Learning Music-Dance Representations Through Explicit-Implicit Rhythm Synchronization

Jiashuo Yu, Junfu Pu, Ying Cheng, Rui Feng, and Ying Shan, Senior Member, IEEE

Abstract—Although audio-visual representation has been proven to be applicable in many downstream tasks, the representation of dancing videos, which is more specific and always accompanied by music with complex auditory contents, remains challenging and uninvestigated. Considering the intrinsic alignment between the cadent movement of the dancer and music rhythm, we introduce MuDaR, a novel Music-Dance Representation learning framework to perform the synchronization of music and dance rhythms both in explicit and implicit ways. Specifically, we derive the dance rhythms based on visual appearance and motion cues inspired by the music rhythm analysis. Then the visual rhythms are temporally aligned with the music counterparts, which are extracted by the amplitude of sound intensity. Meanwhile, we exploit the implicit coherence of rhythms implied in audio and visual streams by contrastive learning. The model learns the joint embedding by predicting the temporal consistency between audio-visual pairs. The music-dance representation, together with the capability of detecting audio and visual rhythms, can further be applied to three downstream tasks: (a) dance classification, (b) music-dance retrieval, and (c) music-dance retargeting. Extensive experiments demonstrate that our proposed framework outperforms other self-supervised methods by a large margin.

Index Terms—Multimodal learning, music and dance, self-supervised learning.

Manuscript received 1 February 2023; revised 4 May 2023; accepted 4 June 2023. Date of publication 9 August 2023; date of current version 31 July 2024. This work was supported by the National Natural Science Foundation of China under Grant 62172101. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Nicu Sebe. Corresponding authors: Junfu Pu; Rui Feng.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by (Name of Review Board or Committee) (IF PROVIDED under Application No. xx, and performed in line with the (Name of Specific Declaration).

Jiashuo Yu was with the ARC Lab, PCG Tencent, Shenzhen 518000, China. He is now with the School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, Shanghai 200438, China (e-mail: jsyu19@fudan.edu.cn).

Junfu Pu is with the Applied Research Center, PCG, Tencent, Shenzhen 518000, China (e-mail: jevinpu@tencent.com).

Ying Cheng is with the Academy for Engineering and Technology, Fudan University, Shanghai 200438, China (e-mail: chengy18@fudan.edu.cn).

Rui Feng is with the School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, Shanghai 200438, China (e-mail: fengrui@fudan.edu.cn).

Ying Shan is with the Applied Research Center, PCG, Tencent, Shenzhen 518000, China (e-mail: yingshan@tencent.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMM.2023.3303690, provided by the authors.

Digital Object Identifier 10.1109/TMM.2023.3303690

Fig. 1. Illustration of a dancing clip with its corresponding audio rhythms. The motion of the dancer is in sync with the music rhythms, which means the starting keypoint of the dancing motion lies on the temporal position of the music onset maximum. The correlations between motion and onset features are used as explicit training signals, which together with the implicit rhythm consistency contribute to self-supervised representation learning.

I. INTRODUCTION

RECENT years have witnessed the rapid growth of dancing video amounts on online video-sharing websites. The demand for automatically processing dancing videos based on content is hence becoming increasingly stronger. In literature, many prior works have achieved promising results on music-dance tasks, e.g., music-driven dance generation [1], [2], [3], [4] and music-dance alignment [5]. However, these methods are usually task-specific and have a strong dependence on labeled data, which restricts their applicability to a large extent.

A more practical way is to leverage the correlation between auditory and visual contents as the proxy to train task-agnostic models without human annotation. The derived representation can further be applied to a variety of downstream tasks with minor modifications. Though recent audio-visual pre-trained models [6], [7], [8], [9], [10], [11], [12], [13] have delivered an impressive performance, these methods are not generalizable for dancing scenarios. The reason is that the visual contents of dancing videos consist of complex motions and characteristics. Meanwhile, dancing music, the auditory counterpart, embodies various attributes like lyrics, rhythms, tempos, etc. These fine-grained features are neglected by existing self-supervised audio-visual methods, leading to the consequence that meticulous conjunctions crossing modalities are not fully probed. As a result, a novel multimodal representation learning strategy is necessitated for dancing videos.

In this article, we focus on the rhythm of dancing videos, which represents the amplitude of music intensity and dancing motions. As illustrated in Fig. 1, dancers tend to rhythmically move by music at the same frequency to make dances more coordinated. Such temporal key moments of dancing motions are
usually in sync with music rhythms, which can be reflected by the audio onset features. Inspired by this insight, we argue that the pattern of dancing motions, termed as the visual rhythms, should be synchronous with music rhythms. This temporal correspondence can further be utilized as the supervised signals. Specifically, the synchronization of music and dance rhythms is conducted both explicitly and implicitly. Since rhythms are implied in the auditory and visual contents, we leverage the consistency between audio and visual streams to perform implicit rhythm alignment. Moreover, the audio onset, which refers to the beginning of a musical note, can be regarded as the music rhythm. In this work, we train the model to learn visual rhythms based on appearance and motion cues, which are expected to be aligned with music rhythms in an explicit way. Finally, the entire model is unified by joint training for multimodal representation as well as visual rhythm extraction, which can further be applied to many downstream applications. We conduct experiments on three practical tasks: dance classification, music-dance retrieval, and music-dance retargeting. The comparison with other self-supervised audio-visual methods verifies the effectiveness of our framework. Our contributions are summarized as follows:

- We devise a novel self-supervised representation learning strategy for dancing videos, which performs the music-dance rhythm synchronization both in explicit and implicit ways.
- We propose a joint music-dance representation and a dance rhythm extractor favorable for music-dance understanding and re-creation tasks.
- Extensive experiments on three downstream tasks, i.e., dance classification, music-dance retrieval, and dance-music retargeting, verify the effectiveness and generalizability of our model on music-dance scenarios.

II. RELATED WORKS

Audio-visual representation learning aims to learn the multisensory embeddings favorable for downstream tasks. Several existing methods [6], [7], [8], [9], [10], [11], [14], [15], [16] focus on the characteristics of auditory and visual streams in a given video. The audio and visual contents in videos are usually corresponding and synchronous, and these natural inter-modality correlations can be utilized as the supervisory signal for large-scale self-supervised training. Apart from such correspondence, some works investigate the discriminative information across the modalities. [12], [17] explore the intrinsic differences between modality-aware semantics and propose a cross-modal deep clustering method to perform cross-modality supervision. [13] propose a cross-modal active contrastive coding strategy to fully explore the diversity between positive and negative samples. Moreover, some methods [18], [19], [20], [21], [22], [23], [24], [25] explore the relationship of visual dynamic motions and audio signals. The correspondence between the dynamic motions of objects and sound sources can be used for self-supervised object detection. In this article, we propose a unified framework to combine these knowledge priors above, where the temporal correspondence is utilized for implicit rhythm synchronization, and the dynamic motions corresponding to music are exploited for explicit rhythm keypoint alignment.

Dance with music is a fundamental and challenging video category, where the movements of dancers are complex and the patterns of music and dance are diverse. Hence, models are required to explore more detailed information by capturing fine-grained features. Attempts in the representation learning field focus on the uni-modal music embeddings [26], [27], [28], [29], [30], [31]. Since music can be handled as a sequence of tokens, an ordinary pipeline is to train pretrained models that have achieved promising performance in natural language processing, such as BERT [32]. However, the embeddings of the visual counterpart: dancing, as well as the joint music-dance representation that captures cross-modal music-dance interactions, are rarely investigated. For the multimodal dance-music applications, many works [1], [2], [3], [4] tackle the problem of music-dance generation, that is, generating a sequence of dancing videos based on the given music. Some methods [5] work on the audio-visual alignment, which tries to align mismatched dancing and music streams by learning frame-level dense correspondence. The most relevant to our work is visual rhythm and beat prediction [33], [34], [35], where visual rhythms and beats are extracted based on dancing motions or dancer skeletons. In our article, we ameliorate visual rhythm predictions by involving more characteristics and further leverage the synchronization of rhythms as a training proxy for multimodal representation learning.

III. APPROACH

As shown in Fig. 2, MuDaR conducts rhythm synchronization via a two-pathway architecture. The explicit pathway temporally aligns the music and dance rhythm keypoints, while the implicit part is implemented by identifying the synchrony between audio and visual streams.

A. Explicit Rhythm Synchronization

1) Music Rhythm Detection: For the music rhythms, a straightforward idea is to utilize the onset feature, which indicates sudden increases in music volumes. Following [36], we first obtain the spectrograms by conducting time-windowed FFT to the audio signal:

\[
X(n, k) = \sum_{q=-N/2}^{N/2} v(hn + q)w(q)e^{-2\pi i q k},
\]  

(1)

where \(X(n, k)\) denotes the \(k\)th frequency bin of time \(n\); \(v(\cdot)\) is the audio signal; \(h\) denotes the hop size; \(w(\cdot)\) is the Hamming window, \(N\) is the time window size.

Then we compute the bin-wise difference between spectrograms to get the spectral flux, which indicates the magnitude of audio signals. The onset envelope, a positive 1D feature, is computed by summing the positive spectral flux:

\[
OE(n) = \sum_{k=-N/2}^{N/2} \max(0, |X(n, k)| - |X_{ref}(n - \mu, k)|),
\]  

(2)

where \(n, k\) denotes the temporal position and bin number, respectively; \(OE(n)\) is the onset envelope at time step \(n\); \(X_{ref}\) denotes the maximum-filtered spectrogram proposed in [36]; \(\mu\) is the time lag.
Finally, we pick the local maximum of the onset envelope to detect the discrete onset feature. The local maximum will be selected as an onset only if it is some threshold above the local average value. The entire rhythm extraction procedure can be derived by the Python package LibRosa [37] with the default hyperparameter settings.

2) **Dance Rhythm Detection:** Optical flow indicates the direction and magnitudes of visual motions. Therefore, we estimate the dense optical flows as the initial step of dance rhythm detection. The lightweight PWC-Net [38] is selected as the backbone optical flow network. To be specific, PWC-Net extracts feature maps in different resolutions via a feature pyramid extractor consisting of several stacking convolution layers, then warps features of the current frame toward the previous frame and constructs a cost volume. Finally, the optical flow is predicted by a multi-layer CNN estimator and refined by a dilated context network.

**Histograms of optical flow:** We utilize the Histogram of Oriented Optical Flow (HOOF) [39], a non-Euclidean feature, to represent motions in a non-linear manifold that is more robust to illumination changes and provides a global description of the motion. In particular, HOOF is computed by the expected Optical Flow (HOOF) network. To highlight the rhythm keypoints in the temporal dimension, we introduce the long-range temporal interactions by the multi-head self-attention mechanism [40]. The refined visual features are put into the visual rhythm estimator as the infusion.

We also argue that the inherent coherence of music-dance signals can be utilized for visual rhythm prediction. The audio-guided spatial-channel attention mechanism [41] is leveraged to explore the relationship between auditory and visual features. We use audio features \( f_a \) extracted by the audio encoder of the implicit rhythm synchronization pathway (which will be introduced in Section III-B), and the audio-guided visual features can be computed as follows,

\[
\begin{align*}
H(n, k) &= \sum_{x,y} M_t(x, y) \mathbb{1}_\theta(P_t(x, y)), \\
\mathbb{1}_\theta(\phi) &= \begin{cases}
1, & \text{if } |\theta - \phi| \leq \frac{2\pi}{B}, \\
0, & \text{otherwise},
\end{cases}
\end{align*}
\]

where \( M_t(x, y) \in \mathbb{R}^{H \times W} \) denotes the magnitude of optical flow in \( t \)-th time, which is computed by \( \sqrt{x^2 + y^2}; P(t) \in \mathbb{R}^{H \times W} \) indicating the angle of optical flow with x-axis in time step \( t \) is calculated by \( \tan^{-1} \frac{y}{x} \), \( \mathbb{1}_\theta(\phi) \) is an indicator function; \( B \) denotes the number of bins.

**RGB injector:** Directly using HOOF for rhythm estimation could lead the prediction to be highly dependent on the quality of optical flow. Considering that the RGB features may also contain contributing information, we propose an RGB injector to infuse visual cues to motion features as an enhancement. To be specific, we use the feature map \( f_{rgb} \) generated by the encoder of PWC-Net in the explicit rhythm synchronization pathway. Then \( f_{rgb} \) is tiled and linearly projected to reduce the dimensionality. To highlight the rhythm keypoints in the temporal dimension, we introduce the long-range temporal interactions by the multi-head self-attention mechanism [40]. The refined visual features are put into the visual rhythm estimator as the infusion.

We use audio features \( f_a \) extracted by the audio encoder of the implicit rhythm synchronization pathway and the audio-guided visual features can be computed as follows,

\[
\begin{align*}
\mathbf{w}_{a,rgb}^{c} &= \sigma(W_1^c \mathbf{U}_1^{c}(\rho_a(f_a \odot U_{rgb}^{c}(f_{rgb})))), \\
f_{a,rgb} &= \sum_{i=1}^{k} w_{a,rgb}^{c,i} f_{rgb}, \\
\mathbf{w}_{a,rgb}^{a} &= \text{softmax} \left( \delta(W_2^a(U_a^{a}(f_a) \odot U_{rgb}^{a}(f_{rgb}))) \right), \\
f_{a,rgb} &= \sum_{i=1}^{k} w_{a,rgb}^{a,i} f_{rgb},
\end{align*}
\]
where $U_{\text{rgb}}^r, U_{\text{rgb}}^f$ are linear layers with non-linearity activation; $W_1, W_2$ are learnable parameters; $\sigma$ indicate the sigmoid function; $\rho$ denotes global average pooling; $\delta$ is the hyperbolic tangent function; $w_{\text{ar}g}^c, w_{\text{a}r\text{bg}}^a$ are the channel and spatial attention map, respectively.

However, the prediction of dance rhythms cannot rely on the corresponding music when applied to downstream tasks. The performance significantly declines when audio signals are available during training while missing in inference. To this end, we refer to Dropout [42], a simple yet effective method which initial goal is to prevent the model from being dependent on specific neurons. Inspired by this paradigm, we propose an audio dropout gate, which randomly drops partial auditory inputs in a mini-batch with constant ratios during training. By doing so, MuDaR can detect visual rhythms both with and without music during inference. The entire procedure can be formulated as:

$$f_{\text{rgb}1}, f_{\text{rgb}2} = \text{AudioDropout} \left( f_{\text{rgb}}, p \right),$$

$$\text{AudioDropout}(f, p) = f\left[ b \ast p; \right]: f\left[ b \ast p \right],$$

$$f_{\text{a}r\text{gb}} = \text{AGVA}(f_{\text{rgb}}, f),$$

$$f_{\text{rgb}2} = \text{Linear} \left( \text{Tile}(f_{\text{rgb}2}) \right),$$

$$f_{\text{a}r\text{gb}} = \text{Concat}(f_{\text{a}r\text{gb}}, f, f_{\text{rgb}2}),$$

$$f_{\text{inj}} = \text{Att}(f_{\text{a}r\text{gb}}, f_{\text{rgb}}, f_{\text{rgb}2}),$$

where $f_{\text{inj}}$ denotes the output of the RGB injector; AGVA denotes the audio-guided visual attention introduced in (5)–(8); Att denotes the multi-head self-attention; $b$ denotes the mini-batch size; $p$ is the audio dropout rate.

**Visual rhythm estimator**: To explore the magnitude of visual appearance, we compute the first-order difference of motion and RGB features, respectively, and combine them via linear projection. Then a linear layer is used for binary classification,

$$p_t^f = \sigma(W_e(U_{\text{mot}} f_{\text{mot}} + U_{\text{inj}} f_{\text{inj}}) + b_e),$$

where $f_{\text{mot}}, f_{\text{inj}}$ are the first-order difference of motion and injected features; $U_{\text{mot}}, U_{\text{inj}}$ are linear layers with ReLU [43]; $\oplus$ denotes concatenation across the channel dimension; $W_e, b_e$ are parameters of classifier; $\sigma$ is the sigmoid function.

**B. Implicit Rhythm Synchronization**

We claim that rhythms of music and dance are implied in the auditory and visual features, and the synchronization of music and dance streams can also be considered as the implicit version of rhythm synchronization. Following [9], [10], [11], [25], we employ Audio-Visual Correspondence (AVC) and Audio-Visual Temporal Synchronization (AVTS) as the pretext tasks for representation learning. Specifically, we leverage the raw videos as the positive samples, while creating asynchronous samples by performing temporal shifts and creating uncorrelated samples by combining visual and audio streams from different videos. Then the model is required to predict whether the assembled sample is synchronous (for AVTS) and corresponding (for AVC) or not, thereby performing self-supervised training.

One problem of this unsupervised paradigm is the construction of *false-negative samples*. For the AVC task, if we randomly sample a sequence of music that is the same as the raw dancing video, the newly-assembled sample is actually music-dance corresponding. To address this problem, we compute the rhythm similarity score of the raw and newly selected audio streams by calculating the coincidence of rhythm positions, which is formulated as follows:

$$s_{\text{rhy}} = \sum_{t=1}^{T} \left| O_t^{\text{pos}} - O_t^{\text{neg}} \right| - \alpha T,$$

where $s_{\text{rhy}}$ is the rhythm similarity score; $O_t^{\text{pos}}$ denotes the positive music onset features; $O_t^{\text{neg}}$ indicates the negative onsets; $T$ denotes the length of dancing videos; $\alpha \in [0, 1]$ is a threshold hyperparameter. The negative sample will be selected only when the score is positive. For the AVTS task, the rhythms of dancing videos are sometimes periodic. Therefore, if we shift the current rhythm point exactly to another point afterward, the audio and visual rhythms will be re-aligned with certain time-lagged. To this end, we conduct additional constraints to prevent the shifted size from being the multiple of rhythm interval. Though the rhythm interval tends to be diverse, most onset peaks are concurrent with music beats, which are distributed with a constant temporal gap. In our work, the eighth note (quaver) is selected as the basic beat unit, and the shifted frames $f_{\text{sft}}$ cannot be the multiple of the frame number identical to an eighth note,

$$f_{\text{sft}} \mod \left( k_{\text{fps}} * \frac{60}{k_{\text{bpm}}} * \frac{1}{2} \right) \neq 0,$$

where $k_{\text{fps}}$ denotes the frame sample rate; $k_{\text{bpm}}$ denotes the number of quarter notes (crotchets) per minute. Since we use the eighth note as the base beat unit, the number of beat units per minute will be $2 * k_{\text{bpm}}/60$. After performing sampling constraints, the implicit rhythm synchronization can be trained in a self-supervised manner. To perform coherent optimization, we leverage the outputs of the visual encoder in the explicit pathway as visual inputs. We also build a similar audio encoder including several 2D convolution layers. Then auditory and visual features are put into $K$ stacking transformer [40] layers, respectively. We do not involve any modality interaction to make our framework generalizable for single-modality downstream tasks. Each transformer layer is constructed by the encoder part of the raw transformer architecture [40], in which features temporally interact to fully explore the long-range temporal correlation. Finally, the triplet loss [44] is used for the refined auditory and visual features, which enlarges the distance between visual features and negative auditory features, while reducing the gap between visual features and positive audio features.

**C. Optimization**

For implicit synchronization, training a binary classifier and adopting the binary cross-entropy loss as the learning objective is a natural choice. However, this method suffers from the difficulty of training convergence. Some prior methods [6], [9] opt for the contrastive loss as a substitution. In this work, the dense optical flow estimation requires high computation costs, thus the performance will be limited when the batch size is small. In this
article, we choose the triplet loss [44] with an offline negative sample strategy. Specifically, we put a pair of positive audio and visual samples into the two-stream implicit pathway. Then we choose the negative audio samples from the whole dataset, which are brought into the audio streams together with the positive audio sample. The triplet loss function tries to make the refined positive audio feature \( f'_{a:pos} \) closer to the refined visual features \( f'_{v} \), while enlarging the distance between negative audio features \( f'_{a:neg} \) and \( f'_{v} \),

\[
L_{im} = \sum_{i=1}^{N} \left( \| f'_{v} - f'_{a:pos} \|_2^2 - \| f'_{v} - f'_{a:neg} \|_2^2 + \alpha \right), \tag{18}
\]

where \( L_{im} \) denotes the implicit loss, \( \alpha \) is the margin hyperparameter; \( [\cdot]_{+} \) denotes picking the positive results, and \( N \) denotes the number of triplet pairs.

For explicit synchronization, the straightforward way of using binary cross-entropy loss may cause overwhelming negative predictions during training due to the unbalanced distribution of rhythm and non-rhythm temporal keypoints. To address this problem, we adopt the focal loss [45], which addresses the class imbalance by introducing a weighting hyperparameter \( \alpha \), and focuses more on hard samples rather than easier ones via the focusing parameter \( \gamma \). The explicit loss \( L_{ex} \) can be formulated as,

\[
L_{ex} = -\alpha_{t}(1 - p_{t}^{e})^{\gamma} \log(p_{t}^{e}), \tag{19}
\]

where \( p_{t} \) is the visual rhythm prediction.

Finally, the optimization is conducted in a joint-training manner, the overall objective function is formulated as,

\[
L = \lambda_{1}L_{im} + \lambda_{2} \frac{1}{N} \sum_{t=1}^{N} L_{ex}, \tag{20}
\]

where \( \lambda_{1}, \lambda_{2} \) are weighted hyperparameters to control the summation of loss terms, \( N \) is the frame number.

IV. EXPERIMENTS

A. Music-Dance Representation Learning

Dataset: MuDaR can be trained by large-scale dancing videos without any annotation. The only requirement for unlabeled data is that dancers should twist their bodies following the rhythms of the music, which is a fundamental rule for dancing. In this article, we collect 194,407 dancing videos from an information feed platform and online video-sharing application. The training data can be separated into four categories: dancing with pop music, hip-hop, modern dance, and K-pop with nearly equal proportions. These category labels are unnecessary both in training and inference.

Implementation details: All dancing videos are cropped to 8 s clips. To remove background noises, we discard the beginning and end parts of each dance, and select 4 seconds before the middle of the video together with 4 seconds after. Visual frames are sampled with a frame rate of 8 fps. Then frames are resized to 256 × 256. To maintain the camera view unchanged, we do not apply any visual transformation for data augmentation. Audios are sampled with a rate of 16 kHz, and we extract Mel-spectrograms, onset envelopes, onset maximum, and beat using the python package LibROSA [37]. For the implicit synchronization pathway, the audio inputs are the concatenation of Mel-spectrograms, beat, and onset envelopes, while the onset maximums are used as the ground truth of rhythms for explicit synchronization. We use Adam [46] optimizer with the initial learning rate of 1e-4. The entire model is trained on 64 NVIDIA Tesla V100 GPUs for 50 epochs with a batch size of 384. For the implicit synchronization, we adopt the curriculum learning strategy as [9]. AVC is used for the first 35 epochs, while the more challenging pretext task AVTS is used for the rest of the training process.

Evaluation of MuDaR: We randomly sample 4,407 videos of the original 194,407 videos for evaluation, while using the remaining 190 k videos for self-supervised training. To fully investigate the performance of the explicit rhythm pathway, we propose three ablated models for comparison. “MuDaR w/ CRF” denotes an ablated model that replaces the binary prediction layer with a BiLSTM-CRF [47] layer, which is composed of a bidirectional long-short memory network [48] and a conditional random field [49] module. This ablated model regards the visual rhythm prediction task as a sequence labeling problem and considers the inner relationship of the consecutive frames. “MuDaR w/ regression” takes the visual rhythm prediction task as a regression problem. Instead of using local maximum over onset envelopes as the audio ground truth, it directly minimizes the distance between model output and the onset envelope curve. The visual rhythms are then generated by picking the local maximum of the model output. “MuDaR w/ BCE” replaces the binary focal loss with the raw binary cross-entropy loss. As shown in Table I, we report recall and precision as the evaluation metrics. Results show that “MuDaR w/ BCE” achieves higher recall compared with raw MuDaR, while obtaining low precision, proving that the cross-entropy loss suffers from the imbalanced class distribution. “MuDaR w/ regression” achieves poor performance on all metrics. We argue that though the temporal position of music and dance rhythms are coincident, it’s illogical to force the intensity of visual rhythms to be identical to the audio rhythms. On the contrary, the full MuDaR model achieves promising results with respect to all metrics, showing the effectiveness of detecting visual rhythms. “MuDaR w/ CRF” performs slightly worse than the full model. This suggests that the BiLSTM-CRF performs unsatisfied when only two unbalanced categories exist for sequence labeling.

| Table I | The Evaluation of Visual Rhythm Prediction Compared With Two Ablated Models |
|---------|--------------------------------------------------|
| Approach | Data Size | Recall | Precision |
| MuDaR w/ BCE | 190k | 82.2% | 37.0% |
| MuDaR w/ regression | 190k | 43.8% | 51.9% |
| MuDaR w/ CRF | 190k | 70.2% | 79.8% |
| MuDaR (full) | 47k | 59.6% | 73.1% |
| MuDaR (full) | 140k | 71.4% | 78.9% |
| MuDaR (full) | 190k | 73.2% | 82.0% |

We also report the performance of our full model with different dataset sizes.
We evaluate the performance of dance classification and pre-trained and Multisensory and AVID-CMA and λ and rhythm S[46]
[11] R[17] [11] [50]
dataset. This dataset contains more than K. We argue that models pre-trained on the generalized audio-visual dataset lack the discrimination of music and dance patterns, hence are unsuitable for the dance-music tasks. Moreover, our model outperforms all self-supervised baselines trained on the music-dance dataset by a large margin, showing the effectiveness of our proposed framework in tackling dance-music tasks. Last but not least, the performance on the full pre-trained dataset (190 k) surpasses the state-of-the-art supervised method, which indicates the rationality of our self-supervised paradigm. We also conduct ablation studies by removing the visual and auditory rhythms. Results show that MuDaR performs slightly worse without rhythm involvement. However, the decline is not significant, revealing that the performance of dance classification mainly depends on the embeddings generated by the implicit synchronization pathway.

B. Dance Classification

**Downstream architecture:** We first evaluate the performance of MuDaR on the task of dance classification. To fully utilize the outputs of MuDaR (audio and visual embeddings, visual rhythms), we concatenate the visual embeddings and rhythms as visual outputs, and the auditory embeddings and onset as audio outputs. Then the audio and visual output features are put into a two-stream classifier. To be specific, audio and visual features are first put into two fully-connected layers with non-linear activation. Then the audio and visual features are concatenated and integrated by temporal pooling. Another two fully-connected layers with non-linearity are used for prediction.

**Dataset:** We evaluate the performance of dance classification on the Let’s Dance [50] dataset. This dataset contains more than 1,400 10-second dancing videos on a variety of 16 dance categories. Since a small part of the videos is unavailable on the Internet, we downloaded 1,262 dancing videos and randomly split the training, validation, and testing set with the proportion of 80%/10%/10% following [50], [51]. We employ the same data preprocessing procedure as the self-supervised representation training.

**Implementation details:** Adam [46] optimizer is used for training with the initial learning rate of 4e-3, which is degraded by 10 after 20 and 50 epochs. The model is trained with a batch size of 64 for 70 epochs.

**Experimental results:** We compare MuDaR with two supervised methods: Temporal Three-Stream CNN [50] and Multimodal Dance Recognition [51]. To find out whether the dance-music representation is really needed, or whether it can be replaced by larger-scale generalized audio-visual pre-trained models, we also conduct experiments with the traditional audio-visual pre-trained models XDC [17] and AVID-CMA [16] by fine-tuning them on the dance classification dataset. Furthermore, we also re-implement three audio-visual self-supervised methods: Multisensory [11], AVTS [9], and LLA [10], which are trained on the same large-scale dancing dataset used for MuDaR to make a fair comparison. Results shown in Table IX indicate that all audio-visual pre-trained models perform worse than our proposed framework pre-trained on the music-dance dataset, even in a far larger-scale dataset (240 K, 2 M, and 65 M vs. 190 K). We argue that models pre-trained on the generalized audio-visual dataset lack the discrimination of music and dance patterns, hence are unsuitable for the dance-music tasks. Moreover, our model outperforms all self-supervised baselines trained on the music-dance dataset by a large margin, showing the effectiveness of our proposed framework in tackling dance-music tasks. Last but not least, the performance on the full pre-trained dataset (190 k) surpasses the state-of-the-art supervised method, which indicates the rationality of our self-supervised paradigm. We also conduct ablation studies by removing the visual and auditory rhythms. Results show that MuDaR performs slightly worse without rhythm involvement. However, the decline is not significant, revealing that the performance of dance classification mainly depends on the embeddings generated by the implicit synchronization pathway.

### Table II

| Approach       | Dataset       | Acc.    | Manner      |
|----------------|---------------|---------|-------------|
| Castro et al.  | /             | 70.2%   | Supervised  |
| MDL [51]       | /             | 77.0%   | Supervised  |
| LLA [10]       | AudioSet (240k) | 66.9%   | S-Sup.+ ft  |
| XDC [17]       | AudioSet (2M)  | 71.1%   | S-Sup.+ ft  |
| XDC [17]       | IG-Kinetics (65M) | 72.3%   | S-Sup.+ ft  |
| AVID-CMA [16]  | AudioSet (2M)  | 75.9%   | S-Sup.+ ft  |
| Multisensory   | M&D (190k)    | 71.4%   | S-Sup.+ ft  |
| AVTS [9]       | M&D (190k)    | 68.3%   | S-Sup.+ ft  |
| LLA [10]       | M&D (190k)    | 73.0%   | S-Sup.+ ft  |
| MuDaR (full)   | M&D (47k)     | 76.1%   | S-Sup.+ ft  |
| MuDaR (full)   | M&D (140k)    | 79.4%   | S-Sup.+ ft  |
| MuDaR w/o rhy  | M&D (190k)    | 80.8%   | S-Sup.+ ft  |
| MuDaR (full)   | M&D (190k)    | 81.7%   | S-Sup.+ ft  |

"S-Sup." Denotes models pretrained on the dataset specified in the second column, then fine-tuned on the dance classification dataset. M&D denotes the music-dance dataset.

### C. Music-Dance Retrieval

**Downstream architecture:** We also evaluate our model on the music-dance retrieval task. For the cross-modal retrieval, we compute the similarity scores of rhythms and embeddings, respectively. To be specific, for the embedding $E_a$, $E_v$ and rhythm $R_a$, $R_v$, we compute the embedding similarity matrix $S_{v}$ consisting of similarity scores of $E_a$ and $E_v$, then compute the rhythm similarity matrix $S_{r}$ via $R_a$ and $R_v$. Finally, the hybrid similarity matrix can be computed by the weighted summation of $S_{v}$ and $S_{r}$:

$$S_{hyb} = \lambda_3 S_{v} + (1 - \lambda_3) S_{r},$$

where $\lambda_3$ is the hyperparameter. The top-K indices of the hybrid matrix are preserved as the retrieval result.

**Dataset and task formulation:** We conduct experiments on the Dance-50 [5] dataset for cross-modal retrieval. Dance-50 contains 50 hours of K-pop dancing videos collected from online video platforms. This dataset is originally used for audio-visual alignment without any category annotation available. To perform music-dance retrieval, we selected 22 pieces of dancing music that appeared with high frequency, where each song corresponds to 15–40 dancing videos from different dancers. We collected 400 labeled videos in total while using another 1833 unlabeled dances as irrelevant data in the retrieval database. Since dances in the Dance-50 dataset are long videos of more than 30 seconds, we pre-process each annotated video by clipping 24 seconds from the beginning. For the unlabelled videos, we separated each video into several 24 s dance clips. In this way, we got 400 labeled and 4,066 unlabeled dancing clips as the retrieval dataset. The objective of this task is to retrieve related dances given the targeted music.

**Experimental results:** We leverage the self-supervised methods Multisensory [11], AVTS [9], and LLA [10] pre-trained on music-dance dataset, and the available XDC [17] and AVID-CMA [16] models pre-trained on traditional audio-visual datasets as baselines. For these baselines, since visual rhythms

---

**TABLE II**

**The Dance Classification Accuracy on the Let’s Dance Dataset**

| Approach   | Dataset       | Acc.    | Manner      |
|------------|---------------|---------|-------------|
| Castro et al. [50] | / | 70.2% | Supervised |
| MDL [51] | / | 77.0% | Supervised |
| LLA [10] | AudioSet (240k) | 66.9% | S-Sup.+ ft |
| XDC [17] | AudioSet (2M) | 71.1% | S-Sup.+ ft |
| XDC [17] | IG-Kinetics (65M) | 72.3% | S-Sup.+ ft |
| AVID-CMA [16] | AudioSet (2M) | 75.9% | S-Sup.+ ft |
| Multisensory | M&D (190k) | 71.4% | S-Sup.+ ft |
| AVTS [9] | M&D (190k) | 68.3% | S-Sup.+ ft |
| LLA [10] | M&D (190k) | 73.0% | S-Sup.+ ft |
| MuDaR (full) | M&D (47k) | 76.1% | S-Sup.+ ft |
| MuDaR (full) | M&D (140k) | 79.4% | S-Sup.+ ft |
| MuDaR w/o rhy | M&D (190k) | 80.8% | S-Sup.+ ft |
| MuDaR (full) | M&D (190k) | 81.7% | S-Sup.+ ft |

"S-Sup." Denotes models pretrained on the dataset specified in the second column, then fine-tuned on the dance classification dataset. M&D denotes the music-dance dataset.
are unavailable, we only use the similarity matrix between embeddings for retrieval. We also investigate the role of visual rhythms via the ablated model “MuDaR w/o rhy”, and we report average top-k retrieval performance (R@K) and top-k precision (P@K) as the evaluation metrics. As shown in Table III, results first reveal that on the music-dance retrieval task, where the downstream task is fulfilled in an unsupervised way, all traditional audio-visual pre-trained models show poor performance and are incapable of retrieving satisfying results due to the semantically similar nature of different dances and songs. Subsequently, MuDaR outperforms all music-dance pre-trained models by a large margin. Especially, our framework outperforms baseline LLA by 15% on R@5 and P@10, and 13% on R@1, which shows the effectiveness on unsupervised downstream tasks. Besides, the performance significantly declines after removing the rhythm information, indicating that rhythms are favorable for cross-modal retrieval.

### D. Music-Dance Retargeting

**Task formulation:** Music-dance retargeting aims to synthesize a new dancing video via two mismatched music and dance clips. This task can be combined with the music-dance retrieval task as an automatic soundtrack generator, which synthesizes rhythm-aligned dancing videos after obtaining related music via cross-modal retrieval. Concretely, given a sequence of visual frames $V$, and a music clip $A$ from different dancing videos, the objective of music-dance retargeting is to combine $V$ and $A$ into a natural new dancing video. This task evaluates the performance of rhythm extraction thoroughly since we conduct the retargeting by warping the visual rhythms $R_v$ into the alignment with the auditory counterpart $R_a$.

**Downstream architecture:** Due to the difference between the temporal lengths of $V$ and $A$, we try to align as many rhythm points as possible via acceleration and shifting. We propose three retargeting patterns: temporal shifting, temporal acceleration, and dynamic time warping. For the temporal shifting, we propose a sliding window, which size is equal to the length of the shorter sequence, and put it on the rhythm of the longer duration. We reserve the cropped clip with the most similar number of rhythm points, then align the starting rhythm points of two sequences. The second pattern is temporal acceleration, where we compute the temporal interval $I$ between two adjacent rhythm points: $I_{k}^{m} = P_{k+1}^{m} - P_{k}^{m}$, where $m \in \{a, v\}$ denotes the modality index, $k$ indicates the rhythm keypoint index, and $P_{k}^{m}$ denotes the temporal position of the $k$th rhythm in the modality $m$. Then we manage to make $I_{k}^{a}$ and $I_{k}^{v}$ equal for all rhythm indices by interpolating and removing visual frames. The third manner is dynamic time warping (DTW) [52], [53], which is capable of selecting the most suitable aligning strategy. Specifically, DTW aims to find the optimal warping path that aligns $R_v$ and $R_a$ with minimum rhythm mismatch cost, which is defined by the accumulated distance between the rhythm keypoints. The cumulative distance $c$ in position $(i, j)$ can be computed using dynamic programming:

$$c(i, j) = d(i, j) + M\{c(i - 1, j - 1), c(i - 1, j), c(i, j - 1)\},$$

$$d(i, j) = |P_i - P_j|,$$

where $M(\cdot)$ denotes the minimize function; $i, j$ are indices of rhythms; $P_i$ denotes the temporal position of the $i$th rhythm point. The warping path determines the detailed frame accelerate strategy with minimum warping costs, thereby resulting in optimal retargeting performance.

**User study:** Since there is no feasible quantitative metric for music-dance retargeting, we conduct a user study to compare our method with the only available baseline method visbeat [34]. We survey 20 experienced dancing viewers, which are required to give judgments to 10 retargeted demos about synthesized fluency, music-dance rhythm consistency, and viewing naturalness. Results shown in Table IV show that our model outperforms the baseline method by a large margin, indicating that our proposed method is capable of generating high-quality retargeted video-music pairs.

**Qualitative results:** We also provide some qualitative results to show the performance of our method as shown in Fig. 3. We compare the raw videos with retargeted videos via temporal acceleration, temporal shifting, and dynamic time warping (DTW), where frames with colored boxes denote visual rhythm positions. All dancing videos are randomly selected from the test set of the representation learning dataset. For the synthetic videos by temporal shifting, the inner rhythm may be mismatched for we only perform alignment on the first rhythm points, which makes the synthetic video uncoordinated. Retargeting videos by temporal acceleration lead parts of the synthetic video extremely fast or...
This is because when the gap between $I_k^v$ and $I_k^a$ is quite large, we need to interpolate or remove a large number of visual frames, thereby seriously affecting viewing fluency. On the contrary, the synthetic video by DTW achieves excellent rhythm alignment compared with videos generated by temporal shifting, and the synthetic frames are more consistent than the temporal acceleration production. To make a qualitative comparison, we also provide some retargeted demos in the supplementary material synthesized by MuDaR and visbeat [34], which detects visual rhythms in a non-deep-learning based manner and retargets via temporal warping. Results show that the synthesized video using MuDaR and our DTW strategy achieves better viewing fluency and rhythm correspondence than the compared baselines. We strongly suggest readers refer to the provided demos since the qualitative comparison is the only possible way to evaluate the performance of our method. We compare the raw videos with retargeted videos via temporal acceleration, temporal shifting, and dynamic time warping (DTW). Results show that videos synthesized by DTW achieve both optimal rhythm alignment and viewing fluency.

E. Ablation Studies

Ablations on MuDaR architecture: We conduct more ablation experiments to investigate the effectiveness of different components, and the results are shown in Table V. We propose several ablated models for music-dance representation learning. “MuDaR w/o injector” denotes removing the entire RGB injector module, “MuDaR w/o sa” denotes RGB injector that only consists of the temporal self-attention. “MuDaR w/o audio-guide” is an ablated model where the temporal self-attention mechanism is removed. “MuDaR w/o HOOF” indicates replacing the HOOF layer with linear projection blocks, which transforms the original optical flow to the features with identical dimensions as the output of the HOOF layer. Results show that our full model outperforms “MuDaR w/o injector” and “MuDaR w/o HOOF” by a large margin, suggesting the necessity of the HOOF layer and the RGB feature infusion. The considerable performance gap between the full model and “MuDaR w/o sa” and “MuDaR w/o audio-guide” also proves the advantage of our proposed modules.

Impact of audio dropout gate: We also conduct additional ablation studies on the audio dropout rate $p$. We set the audio dropout rate $p$ to zero for all experiments. The best results are achieved when $p$ is set to 0.2, indicating the importance of the audio dropout gate in retargeting tasks.

---

**Table V: Ablation Studies on the MuDaR Architectures**

| Approach                | Recall | Precision |
|-------------------------|--------|-----------|
| MuDaR w/o injector     | 64.3%  | 70.8%     |
| MuDaR w/o sa           | 71.4%  | 76.0%     |
| MuDaR w/o audio-guide  | 70.2%  | 75.3%     |
| MuDaR w/o HOOF         | 56.8%  | 67.2%     |
| MuDaR (full)           | 73.2%  | 82.0%     |
dropout rate to different values during training, while only using visual frames as model input for inference. As shown in Table VI, our full model performs slightly better compared with models with \( p = 0.3 \) and \( p = 0.7 \). Moreover, the performance significantly declines when setting \( p = 1 \) due to the unbalanced modality distribution between the training and inference procedure.

**Impact of Transformer architecture:** To fully investigate the performance of the transformer architecture, we perform ablation studies on the dance classification tasks with different implicit synchronization settings. “MuDaR w/ conv” denotes replacing the temporal convolution layers to the convolution layers used in [9]. “MuDaR w/ trans*3” and “MuDaR w/ trans*12” denotes stacking a different number of transformer layers. As shown in Table VII, our model outperforms “MuDaR w/ conv” by 5.2%, showing that the transformer-based architecture achieves better results than the convolutional network. The comparisons among transformer-based architectures show that “MuDaR w/ trans*12” perform worse than our 6-layer version. We argue that the parameter amount of our 6-layer model is enough to serve the dance classification problem, and a larger pre-trained model may lead to the under-fit over the downstream dataset. This also suggests that straightforwardly introducing parameters cannot bring more benefits, and setting layer number 6 balances the classification performance and compute complexity.

**Ablations on training objectives:** As we mentioned above, using InfoNCE loss with a large batch can introduce more negative samples, which brings better performance yet also lead to a large training cost. When training on a Tesla V100 GPU, only 3 video-music pairs can be put into a mini-batch, which means the selection of negative samples is highly limited if we adopt an online negative-picking strategy. We also conduct ablation studies to use the contrastive InfoNCE as training objectives, and the results are shown in Table VIII, which indicates that using InfoNCE with a small batch size results in poor performance.

### Table VI

| Dropout Rate | Recall | Precision |
|--------------|--------|-----------|
| 0            | 71.2%  | 77.5%     |
| 0.3          | 71.8%  | 78.9%     |
| 0.7          | 70.9%  | 76.0%     |
| 1            | 68.1%  | 74.7%     |

| Dropout Rate | Recall | Precision |
|--------------|--------|-----------|
| 0.5 (ours)   | 73.2%  | 82.0%     |

### Table VII

| Approach       | Classification Accuracy |
|----------------|-------------------------|
| MuDaR w/ conv  | 76.5%                   |
| MuDaR w/ trans*3 | 79.8%                |
| MuDaR w/ trans*12 | 80.6%               |
| MuDaR (full)   | 81.7%                   |

### Table VIII

| Approach            | Data Size | Recall | Precision |
|---------------------|-----------|--------|-----------|
| MuDaR w/ InfoNCE    | 190k      | 54.3%  | 67.1%     |
| MuDaR (full)        | 190k      | 73.2%  | 82.0%     |

**Ablations on lambda parameters:** We also provide additional experiments of the weighted hyperparameter \( \lambda_1, \lambda_2 \) in Table IX, where results show that adopting \( \lambda_1 = 1, \lambda_2 = 5 \) achieves the optimal performance.

### Table IX

| \( \lambda_1 + \lambda_2 \) | Recall | Precision |
|-----------------------------|--------|-----------|
| 1+1                         | 69.0%  | 67.4%     |
| 1+3                         | 71.9%  | 68.4%     |
| 1+5                         | 73.2%  | 64.9%     |
| 3+1                         | 78.6%  | 80.9%     |
| 5+1                         | 82.0%  | 87.2%     |

**V. LIMITATION AND FUTURE WORKS**

MuDaR has two main limitations. One is that the dense flow estimation requires high computation costs and GPU usage, further resulting in slow running speed. Moreover, the high memory utilization restricts the batch size of a single GPU, which further limits the performance of self-supervised contrastive learning since the number of negative samples is crucial for the online negative sample strategy. Another limitation is that the explicit visual rhythm prediction module, the major component of MuDaR, is highly relied on the performance of optical flow estimation. This leads the model to behave unsatisfactorily in some specific scenarios, such as multi-dancer videos, where optical flows can be extremely messy. Therefore, taking more visual features, such as motion trajectory, human gesture, and human skeleton into consideration could be a promising way to enhance the robustness and generalization of our model.

In this article, we conduct experiments on three downstream tasks. However, we argue that MuDaR is capable of generalizing to more dance-music applications, such as music-dance realignment [5] and music-inspired dance synthesis [3]. Visual rhythm features can also be utilized to uni-modal dancing tasks, including visual rhythm prediction [33], automatic dance scoring, etc. We will further extend our model to more application scenarios.

### VI. CONCLUSION

In this article, we propose a novel self-supervised Music-Dance Representation learning framework via the synchronization of music and dance rhythms both explicitly and implicitly. We first explicitly extract and align the visual and auditory rhythms of the dancing videos based on the magnitudes of dancer motions and music contents. This part of the framework can also be utilized as a dancing rhythm extractor. We then leverage the contrastive learning strategy to synchronize the auditory and visual streams of the dancing videos, which can also be viewed as an implicit music-dance rhythm synchronization. Our model outperforms other self-supervised methods in downstream tasks by a large margin, verifying the effectiveness of our framework.
from both the video understanding and re-creation perspectives. We hope our work could arouse the generous growth of works on music-dance self-supervised learning, and we believe the paradigm and benchmarks we proposed can contribute to more significant dance-music research.

REFERENCES

[1] W. Zhuang, C. Wang, J. Chai, Y. Wang, M. Shao, and S. Xia, “Music2Dance: DanceNet for music-driven dance generation,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 18, no. 2, pp. 1–21, 2022.
[2] H.-Y. Lee et al., “Dancing to music,” Adv. Neural Inf. Process. Syst., vol. 32, 2019.
[3] X. Guo, Y. Zhao, and J. Li, “DanceNet: Music-inspired dancing video synthesis,” IEEE Trans. Image Process., vol. 30, pp. 5559–5572, 2021.
[4] R. Huang et al., “Dance revolution: Long-term dance generation with music via curriculum learning,” in Proc. Int. Conf. Learn. Representations, 2021.
[5] J. Wang, Z. Fang, and H. Zhao, “Alignnet: A unifying approach to audio-visual alignment,” in Proc. IEEE Winter Conf. Comput. Vis. Appl., 2020, pp. 3309–3317.
[6] J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 251–263.
[7] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proc. IEEE Conf. Comput. Vis., 2017, pp. 609–617.
[8] R. Arandjelovic and A. Zisserman, “Objects that sound,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 435–451.
[9] B. Korbar, D. Tran, and L. Torresani, “Cooperative learning of audio and video models from self-supervised synchronization,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 7774–7785.
[10] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, “Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning,” in Proc. ACM Int. Conf. Multimedia, 2020, pp. 3884–3892.
[11] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 631–648.
[12] D. Hu, F. Nie, and X. Li, “Deep multimodal clustering for unsupervised audiovisual learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern. 2019, pp. 9248–9257.
[13] S. Ma, Z. Zeng, D. McDuff, and Y. Song, “Active contrastive learning of audio-visual video representations,” in Proc. Int. Conf. Learn. Representations, 2021.
[14] J.-B. Alayrac et al., “Self-supervised multimodal versatile networks,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 25–37.
[15] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, “Self-supervised learning of audio-visual objects from video,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 12475–12486.
[16] P. Morgado, N. Vasconcelos, and I. Misra, “Audio-visual instance discrimination with cross-modal agreement,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 12475–12486.
[17] H. Alwassel et al., “Self-supervised learning by cross-modal audio-video clustering,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 9758–9770.
[18] Z. Barzelay and Y. Y. Schechner, “Harmony in motion,” in Proc. IEEE Conf. Comput. Vis. Pattern. 2007, pp. 1–8.
[19] H. Zhao et al., “The sound of pixels,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 1735–1744.
[20] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, “The sound of motions,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 1735–1744.
[21] D. Hu et al., “Discriminative sounding objects localization via self-supervised audiovisual matching,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 10077–10087.
[22] C. Gan, H. Zhao, Z. Pan, D. Cox, and A. Torralba, “Self-supervised moving vehicle tracking with stereo sound,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 7053–7062.
[23] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, “Music gesture for visual sound separation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern. 2020, pp. 10478–10487.
[24] P. Morgado, Y. Li, and N. Vasconcelos, “Learning representations from audio-visual Spatial alignment,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 4733–4744.
[25] T. Afouras, Y. M. Asano, F. Fagan, A. Vedaldi, and F. Metze, “Self-supervised object detection from audio-visual correspondence,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 10575–10586.
[26] H.-H. Wu, M. Fuentes, and J. P. Bello, “Exploring modality-agnostic representations for music classification,” 2021, arXiv:2106.01149.