Harmony-aware Human Motion Synthesis with Music

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 562

Cross-modal media generation has gained lots of attention recently, in which audio-visual harmony is an essential element to consider. In this work, we propose a novel audio-visual harmony evaluation framework that assures accurate alignment between the audio and visual beats. In the proposed framework, the detection of visual beats is redefined in order to make it consistent with the mainstream spectrum-based audio beats, and a saliency weighting mechanism is applied to simulate human attention. Based on such framework, a novel harmony-aware GAN architecture is proposed to tackle the audio-driven human motion synthesis problem. The experimental results indicate that the proposed harmony-aware model generates human motion with high perceptual quality, and it outperforms the start-of-the-art methods by around 20% in audio-visual harmony measurement.

CCS Concepts: • **Computing methodologies** → **Image processing**; **Computer vision**.

Additional Key Words and Phrases: audio-visual harmony, evaluation mechanism, audio-driven motion synthesis

ACM Reference Format:
Anonymous Author(s). 2021. Harmony-aware Human Motion Synthesis with Music. *ACM Trans. Graph.* 1, 1 (January 2021), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Harmony is an essential part of artistic creation. Movie directors trend to produce appealing scenes with songs that enhance emotional expression. When musicians arrange different voice parts in a chorus, they are supposed to consider whether the combination sounds harmonious. Artists pursue harmony in their works to create the senses of beauty and comfort [Moore 1942]. Since professional skills and techniques are required to complete such creative works, to save financial cost and labor, automatic generation is gradually applied to imitate the human creation process by exploiting computational models [Harvey et al. 2020; Lee et al. 2019; Li et al. 2018]. Similar to human work, the machine-based generation needs to obey the rule of harmony in order to produce high-quality results that satisfy human aesthetics.

Machine-based generation is widely used in the tasks of musicing videos [Morgado et al. 2018; Nakamura et al. 1994], speech editing [Chen et al. 2014; Jin et al. 2017], and animation synthesis [Karras et al. 2017; Zhou et al. 2018], where harmony represents the consistent perception of rhythms, emotions or visual appearances in the output subjectively. Handling harmony in those generative tasks means the models should put effort into controlling the consistency between multiple signals, which is shown as the alignment of features explicitly for observation [Bellini et al. 2018] or implicitly in the latent spaces [Theodoridis et al. 2020]. The synchronization for different signal pairs may differ in their relevance so that in the high-related pairs, correlated features are easier to be captured and aligned. In human perception, over 90 percent of sense derives from the stimulus of visual or auditory signals and they interrelate and interact with each other during brain processing [D'Ausilio et al. 2014; Vetter et al. 2014]. Research also reveals that introducing auditory features benefit visual learning [Owens et al. 2016; Yalta et al. 2019] and correspondingly visual priors improve the performance of sound analysis [Ephrat et al. 2018; Khan et al. 2018]. Such strong relationships thus make audio-visual synchronization promising in the cross-domain field [Cardle et al. 2003; Suwajanakorn et al. 2017].

As a typical problem in audio-visual cross-domain generation, the task of audio-driven motion synthesis gains much attention in character animation [Shiratori et al. 2006], video generation [Ren et al. 2020] and choreograph [Ye et al. 2020]. The traditional methods tackle the audio-to-visual generation by retrieving visual clips that share the feature-level similarity with the given music [Lee et al. 2013; Shiratori et al. 2006]. In order to select the visual clip that matches the music best, the need of harmony evaluation emerges [Chu and Tsai 2011; Ho et al. 2013]. With the increasing development in deep learning methods, deep neural networks, that can learn the end-to-end mappings between the audio-visual pairs, has revolutionized and dominated the field. Different from regular motion synthesis [Holden et al. 2016; Ling et al. 2020], when conditioned with music, people are found to be sensitive to the inharmonious synthesized motions, which damages the qualitative evaluation heavily [Lee et al. 2019; Tang et al. 2018]. More and more methods thus start to consider harmony as one of the most important factors that highly influence the quality assessment of cross-domain results [Karras et al. 2017; Suwajanakorn et al. 2017]. However, the feelings of harmony relies on perceptual judgement. This may be challenging to enhance the audio-visual harmony during the network generation. Some of the existing methods [Ahn et al. 2020; Lee et al. 2018; Ren et al. 2020] pursue stronger and deeper network architectures to implicitly harmonize the music and motion rhythms in the feature space. In [Lee et al. 2019; Tang et al. 2018] the auditory beats are pre-computed as prior knowledge to provides additional features for music rhythms in the network learning, which has limited benefits for regularizing harmonious estimations. In the video generation field, visual beats are detected to describe the visual rhythms and correlated with auditory beats for audio-visual synchronization [Bellini et al. 2018; Davis and Agrawala 2018]. [Yalta et al. 2019] then propose a beat-matching loss to constrain the harmony in the generation by performing audio-visual beat alignment. Such method improves the audio-visual synchronization to some degree whereas the depiction of visual beats is not precise enough and lacks the

0730-0301/2021/1-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

consideration of saliency that affects human perception [Mack et al. 1998; Simons and Chabris 1999].

In this paper, we explore the correlations between the audio and visual signals that contributes to the perceptual harmony in human judgements. Following the literature that the audio-visual harmony is highly related to beats [Bellini et al. 2018; Davis and Agrawala 2018; Yalta et al. 2019], we analyze beat extraction that can satisfy the cross-domain consistency and build alignment functions to approximate the real perception. Since the perceptual judgment can be affected by human attention mechanisms, we firstly introduce the attention-based mechanisms into the harmony evaluation framework. The adaptive attentional masks are utilized to depict the visual and auditory saliency in the perceptual evaluation. With such established audio-visual correlations, a novel harmony evaluation mechanism is then proposed to enhance the cross-domain harmony in the audio-to-visual generation by quantifying the synchronization between the audio and the motion sequences. Inspired by the idea of synthesis-by-analysis, to the best of our knowledge we are the first to propose a beat-oriented GAN [Goodfellow et al. 2014] framework to cooperate with the beat-aligned harmony evaluation framework. By training with audio-visual segments from beat-composed music meters, the networks can strengthen the unit-based learning for the mappings between the cross-domain features. Meanwhile, the evaluation framework is incorporated into the generation by forming a harmony-aware hybrid loss function to regularize the synthesized motions to be harmonious with the input audio.

The rest of the paper is organized as follows. We provides a brief review of the current literature for audio-visual evaluation and audio-driven motion synthesis in section 2. The section 3 consists of the details for beat detection and alignment in our proposed harmony evaluation mechanism. In section 4, the architecture of our proposed GAN framework and the harmony-aware hybrid loss function are presented. The implementations and experimental results are demonstrated in section 5. The limitations and future work for our idea are discussed in section 6. Finally we draw our conclusions in section 7.

2 RELATED WORK

2.1 Audio-visual Harmony Analysis

The analysis of audio-visual harmony is highly related to exploring the correlation between the audio and visual features, where beats are greatly focused in the literature as the most popular features that describe the rhythms in both the audio and visual contents. [Chu and Tsai 2011] firstly perform the rhythm-based cross-media alignment for dance videos. Based on the video frames, visual beats are extracted by analyzing the motion trajectory while the music beats are obtained by spectrum. The alignment is applied by shifting and matching between the audio-visual segments. With the collected 3D Kinect data from the dance video, [Ho et al. 2013] detect the visual beats by calculating the changes for the angular and velocity of the joints. They convert the alignment into the path-finding problem between the audio and visual beats by utilizing F-score. Furthermore, inspired by the onset envelope and the tempogram used for detecting audio beats, [Davis and Agrawala 2018] extract visual beats from video frames by calculating the impact envelop and visual tempogram based on the directgram derived from the optical flow. Different from quantifying the distance, [Davis and Agrawala 2018] directly apply warpping to match the audio and visual beats in the spatial-temporal domain. Similar to [Davis and Agrawala 2018], working with skeleton-based movements, [Lee et al. 2019] employ the warpper to make adjustments for the generated motion sequences by aligning the music beats with the offset-based motion beats. Meanwhile, based on such skeleton-based movements, [Yalta et al. 2019] obtain visual beats by detecting the directional changes in the body motions, which is then aligned with the feature-based music beats by computing the entropic distance.

2.2 Audio-driven Motion Synthesis

The traditional methods which tackle the audio-to-motion generation could be broadly divided into three categories: retrieval-based, model-based and learning-based methods. In retrieval-based approaches [Lee et al. 2013; Shiratori et al. 2006], the obtained music is segmented and sent into the recommendation system where a database will provide the best-matched candidate from the pre-captured motion sequences based on the similarity between the extracted features in music and motions. Limited to the richness of the database, the resulting motion sequences could not be consistent and match the music well. Model-based approaches then have been proposed to formulate a more flexible generation by utilizing the Hidden Markov Model (HMM) [Ofli et al. 2010, 2011] or deriving the linear correlation coefficients [Fan et al. 2011] to explore the inner connection between the audio and motions. Soon neural networks lead the learning-based solutions for computer vision tasks. Driven by the large and rich dataset, the networks enhance the performance in capturing deep features and learn the direct mapping between audio and visual signals. [Lee et al. 2018] first train an encoder-decoder based network to translate the auditory features into choreographic skeleton-based poses efficiently. [Tang et al. 2018] further claim that the auto-encoder using long short-term memory (LSTM) blocks could better obtain the temporal features of music in the long term and estimate the consistent human movements smoothly. In order to pursue the robustness in the long-time motion generation, one trend is to assist Recurrent Neural Networks (RNN) [Yalta et al. 2019], which utilizes the previous estimations as the prior knowledge to guide the temporal consistency when generating the next movements. In the other trend, Autoregressive Neural Networks (ARNN) [Ahn et al. 2020] are extended into the existing architectures to improve the efficiency in the sequence-based movement generation. After the Generative Adversarial Networks (GAN) [Goodfellow et al. 2014] has been proposed, it outperforms the previous networks with the synthesis-by-analysis architecture. [Lee et al. 2019] firstly introduce such adversarial architecture to regularize the cross-modal audio-to-visual generative task with multiple discriminators. By encoding and decoding the complex motion sequences into continuous movement units, the model can convert the input music into realistic and natural human motions. Based on the work of [Lee et al. 2019], [Ren et al. 2020] design a perceptual pose loss with the aid of Graph Convolutional Networks (GCN) [Yan et al. 2018]. Different from the classic feature-matching loss performed by VGG-based networks, the trained GCN are claimed to learn the

hierarchical representation of skeleton sequences with the designed spatial-temporal graph. Such skeleton-aware loss in the feature-space has been proved that can regularize the network to estimate high-quality body movements based on the skeleton hierarchy.

3 EXPLORING AUDIO-VISUAL HARMONY

This section discusses the common strategies that assess the audio-visual harmony in the cross-modal generation tasks. Those strategies target the evaluation of harmony as a problem of measuring the rhythmic consistency between the audio and visual sequences. We build our evaluation approach based on such common assumptions and propose a novel attention-aware framework to improve the approximation of human perception in the assessment of audio-visual harmony.

3.1 From Harmony to Beat

Harmony plays an important role in the evaluation of generated cross-modal results. Since the sense of vision and hearing are highly related and affect each other in brain processing, harmony is especially concerned in the tasks of audio-to-visual or visual-to-audio generation [Jin et al. 2017; Tang et al. 2018]. Taking the example of audio-driven human motion synthesis, in the quality assessment the audio-visual harmony is emphasized that the synthesized movements should be rhythmic and harmonious with the music [Lee et al. 2019; Yalta et al. 2019]. In other words, the rhythms in the audio and visual sequences are required to be consistent temporally in order to satisfy the perceptual harmony.

Since the feelings of rhythm rely on subjective human perception, given the audio sequences $A(t)$ and visual sequences $V(t)$ as functions of time t, it's an important topic to approximate the perceptual judgement of harmony into quantitative measurements as:

$$h = H(A(t), V(t)) \tag{1}$$

where H denotes the algorithm that analyzes the harmony between the cross-domain signals and h is a scalar representing the quantified judgement.

In the literature many researchers have made efforts to extract features from $A(t)$ and $V(t)$ to represent and quantify the rhythms. Following the definition of meters in the musical composition, which is a set of indivisible units anchoring the melody, audio beats are tracked to concretize the abstract rhythm in the music sequences [Scheirer 1998]. Similarly on the visual side, visual beats, as the demonstration of changes in body movements, have been proposed to represent the visual rhythms and correlate with the auditory beats [Davis and Agrawala 2018].

Spectrogram analysis is widely used to obtain the musical beats $B_a(t)$ in audio processing. The spectrogram of given audio sequences $A(t)$ could be obtained by the time-windowed Fast Fourier Transform (FFT). With the estimation of amplitude for the spectrogram, the beats are extracted by looking for the distinct amplitude change in the time domain, which could be described as:

$$g(t) = Amp(FFT(A(t))) \tag{2}$$

$$B_a(t) = \begin{cases} 1 & \text{if } g(t) - g(t') > c_1, \forall t' \in \dot{U}_a(t, t_0) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where Amp denotes the function or model that estimates the amplitude of spectrogram. The positive threshold c_1 is set to determine the existence of beats at time t which satisfies $B_a(t) = 1$ compared with any other t' in its punctured neighborhood $\dot{U}_a$ radiused by a pre-defined constant t_0. In the main stream methodologies, amplitude estimation is conducted by deriving the onset strengths from the obtained spectrogram [Davis and Agrawala 2018; McFee et al. 2015]. The music beats $B_a(t) = 1$ then establish on the occurrence of the peak in each onset envelope.

Referring to the rules that detect music beats, the visual beats $B_v(t)$ are similarly extracted based on the analysis of motion trend between visual sequences $V(t)$ and $V(t - t_0)$. When there is change happens drastically in the motion trend, the time t is considered as the occurrence of a beat, which could be depicted as:

$$d(t) = MT(V(t), V(t - t_1)) \tag{4}$$

$$B_v(t) = \begin{cases} 1 & \text{if } d(t) - d(t') > c_2, \forall t' \in \dot{U}_v(t, t_2) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where MT denotes the function or model that estimates the motion trend during t_1 (a pre-defined constant value) and c_2 is a positive value that controls the threshold to obtain the beat at time t where $B_v(t) = 1$ stands in comparison with any other t' in its punctured neighbourhood $\dot{U}_v$ radiused by pre-defined constant t_2. To process the general pixel-based visual signals, the use of optical flow is able to capture the motion trend in the moving events [Bellini et al. 2018; Davis and Agrawala 2018]. With the quantification of optical flow the visual beats could be obtained by deriving the local maximums that denote the obvious changes in movements. When focusing on only the human motion in the such pixel-based signals, the skeleton-driven method has been specially proposed to specify the motion trend for the pure skeleton-based motions extracted from visual signals. Thus, the estimation of motion trend can be converted to analyzing the directions of body movements[Yalta et al. 2019] by joint-based standard derivation. The visual beats $B_v(t) = 1$ thus are defined as the distinct directional changes in the motion sequences.

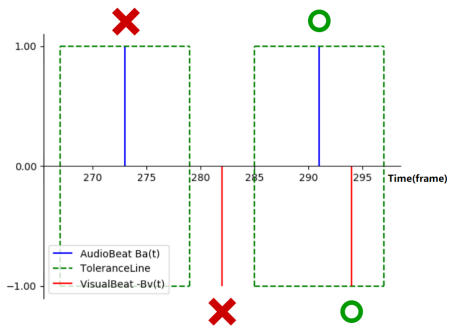

Fig. 1. An example of perceptual unsynchronization between beats. The audio beats are set 1 while the visual beats -1. Given an audio beat, if its nearest unlabeled visual beat is out of the tolerance field, the audio-visual beat pair is considered as unsynchronized in subjective perception, which are labeled as red cross. On the contrary, the synchronized beat pairs are labeled as green circles.

Based on the observed audio and visual beats, a common assumption has been derived to tackle rhythmic consistency that the appearance of every music beat is supposed to synchronize with that of the visual beat and vice versa [Bellini et al. 2018; Davis and Agrawala 2018; Yalta et al. 2019]. Following such assumption, the existing strategies evaluate the quantified audio-visual harmony h by performing the alignment based on the extracted beats, which extend the Eq. (1) as:

$$h = L(f_a(A(t)), f_v(V(t))) = L(B_a(t), B_v(t)) \tag{6}$$

where f_a and f_v denote functions for beat detection in the audio and visual signals respectively and L represents the alignment algorithm.

The existing algorithms L formulate the alignment problem as analyzing the distances between the synchronized beat pairs by warping [Bellini et al. 2018; Davis and Agrawala 2018], cross-entropy or F-score [Sokolova et al. 2006; Yalta et al. 2019], which are effective to align the cross-domain objects. Ideally the extracted audio and visual beats are expected to be synchronized in the ground-truth data, however in practical applications, especially scenarios with human videos, the unsuitable beat detection approaches may cause tremendous distortion for the harmony, which will be demonstrated in section 3.2 with examples.

With the booming of social networks, human related media has become very popular, as indicated in [Statista 2018], 62% of US adults had taken a photograph of themselves and uploaded it to a social media website in 2018. The recent breakthrough in computer vision with deep learning has enabled lots of capabilities and applications, such as real-time human tracking [Habermann et al. 2019; Wang et al. 2020], human motion synthesis and analysis [Holden et al. 2016; Ling et al. 2020], deepfake, and so on, which can bring human video to the next level. Thus, in this work our proposed methodologies are focused on the human video.

3.2 Harmony Distortion in Human Videos

In human videos, given human and their movements as the attention points, the harmony are mainly considered as the alignment between foreground human motion in the visual frames and the associated background music. To better analyze the foreground human motions, the skeletons of human are extracted to represent the human motions in videos [Charles et al. 2016; Cheng et al. 2019; Pavllo et al. 2019]. Hence, harmony has been evaluated between the audio signals and skeleton-based human motions [Lee et al. 2019; Tang et al. 2018; Yalta et al. 2019]. To visualize the subjective evaluation of harmony by objective expressions, based on human reaction time, the tolerance fields [Ho et al. 2013] neighboured with audio beats are set to represent the perceptual judgement of audio-visual harmony in terms of the synchronization between beats. In Fig. 1, an example of unsynchronized audio-visual beats is illustrated, where the processing of beats for audio and visual signals separately did not reach a satisfactory alignment.

To obtain the musical beats, the mainstream approach making use of the onset strength is exploited here, which is commonly seen in the literature for handling general audio sequences [Lee et al. 2019; Tang et al. 2018]. All the musical beats are practically processed by methods in the open-source package LibROSA [McFee et al. 2015], which provides the implementations of the onset-driven beat

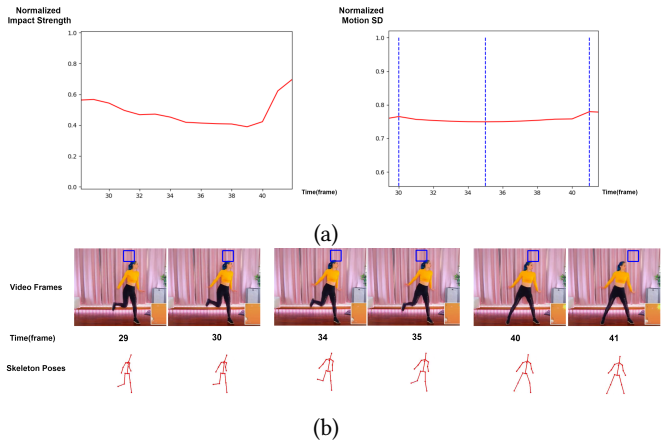

Fig. 2. An Example comparing different strategies for visual beat detection in human videos. (a) The frame-based visual beat detection in [Davis and Agrawala 2018] using optical flow (left) and the skeleton-based beat extraction in [Yalta et al. 2019] using motion standard derivation (right). The obtained beats are set in the blue lines where in the left no beat is detected during the presented time domain. (b) The video frames and corresponding extracted skeleton poses (use model in [Zhou et al. 2017]) from the test case [YouTube:littlesiha 2019]. In video frames, a blue box moving randomly is set to simulate the possible disturbance from the background.

detection for audio signals. By the analysis of Mel spectrogram [Böck and Widmer 2013], the onset strengths are pre-computed to estimate the tempo based on the auto-correlation inside onset envelope [Ellis 2007]. Referring to Eq. (2) and (3) the auditory beat $B_a(t) = 1$ can be explained by the case where there is peak in the onset envelope at t consistent with the obtained tempo. To assemble the valid beats in $B_a(t)$, the position-based beats $\{p_a(b)|b = 1, 2, ..., N\}$ are formed to collect all the positions in time t of occurred N beats that satisfy $B_a(t) = 1$. Simultaneously the corresponding strengths of beats $p_a(b)$ are thus represented with the peak values as $s_a(b)$.

In the visual case, optical flow is often used to extract beats for general frame-based visual signals [Bellini et al. 2018; Davis and Agrawala 2018]. However when fed with human video, this approach does not function effectively compared to the skeleton-based approach. One reason is that optical flow treats the motion of each pixel almost equally where the former method puts much higher weights on the foreground human. As show in Fig. 2, with the distinct occurrences of human movement changes in the video frames perceptually, accordingly the visual beats should be obtained by the beat extraction methods. However, disturbed by possible moving events in the background, the optical flow method has difficulty in detecting beats for foreground human motions while the skeleton-based approach outperforms it significantly in obtaining visual beats that are more consistent with human perception. The joint-wise standard deviation (SD) based visual beat detection proposed in [Yalta et al. 2019] has been used to represent the skeleton-based approach, where the visual beats are detected by estimating the directional changes in the body movements.

In the following trial, we combined the mainstream onset-based musical beat detection with the SD-driven visual beat detection in

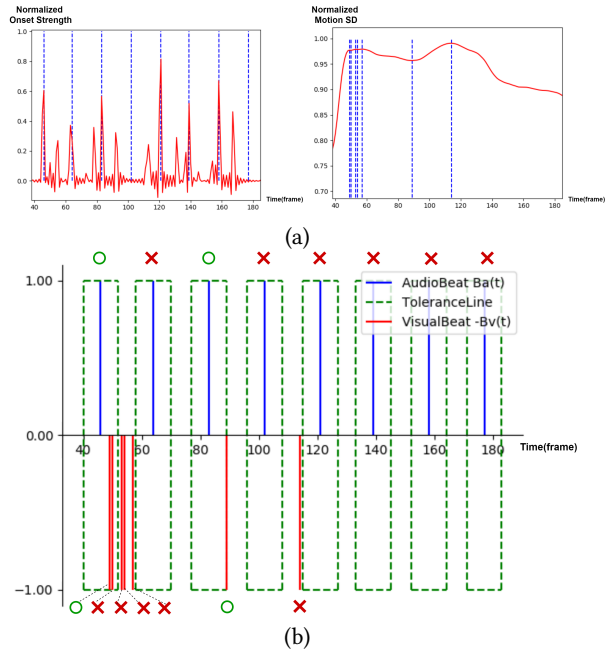

Fig. 3. An illustration of harmony distortion between onset-based audio beats and SD-driven visual beats. (a) Beat extraction in the audio (left) and visual (right) signals. The beats are labeled 1 with blue lines. (b) The unsynchronization between audio-visual beats under perceptual judgements. Since each audio beat should only have one synchronized visual beat, the redundant visual beats will also be labeled as unsynchronized (red cross).

the beat alignment experiment conducted on the dance dataset [Tang et al. 2018]. Since the human reaction time is around 0.25 seconds, the radius of tolerance field for each audio beat is set as 6 frames, with the total duration of 0.24 seconds under 25 fps, to evaluate the audio-visual alignment results. However, the outcome does not work very well. As shown in Fig. 3, the harmony distortion is quite high due to the omission and redundancy. It reveals that, though the SD-driven visual beat detection could basically harmonize with the subjective perception, when referred with onset-based musical beats, such cross-domain audio-visual beat pairs are not consistent with each other. Since the human motion speed, as one of the important components to represent motion trend in the analysis of audio-visual alignment [Ho et al. 2013; Shiratori et al. 2006], has not been effectively considered in the motion SD approach. To cooperate with onset-based audio beats, we propose a novel mechanism considering the velocity of joints in neighboring frames for higher weights to detect visual beats that can satisfy the better beat consistency in the audio-visual alignment.

Given the skeleton-oriented human motion sequences $v_s(t, j)$ with j joints at frame t obtained from $V(t)$, the joint velocity sum $J_v(t)$ is derived by calculating the frame difference as:

$$J_v(t) = \sum_{i=1}^{j} v_s(t, i) - v_s(t-1, i) \tag{7}$$

where i denotes the i^{th} joint.

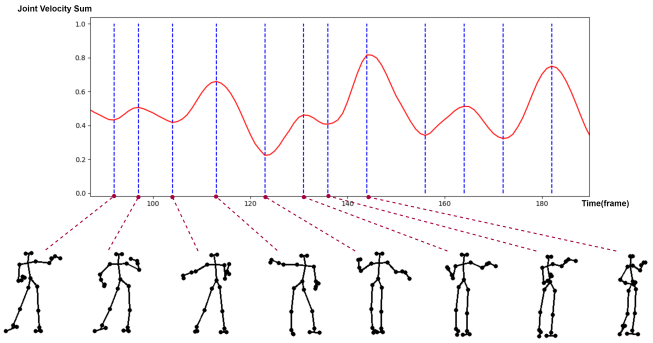

Fig. 4. The relationship between joint velocity sum and evolution of human movements. The approaching of peaks and valleys in the joint velocity sum is usually accompanied with the change of basic human movements.

As discussed in [Ye et al. 2020], dance-like human motions have been proven to be composed of several undividable movement units and those units are highly consistent and correlated with the music beat. In order to define the motion beats that are well-aligned with music beats, the evolution of indivisible movement units (e.g. Hand lift) is mainly focused in the analysis of visual beats in the whole motion sequences.

Fig. 4 demonstrates the correlation between the velocity graph and the real human movements. Our observation reveals that when the change of joint velocity sum is approaching zero, it is usually related to the complement of a single movement unit. Thus, the motion beats can be defined as the peaks or valleys in the velocity graph, where the acceleration equals zero. The Eq. (4) and (5) are then reformed as:

$$\tilde{d}(t) = sign(J_v(t) - J_v(t-1)) \tag{8}$$

$$\tilde{B_v}(t) = \begin{cases} 1 & \text{if } \tilde{d}(t) \times \tilde{d}(t+1) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Similar with the audio case, our position-based visual beats are then formulated as $\{p_v(b)|b = 1, 2, ..., M\}$ for M valid beats satisfying $\tilde{B_v}(t) = 1$ and their strengths are assigned due to the corresponding $J_v(t)$ as $s_v(b)$. Fig. 5 demonstrates that when tested with the same conditions in Fig. 3, such obtained visual beats are basically synchronized with the occurrence of onset-based music beats. Apart from motion speed, due to the lack of consideration for onset-based music beats in [Yalta et al. 2019], comparing Fig. 3(b) and Fig. 5(b) it can be observed that our beat extraction mechanism outperforms the existing skeleton-based approach using motion SD by reducing the omission and redundancy for the beat-wise synchronization.

In addition to the synchronization between beats, another factor that highly influences the perception of audio-visual harmony is our attention mechanisms. Since human attention will be drawn for things that are more "attractive", on the contrary some other things may be overlooked unconsciously in our perception. Thus, to assess the audio-visual harmony close to the real human perception, the attention mechanisms are needed to be introduced in the evaluation framework.

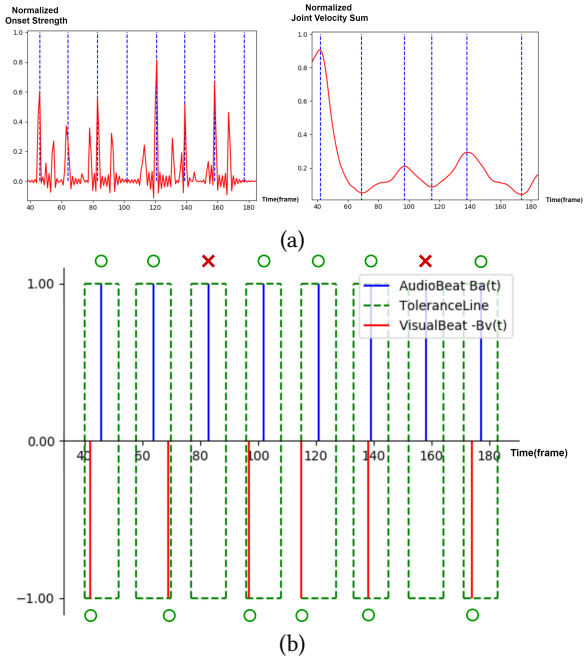

Fig. 5. An illustration of the improved beat-wise synchronization based on our visual beat extraction mechanism. (a) Beat extraction in the audio (left) and visual (right) signals. The beats are labeled 1 with blue lines. (b) Considering joint velocity, the synchronization between audio-visual beats has been improved compared with the motion SD approach(Fig. 3(b)).

3.3 Attention-based Evaluation Framework

The attention mechanisms reveal that unsalient objects are neglected in human perception without any awareness [Mack et al. 1998], which influence both vision and hearing systems [Dehais et al. 2014; Simons and Chabris 1999]. When it comes to the subjective perception of rhythmic harmony, the phenomenon of inattentional blindness and deafness may also affect the judgement based on the saliency distribution in the audio and visual rhythms. In order to approximate the perceptual measurement of harmony, we propose an attention-based evaluation framework to highlight the importance of salient beats, which extends the Eq. (6) as:

$$h = L(W_a(p_a(b)), W_v(p_v(b))) \tag{10}$$

where W_a and W_v denote the attentional weighting masks derived from the audio and visual beat saliency respectively.

3.3.1 Attentional Beat Weighting. Since the salient beats favor the perception of harmony, a weight is assigned to each beat based on its beat saliency to enhance the corresponding attentional impact in the evaluation. The beat saliency is represented by the beat strengths $s_a(b)$ and $s_v(b)$ obtained in Section 3.2 and adaptive weighting masks are constructed by considering the global standard deviation(SD) for the strengths.

In the analysis of auditory saliency, the attentional mask W_a is built as:

$$W_a = sign(s_a(b) - SD(s_a(b)) \times \lambda_1) \tag{11}$$

where λ_1 denotes a constant scale factor to adjust the audio saliency threshold.

Different from processing the mask for audio beats, in the visual case the motion beats are extracted from not only the peaks but also the valleys of the joint velocity sum, which means the direct comparison with standard deviation is not applicable for analyzing the visual saliency. Therefore the peak-to-valley difference is utilized to define the visual saliency strength for detecting the appearances of high-impact visual beats, which is shown as:

$$R(b) = |s_v(b) - s_v(b-1)|, b = 2, ..., M \tag{12}$$

where R(b) denotes the peak-to-valley difference for each beat.

The visual saliency mask W_v is then offered by utilizing global standard deviation(SD) as:

$$W_v = sign(s_v(b) - SD(R(b)) \times \lambda_2) \tag{13}$$

where λ_2 denotes a constant scale factor to adjust the visual saliency threshold.

By applying the weighting masks W_a and W_v on the beats $p_a(b)$ and $p_v(b)$ respectively, we obtain the attentional beats $p'_a(b)$, $p'_v(b)$ by extracting the positive results from $W_a(p_a(b))$ and $W_v(p_v(b))$. The corresponding beat strengths for the attentional beats are similarly defined as $s'_a(b)$ and $s'_v(b)$.

3.3.2 Aligning the Cross-domain Beats. The harmonious feeling in audio-visual human perception can be described as fuzzy measurement, which derives from the way that our brains recognize sensory signals [Bay and Usakli 2003; Picot et al. 2011]. To handle the beat alignment, the existing warping method [Davis and Agrawala 2018] directly adjust the strength curve for visual beats to fit that of audio beats by applying compensations. Similarly, in [Yalta et al. 2019] the contrastive difference has been constructed by calculating cross-entropy distance between the auditory amplitude and motion labels. Since our brain has limitations for recognizing the signals in precise amplitude, such strength-based fine mappings between audio-visual beats are not consistent with the real perception.

Inspired by the binary labels given to present whether the audio beats and visual beats are synchronize in the time domain(e.g. Fig. 1), we propose to construct hitting scores by counting the "good" labels in the audio and visual domain to represent whether the beats are well-aligned in the whole sequences fuzzily. To balance the audio-visual perception, the F-score method is performed to fuse the cross-domain scores for the final judgement.

Beginning with the selected high-saliency N audio beats $p'_a(b)$ and M visual beats $p'_v(b)$ in the Section 3.3.1, the Eq. (10) is reformed by using the F-score measurement as:

$$h = L(p'_a(b), p'_v(b)) = F_s(E(p'_a(b)), E(p'_v(b))) \tag{14}$$

where E denotes the algorithm obtaining the hitting score in both audio and visual domain and F_s represents the F-score measurement.

With the observation that there is a delay between visual perception and brain-processed recognition [Ho et al. 2013], the assumption could be made that the beat can be considered to be hit as long as the time interval between this beat and the nearest cross-domain beat is less than the human-reaction delay. In this way, a fuzzy interval-based judgement could be made for measuring the

alignment, instead of depending on precise strength-based mappings. As the synchronized beats appear in audio-visual pairs, the musical beats could be seen as anchors in the analysis of hitting. In order to obtain the interval, the position matrix $Z(b_a, b_v)$ is built by repeating the M visual beat $p'_v(b)$ for N times as

$$\forall b_a, Z(b_a, b_v) = Z(b_v) = p'_v(b_v) \tag{15}$$

where $b_a = 1, 2, ..., N$ and $b_v = 1, 2, ..., M$.

The column-wise audio-visual interval based on $Z(b_a, b_v)$ is computed by subtracting $p'_a(b)$ absolutely:

$$D(b_a, b_v) = |Z(b_a, b_v) - p'_a(b_a)| \tag{16}$$

Then the judgement of whether the musical beat is hit could be obtained by comparing its minimum audio-visual interval rowwisely with the pre-defined reacting delay as:

$$T(b_a) = min(D(b_a, b_v)) \tag{17}$$

$$Hp(b_a) = \begin{cases} 1 & \text{if } T(b_a) \leq T_{delay} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where T_{delay} is a constant frame time and $Hp(b_a) = 1$ denotes there exists a synchronized audio-visual beat pairs.

Finally, the hitting score h_s can be derived by performing weighted sum of all the hitting points as

$$h_s = \sum_{b_a=1}^{N} Hp(b_a) \times s'_a(b_a) \tag{19}$$

Considering the normalization for the total numbers of music beats and motion beats, we form the hitting score for audio harmony as $h_a = \frac{h_s}{N}$ and visual harmony score $h_v = \frac{h_s}{M}$. However, the correlation between h_a and h_v differs from sources to sources. For instance, given a specific input audio-visual sequences, the obtained h_a may be higher than h_v but the contrary observation can be obtained for another input sequences. In order to balance between the audio-visual scores, inspired by F-score [Sokolova et al. 2006] the final audio-visual harmony h is obtained by performing the harmonic mean, which reform the Eq. (14) as:

$$h = F_s(h_a, h_v) = \frac{(1 + \beta^2) h_v h_a}{\beta^2 h_v + h_a} \tag{20}$$

where β is a pre-defined constant. Therefore Eq. (20) can be transformed into the function of h_s as:

$$h = \frac{(1 + \beta^2) h_s}{N\beta^2 + M} \tag{21}$$

Implied by Eq. (20)-(21), the quantification of audio-visual harmony in the evaluation can be suggested as:

Lemma 3.1. *Given an audio clip with N obtained attentional musical beats and a visual clip with M visual beats, the quantified audio-visual harmony can be uniquely determined by h_s.*

If our quantified audio-visual harmony can represent the assessment close to the subjective perception, we could consider introducing such an evaluation framework as the regularization to improve the learning of our network.

4 LEARNING THE HARMONY-AWARE AUDIO-DRIVEN MOTION SYNTHESIS

To show the practicality of the proposed beat-based harmony mechanisms in Section 3, the analysis of audio-visual rhythmic harmony can be adapted to the task of audio-driven human motion synthesis. In order to generate motions that are harmonious with musical rhythms, in this section we propose a beat-oriented GAN framework and incorporate the harmony loss as one of the components in the hybrid loss function to regularize the consistency of audio-visual beats in the generated sequence pairs.

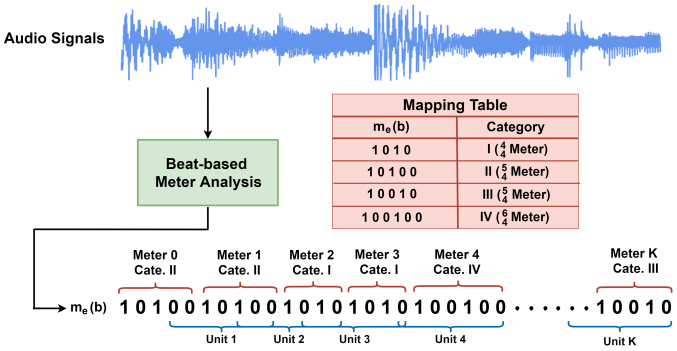

Fig. 6. The extraction of meter unions based on the obtained audio beats.

4.1 Beat-oriented Adversarial Learning Framework

When it comes to introducing musical beats into the audio-to-motion deep learning frameworks, the literature focuses mostly on the primitive beat information such as strengths and time occurrences, forming the prior knowledge [Tang et al. 2018], loss function [Yalta et al. 2019] or post-processing modulation [Lee et al. 2019] to benefit the understanding of the rhythms in the input audio sequences during the generation of human motions. However, the high-level interconnections between the auditory beats, which are closely correlated with the music rhythms, are rarely perceived in the previous methods. The melody of music can be represented by the composition of multiple self-identical segments, denoted as loops [Duffell 2005]. To quantify the specific rhythms in each loop, composers utilize music meters to measure such regularly recurring patterns [Cooper et al. 1963], depicted by a set of strong and weak beats [Benward 2014]. In order to enhance the learning of music rhythms based on the composition units, we group the obtained audio beats into combinations and map them to the corresponding types of meters. Our motion synthesis model is trained and tested on such beat-composed meter units.

As shown in Fig. 6, given the audio sequences $A(t)$ the music beats $p_a(b)$ and their strengths $s_a(b)$ can be obtained. Whether the beat is strong or weak could be determined by comparing the strength with its previous beat as:

$$m_e(b) = \begin{cases} 1 & \text{if } s_a(b) > s_a(b-1) \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where $m_e(b) = 1$ means the beat is a strong beat and 0 is a weak beat. Exampled with the quarter note [Acuff and Evridge 1906],

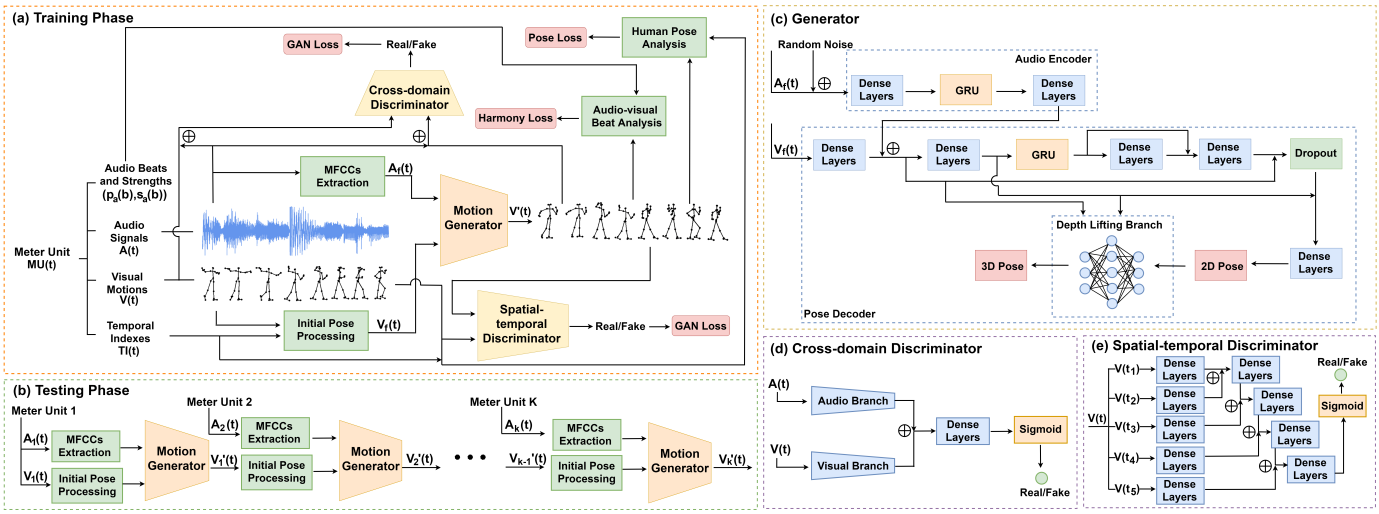

Fig. 7. The overview of the whole framework, where the training and testing process are demonstrated with detailed network structures used in our models. The addition symbol in the figure denotes the concatenation operation.

strong-weak beat combinations are mapped into 3 meter types: $\frac{4}{4}$, $\frac{5}{4}$ and $\frac{6}{4}$ in 4 categories totally [Read 1964]. Taking care of the transitions between meters, we add several beats in the last meter as prior knowledge and form the unit into the unified beat length to describe the flow of musical rhythms. Similarly in the human motion sequences, if harmonious with the given auditory rhythms, it can show regular recurring movement units related to the music beats [Lee et al. 2019; Ye et al. 2020]. It could be assumed that the correlation exists between such movement units and our obtained auditory units. Thus the audio-visual clips are segmented based on the defined meter units temporally as input to strengthen the learning of beat-driven cross-domain unit mapping in our deep model, which can indirectly benefit the audio-visual harmony for the generation.

Since GANs [Goodfellow et al. 2014] show their outstanding power in visual generation tasks, they are also popular to be used in cross-domain generation. Therefore to handle the music-driven motion synthesis, we utilize the adversarial framework that can inherently introduce the audio features in the generation of human motion. Fig. 7(a) demonstrates the overview of the whole training process. For each meter unit, the start time and end time are recorded as $MU(t), t \in [t_{start}, t_{end}]$. According to the time records, the corresponding audio and motion sequences $A(t)$ and $V(t)$ are obtained as the ground-truth pairs, including the music beats $p_a(b)$ and their strengths $s_a(b)$. Meanwhile, the temporal indexes $TI(t)$ is formed in binary to denote the separation between the current meter and the last one, where $TI(t)$ is set to 1 if the time t belongs to the priors from the last meter. For $A(t)$, the features of Mel Frequency Cepstral Coefficients (MFCCs) are extracted as the auditory input $A_f(t)$. Different from existing deep models [Lee et al. 2019; Ren et al. 2020; Tang et al. 2018], instead of only fed with auditory features, we process initial poses $V_f(t)$ as visual input based on $TI(t)$ and $V(t)$ by

$$V_f(t) = \begin{cases} V(t) & \text{if } TI(t) = 1 \\ \frac{\sum_t V(t) \times TI(t)}{\sum_t TI(t)} & \text{otherwise} \end{cases} \quad (23)$$

where the visual features $V_f(t)$ keep the movements from the last meter as priors and use the mean pose as initialization for the current meter. This allows the network to enhance the temporal consistency in the synthesis of human motion. Inputted with audio-visual features, the structure of our generator G can be summarized as $G(A_f(t), V_f(t)) = V'(t)$, where $V'(t)$ denotes the generated motion sequences. Based on this generator structure, in the testing phase $V_f(t)$ are processed from the previously synthesized motions, which contributes to the generation of consistent human motions recurrently with meter units for audio clips in random duration (See Fig. 7(b)). To supervise the reality of produced human motions, apart from the cross-domain discriminator, which is widely used to control the validity of feature conversion in different domains, a spatial-temporal pose discriminator is utilized to judge realistic movements both spatially and temporally. In addition to the GAN losses provided by discriminators, we propose to incorporate the multi-space pose loss and beat-driven harmony loss based on the attentional harmony mechanism as the regularization for our generator. Such harmony-aware hybrid loss functions can guide the generator to output human-like motion sequences that are harmonious with the given music by constraining the consistency between the pre-computed audio beats and visual beats extracted from the generated motions.

4.2 Network Architecture

In the tasks of audio-driven motion synthesis, the network is fed with the input of music sequences or extracted auditory features to generate the visual motion sequences. Due to the difficulty of cross-domain synthesis, it's always a problem to encourage effective feature transformation in the architecture. To solve this problem,

the encoder-decoder structure has been considered to handle the translation between sequences to sequences. Taken consideration the chronological order in the input and output sequences, recurrent structures are introduced into the architecture of encoder and decoder in order to obtain features considering temporal correlations. The Gated Recurrent Units (GRUs) [Chung et al. 2014]), as a typical structure of RNN, has been confirmed to outperform the common Long Short Term Memory (LSTM) structure in sequence learning for its fewer parameters and cheaper computation. Similar to previous work [Lee et al. 2019; Ren et al. 2020], our music-to-motion generator G consists of a GRU-based audio encoder and pose decoder, shown in Fig 7(c).

In addition, different from analyzing only the audio features outputted from the encoder in the decoding of poses, we concatenate our initial pose features with the audio features to enhance cross-domain learning for our decoder. The skip connections are also applied to intentionally add those audio-visual features into the future layers. Since our generator G is aimed to produce human motions in 3D poses, it's more difficult to accurately estimate the additional depth dimension compared with synthesizing 2D motions. Inspired by the 2D-to-3D lifting for the depth analysis in image and video processing [Kopf et al. 2020; Luo et al. 2020; Tome et al. 2017], we estimate the 2D poses first and construct the depth lifting branch to produce the 3D poses based on the 2D estimation. Taking advantage of the similarity between the 2D and 3D poses, the depth can be efficiently generated.

In the music-to-motion synthesis, not only the consistency of content style between the generated human movements and target audio sequences is needed to be supervised, but also the reality of synthesized human motions. Thus a cross-domain discriminator D_{cd} and a spatial-temporal discriminator D_{st} are built to guide the network to learn the global content consistency between the audio-visual pairs and the targeted pose flow in the spatial and temporal domain, respectively.

In the cross-domain discriminator D_{cd}, for any audio-visual pair $(A(t), V(t))$, a two-branch classification network is leveraged to judge the global style consistency. After the extraction of the audio and visual features separately, we concatenate them together and classify the similarity based on obtained audio-visual features. Such D_{cd} can improve the reasonable cross-domain translation for our generator G.

In video generation, methods constrain the temporal consistency of generated video frames by analyzing the temporally concatenated spatial features [Chu et al. 2020; Lucas et al. 2019]. Inspired by those methods, for penalizing the unrealistic produced motions, such as distorted human poses and unnatural transition between movements, the spatial-temporal discriminator D_{st} is constructed by applying a temporal progressing network. The input of motion sequences $V(t)$ are segmented evenly into 5 parts based on the time duration. By repeating the feature extraction and the concatenation of obtained spatial features progressively in chronological order, D_{st} can lead the generator G to understand the spatial-temporal relationship of the human motions in the ground truth data.

4.3 Loss Functions

Our hybrid loss function for the generator consists of 3 parts: the pose loss, the harmony loss and the GAN loss.

A multi-space pose loss is employed to regularize the realism of the estimated human movements. For distribution space, the KL loss function is applied based on the intermediate results of 2D poses in the generation process, shown as:

$$\mathcal{L}_{kl} = KL(\mathbb{P}(V_{2d}(t))||\mathbb{P}(V'_{2d}(t)) \tag{24}$$

where $\mathbb{P}$ denotes the operation that transforms the ground-truth 2D motion sequences $V_{2d}(t)$ and the intermediate output $V'_{2d}(t)$ to the probability distribution.

In pixel space, a Charbonnier-based MSE loss is established to constrain the generation of the 3D poses as:

$$\mathcal{L}_{mse} = \sum_t \sqrt{W_{tp}(t)(V(t) - G(A_f(t), V_f(t)))^2 + \epsilon^2} \tag{25}$$

where ϵ is a positive constant close to zero to soothe the gradient vanishing in training. A weight mask $W_{tp}(t) = \frac{(2-TI(t))}{2}$ is applied based on the temporal index $TI(t)$ to guide the network focus more on the generation of motions for the current meter.

As VGG networks [Simonyan and Zisserman 2014] has been widely used to assist the generation of visual features consistent with human perception [Johnson et al. 2016; Lucas et al. 2019], the Charbonnier-based VGG loss is also performed to regularize our produced human motion in the deep feature space by:

$$\mathcal{L}_{feat} = \sum_t \sqrt{(VGG(V(t)) - VGG(G(A_f(t), V_f(t))))^2 + \epsilon^2} \tag{26}$$

We assume such feature-space pose loss can capture the deep features for the motion flow and regularize the flow in the synthesized motions to be consistent with the ground truth.

Following the Lemma 3.1 discussed in Section 3.3.2, the harmony between the audio and human motion sequences can be determined by evaluating the audio-visual beat consistency, which is uniquely dependent on the hitting score h_s. Thus the harmony loss is created by formulating the function:

$$\begin{aligned}\mathcal{L}_{harmo} = -HS(p'_a(b), s'_a(b), VB(G(A_f(t), V_f(t)))) \\ +\sqrt{|M - N|}\end{aligned} \tag{27}$$

where VB denotes the extraction of attentional visual beats with the corresponding beat strengths based on the estimated human motion sequences from the generator. Such results are then sent to calculate the hitting score with the pre-computed $p'_a(b)$ and $s'_a(b)$ by HS (see Section 3.2 and 3.3 for more details). Apart from minimizing the negative hitting score, the over-frequent visual beats is penalized by adding a L1 distance comparing the number of visual beats M with N audio beats.

The GANs can learn to generate output based on the distribution in the given data during the adversarial training by solving the min-max problem:

$$\min_\theta \max_\phi \mathcal{L}_{adv}(\phi, \theta) = \mathbb{E}_x[\log D_\phi(x)] + \mathbb{E}_Y[\log(1 - D_\phi(G_\theta(Y)))] \tag{28}$$

where the ϕ and θ denote the parameters for the discriminator and generator respectively. The x represents the ground truth data while Y is the input for the generator.

Thus, our discriminator D_{cd} and D_{st} try to distinguish between the real and fake through maximizing the loss:

$$\begin{aligned}\mathcal{L}_{dcd} = \mathbb{E}[\log(1 - D_{cd}(A(t), G(A_f(t), V_f(t))))] \\ +\mathbb{E}[\log D_{cd}(A(t), V(t))]\end{aligned} \tag{29}$$

$$\mathcal{L}_{dst} = \mathbb{E}[\log D_{st}(V(t))] + \mathbb{E}[\log(1 - D_{st}(G(A_f(t), V_f(t))))] \tag{30}$$

On the contrary, our generator attempts to fool the discriminators by minimizing the function:

$$\begin{aligned}\mathcal{L}_{gan} = \mathbb{E}[-\log D_{cd}(A(t), G(A_f(t), V_f(t)))] + \\ \mathbb{E}[-\log D_{st}(G(A_f(t), V_f(t)))]\end{aligned} \tag{31}$$

In summary, combining all the loss functions above, finally the loss function for the generator can be formulated as:

$$\begin{aligned}\mathcal{L}_{total} = \lambda_{kl}\mathcal{L}_{kl} + \lambda_{mse}\mathcal{L}_{mse} + \lambda_{feat}\mathcal{L}_{feat} \\ +\lambda_{harmo}\mathcal{L}_{harmo} + \lambda_{gan}\mathcal{L}_{gan}\end{aligned} \tag{32}$$

where the λ denote the corresponding weight for each loss component.

Such harmony-aware hybrid loss function is employed in the training of our model in order to estimate natural human movements harmonized with the target audio sequences. In the next section, we provide the experimental details and compare our results with other state-of-the-art models.

5 EXPERIMENTS AND RESULTS

In this section, we describe the datasets and the implementation details used in training our final model, which is named HarmoGAN, for the task of audio-driven motion synthesis. The harmony of the resulting audio-visual pairs is assessed by using our harmony evaluation mechanism. To analyze the effectiveness of the proposed harmony loss, we conduct the ablation studies on the self-created testing dataset and compare our HarmoGAN with the start-of-the-art dance generation models based on the public music dataset quantitatively and qualitatively.

5.1 Dataset Processing

The dance dataset released in [Tang et al. 2018] is utilized to train our HarmoGAN, which includes 61 sequences of dancing videos performed by the professional dancer totaling 94 minutes and 907,200 frames in 25 fps. It provides the 3D human body keypoints with 21 joints collected from wearable devices and the corresponding audio tracks. The dance dataset contains four typical types of dance: cha-cha, rumba, tango and waltz. To save the memory cost, all the videos are resampled to 15 fps to create our own dataset. We obtain 2014 clips of concatenated audio-visual input features with the corresponding target poses are obtained from the whole dance dataset, where 214 of them are selected randomly as the self-created testing data and the rest are used for model training. All the functions that handle the extraction of musical features can be found in the Librosa [McFee et al. 2015] package.

To evaluate the harmony between the audio and synthesized audio-driven motion sequences, we test the models based on the ballroom music dataset [Gouyon et al. 2006]. The ballroom dataset extracts 698 background music in the duration of 30 seconds per clip from the online dance videos. It contains music for 7 types of dance: cha-cha, jive, quickstep, rumba, samba, tango and waltz. In each dance category, 6 audio sequences are randomly picked to form the testing dataset. The mechanism proposed in Section 3 is employed to quantify the audio-visual harmony with the use of Librosa package [McFee et al. 2015] to obtain information of auditory beats.

5.2 Implementation Details

Our HarmoGAN is implemented in PyTorch. The generator is first pretrained to prepare a reasonable initialization for the following GAN training. The pretraining ends at 225 epochs with the use of $\mathcal{L}_{pretrain}$ where $\mathcal{L}_{pretrain} = 0.1\mathcal{L}_{kl} + \mathcal{L}_{mse}$. The Adam optimizer [Kingma and Ba 2014] is utilized with batch size of 10. The initial learning rate is set to 0.001 and get decreased every 50 epochs by multiplying the factors in the order of [0.5,0.2,0.2,0.5]. Initialized with the pretrained model, GAN training is started with both the generator and discriminator networks. The weights of loss components in the hybrid loss function for our generator are set as follows: $\lambda_{kl} = 0.0001, \lambda_{mse} = \lambda_{feat} = \lambda_{gan} = 0.001, \lambda_{harmo} = 1$. The weight decay is set as 0.001 for the discriminators and 0.0001 for the generator. The learning rates for all the networks are initialized as 0.0001 and divided by 2 and 5 alternatively every 5 epochs. The optimizer and batch size are kept as same as in the pretraining. After 45 epochs of adversarial training, we find that the convergence is achieved to obtain our final HarmoGAN. It only takes 53 minutes to finish the whole training process based on the NVIDIA TITAN V GPU, which is fairly efficient.

For our proposed harmony evaluation mechanism, the constant factors λ_1 and λ_2 in Eq. (11), (13) are set as 0.1 and 1 respectively to obtain the attentional saliency. Meanwhile, the reaction delay is defined as 0.25 seconds, shown as $T_{delay} = 3.75$ frames in Eq. (18) under 15 fps. When evaluating the quantified audio-visual harmony, the β of F-score in Eq. (20) is set as 2 in order to focus more on the hit rate of audio beats.

5.3 Human Perception for Audio-visual Harmony

To confirm the assumption that the occurrence of inharmony in the audio-visual objects can be observed by human perception, a user study is conducted to test whether the participants are sensitive to the inharmonious audio-visual clips. We collect 20 dance videos which consist of 10 harmonious contents from the ground truth in the dance dataset [Tang et al. 2018] and 10 inharmonious clips created by permuting the audio or visual sequences. The invited 10 participants are required to watch the whole 20 videos and provide the perceptual harmony evaluation by picking up all the sequences that are considered as inharmony.

Fig. 8 illustrates the results of the user study, where 78% of the inharmonious videos have been accurately selected by the participants. Due to the lack of background knowledge for professional dancing, participants have difficulty in distinguishing all the inharmonious clips from the harmonious ones. Overall, we can conclude that the audio-visual harmony affects human perception and in

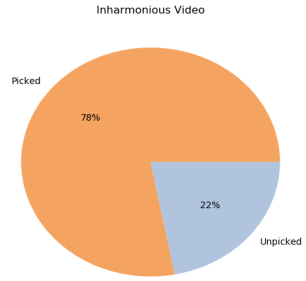

Fig. 8. The percentage of how many inharmonious videos have been accurately picked up.

most cases the occurrence of inharmony can be correctly observed and judged by perception.

5.4 Evaluation of Motion Generation

Before analyzing the performance of harmonization for the model, at first our HarmoGAN is supposed to show reasonable ability to synthesize natural motion flows based on human skeletons. To evaluate the motion generation, we test our HarmoGAN on the self-created testing dataset obtained from the dance dataset [Tang et al. 2018], which can provide ground-truth dance movements performed by the real human dancer. The Fréchet Inception Distance (FID) metric [Heusel et al. 2017] is utilized to measure the perceptual distance between the estimated motion sequences and the human ground truth. As there exists no standard for extracting features in pose sequences, in our work the VGG network [Simonyan and Zisserman 2014] is employed to obtain pose features for measuring FID. The average results are shown in Fig. 9. Compared with the references from the dance generation models [Lee et al. 2019] and [Ren et al. 2020] shown in Fig. 9, it can be implied that our model is competitive with those state-of-the-art models that can learn to synthesize human motions sharing high feature-space similarity with the real human movements in the training dataset.

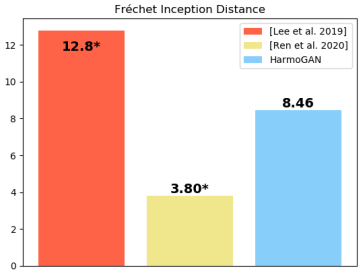

Fig. 9. The results of average FID between the generated motion sequences and the human ground truth. The lower value is better. The FID results of [Lee et al. 2019] and [Ren et al. 2020] are obtained from their papers, which are marked with stars as they are compared with their own ground truth.

Meanwhile, in Fig. 10, an example of motion sequence pairs is presented for the qualitative evaluation. The sequences of human movements synthesized by our HarmoGAN model show similar motion flows compared with the human ground truth.

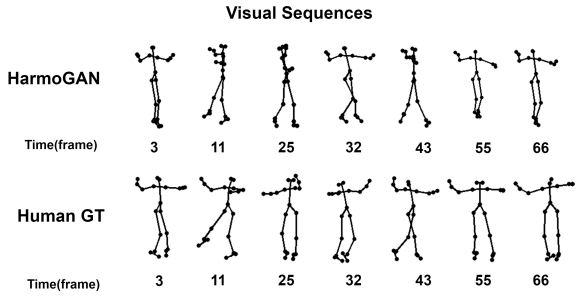

Fig. 10. The qualitative example from the dance dataset.

5.5 Evaluation of Harmony

5.5.1 Loss Ablation Study. To analyze the enhancement of audio-visual harmony after introducing the harmony loss into the network training, we conduct the ablation study to evaluate the performance of our HarmoGAN with its variant without the use of $\mathcal{L}_{harmo}$ on the self-created testing dataset, which contains relevant initial poses for the generation of motion sequences. Given the pre-computed audio beats from the music sequences, the harmony can be assessed by analyzing the audio-visual beat consistency based on the estimated human movements.

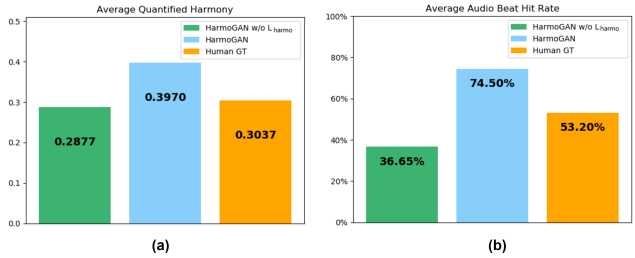

Fig. 11. The performance of audio-visual harmony tested on the self-created testing dataset with the ground truth from the real dancer and results from HarmoGAN and its variant. (a) The results for our proposed evaluation mechanism. (b) The results for the hit rate of audio beats in the music sequences. Both of the metrics are the higher the better.

Apart from the quantified harmony derived from our proposed mechanism shown in Fig. 11(a), the hit rate, which is popular in demonstrating harmonization [Lee et al. 2019], is also calculated to evaluate the performance. In Fig. 11(b) the hit rate for music beats is presented by computing the percentages of music beats that have been hit by the visual beat within the duration of the reaction delay. It can be seen that the human dancer basically hits half the music beats, which is a reasonable result considering the limited accuracy for data acquisition when obtaining motion sequences [Lee et al. 2019]. Based on Fig. 11, it's obvious that with the incorporation of $\mathcal{L}_{harmo}$ the performance of audio-visual harmony boosts compared with the variant without $\mathcal{L}_{harmo}$ and surpasses the real dancer

significantly. At the same time, it implies the $\mathcal{L}_{harmo}$ requires the "privilege" to edit the motion sequences for harmonization, even though different from the ground truth, which can explain the relatively weak performance shown for the baseline variant without the use of $\mathcal{L}_{harmo}$ as it is not designed to strictly simulate the ground truth. In all, the test on our self-created dataset demonstrate the outstanding performance of our HarmoGAN which can achieve improved audio-visual harmony between the given music sequences and the generated motions under the assistance of the harmony loss.

5.5.2 Quantitative Comparison with State-of-the-Art Deep Models. To further assess the ability of harmonization in our model, in addition to the variant we compare our HarmoGAN against the two other powerful GAN-based state-of-the-art models [Lee et al. 2019], [Ren et al. 2020] for audio-driven motion synthesis. For a fair comparison, all the models are tested on the Ballroom dataset [Gouyon et al. 2006], which is a public music dataset only providing background music for various dance types. The 42 clips of 6-second audio tracks are randomly collected from the Ballroom dataset as our testing dataset. Without any given ground-truth human movement, motion sequences in our training dataset are selected as the initial poses to generate the dance sequences.

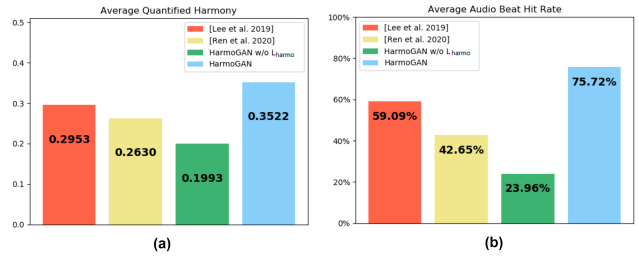

Fig. 12. The performance of audio-visual harmony for the state-of-the-art models tested on the Ballroom dataset. (a) The results for our proposed evaluation mechanism. (b) The results for the hit rate of audio beats in the music sequences. Both of the metrics are the higher the better.

In Fig. 12, the results of quantified harmony and the hit rate of music beats are demonstrated by computing the average results of 7 types of dance music. It shows that our HarmoGAN outperforms the other models distinctly in both metrics.

We also present the detailed evaluation results for each dance type, shown in Table 1 and 2. Compared with our baseline HarmoGAN without the use of $\mathcal{L}_{harmo}$, the assistance of spatial-temporal GCN in [Ren et al. 2020] may intrinsically benefit the harmonization by regularizing the hierarchical representations of skeletons in the generation of motion sequences. However, such improvement lacks robustness and is highly affected by the bias in the training dataset. The post-processing beat warper in [Lee et al. 2019] can relatively lift the performance evenly but are still limited. In comparison with the other models, our HarmoGAN can directly produce distinct and robust improvement for the audio-visual harmony that is independent of the dance types.

In addition, the cost of the tested models is analyzed based on the number of model parameters and training pairs. The number of parameters for generator in the HarmoGAN is closer to [Ren et al. 2020] and half of [Lee et al. 2019] while [Lee et al. 2019] require a 10-times larger training dataset for obtaining the final model. Thus, considering the results of harmony evaluation for each model, it reveals that our HarmoGAN can improve the performance efficiently without increasing too much cost in both the training and testing phase.

5.6 Qualitative Comparison with State-of-the-Art Deep Models

To evaluate the audio-visual harmony qualitatively, we synthesize dance videos by combining the audio sequences and the generated motions from the tested models. Then the user study is conducted to compare the perceptual harmony for the synthesized videos. 12 unprofessional participants are invited to watch the video pairs from the different 3 models. They are asked to vote which is better in terms of the audio-visual harmony blindly.

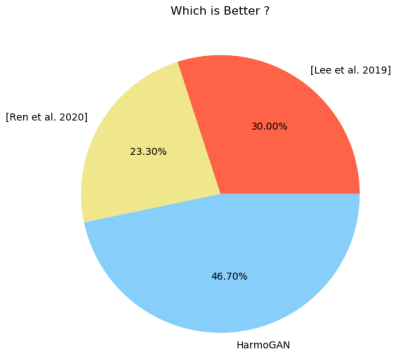

Fig. 13. 30 video pairs are created by putting the dance videos from the three models side by side. The participants are required to choose the one that agrees more with the perceptual harmony based on the assigned pair.

As shown in Fig. 13, it can be concluded that our model performs best for the perceptual harmony, which is consistent with the results from the quantitative metrics. We can also verify the assumption made before that, our proposed harmony evaluation mechanism is able to accurately reflect the perceptual audio-visual harmony to some degree.

As an example of qualitative evaluation, in Fig. 14, the generated motion sequences from the tested models are demonstrated based on the Ballroom dataset. Given the tracked audio beats, in the movements produced by [Ren et al. 2020], the distinct visual beats can hardly be perceived from the slight body swings, let alone the audio-visual consistency. When it comes to the visual results estimated from [Lee et al. 2019], reasonable visual beats can be perceived with the observation of changes between movements. However, such changes are relatively even in the whole sequences, which may suffer from the inattentional blindness and result in the perceptual inconsistency between audio-visual beats due to the misjudgments. By comparison, with the regularization of the harmony loss our HarmoGAN is able to produce distinct changes in motions close to the occurrences of the music beats in order to provide visual beats

Table 1. Quantified Harmony for 7 Types of Dance Music

Model Name	Cha-cha	Jive	Quickstep	Rumba	Samba	Tango	Waltz
[Lee et al. 2019]	0.3509	0.3359	0.2773	0.2862	0.2805	0.2657	0.2704
[Ren et al. 2020]	0.2759	0.3321	0.1625	0.2154	0.2761	0.2671	0.3122
HarmoGAN w/o $\mathcal{L}_{harmo}$	0.1983	0.2337	0.1511	0.2104	0.2012	0.1929	0.2076
HarmoGAN	**0.4097**	**0.3995**	**0.3199**	**0.3495**	**0.3468**	**0.2948**	**0.3455**

Table 2. Audio Beat Hit Rate for 7 Types of Dance Music

Model Name	Cha-cha	Jive	Quickstep	Rumba	Samba	Tango	Waltz
[Lee et al. 2019]	58.10%	59.92%	53.03%	64.91%	64.93%	58.46%	54.30%
[Ren et al. 2020]	28.13%	51.03%	33.33%	41.53%	49.07%	42.84%	52.59%
HarmoGAN w/o $\mathcal{L}_{harmo}$	22.70%	23.97%	18.18%	25.33%	29.70%	21.48%	26.36%
HarmoGAN	**74.60%**	**75.65%**	**71.22%**	**77.01%**	**83.69%**	**74.98%**	**72.90%**

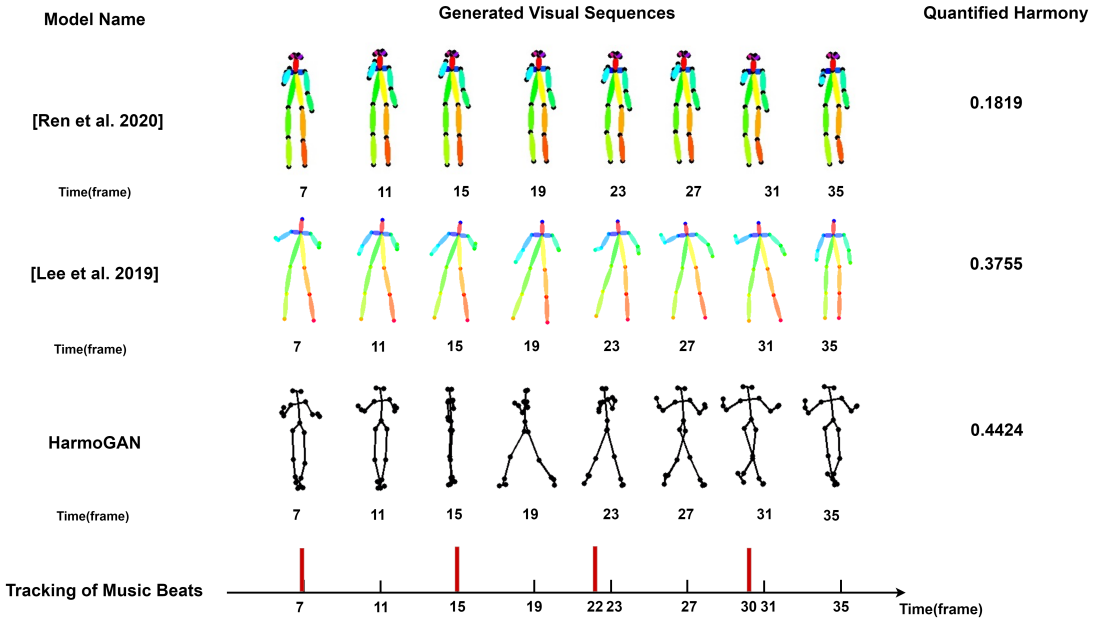

Fig. 14. The example for qualitative evaluation. The generated motion sequences are presented with the tracked audio beats to demonstrate the audio-visual harmony based on different models. The time interval is taken as 4 frames, which is in the extent of human perceptual domain under 15 fps. The occurrences of the music beats are labeled with red bars.

that can draw enough attention to favor the harmony evaluation based on the human perception.

6 LIMITATIONS AND FUTURE WORK

Based on our evaluation mechanism, as long as each audio beat is hit by a visual beat, the audio-visual harmony is considered high in the sequence pairs. Therefore, the generator may try to synthesize over-frequent visual beats to fool the metric. Even though currently a regularizer is set to penalize the total number of visual beats, it's still difficult to balance between the movement diversity and the over-frequency for visual beats. Since our visual beats are derived from the velocity sum for joints, it could be possible to regularize the diversity and frequency based on analyzing the joint sum. Also, it's always the topic to enhance the realism and temporal consistency for our produced motion sequences.

Meanwhile, as we mentioned before, human video is seen as the mainstream target when it comes to the visual generation. Our harmony evaluation mechanism can be performed on the human video by utilizing the image-to-pose networks. However, the visual sequences generated from HarmoGAN are based on the human skeleton, instead of the video frames. To harmonize the human video, we can generate the harmonious skeleton-based motion sequences first, and then edit the video based on the estimated pose priors. Since video editing is popular for facial emotion, speech and gesture, it's promising to build a multi-stage or end-to-end system to perform our proposed audio-visual harmonization based on video frames.

7 CONCLUSION

In this paper, we propose a novel evaluation mechanism to quantify the harmony between audio and visual sequences by analyzing beat consistency. We create a harmony loss based on the mechanism and incorporate such loss into the model to tackle the task of audio-driven motion synthesis. The experimental results show that our harmony loss provides significant improvement for audio-visual harmony in the generation, where the effectiveness of the proposed harmony evaluation framework has been verified.

REFERENCES

James W Acuff and William D Evridge. 1906. The SDN Theory of Music: Rudiments. (1906).

Hyemin Ahn, Jaehun Kim, Kihyun Kim, and Songhwai Oh. 2020. Generative Autoregressive Networks for 3D Dancing Move Synthesis From Music. *IEEE Robotics and Automation Letters* 5, 2 (2020), 3500–3507.

Omer Faruk Bay and Ali Bülent Usakli. 2003. Survey of fuzzy logic applications in brain-related researches. *Journal of Medical Systems* 27, 2 (2003), 215–223.

Rachele Bellini, Yanir Kleiman, and Daniel Cohen-Or. 2018. Dance to the beat: Synchronizing motion to audio. *Computational Visual Media* 4, 3 (2018), 197–208.

Bruce Benward. 2014. *Music in Theory and Practice Volume 1*. Vol. 1. McGraw-Hill Higher Education.

Sebastian Böck and Gerhard Widmer. 2013. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, Vol. 7.

Marc Cardle, Stephen Brooks, Ziv Bar-Joseph, and Peter Robinson. 2003. Sound-by-numbers: motion-driven sound synthesis.. In *Symposium on Computer Animation*. 349–356.

James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. 2016. Personalizing human video pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3063–3072.

Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai. 2014. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 12 (2014), 1859–1872.

Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. 2019. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE International Conference on Computer Vision*. 723–732.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 75–1.

Wei-Ta Chu and Shang-Yin Tsai. 2011. Rhythm of motion extraction and rhythm-based cross-media alignment for dance videos. *IEEE Transactions on Multimedia* 14, 1 (2011), 129–141.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

Grosvenor W Cooper, Grosvenor Cooper, and Leonard B Meyer. 1963. *The rhythmic structure of music*. University of Chicago Press.

Abe Davis and Maneesh Agrawala. 2018. Visual Rhythm and Beat. *ACM Trans. Graph.* 37, 4 (2018), 122–1.

Frédéric Dehais, Mickaël Causse, François Vachon, Nicolas Régis, Eric Menant, and Sébastien Tremblay. 2014. Failure to detect critical auditory alerts in the cockpit: Evidence for inattentional deafness. *Human factors* 56, 4 (2014), 631–644.

Daniel Duffell. 2005. *Making Music with Samples: Tips, Techniques & 600+ Ready-to-use Samples*. Hal Leonard Corporation.

Alessandro D'Ausilio, Eleonora Bartoli, Laura Maffongelli, Jeffrey James Berry, and Luciano Fadiga. 2014. Vision of tongue movements bias auditory speech perception. *Neuropsychologia* 63 (2014), 85–91.

Daniel PW Ellis. 2007. Beat tracking by dynamic programming. *Journal of New Music Research* 36, 1 (2007), 51–60.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. 37, 4, Article 112 (July 2018), 11 pages.

Rukun Fan, Songhua Xu, and Weidong Geng. 2011. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics* 18, 3 (2011), 501–515.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).

Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. 2006. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 5 (2006), 1832–1844.

Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2019. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)* 38, 2 (2019), 1–17.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. 2020. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 60–1.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500* (2017).

Chieh Ho, Wei-Tze Tsai, Keng-Sheng Lin, and Homer H Chen. 2013. Extraction and alignment evaluation of motion beats for street dance. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2429–2433.

Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.

Zeyu Jin, Gautham J Mysore, Stephen Diverdi, Jingwan Lu, and Adam Finkelstein. 2017. Voco: Text-based insertion and replacement in audio narration. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

Faheem Ullah Khan, Ben P Milner, and Thomas Le Cornu. 2018. Using visual speech information in masking methods for audio speaker separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 10 (2018), 1742–1754.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. 2020. One shot 3D photography. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 76–1.

Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to music. In *Advances in Neural Information Processing Systems*. 3586–3596.

Juheon Lee, Seohyun Kim, and Kyogu Lee. 2018. Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network. *arXiv preprint arXiv:1811.00818* (2018).

Minho Lee, Kyogu Lee, and Jaeheung Park. 2013. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications* 62, 3 (2013), 895–912.

Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. 2018. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–12.

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. 2020. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 40–1.

Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. 2019. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing* 28, 7 (2019), 3312–3327.

Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. 2020. Consistent Video Depth Estimation. *ACM Trans. Graph.* 39, 4, Article 71 (July 2020), 13 pages. https://doi.org/10.1145/3386569.3392377

Arien Mack, Irvin Rock, et al. 1998. *Inattentional blindness*. MIT press.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. 18–25.

Jared S Moore. 1942. Beauty as harmony. *The Journal of Aesthetics and Art Criticism* 2, 7 (1942), 40–2.

Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. 2018. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*. 362–372.

Jun-Ichi Nakamura, Tetsuya Kaku, Kyungsil Hyun, Tsukasa Noma, and Sho Yoshida. 1994. Automatic background music generation based on actors' mood and motions. *The Journal of Visualization and Computer Animation* 5, 4 (1994), 247–264.

Ferda Ofli, Engin Erzin, Yucel Yemez, and A Murat Tekalp. 2010. Multi-modal analysis of dance performances for music-driven choreography synthesis. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2466–2469.

Ferda Ofli, Engin Erzin, Yücel Yemez, and A Murat Tekalp. 2011. Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia* 14, 3 (2011), 747–759.

Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. 2016. Ambient sound provides supervision for visual learning. In *European conference on computer vision*. Springer, 801–816.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7753–7762.

Antoine Picot, Sylvie Charbonnier, and Alice Caplier. 2011. On-line detection of drowsiness using brain and visual information. *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans* 42, 3 (2011), 764–775.

Gardner Read. 1964. *Music notation: a manual of modern practice*. Technical Report.

Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. 2020. Self-supervised Dance Video Synthesis Conditioned on Music. In *Proceedings of the 28th ACM International Conference on Multimedia*. 46–54.

Eric D Scheirer. 1998. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America* 103, 1 (1998), 588–601.

Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. 2006. Dancing-to-music character animation. In *Computer Graphics Forum*, Vol. 25. Wiley Online Library, 449–458.

Daniel J Simons and Christopher F Chabris. 1999. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *perception* 28, 9 (1999), 1059–1074.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*. Springer, 1015–1021.

Statista. 2018. *Share of adults in the United States who have ever taken a selfie as of August 2018, by age group*. https://www.statista.com/statistics/304861/us-adults-shared-selfie-generation/

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.

Taoran Tang, Jia Jia, and Hanyang Mao. 2018. Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*. 1598–1606.

Thomas Theodoridis, Theocharis Chatzis, Vassilios Solachidis, Kosmas Dimitropoulos, and Petros Daras. 2020. Cross-Modal Variational Alignment of Latent Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 960–961.

Denis Tome, Chris Russell, and Lourdes Agapito. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2500–2509.

Petra Vetter, Fraser W Smith, and Lars Muckli. 2014. Decoding sound and imagery content in early visual cortex. *Current Biology* 24, 11 (2014), 1256–1262.

Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. 2020. RGB2Hands: real-time tracking of 3D hand interactions from monocular RGB video. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

Nelson Yalta, Shinji Watanabe, Kazuhiro Nakadai, and Tetsuya Ogata. 2019. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. 2020. ChoreoNet: Towards Music to Dance Synthesis with Choreographic Action Unit. In *Proceedings of the 28th ACM International Conference on Multimedia*. 744–752.

YouTube:littlesiha. 2019. *Bassa Sababa - Netta - Just Dance 2020*. https://www.youtube.com/watch?v=Wdtau3nUPy0

Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*. 398–407.

Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–10.