO R., CUI J., ZHAO W., LI S., HAO A.: Hand-by-hand mentor: An ar based training system for piano performance. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2021), pp. 436–437. doi:10.1109/VRW52623.2021.00100. 18
- [GDDP*08] GOEBL W., DIXON S., DE POLI G., FRIBERG A., BRESIN R., WIDMER G.: Sense in expressive music performance: Data acquisition, computational studies, and models. *Sound to sense-sense to sound: A state of the art in sound and music computing* (2008), 195–242. 15
- [GEF*17] GEMMEKE J. F., ELLIS D. P. W., FREEDMAN D., JANSEN A., LAWRENCE W., MOORE R. C., PLAKAL M., RITTER M.: Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), pp. 776–780. doi:10.1109/ICASSP.2017.7952261. 7
- [Ger] GERRY L. J.: Motion capture of pianists - RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion — uio.no. <https://www.uio.no/ritmo/english/news-and-events/blog/2019/gerry/>. [Accessed: October 2023]. 11

- [GFH*23] GHORBANI S., FERSTL Y., HOLDEN D., TROJE N. F., CARBONNEAU M.-A.: Zeroeggs: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum* 42 (2023), 206–216. doi:10.48550/arXiv.2209.07556. 17
- [GHNO02] GOTO M., HASHIGUCHI H., NISHIMURA T., OKA R.: Rwc music database: Popular, classical and jazz music databases. In *International Society for Music Information Retrieval Conference* (2002), vol. 2, pp. 287–288. 6, 7
- [GHNO03] GOTO M., HASHIGUCHI H., NISHIMURA T., OKA R.: RWC music database: Music genre database and musical instrument sound database. In *Proceedings of the 4th International Conference on Music Information Retrieval* (2003), ISMIR'03'. 6
- [GMK*19] GOPALAKRISHNAN A., MALI A., KIFER D., GILES L., ORORBIA A. G.: A neural temporal model for human motion prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, jun 2019), IEEE Computer Society, pp. 12108–12117. doi:10.1109/CVPR.2019.01239. 17
- [Goo23] GOOGLE: Piano scribe, 2023. Accessed: October 2023. URL: <https://piano-scribe.glitch.me>. 11
- [GPKT10] GANAPATHI V., PLAGEMANN C., KOLLER D., THRUN S.: Real time motion capture using a single time-of-flight camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), pp. 755–762. doi:10.1109/CVPR.2010.5540141. 12
- [GR06] GILLET O., RICHARD G.: Enst-drums: an extensive audio-visual database for drum signals processing. In *Proceedings of the 7th International Conference on Music Information Retrieval* (2006), ISMIR, pp. 156–159. 8, 9
- [GRE*19] GILICK J., ROBERTS A., ENGEL J., ECK D., BAMMAN D.: Learning to groove with inverse sequence transformations. In *International Conference on Machine Learning* (2019), PMLR, pp. 2269–2279. doi:10.48550/arXiv.1905.06118. 7
- [GSdA*09] GALL J., STOLL C., DE AGUIAR E., THEOBALT C., ROSENHAHN B., SEIDEL H.-P.: Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 1746–1753. doi:10.1109/CVPR.2009.5206755. 12
- [HAF*16] HUANG C.-H., ALLAIN B., FRANCO J.-S., NAVAB N., ILIC S., BOYER E.: Volumetric 3d tracking by detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 3862–3870. doi:10.1109/CVPR.2016.419. 12
- [Har23] HARRER S.: Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine* 90 (2023). 20
- [HBKD06] HERRERA-BOYER P., KLAURI A., DAVY M.: *Automatic Classification of Pitched Musical Instrument Sounds*. Springer US", Boston, MA, 2006, pp. 163–200. doi:10.1007/0-387-32845-9_6. 15
- [HCD*21] HUNG H.-T., CHING J., DOH S., KIM N., NAM J., YANG Y.-H.: EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In *Proc. Int. Society for Music Information Retrieval Conf.* (2021). doi:10.48550/arXiv.2108.01374. 7, 8
- [HCE*17] HERSHEY S., CHAUDHURI S., ELLIS D. P. W., GEMMEKE J. F., JANSEN A., MOORE C., PLAKAL M., PLATT D., SAUROUS R. A., SEYBOLD B., SLANEY M., WEISS R., WILSON K.: Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP. IEEE*, 2017. 8, 9
- [HCR*19] HUANG C.-Z. A., COUJMAN S., ROBERTS A., COURVILLE A., ECK D.: Counterpoint by convolution. *arXiv preprint arXiv:1903.07227* (2019). 7
- [HCW*17] HOPPER L., CHAN C., WIJSMAN S., ACKLAND T., VISENTIN P., ALDERSON J.: Torso and bowing arm three-dimensional joint kinematics of elite cellists: clinical and pedagogical implications for practice. *Medical Problems of Performing Artists* 32, 2 (2017), 85–93. doi:10.21091/mppa.2017.2015. 15
- [HHR*19] HUANG C.-Z. A., HAWTHORNE C., ROBERTS A., DINCULESCU M., WEXLER J., HONG L., HOWCROFT J.: The Bach Doodle: Approachable music composition with machine learning at scale. In *International Society for Music Information Retrieval (ISMIR)* (2019). 7
- [HKS12] HARGREAVES S., KLAURI A., SANDLER M.: Structural segmentation of multitrack audio. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 10 (2012), 2637–2647. doi:10.1109/TASL.2012.2209419. 6, 7
- [Hol18] HOLDEN D.: Robust solving of optical motion capture data by denoising. *ACM Trans. Graph.* 37, 4 (jul 2018). doi:10.1145/3197517.3201302. 11
- [HSR*18] HAWTHORNE C., STASYUK A., ROBERTS A., SIMON I., HUANG C.-Z. A., DIELEMAN S., ELSER E., ENGEL J., ECK D.: Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv preprint arXiv:1810.12247* (2018). doi:10.48550/arXiv.1810.12247. 6, 7
- [HTHM22] HIRATA A., TANAKA K., HAMANAKA M., MORISHIMA S.: Audio-driven violin performance animation with clear fingering and bowing. In *ACM SIGGRAPH 2022 Posters* (New York, NY, USA, 2022), SIGGRAPH '22, Association for Computing Machinery. doi:10.1145/3532719.3543240. 1, 18
- [HTSM21] HIRATA A., TANAKA K., SHIMAMURA R., MORISHIMA S.: Bowing-net: Motion generation for string instruments based on bowing information. In *ACM SIGGRAPH 2021 Posters* (New York, NY, USA, 2021), SIGGRAPH '21, Association for Computing Machinery. doi:10.1145/3450618.3469170. 18
- [HW23] HYODO H., WADA T.: 3d motion analysis of kick start motion-effects of upper limb movements on the body when leaving the platform. *Electronics and Communications in Japan* (2023), e12400. doi:10.1002/ecj.12400. 15
- [HXZ*19] HABERMANN M., XU W., ZOLLHÖFER M., PONS-MOLL G., THEOBALT C.: LiveCap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)* 38, 2 (mar 2019). doi:10.1145/3311970. 12
- [HYXC15] HONG C., YU J., XIE Y., CHEN X.: Multi-view deep learning for image-based pose recovery. In *2015 IEEE 16th International Conference on Communication Technology (ICCT)* (2015), pp. 897–902. doi:10.1109/ICCT.2015.7399969. 12
- [HZZ*21] HU W., ZHANG C., ZHAN F., ZHANG L., WONG T.-T.: Conditional directed graph convolution for 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia* (New York, NY, USA, 2021), MM '21, Association for Computing Machinery, p. 602–611. doi:10.1145/3474085.3475219. 12
- [IBP*18] ISLAN M., BLAYA F., PEDRO P. S., D'AMATO R., URQUIJO E. L., JUANES J. A.: Analysis and fem simulation methodology of dynamic behavior of human rotator cuff in repetitive routines: Musician case study. *Journal of medical systems* 42 (2018), 1–10. doi:10.1007/s10916-018-0908-7. 16
- [JLT*22] JU Z., LU P., TAN X., WANG R., ZHANG C., WU S., ZHANG K., LI X., QIN T., LIU T.-Y.: Telemelody: Lyric-to-melody generation with a template-based two-stage method, 2022. *arXiv:2109.09617*. 10
- [JYL23] JI S., YANG X., LUO J.: A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Comput. Surv.* 56, 1 (aug 2023). doi:10.1145/3597493. 10
- [KCMT00] KIM J., CORDIER F., MAGNENAT-THALMANN N.: Neural network-based violinist's hand animation. In *Proceedings Computer Graphics International 2000* (2000), pp. 37–41. doi:10.1109/CGI.2000.852318. 18
- [KMO*09] KUGIMOTO N., MIYAZONO R., OMORI K., FUJIMURA T., FURUYA S., KATAYOSE H., MIWA H., NAGATA N.: Cg animation for

- piano performance. In *SIGGRAPH '09: Posters* (New York, NY, USA, 2009), SIGGRAPH '09, Association for Computing Machinery. doi:10.1145/1599301.1599304. 13, 18
- [KPJ*23] KIM K., PARK M., JOUNG H., CHAE Y., HONG Y., GO S., LEE K.: Show me the instruments: Musical instrument retrieval from mixture audio. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), pp. 1–5. doi:10.1109/ICASSP49357.2023.10097162. 6, 7
- [KRI2] KAMINSKAS M., RICCI F.: Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review* 6, 2 (2012), 89–119. doi:10.1016/j.cosrev.2012.04.002. 15
- [KS20] KAO H.-K., SU L.: Temporally guided music-to-body-movement generation. In *Proceedings of the 28th ACM International Conference on Multimedia* (New York, NY, USA, 2020), MM '20, Association for Computing Machinery, p. 147–155. doi:10.1145/3394171.3413848. 18
- [LAPG19] LERCH A., ARTHUR C., PATI A., GURURANI S.: Music performance analysis: A survey. *arXiv preprint arXiv:1907.00178* (2019). doi:10.48550/arXiv.1907.00178. 15
- [LAPG21] LERCH A., ARTHUR C., PATI A., GURURANI S.: An interdisciplinary review of music performance analysis. *arXiv preprint arXiv:2104.09018* (2021). doi:10.5334/tismir.53. 15
- [LAR*14] LEWIS J. P., ANJYO K., RHEE T., ZHANG M., PIGHIN F. H., DENG Z.: Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)* 1, 8 (2014), 2. doi:10.2312/egst.20141042. 14
- [LAZ*22] LI P., ABERMAN K., ZHANG Z., HANOCKA R., SORKINE-HORNUNG O.: Ganimot: Neural motion synthesis from a single sequence. *ACM Trans. Graph.* 41, 4 (jul 2022). doi:10.1145/3528223.3530157. 19
- [LC10] LOU H., CHAI J.: Example-based human motion denoising. *IEEE Transactions on Visualization and Computer Graphics* 16, 5 (Sept. 2010), 870–879. 11
- [LDSR*20] LONGO L., DI STADIO A., RALLI M., MARINUCCI I., RUOPPOLO G., DIPIETRO L., DE VINCENZIIS M., GRECO A.: Voice parameter changes in professional musician-singers singing with and without an instrument: The effect of body posture. *Folia Phoniatrica et Logopaedica* 72, 4 (2020), 309–315. doi:10.1159/000501202. 16
- [Leg20] LEGEND J.: John legend live - a night for “bigger love” presented by wave, 2020. URL: <https://www.youtube.com/watch?v=eGy6419Yuuw>. 3
- [Ler12] LERCH A.: Music performance analysis. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics 1* (2012), 169–179. doi:10.1002/9781118393550.ch10. 15
- [LGS*13] LIU Y., GALL J., STOLL C., DAI Q., SEIDEL H.-P., THEOBALT C.: Markerless motion capture of multiple characters using multiview image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 11 (2013), 2720–2735. doi:10.1109/TPAMI.2013.47. 12
- [LGS15] LIEM C. C., GÓMEZ E., SCHEDL M.: Phenix: Innovating the classical music experience. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (2015), IEEE, pp. 1–4. doi:10.1109/ICMEW.2015.7169835. 2, 15
- [Li22] LI W.: Musical instrument performance in augmented virtuality. In *Proceedings of the 6th International Conference on Digital Signal Processing* (New York, NY, USA, 2022), ICDSP '22, Association for Computing Machinery, p. 91–97. doi:10.1145/3529570.3529586. 19
- [LKT*20] LIN Y.-J., KAO H.-K., TSENG Y.-C., TSAI M., SU L.: A human-computer duet system for music performance. In *Proceedings of the 28th ACM International Conference on Multimedia* (10 2020), ACM, pp. 772–780. doi:10.1145/3394171.3413921. 18
- [LL21] LI K., LI W.: MusicTXT: A Text-based Interface for Music Notation. In *Proceedings of the 11th Workshop on Ubiquitous Music (UbiMus 2021)* (Matosinhos, Portugal, Sept. 2021), Proceedings of the 11th Workshop on Ubiquitous Music (UbiMus 2021), g-ubimus, pp. 62–71. URL: <https://hal.science/hal-03398727>. 11
- [LLD*19] LI B., LIU X., DINESH K., DUAN Z., SHARMA G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia* 21, 2 (2019), 522–535. doi:10.1109/TMM.2018.2856090. 5, 8, 9, 14
- [LLH*20] LIU J.-W., LIN H.-Y., HUANG Y.-F., KAO H.-K., SU L.: Body movement generation for expressive violin performance applying neural networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), pp. 3787–3791. doi:10.1109/ICASSP40776.2020.9054463. 18
- [LMD18] LI B., MAEZAWA A., DUAN Z.: Skeleton plays piano: On-line generation of pianist body movements from midi performance. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR* (2018), pp. 218–224. doi:10.5281/zenodo.1492387. 18
- [LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: A skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (oct 2015). doi:10.1145/2816795.2818013. 4
- [LSG*11] LIU Y., STOLL C., GALL J., SEIDEL H.-P., THEOBALT C.: Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR 2011* (2011), pp. 1249–1256. doi:10.1109/CVPR.2011.5995424. 12
- [LTL*23] LV A., TAN X., LU P., YE W., ZHANG S., BIAN J., YAN R.: Getmusic: Generating any music tracks with a unified representation and diffusion framework, 2023. *arXiv:2305.10841*. 9
- [LTN*19] LUGARESI C., TANG J., NASH H., MCCLANAHAN C., UBOWEJA E., HAYS M., ZHANG F., CHANG C.-L., YONG M. G., LEE J., ET AL.: Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019). doi:10.48550/arXiv.1906.08172. 12, 13
- [LTY*22] LU P., TAN X., YU B., QIN T., ZHAO S., LIU T.-Y.: Meloform: Generating melody with musical form based on expert systems and neural networks, 2022. *arXiv:2208.14345*. 10
- [LXK*23] LU P., XU X., KANG C., YU B., XING C., TAN X., BIAN J.: Muscoco: Generating symbolic music from text, 2023. *arXiv:2306.00110*. 9
- [LYL*19] LEE H.-Y., YANG X., LIU M.-Y., WANG T.-C., LU Y.-D., YANG M.-H., KAUTZ J.: Dancing to music. In *Advances in Neural Information Processing Systems* (2019), Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., (Eds.), vol. 32, Curran Associates, Inc. 17
- [LYRK21] LI R., YANG S., ROSS D. A., KANAZAWA A.: Learn to dance with AIST++: Music conditioned 3d dance generation. In *Proceedings of IEEE International Conference on Computer Vision* (2021), ICCV. 17
- [Mat23a] MATLA S.: Music production software: The definitive guide, 2023. Accessed: November 2023. URL: <https://www.edmprod.com/music-production-software/>. 11
- [Mat23b] MATTHIS J.: FREEMOCAP. <https://freemocap.org/about-us.html>, 2023. Accessed: August 2023. 12, 13
- [MCOB*16] MIRON M., CARABIAS-ORTI J. J., BOSCH J. J., GÓMEZ E., JANER J., ET AL.: Score-informed source separation for multichannel orchestral recordings. *Journal of Electrical and Computer Engineering* 2016 (2016). doi:10.1155/2016/8363507. 6, 7
- [Met22] META: Catch foo fighters in vr: Horizon venues concert to air february 13 after the big game, 2022. URL: <https://www.meta.com/blog/quest/catch-foo-fighters-in-vr-horizon-venues-concert-to-air-february-13-after-the-big-game/>. 3

- [Met23a] META: Horizon venues, 2023. URL: <https://www.meta.com/experiences/3002729676463989/>. 3
- [Met23b] META M.: Manus Metagloves - Home. <https://www.manus-meta.com>, 2023. Accessed: October 2023. 13
- [MHK06] MOESLUND T. B., HILTON A., KRÜGER V.: A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* 104, 2 (nov 2006), 90–126. doi:10.1016/j.cviu.2006.08.002. 11
- [Mic23] MICROSOFT KINECT: KINECT Mocap. <https://learn.microsoft.com/de-de/shows/2p-start/diy-motion-capture-kinect-2-unity-cinema-mocap>, 2023. Accessed: August 2023. 12
- [MIDa] MIDI ASSOCIATION: About midi-part 3:midi messages. Accessed: October 2023. URL: <https://www.midi.org/midi-articles/about-midi-part-3-midi-messages>. 4
- [MIDb] MIDI ASSOCIATION: Opensoundcontrol. Accessed: October 2023. URL: <https://ccrma.stanford.edu/groups/osc/index.html>. 4
- [MIR23] MIREX: MIREX - Home. https://www.music-ir.org/mirex/wiki/MIREX_HOME, 2023. Accessed: August 2023. 6
- [MLMG11] MAYOR O., LLOP J., MAESTRE GÓMEZ E.: Repovizz: a multi-modal on-line database and browsing tool for music performance research. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*; Miami; 2011 Oct. 24–28. (2011), International Society for Music Information Retrieval (ISMIR). 5, 9
- [MM*01] MEREDITH M., MADDOCK S., ET AL.: Motion capture file formats explained. *Department of Computer Science, University of Sheffield 211* (2001), 241–244. 4
- [Mov23a] MOVELLA: The story behind the virtual concert experience of twenty one pilots in the roblox metaverse using xsens motion capture technology., 2023. URL: <https://www.movella.com/resources/cases/the-story-behind-the-virtual-concert-experience-of-twenty-one-pilots-in-the-roblox-metaverse-using-xsens-motion-capture-technology>. 3
- [Mov23b] MOVELLA: Xsens - Metagloves. <https://www.movella.com/products/motion-capture/xsens-metagloves-by-manus>, 2023. Accessed: October 2023. 13
- [Mov23c] MOVELLA: XSENS - Movella. <https://www.movella.com/>, 2023. Accessed: August 2023. 12
- [MRL*15] MCFEE B., RAFFEL C., LIANG D., ELLIS D. P., MCVICAR M., BATTENBERG E., NIETO O.: librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (2015), vol. 8, pp. 18–25. 11
- [MRPM14] MARCHINI M., RAMIREZ R., PAPIOTIS P., MAESTRE E.: 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. doi:10.1080/09298215.2014.922999. 8, 9
- [MSH20] MONTESINOS J. F., SLIZOVSKAIA O., HARO G.: Solos: A dataset for audio-visual music analysis. In *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)* (2020), pp. 1–6. doi:10.1109/MMSP48831.2020.9287124. 8, 9
- [MSM*20] MEHTA D., SOTNYCHENKO O., MUELLER F., XU W., ELGHARIB M., FUA P., SEIDEL H.-P., RHODIN H., PONS-MOLL G., THEOBALT C.: Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Trans. Graph.* 39, 4 (aug 2020). doi:10.1145/3386569.3392410. 12
- [Mul23] MULTIMEDIA I.: The ultra-compact professional audio/midi interface for all your gear, 2023. Accessed: August 2023. URL: <https://www.ikmultimedia.com/products/irigproio/>. 10
- [MWSLR19] MANILOW E., WICHERN G., SEETHARAMAN P., LE ROUX J.: Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2019), IEEE. 7, 8
- [Neu23] NEURON P.: Perception Neuron Studio Gloves. <https://neuronmocap.com/products/perception-neuron-studio-gloves>, 2023. Accessed: October 2023. 13
- [OERF*16] ORTS-ESCOLANO S., RHEMANN C., FANELLO S., CHANG W., KOWDLE A., DEGTAREV Y., KIM D., DAVIDSON P. L., KHAMIS S., DOU M., TANKOVICH V., LOOP C., CAI Q., CHOU P. A., MENICKEN S., VALENTIN J., PRADEEP V., WANG S., KANG S. B., KOHLI P., LUTCHYN Y., KESKIN C., IZADI S.: Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (New York, NY, USA, 2016), UIST '16, Association for Computing Machinery, p. 741–754. doi:10.1145/2984511.2984517. 12
- [OMB*18] OHLENDORF D., MAURER C., BOLENDER E., KOCIS V., SONG M., GRONEBERG D. A.: Influence of ergonomic layout of musician chairs on posture and seat pressure in musicians of different playing levels. *PLoS one* 13, 12 (2018), e0208758. doi:10.1371/journal.pone.0208758. 16
- [OMC*18] OHLENDORF D., MARX J., CLASEN K., WANKE E. M., KOPP S., GRONEBERG D. A., UIBEL S.: Comparison between the musician-specific seating position of high string bow players and their habitual seating position—a video raster stereographic study of the dorsal upper body posture. *Journal of Occupational Medicine and Toxicology* 13, 1 (2018), 1–8. doi:10.1186/s12995-018-0217-6. 16
- [Opt] OPTITRACK: Unreal engine: Optitrack incamera vfx. Accessed: November 2023. URL: <https://docs.optitrack.com/v/v2.3/virtual-production/unreal-engine-optitrack-incamera-vfx>. 15
- [Opt23a] OPTITRACK: OptiTrack - Hand Kit. <https://optitrack.com/accessories/markers/>, 2023. Accessed: October 2023. 13
- [Opt23b] OPTITRACK: OptiTrack - Home. <https://www.optitrack.com/>, 2023. Accessed: August 2023. 11
- [Ori06] ORIO N.: Music retrieval: A tutorial and review. *Foundations and Trends® in Information Retrieval* 1, 1 (2006), 1–90. doi:10.1561/1500000002. 15
- [PBDL23] PERRY D., BIVINS T., DEHAAN B., LI W.: Procedural rhythm game generation in virtual reality. In *Proceedings of the 2023 6th International Conference on Image and Graphics Processing* (New York, NY, USA, 2023), ICIGP '23, Association for Computing Machinery, p. 218–222. doi:10.1145/3582649.3582664. 15
- [PCAW16] PEREZ-CARRILLO A., ARCOS J.-L., WANDERLEY M.: Estimation of guitar fingering and plucking controls based on multimodal analysis of motion, audio and musical score. In *11th International Symposium on Music, Mind, and Embodiment* (2016), Springer, pp. 71–87. doi:10.1007/978-3-319-46282-0_5. 1, 8, 9
- [PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 10967–10977. doi:10.1109/CVPR.2019.01123. 12
- [PE07] POLINER G. E., ELLIS D. P. W.: A discriminative model for polyphonic piano transcription. *EURASIP J. Adv. Signal Process* 2007, 1 (jan 2007), 154. doi:10.1155/2007/48317. 6, 7
- [PGA18] PAVLLO D., GRANGIER D., AULI M.: QuaterNet: A Quaternion-based Recurrent Model for Human Motion. *arXiv e-prints* (May 2018), arXiv:1805.06485. arXiv:1805.06485, doi:10.48550/arXiv.1805.06485. 4
- [Pha23a] PHASESPACE: PhaseSpace - Gloves. <https://www.phasespace.com/gloves.html>, 2023. Accessed: October 2023. 13
- [Pha23b] PHASESPACE: PhaseSpace - Home. www.phasespace.com, 2023. Accessed: August 2023. 11

- [PHG*18] PIATEK S., HARTMANN J., GÜNTHER P., ADOLF D., SEIDEL E. J.: Influence of different instrument carrying systems on the kinematics of the spine of saxophonists. *Medical Problems of Performing Artists* 33, 4 (2018), 251–257. doi:10.21091/mppa.2018.4037. 16
- [PLHW15] PENG S.-J., HE G.-F., LIU X., WANG H.-Z.: Hierarchical block-based incomplete human mocap data recovery using adaptive non-negative matrix factorization. *Computer & Graphics* 49, C (June 2015), 10–23. 11
- [PKH*12] PARK K.-N., KWON O.-Y., HA S.-M., KIM S.-J., CHOI H.-J., WEON J.-H.: Comparison of electromyographic activity and range of neck motion in violin students with and without neck pain during playing. *Medical problems of performing artists* 27, 4 (2012), 188–192. doi:10.21091/mppa.2012.4035. 16
- [Pla23] PLASK: Plask - home. <https://plask.ai>, 2023. Accessed: August 2023. 12
- [PMPCM14] PAPIOTIS P., MARCHINI M., PEREZ-CARRILLO A., MAESTRE E.: Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data. *Frontiers in Psychology* 5 (2014). doi:10.3389/fpsyg.2014.00963. 8, 9
- [PMTS*15] PONS-MOLL G., TAYLOR J., SHOTTON J., HERTZMANN A., FITZGIBBON A.: Metric regression forests for correspondence estimation. *International Journal of Computer Vision* 113 (2015), 163–175. doi:https://doi.org/10.1007/s11263-015-0818-9. 12
- [PPHB18] PAVLLO D., PORSSUT T., HERBELIN B., BOULIC R.: Real-time marker-based finger tracking with neural networks. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Los Alamitos, CA, USA, mar 2018), IEEE Computer Society, pp. 651–652. doi:10.1109/VR.2018.8446173. 13
- [PPL08] PÄTYNEN J., PULKKI V., LOKKI T.: Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica* 94, 6 (2008), 856–865. doi:10.3813/AAA.918104. 6, 7
- [PRE23] PREMIERE: Performing arts in a new era, 2023. Accessed: October 2023. URL: <https://premiere-project.eu/>. 2
- [PYA*23] PONTON J. L., YUN H., ARISTIDOU A., ANDUJAR C., PELECHANO N.: SparsePoser: Real-time full-body motion reconstruction from sparse data. *ACM Trans. Graph.* 43, 1 (oct 2023). doi:10.1145/3625264. 12
- [Qua23] QUALISYS: Qualisys - Home. <https://www.qualisys.com/>, 2023. Accessed: August 2023. 11
- [Raf16] RAFFEL C.: *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, Columbia University, 2016. 6, 8, 9
- [RCBF07] RABUFFETTI M., CONVERTI R. M., BOCCARDI S., FERRARIN M.: Tuning of the violin–performer interface: an experimental study about the effects of shoulder rest variations on playing kinematics. *Medical Problems of Performing Artists* 22, 2 (2007), 58–66. doi:10.21091/mppa.2007.2013. 15
- [Rea23] REALLUSION: Faceware facial mocap, 2023. Accessed: August 2023. URL: <https://mocap.reallusion.com/iClone-faceware-mocap/>. 14
- [RHT*20] REN Y., HE J., TAN X., QIN T., ZHAO Z., LIU T.-Y.: Popmag: Pop music accompaniment generation, 2020. *arXiv:2008.07703*. 9
- [RKES21] RASTGOO R., KIANI K., ESCALERA S., SABOKROU M.: Sign language production: A review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2021), pp. 3446–3456. doi:10.1109/CVPRW53098.2021.00384. 17
- [RLS*17] RAFI Z., LIUTKUS A., STÖTER F.-R., MIMILAKIS S. I., BITTNER R.: The MUSDB18 corpus for music separation, Dec. 2017. doi:10.5281/zenodo.1117372. 7, 8
- [Rob23] ROBLOX: Roblox, 2023. URL: <https://www.roblox.com>. 3
- [Rok23a] ROKOKO: Rokoko - face capture, 2023. Accessed: August 2023. URL: <https://www.rokoko.com/products/face-capture>. 14
- [Rok23b] ROKOKO: Rokoko - home, 2023. Accessed: August 2023. URL: www.rokoko.com. 12
- [Rok23c] ROKOKO: Rokoko - smartlgoves, 2023. Accessed: October 2023. URL: <https://www.rokoko.com/products/smartlgloves>. 13
- [Rus80] RUSSELL J.: A circumplex model of affect. *Journal of Personality and Social Psychology* 39 (12 1980), 1161–1178. doi:10.1037/h0077714. 8
- [SAA*20] SHI M., ABERMAN K., ARISTIDOU A., KOMURA T., LISCHINSKI D., COHEN-OR D., CHEN B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Trans. Graph.* 40, 1 (sep 2020). doi:10.1145/3407659. 12
- [Sar16] SARRAZIN N.: *Music and the Child*. Open SUNY Textbooks, 2016. 4
- [SCTO17] SARASÚA A., CARAMIAUX B., TANAKA A., ORTIZ M.: Datasets for the analysis of expressive musical gestures. In *Proceedings of the 4th International Conference on Movement Computing* (New York, NY, USA, 2017), MOCO '17, Association for Computing Machinery. doi:10.1145/3077981.3078032. 8, 9
- [SDB*12] SHEN W., DENG K., BAI X., LEYVAND T., GUO B., TU Z.: Exemplar-based human action pose correction and tagging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2012), CVPR '12, pp. 1784–1791. 11
- [SDSKS18] SHLIZERMAN E., DERY L., SCHOEN H., KEMELMACHER-SHLIZERMAN I.: Audio to body dynamics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7574–7583. doi:10.1109/CVPR.2018.00790. 1, 18
- [SE00] SEKIGUCHI H., EIHO S.: Generating the human piano performance in virtual space. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000* (2000), vol. 4, pp. 477–481 vol.4. doi:10.1109/ICPR.2000.902961. 1, 18
- [SEHS21] SIPHOCLY N. N. J., EL-HORBATY E.-S. M., SALEM A.-B. M.: Top 10 artificial intelligence algorithms in computer music composition. *International Journal of Computing and Digital Systems* 10 (2021), 373–394. URL: <https://api.semanticscholar.org/CorpusID:233366278>. 10
- [SFH*22] SHRESTHA S., FERMÜLLER C., HUANG T., WIN P. T., ZUKERMAN A., PARAMESHWARA C. M., ALOIMONOS Y.: Aimusicguru: Music assisted human pose correction. *arXiv preprint arXiv:2203.12829* (2022). doi:10.48550/arXiv.2203.12829. 8, 9, 18
- [SG12] SALAMON J., GOMEZ E.: Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 6 (2012), 1759–1770. doi:10.1109/TASL.2012.2188515. 15
- [SGLH*20] SLATER M., GONZALEZ-LIENCRE S., HAGGARD P., VINKERS C., GREGORY-CLARKE R., JELLEY S., WATSON Z., BREEN G., SCHWARZ R., STEPTOE W., ET AL.: The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality* 1 (2020), 1. 21
- [SGU*14] SCHEDL M., GÓMEZ E., URBANO J., ET AL.: *Music information retrieval: Recent developments and applications*, vol. 8. Now Publishers, Inc., 2014. doi:10.1561/15000000042. 15
- [SHA23a] SHARESPACE: 2023. Accessed: October 2023. URL: <https://sharespace.eu/>. 2
- [Sha23b] SHAZAM: Shazam, 2023. Accessed: October 2023. URL: <https://www.shazam.com/>. 15
- [Sic14] SICILIA M.-A.: *Handbook of Metadata, Semantics and Ontologies*. WORLD SCIENTIFIC, 2014. doi:10.1142/7077. 5
- [SKJS23] SCHNEIDER F., KAMAL O., JIN Z., SCHÖLKOPF B.: Moûsai: Text-to-music generation with long-context latent diffusion, 2023. *arXiv:2301.11757*. 9

- [SL] (SPI) S. P. I., LTD L.: Alembic — alembic.io. <https://www.alembic.io/>. [Accessed: October 2023]. 5
- [SLS*06] SAKAI N., LIU M. C., SU F.-C., BISHOP A. T., AN K.-N.: Hand span and digital motion on the keyboard: concerns of overuse syndrome in musicians. *The Journal of hand surgery* 31, 5 (2006), 830–835. doi:10.1016/j.jhsa.2006.02.009. 15
- [SMB*13] SERRA X., MAGAS M., BENETOS E., CHUDY M., DIXON S., FLEXER A., GOMEZ E., GOUYON F., HERRERA P., JORDA S., PAYTUVI O., PEETERS G., SCHLÜTER J., VINET H., WIDMER G.: *Roadmap for Music Information ReSearch*. MIREs Consortium, London, UK, 2013. URL: <http://mires.eecs.qmul.ac.uk/about.html>. 5
- [SNA19] SIMONETTA F., NTALAMPIRAS S., AVANZINI F.: Multimodal music information processing and retrieval: Survey and future challenges. In *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)* (2019), pp. 10–18. doi:10.1109/MMRP.2019.00012. 15
- [Spo23] SPOTIFY: Basic pitch, 2023. Accessed: October 2023. URL: <https://basicpitch.spotify.com>. 11
- [SS21] SHIRAI T., SAKO S.: 3d skeleton motion generation of double bass from musical score. In *15th International Symposium on Computer Music Multidisciplinary Research* (2021), CMMR. 18
- [SSK*13] SHOTTON J., SHARP T., KIPMAN A., FITZGIBBON A., FINOCCHIO M., BLAKE A., COOK M., MOORE R.: Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (jan 2013), 116–124. doi:10.1145/2398356.2398381. 12
- [SST*20] SHENG Z., SONG K., TAN X., REN Y., YE W., ZHANG S., QIN T.: Songmass: Automatic song writing with pre-training and alignment constraint, 2020. arXiv:2012.05168. 10
- [Ste23] STEINBERG: Cubase, 2023. Accessed: October 2023. URL: <https://www.steinberg.net/cubase/>. 11
- [Stu23] STUDIO F.: Fl studio, 2023. Accessed: October 2023. URL: <https://www.image-line.com>. 11
- [Sui23a] SUIT T.: Teslaglove dev kit, 2023. URL: <https://teslasuit.io/products/teslaglove/>. 13
- [Sui23b] SUIT T.: Teslasuit, 2023. URL: <https://teslasuit.io>. 12
- [SV03] SHAN G., VISENTIN P.: A quantitative three-dimensional analysis of arm kinematics in violin performance. *Medical problems of performing artists* 18, 1 (2003), 3–10. doi:10.21091/mppa.2003.1002. 15
- [SWC*21] SUN G., WONG Y., CHENG Z., KANKANHALLI M. S., GENG W., LI X.: Deepdance: Music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia* 23 (2021), 497–509. doi:10.1109/TMM.2020.2981989. 17
- [SWEN14] SPAHN C., WASMER C., EICKHOFF F., NUSSECK M.: Comparing violinists' body movements while standing, sitting, and in sitting orientations to the right or left of a music stand. *Medical problems of performing artists* 29, 2 (June 2014), 86–93. doi:10.21091/mppa.2014.2019. 15
- [SY16] SU L., YANG Y.-H.: Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *11th International Symposium on Music, Mind, and Embodiment* (2016), Springer, pp. 309–321. doi:10.1007/978-3-319-46282-0_20. 6, 7
- [Sys23] SYSTEMS V. M.: Vicon - Home. www.vicon.com, 2023. Accessed: August 2023. 11, 13
- [TC02] TZANETAKIS G., COOK P.: Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10, 5 (2002), 293–302. doi:10.1109/TSA.2002.800560. 15
- [THK17] THICKSTUN J., HARCHAOUI Z., KAKADE S. M.: Learning features of music from scratch. In *International Conference on Learning Representations (ICLR)* (2017). 6, 7
- [TRG*23] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., BERMANO A. H.: Human motion diffusion model. In *The Eleventh International Conference on Learning Representations* (2023), ICLR. 19
- [TSSF12] TAYLOR J., SHOTTON J., SHARP T., FITZGIBBON A.: The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 103–110. doi:10.1109/CVPR.2012.6247664. 12
- [TWV05] TYPKE R., WIERING F., VELTKAMP R. C.: A survey of music information retrieval systems. In *Proceedings of the 6th International Conference on Music Information Retrieval* (2005), pp. 153–160. doi:10.5281/zenodo.1417383. 15
- [Ult23] ULTRALEAP: Leap motion controller 2, 2023. Accessed: October 2023. URL: <https://leap2.ultraLeap.com/leap-motion-controller-2/>. 13
- [VBMP08] VLASIC D., BARAN I., MATUSIK W., POPOVIĆ J.: Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.* 27, 3 (aug 2008), 1–9. doi:10.1145/1360612.1360696. 12
- [vdKMR18] VAN DER KRUK E., M. REIJNE M.: Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European Journal of Sport Science* 18, 6 (2018), 806–819. doi:10.1080/17461391.2018.1463397. 11
- [VGRGAG20] VALENZUELA-GÓMEZ S.-A., REY-GALINDO J.-A., ACEVES-GONZALEZ C.: Analyzing working conditions for classical guitarists: Design guidelines for new supports and guitar positioning. *Work* 65, 4 (2020), 891–901. doi:10.3233/WOR-203140. 16
- [VIN08] VINEYS M.: Mtg mass database. <http://www.mtg.upf.edu/static/mass/resources> (2008). 7
- [VKV*17] VOLPE G., KOLYKHALOVA K., VOLTA E., GHISIO S., WADDELL G., ALBORNO P., PIANA S., CANEPA C., RAMIREZ-MELENDZ R.: A multimodal corpus for technology-enhanced learning of violin playing. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* (New York, NY, USA, 2017), CHIItaly '17, Association for Computing Machinery. doi:10.1145/3125571.3125588. 8, 9
- [VPGMRGM23] VILCHIS C., PEREZ-GUERRERO C., MENDEZ-RUIZ M., GONZALEZ-MENDOZA M.: A survey on the pipeline evolution of facial capture and tracking for digital humans. *Multimedia Syst.* 29, 4 (apr 2023), 1917–1940. doi:10.1007/s00530-023-01081-2. 14
- [W*03] WANG A., ET AL.: An industrial strength audio search algorithm. In *Proceedings of the 4th International Conference on Music Information Retrieval* (2003), vol. 2003, Washington, DC, pp. 7–13. 15
- [Wav] WAVE: Wave. URL: <https://wavexr.com/past-waves/>. 3
- [WCJ*20] WANG Z., CHEN K., JIANG J., ZHANG Y., XU M., DAI S., BIN G., XIA G.: Pop909: A pop-song dataset for music arrangement generation. In *Proceedings of 21st International Conference on Music Information Retrieval, ISMIR* (2020). 6, 7
- [WF02] WELCH G., FOXLIN E.: Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Comput. Graph. Appl.* 22, 6 (nov 2002), 24–38. doi:10.1109/MCG.2002.1046626. 11
- [WF15] WINGES S. A., FURUYA S.: Distinct digit kinematics by professional and amateur pianists. *Neuroscience* 284 (2015), 643–652. doi:10.1016/j.neuroscience.2014.10.041. 15
- [WLL*20] WANG Y., LIANG W., LI W., LI D., YU L.-F.: Scene-aware background music synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia* (New York, NY, USA, 2020), MM '20, Association for Computing Machinery, p. 1162–1170. doi:10.1145/3394171.3413894. 10
- [WMB*19] WOLF E., MÖLLER D., BALLEMBERGER N., MORISSE K., ZALPOUR K.: Marker-based method for analyzing the three-dimensional upper body kinematics of violinists and violists: development and clinical feasibility. *Medical Problems of Performing Artists* 34, 4 (2019), 179–190. doi:10.21091/mppa.2019.4029. 15

- [WMR*21] WEIDINGER L., MELLOR J., RAUH M., GRIFFIN C., UESATO J., HUANG P.-S., CHENG M., GLAESE M., BALLE B., KASIRZADEH A., ET AL.: Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021). [20](#)
- [WWS*15] WHEATLAND N., WANG Y., SONG H., NEFF M., ZORDAN V., JÖRG S.: State of the art in hand and finger modeling and animation. *Computer Graphics Forum* 34, 2 (2015), 735–760. [doi:10.1111/cg.f.12595](#). [11](#)
- [WZC12] WEI X., ZHANG P., CHAI J.: Accurate realtime full-body motion capture using a single depth camera. *ACM Trans. Graph.* 31, 6 (nov 2012). [doi:10.1145/2366145.2366207](#). [12](#)
- [WZHC22] WEN L., ZHOU J., HUANG W., CHEN F.: A survey of facial capture for virtual reality. *IEEE Access* 10 (2022), 6042–6052. [doi:10.1109/ACCESS.2021.3138200](#). [14](#)
- [XBP*18] XI Q., BITTNER R. M., PAUWELS J., YE X., BELLO J. P.: Guitarset: A dataset for guitar transcription. In *International Society for Music Information Retrieval Conference* (2018). URL: <https://api.semanticscholar.org/CorpusID:53875945>. [6, 7](#)
- [XCZ*18] XU W., CHATTERJEE A., ZOLLHÖFER M., RHODIN H., MEHTA D., SEIDEL H.-P., THEOBALT C.: Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)* 37, 2 (may 2018). [doi:10.1145/3181973](#). [12](#)
- [XGL*17] XIA S., GAO L., LAI Y.-K., YUAN M.-Z., CHAI J.: A survey on human performance capture and animation. *Journal of Computer Science and Technology* 32 (2017), 536–554. [doi:10.1007/s11390-017-1742-y](#). [12](#)
- [XLW*22] XU H., LUO Y., WANG S., DARRELL T., CALANDRA R.: Towards learning to play piano with dexterous hands and touch. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2022), pp. 10410–10416. [doi:10.1109/IROS47612.2022.9981221](#). [18](#)
- [XSe] XSENS: Ltc timecode in mvn. Accessed: November 2023. URL: <https://base.movella.com/s/article/LTC-Timecode-in-MVN>. [15](#)
- [XSW*21] XUE L., SONG K., WU D., TAN X., ZHANG N. L., QIN T., ZHANG W.-Q., LIU T.-Y.: Deeprapper: Neural rap generation with rhyme and rhythm modeling. 2021. [arXiv:2107.01875](#). [10](#)
- [YGTW15] YU J., GUO Y., TAO D., WAN J.: Human pose recovery by supervised spectral embedding. *Neurocomputing* 166 (2015), 301–308. [doi:https://doi.org/10.1016/j.neucom.2015.04.005](#). [12](#)
- [YKKG09] YAGISAN N., KARABORK H., GOKTEPE A., KARALEZLI N.: Evaluation of three-dimensional motion analysis of the upper right limb movements in the bowing arm of violinists through a digital photogrammetric method. *Medical Problems of Performing Artists* 24, 4 (2009), 181–184. [doi:10.21091/mppa.2009.4036](#). [16](#)
- [YLH*12] YE G., LIU Y., HASLER N., JI X., DAI Q., THEOBALT C.: Performance capture of interacting characters with handheld kinects. In *Computer Vision – ECCV 2012* (Berlin, Heidelberg, 2012), Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C., (Eds.), Springer Berlin Heidelberg, pp. 828–841. [doi:10.1007/978-3-642-33709-3_59](#). [12](#)
- [YLW*22] YU B., LU P., WANG R., HU W., TAN X., YE W., ZHANG S., QIN T., LIU T.-Y.: Museformer: Transformer with fine- and coarse-grained attention for music generation, 2022. [arXiv:2210.10349](#). [10](#)
- [YSD*16] YE M., SHEN Y., DU C., PAN Z., YANG R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1517–1532. [doi:10.1109/TPAMI.2016.2557783](#). [12](#)
- [YSI*23] YUAN Y., SONG J., IQBAL U., VAHDAT A., KAUTZ J.: PhysDiff: Physics-guided human motion diffusion model. In *IEEE International Conference on Computer Vision* (October 2023), ICCV. [19](#)
- [YUS*10] YAMAMOTO K., UEDA E., SUENAGA T., TAKEMURA K., TAKAMATSU J., OGASAWARA T.: Generating natural hand motion in playing a piano. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2010), pp. 3513–3518. [doi:10.1109/IRIOS.2010.5650193](#). [18](#)
- [YJW*20] YE Z., WU H., JIA J., BU Y., CHEN W., MENG F., WANG Y.: ChoreoNet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia* (New York, NY, USA, 2020), MM '20, Association for Computing Machinery, p. 744–752. [doi:10.1145/3394171.3414005](#). [17](#)
- [YZH*22] YI X., ZHOU Y., HABERMANN M., SHIMADA S., GOLYANIK V., THEOBALT C., XU F.: Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, June 2022), IEEE, pp. 13157–13168. [doi:10.1109/CVPR52688.2022.01282](#). [12](#)
- [YYZ*20] YU T., ZHAO J., ZHENG Z., GUO K., DAI Q., LI H., PONS-MOLL G., LIU Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 10 (oct 2020), 2523–2539. [doi:10.1109/TPAMI.2019.2928296](#). [12](#)
- [ZBL*19] ZHOU Y., BARNES C., LU J., YANG J., LI H.: On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (06 2019), pp. 5738–5746. [doi:10.1109/CVPR.2019.00589](#). [4](#)
- [Zie23] ZIELINSKY G.: How to mic a cello, 2023. Accessed: August 2023. URL: <https://www.remic.dk/academy/article/how-to-mic-a-cello/>. [11](#)
- [ZLAH23] ZHANG Z., LIU R., ABERMAN K., HANOCKA R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis, 2023. [arXiv:2307.15042](#). [19](#)
- [ZLZ*21] ZHU H., LI Y., ZHU F., ZHENG A., HE R.: Let's play music: Audio-driven performance video generation. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), pp. 3574–3581. [doi:10.1109/ICPR48806.2021.9412698](#). [19](#)
- [ZLZ*23] ZHOU Q., LI M., ZENG Q., ARISTIDOU A., ZHANG X., CHEN L., TU C.: Let's all dance: Enhancing amateur dance motions. *Computational Visual Media* 9, 3 (sep 2023), 531–550. [doi:10.1007/s41095-022-0292-6](#). [17](#)
- [ZPT*19] ZHAO L., PENG X., TIAN Y., KAPADIA M., METAXAS D. N.: Semantic graph convolutional networks for 3d human pose regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 3420–3430. [doi:10.1109/CVPR.2019.00354](#). [12](#)
- [ZRH13] ZHU Y., RAMAKRISHNAN A. S., HAMANN B., NEFF M.: A system for automatic animation of piano performances. *Computer Animation and Virtual Worlds* 24 (9 2013), 445–457. [doi:10.1002/cav.1477](#). [18](#)
- [ZSQJ12] ZHANG Y. C., SEÁGHDA D. O., QUERCIA D., JAMBOR T.: Auralist: Introducing serendipity into music recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2012), WSDM '12, Association for Computing Machinery, p. 13–22. [doi:10.1145/2124295.2124300](#). [15](#)
- [ZWC*22] ZHUANG W., WANG C., CHAI J., WANG Y., SHAO M., XIA S.: Music2dance: Dancenet for music-driven dance generation. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 2 (feb 2022). [doi:10.1145/3485664](#). [17](#)
- [ZWS*23] ZAKKA K., WU P., SMITH L., GILEADI N., HOWELL T., PENG X. B., SINGH S., TASSA Y., FLORENCE P., ZENG A., ABBEEL P.: Robopianist: Dexterous piano playing with deep reinforcement learning. In *Conference on Robot Learning (CoRL)* (2023). [doi:https://doi.org/10.48550/arXiv.2304.04150](#). [18](#)