学号 1706010120

年级 2017

河海大學

本科毕业论文

基于动态频域分解与自监督跨模态感知的乐队指挥动作生成

专　　业 计算机科学与技术

姓　　名 陈德龙

指导教师 刘　凡

评 阅 人 徐媛媛

２０２１年６月

中国　南京

BACHELOR'S DEGREE THESIS OF HOHAI UNIVERSITY

Music-driven Conducting Motion Generation based on Motion Decomposition and Self-supervised Cross-modal Perceptual Loss

College : College of Computer and Information

Subject : Computer Science and Technology

Name : Delong Chen

Directed by : Fan Liu

NANJING CHINA

郑 重 声 明

本人呈交的毕业论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本设计（论文）的研究成果不包含他人享有著作权的内容。对本设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本设计（论文）的知识产权归属于培养单位。

本人签名：_____ 日期：_____

摘　要

音乐与人体动作之间的内在关联性一直以来都在被广泛研究。最近，许多学者成功地使用深度学习模型进行了舞蹈动作或乐器演奏动作的生成，但很少有人关注乐队指挥的动作。本文聚焦于这一任务，以音乐为条件控制信号，生成与之节奏同步且语义相关的指挥动作。具体地，本文首先提出动作动态频域分解，以音乐节奏为依据将指挥动作分解为高频分量与低频分量，突破了现有动作分解方法中连贯性与协调性不可兼得的局限。随后，本文融合自监督学习与感知损失两大新兴技术，实现了跨模态的指挥动作感知，并在此基础上使用对抗损失与提出的同步损失训练指挥动作生成模型。为了提供可靠的数据支撑，本文还基于目标检测与姿态估计算法从在线视频平台收集并构建了一个大规模指挥动作数据集。在数据集上的实验证明了本文提出的方法可以生成自然、美观、多样、且与音乐同步的指挥动作。

关键词：感知损失；生成对抗网络；自监督学习；音乐-动作同步性学习；乐队指挥

ABSTRACT

Music-motion correlation attracts much attention. Many recent works focus on the motion generations for dancers and musicians, but few works for the conductors. In this paper, we concentrate on the music-driven conducting motion generation approach, which aims to generate the conducting motions according to a piece of music. Specifically, we first propose a motion decomposition method to represent the macro- and micro- motions by decomposing the movements of articulations in the temporal frequency domain. The low- and high- frequency signals are utilized to represent the macro- and micro- motions, respectively. We then feed the signals to a two-branch model and associate each branch with a specific kind of motion. The composition of the two branches produces the final conducting motion. Finally, to train an effective model, we propose an noval sync loss, where the perceptual features are learned from the contrastive correlations between the music and the conducting motions. We also build a large-scale dataset on conducting motions namely *ConductorMotion100*. The extensive experiments demonstrate that our proposed approach achieves an impressive performance in generating the conducting motions.

Key words: perceptual loss; adversarial learning; music motion synchronization; orchestral conductor

目　录

第一章　绪论

1.1　研究背景

指挥是交响乐团的灵魂。自中世纪欧洲教堂唱诗班到二十一世纪的现代音乐，指挥技术与艺术不断发展，已经成为一门内容丰富的学科[1]。指挥的肢体语言复杂多变[2]，需要在乐团演奏时实时地传达节拍、力度、情感、演奏法等多种信息[3]，且同时保持一定的风格与美感。近年来,随着深度学习算法理论的发展与计算性能的飞速提升，人工智能领域的学者已经成功地对多种人类艺术进行建模与学习。深度学习已经能生成包括诗歌艺术、绘画艺术、音乐艺术、舞蹈艺术在内的多种人类艺术形式。

然而，学界对于指挥艺术的建模研究还比较初步，且主要面向判别类的任务，例如节拍跟踪、拍式识别、演奏法识别、情感识别等。对于生成式任务，即音乐驱动的指挥动作生成任务，Wang 等人[4] 在 2003 年提出了首个指挥动作生成方法。随后，几种基于规则的生成方法[5][6][8]陆续被提出，但这些方法无法灵活地学习真实指挥动作的内在规律，导致生成动作重复性强，多样性差。Dansereau 等人[9]在 2013 年提出了一种基于机器学习的指挥动作预测方法以应对云合奏中的网络延迟问题，但该方法仅能向前预测很短的时间。音乐驱动的指挥动作生成任务可以归入人体动作的条件生成的范畴。其主要包括由语音、音乐等音频模态的条件控制信号生成人体的说话[10][11]，舞蹈[12]或乐器演奏[13]的姿态。由于同为音频模态到姿态模态的生成任务，研究音乐驱动的指挥动作生成时也有必要借鉴人体动作条件生成领域的研究成果并参考其现有方法的不足。

1.2　国内外研究现状

1.2.1　基于深度学习的人体动作生成

基于深度学习的人体动作条件生成指的是以音频模态的控制信号为条件，通过深度学习方法生成与之同步且语义相关的人体动作，包括语音动作生成（speech gesture generation）[28]，说话人脸生成（talking face generation）[34],音乐驱动的舞蹈生成（music-driven dance generation）[12]，以及音乐驱动的乐器演奏动作生成（music-dirven instrument playing motion generation）[13]等多种应用。完

成这类任务需要同时从音频与动作两个模态中进行学习，并试图建立两者之间复杂的依赖关系。更重要的是，给定音频条件，可以对应着多种合理的人体动作。也就是说，从音频到人体动作的生成是一个病态（ill-posed）问题，是一对多的生成任务。这些因素都增加了人体动作的条件生成任务的难度。在深度学习兴起之前，人体动作的条件生成主要通过基于检索的方法完成，即在构建的音频-动作数据库里检索与给定音频条件最相似的样本作为模型输出。这类方法生成动作的多样性差，时间复杂度高，且受数据库规模的严重制约。近年来，深度生成式模型不断发展，并陆续地被应用于人体动作条件生成任务，取得了一定的成功。接下来，本文将对这些方法进行介绍与分析。

（1） 基于确定性模型的生成方法

确定性模型是最简单直接的人体动作条件生成方法。这类方法将音频输入至生成器中，计算生成动作与真实动作在样本空间中的欧氏距离（L_1或L_2）作为损失函数来指导模型的学习。换句话说，这类方法将人体动作的条件生成问题建模为一个回归问题。给定音频控制条件，模型生成的结果是唯一的，因此被称为确定性模型。在文献[20]中，Yalta 等人使用二维卷积神经网络（Convolutional Neural Network，CNN）提取音频频谱特征，随后使用长短期记忆网络（Long Short Term Memory，LSTM）学习动作与音频特征之间的时间依赖性。随后在文献[31]中，作者增加了一项对比损失（contrastive loss）使生成的动作与音频同步，但这样的限制也使得模型生成的动作高度重复，缺乏多样性。Li 等人[27]也采用了类似的 CNN-LSTM 模型架构，但该方法使用 MIDI 格式的音频作为输入，因此拓展性较低。此外，也有许多研究人员也尝试仅使用 LSTM 模型或门控循环单元（Gated Recurrent Unit，GRU[123]）进行生成[24][37][33][25][19][26][21]。但由于所采用的梅尔倒谱系数（Mel-scaleFrequency Cepstral Coefficients，MFCC）等音频特征在时间上是不平滑的，这类方法生成的动作往往有着难以消除的抖动现象。此外，由于训练集与测试集的差异（train-test divergence）导致的误差积累，LSTM 模型在测试时生成的动作经常在一定时间步后趋于不动[43][12][45]。此外，在这一任务上 LSTM 模型的训练也十分不稳定[24]。相比之下，基于卷积神经网络的生成模型[36][32]则可以在一定程度上避免这些问题。此外，最近兴起的 Transformer 模型[122]也被应

用在这一任务上[43][44]，并取得了令人印象深刻的生成效果，但他们也仍都是确定性模型，无法从给定音频输入生成多样的动作。

（2） 基于概率模型的生成方法

近年来，研究人员们逐渐意识到确定性模型与回归损失在人体动作的条件生成这一病态生成任务上难以避免的缺陷，因此基于概率模型的方法吸引了越来越多的注意力。与确定性模型不同，概率模型试图对真实人体动作的条件分布进行建模。这使得模型不必严格生成与真实样本完全一致的动作，在一定程度上缓解了回归损失带来的过渡平滑的问题[28]。在近几年中被广泛研究的多种深度生成模型，例如变分自编码器、生成对抗网络、流模型等，都被成功地应用在了这一任务上。然而，绝大部分基于概率模型的生成方法仍然保留着回归损失[45][35][29][23][30][42][34]。本文认为，回归损失与对抗损失是冲突的：由于需要权衡数据集中的不同样本，回归损失的全局最优解是过渡平滑的生成结果，而判别器可以轻松的识别生成结果中过渡平滑的特征，从而产生与回归损失相冲突的梯度。现有方法中，只有文献[22]和[28]完全避免了回归损失的使用。然而，文献[22]面向的是较为简单的头部姿态生成任务（仅输出头部朝向的 3 维向量，无需学习关键点之间的空间关系），而文献[28]引入了先验的“姿态阶段”（gesture phrase）监督信息，带来了额外的数据标注需求，也限制了该方法的应用场景。

（3） 基于动作分解的生成方法

动作分解是人体动作条件生成领域内另一个值得注意的研究趋势。其中，舞蹈动作的时域分解是最常用的分解方法。这类方法首先将复杂的舞蹈动作在以节拍为依据分为基础的分解动作单元，再学习如何对这些单元进行排列组合。这种动作单元在不同的文献中被冠以不同的名称，例如 Choreographic Dance Units (CAUs) [38]，Pose Fragments [39]，Dnce Phrases [40]等。然而，这些方法中基础动作单元是固定而不可学习的，且往往需要专业的编舞人员定义。这一方面引入了额外的数据标注工作量，另一方面也限制了生成模型的多样性（当所标注的基础动作单元较少时，这类方法便退化为基于检索的方法）。该问题的解决方法在于引入可学习的基础动作单元：Lee 等人[12]和 Li 等人[46]分别令模型首先学习低层次的舞步单元（Dance Units）[12]或关键姿态（Key Pose）[46]，再训练高层次的模型来编排学得的舞步单元或填充关键姿态之间的动作。这类方法的缺陷在于，其第

一阶段的学习样本是通过自动音乐节拍检测来分割得到的，而节拍检测的误差以及真实数据中音乐与舞蹈的细微错位都会限制模型的学习效果。与以上时域分解方法相对应，还有一些学者提出了空间域的动作分解。在文献[41][13]中，研究人员将小提琴家的演奏姿态数据分为左手、右手以及其余身体部分，并使用不同模型分别学习这些动作的生成。然而，这类方法的一个潜在假设是人体动作的不同部分是相互独立的，而事实并非如此。因此这一缺陷会使得该类方法生成的动作缺乏协调性。

（4） 发展动态与分析

在生成模型架构的选择上，从 GRU，LSTM，CNN-LSTM 到最近的 Transformer 模型，研究人员进行了许多不同的尝试。此外，领域内最值得注意的演变在于研究重点逐渐从确定性模型转向概率模型。本文认为，在样本空间中存在着三个概率分布：生成动作的分布P_G，真实动作的分布P_{data}以及真实动作关于音频控制信号的条件分布P_c。借鉴生成对抗网络（Generative Adversarial Net，GAN[97]）等模型在图像生成领域的成功经验，现有的概率模型能够较好地将P_G拉向P_{data}。然而，如何施加合适的约束，使得P_G同时符合P_c仍然是一个开放的问题。如前文所述，现有的概率模型难以避免使用回归损失。其原因可能在于这些方法没有找到合适的途径来为模型施加与同步性的监督约束，以使得P_G趋向于P_c。因此，在人体动作的条件生成问题上，如何施加使得生成动作与输入条件相适应的监督约束，是生成高质量动作的关键点与难点。

1.2.2 乐队指挥动作感知

乐队指挥动作感知指的是通过计算模型从指挥动作中提取语义信息，与舞蹈信息检索（Dance Information Retrieval, DIR）[63]类似，并与音乐信息检索（Music Information Retrieval, MIR）[48][58] 领域高度相关。乐队指挥动作感知技术的研究最早可以追溯到 20 世纪 80 年代[47]。几十年来，随着计算机算力的增强与机器学习理论方法的进步，乐队指挥动作感知技术不断发展，并受到了越来越多学者的关注。其包含的任务主要包括指挥节拍跟踪（beat tracking）[52][51]、拍式识别（rhythmic pattern recognition）[53][60]、演奏法识别（articulation recognition）[51][59][7][55][54]、情感识别（sentiment recognition）[56]等。乐队指挥动作感知算法的一般流程是先将原始的指挥动作信号转换为动作特征，再结合分类或回归算法来

完成目标任务，其关键点在于如何有效提取指挥动作中的高层次语义信息。现有的动作特征提取方法主要包括基于力度特征的方法[52][55][51]与基于机器学习的方法[7][54][56][53][60]。乐队指挥动作感知技术有着广泛的应用场景，例如基于节拍跟踪与虚拟现实（Virtual Reality, VR）技术的交互式的虚拟乐团[62][49][57][50][61]，以及本文的基于指挥动作生成与姿态迁移的指挥视频生成等。

（1） 基于力度特征的乐队指挥动作感知

由于指挥动作的力度与节拍、演奏法、以及音乐情感的强烈程度高度相关，指挥动作的速度、加速度信息可以有效地被用作指挥动作的特征表示。例如，Sarasua 等人[52]使用指挥的左右手加速度及其垂直分量作为动作特征，设计了一个基于阈值的判定规则来进行指挥节拍跟踪。随后在文献[7]中，作者引入高斯混合模型（Gaussian Mixture Model, GMM）根据力度特征进行演奏法识别。类似地，Lee 等人在文献[54]中使用 K-Means 聚类算法来识别指挥动作的情感信息并进行可视化。同时，在文献[55]中作者还发现使用动作速度与加速度的均值与方差作为特征，结合朴素贝叶斯分类器，也可以有效地完成完成演奏法识别。Cosentino 等人[51]提出使用主成分分析（Principal Component Analysis, PCA）对指挥动作的加速度信息降维以完成动作的节拍与乐曲速度的识别。以上这些基于力度特征的方法的可解释性强，但仅使用力度特征会遗失指挥动作的空间信息。此外，由于不具备学习能力，这类方法的准确率与泛化能力较差。

（2） 基于机器学习乐队指挥动作感知

相比之下，引入机器学习能够大大提升模型提取更高层次的语义特征的能力，并更有效地学习指挥动作的内在规律，从而完成更为复杂的乐队指挥动作感知任务。这类方法包括基于隐含马尔可夫模型（Hidden Markov Model, HMM）的方法[56]、基于动态时间规整（Dynamic Time Warping, DTW）的方法[53][60]、LSTM 的方法[59]等。Karipidou 等人在文献[56]中构建了一个在不同情感下指挥相同乐曲的指挥动作数据集，并利用 HMM 模型识别出动作中蕴含的不同情感。Schramm 等人[53]与 ChinShyurng 等人[60]将 DTW 分类器用于分辨指挥动作中的 2/4，3/4，4/4 拍式。此外，在文献[59]中，在众多时间序列预测任务上取得了广泛成功的 LSTM 也被应用于识别指挥动作中音乐结构、力度与演奏法信息。

1.2.3 音乐驱动的指挥动作生成

与以上乐队指挥动作感知任务相比，音乐驱动的指挥动作生成的研究则相对不足。在 2003 年，Wang 等人[4]提出了首个音乐驱动的指挥动作生成方法。该方法设计了一个基于核的 HMM 模型，从音高、响度与节拍三种特征预测指挥动作。几种基于规则的生成方法[5][6][8]陆续被提出，但他们生成动作的多样性较差。Dansereau 等人[9]提出了一种基于机器学习的指挥动作预测方法以应对云合奏中的网络延迟问题，但该方法仅能向前预测很短的时间。现有的这几种音乐驱动的指挥动作生成方法的生成效果如图 1.1 所示。目前，近年来在各种生成任务中广泛成功的深度学习方法尚未被应用至音乐驱动的指挥动作生成任务上。

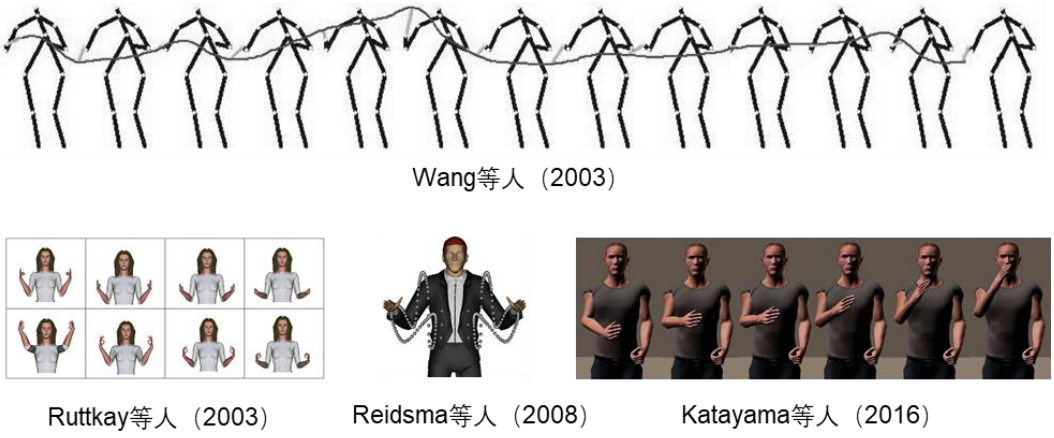


图 1.1 现有的四种音乐驱动的指挥动作生成方法效果图

纵观 40 年来人工智能领域对乐队指挥动作的研究进展，可以总结出以下发展趋势与不足。1）乐队指挥动作感知中使用的特征提取技术逐渐从单一力度特征转变为基于学习的特征。众多机器学习算法陆续被不同学者尝试应用于这一任务。然而由于指挥动作是包含不同频率成分的连续信号，基于卷积的信号处理方法以及卷积神经网络理应在乐队指挥动作感知任务上取得成功，但目前基于卷积的 R 方法研究还很少；2）在深度学习领域，跨模态条件生成任务（例如语音-动作生成、音乐-舞蹈生成）在近年来取得了较大的进展，这些方法在指挥动作生成任务上应当有着很大的潜力，但目前尚没有基于深度学习的指挥动作生成方法被提出；3）现有乐队指挥动作数据集的收集往往基于较为昂贵，使用不便且难以推广的设备，例如动作捕捉系统（motion capture equipment）、惯性测量单元（Inertial Measurement Units, IMU）以及深度相机或 RGB-D 相机等。此外，指挥动作数据

集的标注（例如对节拍、演奏法、情感等）也十分费力，且需要专业的音乐知识。这两个原因限制了更大规模乐队指挥数据集的构建，也限制了机器学习或深度学习模型在面向乐队指挥动作的任务上进一步的发展与应用。

1.3 本文研究内容

本文的研究面向指挥动作的条件生成这一任务，覆盖了乐队指挥动作感知、人体动作生成、感知损失、跨模态自监督学习等多个前沿领域，涉及了生成对抗网络、对比学习、音乐信息检索、姿态估计、目标检测、姿态迁移等多种深度学习算法。本文首先提出动作动态频域分解，以音乐节奏为依据将指挥动作分解为高频分量与低频分量，突破了现有动作分解方法中连贯性与协调性不可兼得的局限。随后，本文融合自监督学习与感知损失两大新兴技术，实现了跨模态的指挥动作感知，并在此基础上使用对抗-感知损失训练指挥动作生成模型。此外，本文还通过收集互联网中大量的指挥视角演出录像视频，使用目标检测、姿态估计等技术，以较低的成本高效地构建了大规模的指挥动作数据集。具体地，本文本文的研究内容可以总结为以下三点：

（1）研究指挥动作的动态频域分解：针对时域与空间域动作分解对于分解后子序列相互独立这一不合理假设，从指挥动作在频域可分的特性出发，以音乐节奏为依据动态地分解指挥动作，在有效减轻生成模型学习难度的同时，保证了指挥动作在时间域与空间域的上下文信息不丢失，突破动作分解中连贯性与协调性不可兼得的局限；

（2）研究基于自监督学习的跨模态动作感知：跨模态自监督学习的下游任务目前还局限于分类与识别，而当前感知损失网络包括分类、重建、对抗在内的三种预训练任务各有不足。本文开创性地将这两种方法的优势结合，以跨模态自监督学习作为感知损失网络的预训练任务，实现了高度适用于生成任务的跨模态动作感知；

（3）研究基于同步损失与对抗损失的指挥动作生成：现有的指挥动作生成方法仅有基于规则或基于机器学习的方法，而本文提出了首个基于深度学习的方法，能够生成更加真实，连贯，多样，且与音乐紧密同步的指挥动作。此外，现有基于概率模型的生成方法往往因为缺少合适的同步性条件监督约束方法而保留了

回归损失。本文提出的方法可以在不使用回归损失的情况下有效地施加该约束，且多样性不随训练集规模的增加而降低。

1.4 本文组织架构

本文组织架构如下：第一章为绪论，介绍研究背景、相关领域的研究现状以及本文研究内容。第二章介绍并分析与本文提出方法密切相关的跨模态自监督学习（第 2.1 节）、生成对抗网络（第 2.2 节）与感知损失技术（第 2.3 节）。第三章介绍动态频域分解（第 3.1 节）、基于自监督学习的跨模态动作感知（第 3.2 节）这两个本文提出的核心方法。第四章介绍本文构建大规模指挥动作数据集 *ConductorMotion100* 的过程。的在第五章，本文通过实验验证了所提出方法的有效性。最后，第六章回顾本文的工作内容，总结本文研究的不足之处指出并提出未来工作的方向。

第二章　相关技术

2.1　跨模态自监督学习

自监督学习（self-supervised learning）指在不依靠人为给定数据标注的情况下，模型自动归纳数据规律的学习方式。自监督学习可以分为生成式与判别式两种[84]，本文主要关注判别式的自监督学习。Kaiming He 指出[90]，自监督学习实际上就是无监督学习（unsupervised learning）[14]，其“自”的含义在于在训练过程中模型自发地生成正负样本及其标签。自监督学习在近年来受到了深度学习学界的重视，其主要原因在于它可以克服传统的监督学习（supervised learning）对于人工数据标注的依赖，而转以利用大规模的无标注数据学得高质量的数据表示[17]。凭借这样的优点，最近涌现出了一批受到广泛关注的自监督模型，其中具有代表性的包括 MOCO[90]、SimCLR[91]、BERT[83]等。

以上的自监督学习方法都局限于单一模态，即图像或文本。近年来，也有许多学者意识到互联网中广泛存在的多模态数据的巨大价值，并提出了许多跨模态的自监督学习方法。与单模态自监督学习不同，跨模态的自监督学习中两个模态的特征表示互相指导对方的学习，能从数据中挖掘到更丰富的信息。基于深度学习的跨模态自监督方法开创性的工作是 Zisserman 等人[77]于 2017 年提出的L^3Net（Look，Listen and Learn）。L^3Net 通过设计的听视觉相关性学习（Audio-Visual Correspondence learning, AVC）来判断输入的音频-图像对是否同属于同一个视频样本，进而学习到两个模态高质量特征表式。AVC 是前置任务（pretext task）的一种，这类前置任务本身不具备应用价值，但其学得的特征表示可以应用于不同的下游任务（downstream tasks）。当把经过前置任务 AVC 学习训练的L^3Net用于图像分类、音频分类等任务时，其准确率甚至可以超过监督学习的方法[77]。

随着L^3Net的提出，近年来跨模态自监督学习的理论与方法迅猛发展。在 L^3Net 的基础上，Cramer 等人[81]对其音频表示、数据集、训练数据量等因素进行了较为完善的消融实验，进一步提升了L^3Net的特征表示的在下游任务上的性能。随后，Zisserman 等人与 Owens 等人在 2018 年同时证明了 AVC 学习在发声物体识别任务上的潜力[80][78]。在[82]中，Verma 等人设计了一个类似 AVC 的机制来将音乐与图像嵌入一个共享的语义空间，以学习两者之间情感属性的关联。Cheng 等人[85]将协同注意力机制引入L^3Net，增加了两个模态之间信息交互的渠

道。Korbar 等人引入音频与视频在时间上的相关性[79]，设计了听视觉同步性学习（Audio-Visual Temporal Synchronization learning, AVTS）机制。同时作者还发现，相较于 AVC 中的二分类交叉熵损失，使用对比损失可以使模型的训练更加稳定。最近，Zisserman 等人在文献[86]中还加入了文本模态，并将音频、图像、文本这三个模态的样本嵌入到一个共享的特征空间。在文献[89]中，Korbar 等人提出跨模态深度聚类（Cross-Modal Deep Clustering, XDC），使用训练过程中的特征聚类信息作为伪标签来训练模型。类似的聚类思想在文献[87]中也有体现：作者利用多模态特征的相似性信息来选择学习价值更高的正负样本对，并避免错误负采样（false nagative sampling）问题。在文献[88]中，Patrick 等人对跨模态自监督学习理论方法在近年来的进展进行了全面的总结与形式化的归纳，并将负样本的采样方式归纳为通用数据变换框架（Generalized Data Transformations, GDT）。根据实验，基于 GDT 的方法在多个种类的数据集上达到了当今最先进的性能。

2.2 生成对抗网络

生成对抗网络（Generative Adversarial Nets，GAN）[97]的理论与应用研究[18]是近年来深度学习领域最引人注目的进展之一。GAN 的核心思想是利用一对神经网络（生成器G与判别器D）之间的对抗来无监督地拟合数据分布。其中，判别器需要尽可能地区分生成器输出的分布P_G与真实数据分布P_{data}，而生成器的任务在于欺骗判别器，将高斯分布的噪音$P_z(z)$映射至P_G。当两者达到纳什均衡时则有$P_G = P_{data}$。GAN 的目标函数$V(D,G)$可以定义为：

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}}[\log D(x)] + \mathbb{E}_{z\sim P_z}[\log(1-D(G(z)))] \quad (2.1)$$

如 Arjovsky 等人在文献[66]中所述，优化$V(D,G)$等价最小化P_G与P_{data}之间的 JS（Jensen-Shannon）散度。可以得出当P_G与P_{data}不存在重合时，判别器产生的梯度为零，这是导致 GAN 训练不稳定的重要原因。为了解决这样的问题，近年来学者们对于P_G与P_{data}之间的距离度量进行了不同的尝试。其中 Arjovsky 等人提出基于 Wasserstein 距离的 GAN（WGAN）是最有效的途径之一。Wasserstein 距离又被称为推土机（Earth-Mover，EM）距离，其定义为定义为：$W(P_G,P_{data}) = \inf_{\gamma\sim\prod(P_{data},P_G)} \mathbb{E}_{x\sim\gamma}[\|x-y\|]$，可以理解为把一个概率分布移动至另一个概率分布在最优路径规划下的最小消耗。无论P_G与P_{data}距离多远，是否重合，基于

Wasserstein 距离的判别器都可以提供有效的梯度。WGAN 其核心思想在于利用判别器估计P_G与P_{data}之间的 Wasserstein 距离。使用判别器估计 Wasserstein 距离的前提条件是判别器是一个 Lipschitz 连续的函数，即存在一个常数$K \geq 0$使得定义域内对于任意输入的样本x_1, x_2都有 $|D(x_1) - D(x_2)| \leq K|x_1 - x_2|$，即要求判别器产生的梯度有界。在文献[95]中 Arjovsky 等人提出通过对判别器的参数进行梯度裁剪（gradient cliping）来施加 Lipschitz 限制。而在文献[96]中，Gulrajani 通过实验证明梯度裁剪会导致判别器的参数分布在两个极端，同时梯度裁剪的阈值设定不当时很容易导致梯度爆炸。而梯度惩罚（gradient penalty）可以通过增加正则项来完成P_G与P_{data}及其中间区域的 Lipschitz 限制，从而避免这样的问题。具体地，梯度惩罚是通过在P_G与P_{data}之间插值得到的分布$P_{\hat{x}}$中采样得到样本x，进而令判别器对x产生的梯度进行限制。具体地，基于梯度惩罚的 WGAN 的目标函数为：

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}}[D(x)] - \mathbb{E}_{x\sim P_G}[D(x)] + \mathbb{E}_{x\sim P_{\hat{x}}}\left[\|\nabla_x D(x)\|_p - 1\right]^2 \tag{2.2}$$

2.3 感知损失

感知损失（perceptual loss）[70]是面向生成任务的一种损失函数。与传统的在样本空间进行欧式距离度量的L_1或L_2损失不同，感知损失度量的是生成样本与真实样本在特征空间中的距离。这一特征空间是通过预训练的卷积神经网络所得到的，该网络也被称为感知损失网络（perceptual loss network）。具体地，给定生成器G，其任务是根据条件x生成样本$\hat{y}$，并使其与真实样本y尽可能地语义性相似。将感知损失网络记为φ。则在这样的任务上，传统的 MSE 损失L_{MSE}与感知损失L_{per}的定义为：

$$L_{per} = ||\langle\varphi(y)\rangle - \langle\varphi(\hat{y})\rangle||_2 = \frac{1}{n}\sum_{i=1}^{n}\omega_i||hy_i - h\hat{y}_i||_2 \tag{2.4}$$

其中，$\langle\varphi(y)\rangle = [hy_1, hy_2, hy_3, \dots, hy_n]$,表示预训练的感知损失网络$\varphi$提取到的不同层的特征图，$\omega_1, \omega_2, \omega_3, \dots, \omega_n$是为各层特征图指定的权重。目前，学界对于$\varphi$的最广泛的选择是在 ImageNet[68]数据集上预训练的用于图像分类的

VGGNet[69]。在图像分类任务的训练中，神经网络需要面对复杂多样的类内变化，学习同类样本的共同特征，从而完成语义归纳与类别预测。因此，经过图像分类任务训练的神经网络能将在样本空间中距离较远但语义相近的样本对映射至特征空间中互相靠近的位置。其原理如图 2.1 所示：卷积神经网络可以把感兴趣特征投影至激活函数（ReLU）大于零的范围，而把不感兴趣的特征投射到小于零的范围，并由此逐层地筛选掉无用信息[72]，保留有用信息并完成抽象化。卷积神经网络这样的特征选择的机制，使其在被用作感知损失网络时可以对生成样本进行有选择的约束——即仅约束样本特定部分的语义属性，而忽略其余不重要的部分的影响。 感知损失不寻求让生成样本与真实样本在样本空间完全一致，而只要求两者在特征空间中“感知”起来相似[71]。这样的特性使感知损失避免了 MSE 损失的多种缺陷[67]，并拥有了应对多种一对多（病态）生成问题的潜力，例如图像风格迁移[70]、图像超分辨率[75]、低剂量 CT 去噪[76]、语音频带扩展[16]、遥感图像全色锐化[15]等。

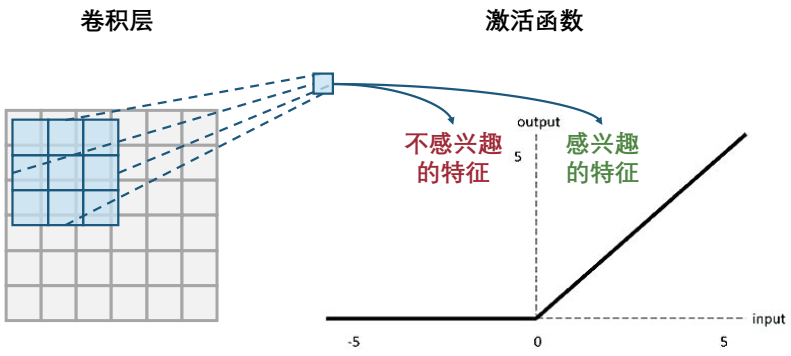


图 2.1 卷积神经网络的特征选择机制示意图

感知损失于 2016 年被 Johnson 等人提出[70]。在过去的 5 年中，感知损失的相关研究取得了一定的进展。尤其是在图像超分辨率任务上，随着感知损失的引入，超分辨率模型的性能有了大幅的提高。也有一些学者从不同角度提出对感知损失进行改进。例如，受人类视觉系统（Human Visual System, HVS）启发，Tariq 等人[73]提出根据图像中的频域信息生成注意力图以重新分配感知损失的权重，使模型更重视对于图像感知效果更重要的部分，例如毛发等细节区域。Rad 等人[74]提出目标感知损失（targeted perceptual loss），通过区分图像中的背景

（background）、边缘（boundary）与目标(object)，为图像的不同语义区域使用不同形式的感知损失。

感知损失研究的另一个方向是感知损失网络的选择。Tej 等人[75]指出，在使用传统的基于 ImageNet 预训练 VGGNet 的感知损失进行图像超分辨率时，会导致出现不自然的图像细节。作者认为这是由 VGGNet 的预训练目标与图像超分辨率目标的不匹配所造成的，并提出使用生成对抗网络中的判别器作为感知损失网络。类似地，面向低剂量 CT 去噪问题，Li 等人[76]指出，在自然场景的图像分类数据集 ImageNet 上训练的网络不适合提取 CT 图像中的语义信息。因此，作者在 CT 数据集上训练了一个自编码器作为感知损失网络。

根据预训练任务内容，不难分析各类感知损失网络各自对何种特征感兴趣。分类任务使网络提取的是类别相关特征，但这样的特征难以完成跨模态的条件约束。例如基于舞蹈种类分类（区分芭蕾舞、流行舞、桑巴舞等）预训练的网络提取到的高层语义特征中，同类舞蹈动作特征间距很近。这会导致模型仅需要生成风格相符的舞蹈动作，而不需要使其与音乐同步。由于这样的原因，文献[35]中作者仅使用感知损失网络的浅层特征计算感知损失，这违背了感知损失衡量高层次语义属性差异的初衷，也限制了模型的性能。自编码任务使网络提取到的特征需要包含尽可能多样本信息。自编码重建任务训练得越好，感知损失向 MSE 损失的退化程度就越高。对抗判别任务使网络提取与真实样本与生成样本的判别性特征，缺乏语义性。综上，现有的三类感知损失网络都有着各自的局限性。因此，有必要探寻一种新的感知损失网络预训练方式，提取高质量的特征以完成准确的监督约束。本文关于感知损失的贡献也在于这一研究方向：在分类任务[70]、判别任务[75]、重建任务[76]之外，本文提出将跨模态的自监督学习任务[79]作为感知损失网络φ的预训练任务。

第三章　基于动态频域分解与自监督跨模态感知的指挥动作生成

给定包含N个样本的数据集$D = \{(X_i, Y_i)\}_{i=1}^{N}$,其中$X_i = \{x_t\}_{t=1}^{T}$与$Y_i = \{y_t\}_{t=1}^{T}$为时长为$T$的音频特征序列与指挥动作序列样本，$x_t$和$y_t$分别为第$t$个时间步上的$p$维音频特征$x_t \in R^p$和$q$个关键点的 2 维骨架坐标$y_t \in R^{2q}$。本文的任务是在数据集$D$上训练一个映射$G: R^{T\times P} \to R^{T\times 2q}$，生成对应的指挥动作序列$\hat{Y} = G(X)$。本文提出动作动态频域分解方法与自监督跨模态感知方法来共同完成这一任务。

3.1　动作动态频域分解

由于在指挥动作中同时叠加着关于节拍、演奏法、力度以及音乐情感等信息[59]，单个学习模型难以同时兼顾。这一的问题在舞蹈生成与乐器演奏动作生成任务上也同样存在。现有的解决这一问题的主要方法是动作分解，包括时域分解[38][39][40][12][46]与空间域分解[41][13]。这些方法的核心动机是试图降低人体动作的复杂性，通过动作分解将单个困难的学习任务转化为多个较简单的学习任务，从而提高模型的学习效果。如图 3.1 所示，时域分解与空间域分解将原始动作序列沿时间轴或空间轴的方向分割成多个子序列。然而，这些子序列一经分割之后便互相独立，子序列之间的上下文关系也随着之丢失。时域中的上下文关系代表着动作的连贯性，空间域中的上下文关系代表着动作的协调性。因此，基于时间分解的方法生成的动作往往协调但不连贯，而基于空间域分解的方法则连贯但不协调。

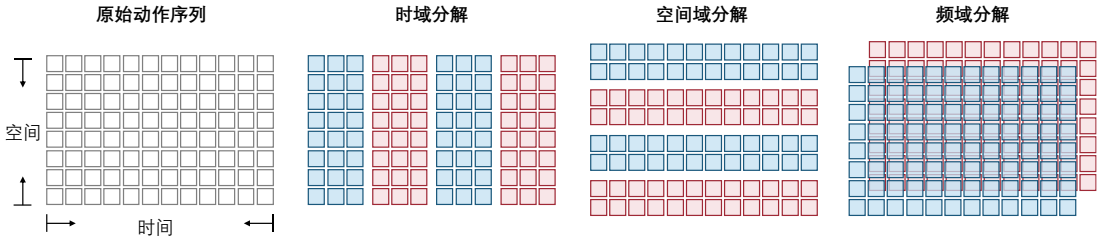


图 3.1 时域分解、空间域分解与频域分解示意图

本文提出动作的动态频域分解，在同时保留时域与空间域的上下文信息的前提下，将复杂的指挥动作分解为相互独立的两个动作分量。考虑到与现有方法所面向的舞蹈动作或乐器演奏动作不同，，本文将指挥动作序列看作由高频分量与低频分量叠加而成的多维稳定信号，其中高频分量包含了幅值较小但频率较高的

节拍、力度等信息，而低频分量包含了幅值较大但频率较低的情感以及身体朝向转动等信息。将原始动作序列Y_i分解得到的高频与低频分量分别记为$Y_{high,i}$，$Y_{low,i}$ 则有：

$$Y_i = Y_{high,i} + Y_{low,i} \tag{3.1}$$

为了对指挥动作进行这样的频域分解，最简单直接的方法是寻找一组合适的频率阈值，从而得到高通滤波器与低通滤波器并对原始动作序列进行小波分解，得到高频分量与低频分量。然而，这两种动作分量在频域分布的界限是随时间改变的，当音乐节奏较慢时该界限会下降，而音乐节奏较快时该界限会上升。而在不同乐曲之间音乐节奏变化幅度很大，很难为所有样本找到一个普遍适用的频率阈值。所以，基于固定阈值的方法并不能够确保分解后的动作成分之间相互独立：在高频分量中往往也会包含幅值较大而频率较低成分，而低频分量中会包含幅值较小而频率较高成分。

本文提出动态频域分解（**D**ynamic **F**requency-domain **D**ecomposition, DFD）以解决这一问题。不同于基于固定阈值的方法，动态频域分解根据音乐节奏自适应地确定用于频域分解的阈值。具体地，对于音频特征序列X_i将指挥动作序列Y_i，首先将两者切分为k个片段$X_i = [X_i^1, X_i^2, X_i^3, \dots, X_i^k,]$，$Y_i = [Y_i^1, Y_i^2, Y_i^3, \dots, Y_i^k,]$。对于每一个片段的样本对$(X_i^j, Y_i^j)$，使用音乐节奏估计算法从$X_j^k$得到对应的节奏值$tempo_j^k$（以 BPM 为单位），再根据下式确定对应于该片段的频率阈值f_i^j（以 Hz 为单位）：

$$f_i^j = \frac{1}{2} tempo_i^j / 60 \tag{3.2}$$

计算低通滤波器与高通滤波器的归一化截止频率$W\hat{}_i$：

$$W_i^j = 2 \cdot f_i^j / \mathrm{sr} \tag{3.3}$$

其中，sr为动作信号的采样频率。根据计算得到的归一化截止频率W_i^j，构造一个N阶的巴特沃斯低通滤波器，并对Y_i^j每一个维度上的动作信号进行滤波，其结果为动作的低频分量，记为$Y_{low,i}^j$。高频分量$Y_{high,i}^j$则由原始动作减去低频分量得到：

$$Y_{high,i}^{j} = Y_{i}^{j} - Y_{low,i}^{j} \tag{3.4}$$

最后，将得到的k个片段的分解动作拼接，由此就能够根据音乐节奏将指挥动作分解为高频分量与低频分量。动作动态频域分解的算法流程图下：

输入： 音频特征序列X_i;
维数为d_y，采样率为sr的指挥动作序列Y_i;
分割后每个片段的长度l;
滤波器阶数N;

过程：

1: $k = \text{len}(Y_i)/l$，k向上取整;
2: 将X_i和Y_i切分为k个片段$X_i = [X_i^1, X_i^2, X_i^3, \dots, X_i^k,]$，$Y_i = [Y_i^1, Y_i^2, Y_i^3, \dots, Y_i^k,]$;
3: **for** $j = 1, 2, \dots k$ **do**
4: 使用节拍估计算法估计音频X_j^k的节奏值$tempo_j^k$;
5: $f_i^j = \frac{1}{2} tempo_i^j / 60$;
6: $W_i^j = 2 \cdot f_i^j / \text{sr}$;
7: 以W_i^j为归一化截止频率构造N阶的巴特沃斯低通滤波器filter_i^j;
8: **for** $d = 1, 2, \dots d_y$ **do**
9: $Y_{low,i}^j[d] = \text{filter}_i^j(Y_j^k[d])$;
10: $Y_{high,i}^j[d] = Y_i^j[d] - Y_{low,i}^j[d]$;
11: **end for**
12: **end for**
13: 拼接$Y_{low,i} = [Y_{low,i}^1, Y_{low,i}^2, Y_{low,i}^3, \dots, Y_{low,i}^k,]$;
14: 拼接$Y_{high,i} = [Y_{high,i}^1, Y_{high,i}^2, Y_{high,i}^3, \dots, Y_{high,i}^k,]$;

输出： 高频分量$Y_{high,i}$;
低频分量$Y_{low,i}$;

经过动作动态频域分解后，数据集D便可从$D = \{(X_i, Y_i)\}_{i=1}^N$转化为$D = \{(X_i, Y_{high,i}, Y_{low,i})\}_{i=1}^N$。随后，为两个动作成分独立地训练两个生成器$G_{high}$与$G_{low}$，将两者生成结果相加，得到最终的模型输出$\hat{Y}$。在本文方法的余下部分中，对于高频分量与低频分量的学习方式完全相同。为了简便，在下文中将省略G的脚标。

3.2 基于自监督跨模态感知的指挥动作对抗生成

3.2.1 方法概述

对于音乐驱动的指挥动作生成这一任务，最直接的方法是使用L_1或L_2损失拟合真实动作。但对于同一个音乐片段，数据集里可能存在着多个不同的指挥动作，而L_1或L_2损失的最优解是这些不同指挥动作的平均值，这一平均值往往是幅值很小的过度平滑（over-smooth）的动作。为了解决这一问题，本文从概率视角出发，寻求施加合适的约束，以使生成的指挥动作序列$\hat{Y}$在样本空间构成的分布P_G尽可能地趋近于真实动作的分布P_{data}（看起来真实自然），同时趋近于真实动作关于音乐的条件概率分布P_c（与音乐紧密同步）。其中，$P_G \rightarrow P_{data}$的约束可以直接通过生成对抗网络的形式来实现。而对于$P_G \rightarrow P_c$的约束本文试图使用感知损失实现。其中的关键问题在于感知损失网络的选择。本文认为音乐相关的动作特征最适合作为感知损失选择性约束的内容，受到近年来跨模态自监督学习方法[77][79]的启发，本文提出音频-动作同步性学习（**M**usic-**M**otion **S**ynchronization learning, M^2S）作为感知损失网络的预训练任务。

因此，如图 3.2 所示，本文方法的学习流程将分为两个阶段：对比学习阶段与生成学习阶段。在对比学习阶段，通过M^2S学习训练一个两分支的网络M^2SNet。在生成学习阶段，构建一个带有四个模块的M^2SGAN，并将M^2SNet中训练好的动作编码器以感知损失的方式施加同步性约束$P_G \rightarrow P_c$。为了与传统的基于分类等预训练任务的感知损失区别，本文将这一损失命名为同步损失（sync loss）。同时，M^2SNet的音乐编码器也迁移至M^2SGAN，为生成器提供高质量语义性的音乐特征。

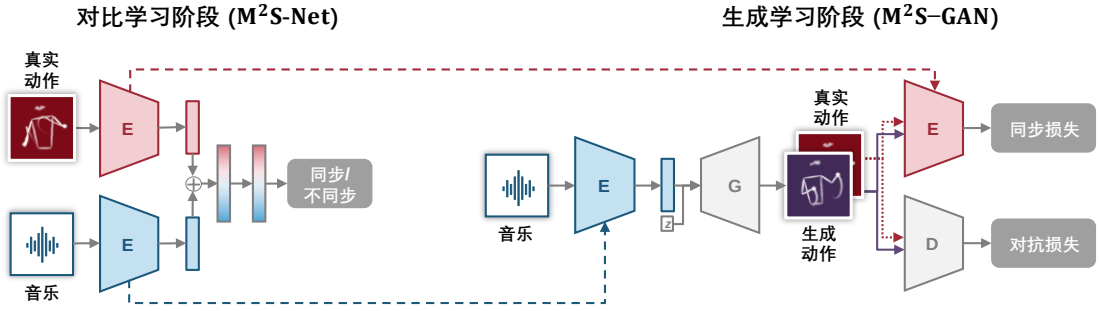


图 3.2 对比学习阶段与生成学习阶段的关系示意图

3.2.2 网络结构

本文提出的方法共涉及四个神经网络：音乐编码器E_{music}、动作编码器E_{motion}、生成器G与判别器D。M^{2}SNet由E_{music}、E_{motion}以及三个全连接层构成。E_{music}与E_{motion}的输出经拼接后输入全连接层，最后一个全连接层输出一个在(0,1)之间的标量，代表网络对于输入样本对同步性的预测。M^{2}SGAN则由全部四个神经网络组成。E_{music}提取音乐特征输入G，G生成指挥动作，并传递给E_{motion}与D分别计算同步损失与对抗损失。接下来，将结合图 3.3 详述每一个网络的内部结构。

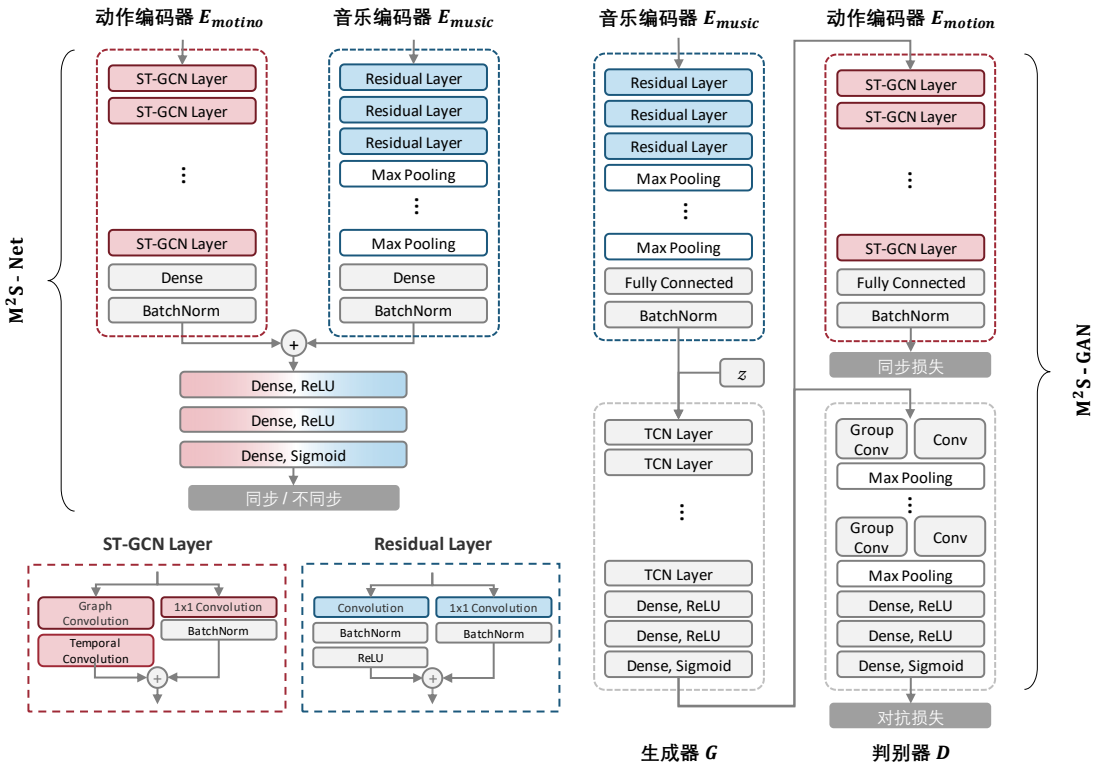


图 3.3 M^{2}SNet与M^{2}SGAN的网络结构

（1） 音乐编码器

音乐编码器E_{music}从 2 维的梅尔频谱图中提取音乐特征。E_{music}包含三个组，每组由三个残差层（Residual Layer）和一个池化层组成。残差层中包含卷积核为 3×3 的卷积层、批归一化层（Batch Normalization）、ReLU 激活函数与基于 1×1 卷积的残差连接。池化层将特征图在时间维度与频率维度降采样。其中，由于音频的采样率是指挥动作采样率的三倍，在时间维度上仅进行一次× 3的降采样。在降采样前，E_{music}的每个卷积层有 16 个通道，采样后则增至 32 个。

（2） 动作编码器

动作编码器E_{motion}需要同时从时间与空间两个角度对指挥动作进行分析，因此，本文采用时空图卷积神经网络（Spatial Temporal Graph Convolution Network, ST-GCN）[124]作为动作编码器。ST-GCN 原本是为姿态分类任务设计的，因此可以很好地完成指挥动作特征提取的任务。E_{motion}包含 10 个 ST-GCN 层，每层中图卷积层与时间卷积层分别提取空间特征与时间特征。类似地，1×1 卷积在每一层构造残差连接。E_{motion}不进行任何降采样，因此其输入与输出保持同样的采样率。E_{motion}的每一个卷积层都带有 32 个通道。

（3） 生成器

生成器G根据音乐编码E_{music}器提取到的特征生成指挥动作。文献[92]发现 Temporal Convolution Network (TCN)可以达到与 LSTM 相似的性能。同时， TCN 的训练速度相较于 LSTM 有着明显的提升。因此，本文采用 TCN 作为生成器的网络结构。该网络包含 5 个带有残差连接（residual connection）[93]的空洞卷积（dilated convolution）[94]层。原始的 TCN（即 WaveNet [94]）中的卷积层为因果卷积，但在指挥动作生成任务中模型需要学习双向的依赖，因此本文将因果卷积改为普通卷积。G的每一个卷积层都带有 64 个通道。

（4） 判别器

判别器D需要区分生成器G生成的指挥动作与真实的指挥动作。D包含两分支的结构，一个分支是分组的一维卷积，用于单独提取动作每一个维度的特征。另一个分支是常规的一维卷积，用于提取整体的空间特征。两个分支的输出最终被拼接并传入全连接层。实验表明，降采样对于提升D的性能至关重要。最终输出的采样率由 30Hz 降为 2.5Hz。D的每一个卷积层都带有 32 个通道。

3.2.3 损失函数

（1） 对比学习阶段

在对比学习阶段，本文采用二值交叉熵损失L_{CE}与对比损失L_{CT}来训练M^2SNet。其中，L_{CE}要求网络能够正确判断输入的音频-动作样本对是正样本还是负样本，而L_{CT}则要求M^2SNet的E_{music}与E_{motion}从正样本对中提取到的特征尽可能相近，而负样本对中提取到的特征尽可能疏远。同时，L_{CT}还能帮助去除E_{music}与E_{motion}

提取到的特征中的噪音。此外，L_{CT}还可以使训练过程更加稳定[79]。L_{CE}与L_{CT}的定义如下：

$$L_{CE} = \frac{1}{M}\sum_{j,k=1}^{M} c_{jk}\log\left[fc\left(E_{music}(X_j) \oplus E_{motion}(Y_k)\right)\right] + \left(1 - c_{jk}\right)\log\left[1 - fc\left(E_{music}(X_j) \oplus E_{motion}(Y_k)\right)\right] \tag{3.5}$$

$$L_{CT} = \frac{1}{M}\sum_{j,k=1}^{M} c_{jk}\left\|E_{music}(X_j) - E_{motion}(Y_k)\right\|_2^2 + \left(1 - c_{jk}\right) max\left[1 - \left\|E_{music}(X_j) - E_{motion}(Y_k)\right\|_2, 0\right]^2 \tag{3.6}$$

其中，(X_j, Y_k)是从数据集D中采样得到的样本对，c_{jk}是该样本对的标签，为正样本对（$j = k$）时$c_{jk} = 1$，负样本对（$j \neq k$）时$c_{jk} = 0$。fc表示$\mathrm{M^2SNet}$的全连接层，$\oplus$表示特征拼接操作。在对比学习阶段$\mathrm{M^2SNet}$的损失函数$L_{\mathrm{M^2SNet}}$的定义如下：

$$L_{\mathrm{M^2SNet}} = L_{CE} + L_{CT} \tag{3.7}$$

（2） 生成学习阶段

在生成学习阶段，生成器根据音乐编码器E_{music}提取到的音乐特征与采样于正态分布的噪音z生成指挥动作$\hat{Y} = G(E_{music}(X), z)$，并使生成样本的分布$P_G$同时趋向于真实动作分布$P_G \to P_{data}$以及动作关于音乐的条件分布$P_G \to P_c$ 。生成器G的损失函数L_G的定义如下所示：

$$L_G = \lambda_{sync}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\omega_k\left\|\langle E_{motion}(Y_i)\rangle_k - \langle E_{motion}(G(X_i))\rangle_k\right\|_2^2 - \lambda_{adv}\frac{1}{N}\sum_{i=1}^{N} D(G(X_i)) \tag{3.8}$$

其中，E_{motion}是感知损失网络，即基于$\mathrm{M^2S}$学习的动作编码器，$\langle E_{mo}(Y_i)\rangle_k$是$E_{motion}$从动作序列$Y_i$上提取到的第$k$层特征，而$\omega_k$是对应于第$k$层的权重。$D$是判别器。$\lambda_{sync}$，$\lambda_{adv}$分别为感知损失与对抗损失的权重。

判别器D的任务是判断生成指挥动作序列的真实程度，并为生成器提供准确的梯度以使$P_G \to P_{data}$。本文使用基于 Wasserstein GAN (WGAN) [95][96]的判别器。WGAN 估计的是P_G与P_{data}之间的 Wasserstein 距离，与原始的 GAN[97]相比，这

种方法在P_G与P_{data}相距较远时也能提供有效的梯度，从而有效缓解了 GAN 训练不稳定与模式崩塌（mode collapse）的问题。判别器的损失函数L_D为：

$$L_D = \frac{1}{N}\sum_{i=1}^{N}\left[D\left(G(X_i)\right) - D(Y_i)\right] + \omega_{GP}\mathbb{E}_{\hat{x}}[\|\nabla_{\hat{x}} - 1\|_2] \tag{3.9}$$

其中，第二项为梯度惩罚（Gradient Penalty, GP）项，ω_{GP}是该项的权重。$\hat{x}$是在P_G与P_{data}之间随机插值采样得到的动作序列，用于在P_G与P_{data}之间施加 Lipschitz 限制。

3.2.4 负样本采样策略

在M²S学习的过程中，模型自动地从数据集中生成同步的音乐-动作正样本对，错位的负样本对，以及对应的二值标签c_i^j。根据 Korbar 等人[79]所述，负样本采样策略可以分为 Easy Negatives，Hard Negatives 以及 Super-hard Negatives 三种。接下来，结合图 3.4 所示，将根据本文面向的任务，对这三种负样本进行具体定义：

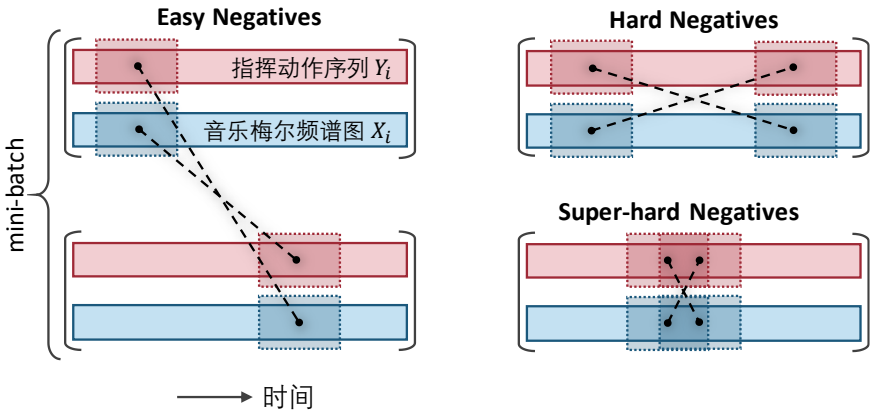


图 3.4 三种负样本对采样策略示意图

- Easy Negatives 指随机采样于同一个 mini-batch 中的不同的乐曲。这种负样本是跨模态自监督学习领域中使用最广泛的一种。有学者指出，这种负样本鼓励模型学习跨模态的语义相关性。
- Hard Negatives 指随机采样于同一个 mini-batch 中的相同的乐曲，但采样的负样本对之间不存在重合。本文中，令负样本之间的时间间隔大于 10 秒。与 Easy Negatives 相比，Hard Negatives 引入了对与同步性的学习，这要求模型需要同时关注跨模态的同步性。

- Super-hard Negatives 指随机采样于同一个 mini-batch 中的相同的乐曲，且采样的负样本对之间存在一定程度的重合。具体地，本文令负样本之间的时间间隔在 0.5 秒~5 秒之间。与 Hard Negatives 相比，这样较短的时间间隔将无法容纳语义差异，模型将仅学习时间上的同步性。

本文采用 Hard Negatives 策略进行负样本采样。其原因在于，与传统的多模态自监督学习面向的无限制网络视频数据相比，M^2S学习所面向的音乐-指挥动作数据是较细粒度的数据。本文发现这样的细粒度属性导致 Easy Negatives 策略下错误负采样（False Negative Sampling）的概率上升。在 Easy Negatives 策略下进行M^2S学习时，模型会枚举数据集内所有可能的音乐-动作样本对组合，其中大量的错误负采样会导致模型试图分离各个曲目的身份。相比而言，Hard Negatives 与 Super-hard Negatives 则消除了这种跨曲目错误负采样的可能性，从而提高了模型学习时的稳定性。Hard Negatives 鼓励模型同时学习有语义相关性与时间同步性，而 Super-hard Negatives 只关注时间同步性。因此，Hard Negatives 应是M^2S学习中的最佳负采样策略。本文将在实验部分验证这一观点。

第四章　数据准备

由于现有的指挥动作数据集规模都较小，无法满足深度生成式模型的训练需求。因此，本文构建了一个大规模的指挥动作数据集 *ConductorMotion100*。如图 4.1 所示，构建 *ConductorMotion100* 数据集时，首先从网络视频平台收集交响音乐会的指挥视角录像视频，再分别对视频中的指挥姿态与音乐特征进行提取。提取到的数据经归一化后形成 *ConductorMotion100* 数据集。

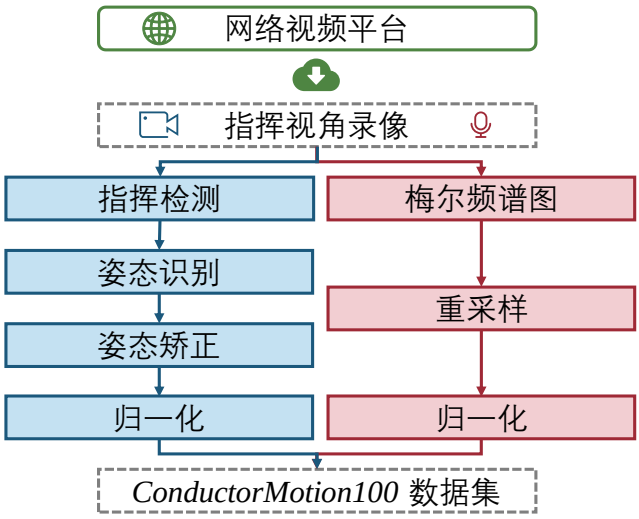


图 4.1 *ConductorMotion100* 数据集构建流程示意图

4.1　数据收集

本文从网络视频平台 bilibili.com 与 youtube.com 中爬取大量的指挥视角演出录像视频。视频的选择标准包括 1）面对指挥正面；2）镜头稳定不动；3）视频时长大于等于 5 分钟。指挥的乐曲覆盖了古典时期、浪漫主义时期以及现代主义时期的音乐。在爬取视频时，总是选择可用的最高视频分辨率。本文将所有视频都转化为 30fps 的采样率。收集到的视频标题构成的词云如图 4.2 所示，出现最频繁的词包括“交响曲（Symphony）”、“乐章（Movement）”、“指挥（Conductor）”、“视角（View）”、“协奏曲（Concerto）”、“组曲（Suite）”等。

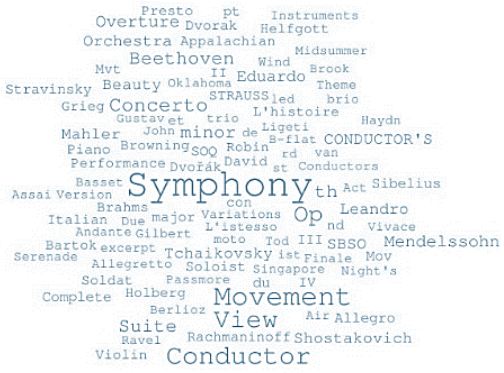


图 4.2 从网络视频平台中收集到视频的标题构成的词云

如**表 1** 所示，*ConductorMotion100* 数据集的规模超过了现有的绝大部分指挥动作数据集、音乐舞蹈数据集以及音乐节拍检测数据集。

表 1 *ConductorMotion100* 数据集与现有相关数据集的规模对比。

数据集类型	年份	数据集		时长（分钟）	类型平均时长
节拍检测	2006	Ballroom 数据集	[107]	357	479 分钟（7.9 小时）
	2009	Beatles 数据集	[108]	489	
	2004	Hainsworth 数据集	[109]	199	
	2005	Simac 数据集	[110]	198	
	2012	SMC 数据集	[111]	145	
	2012	HJDB 数据集	[112]	199	
	2012	ACM Mirum 数据集	[113]	905	
	2015	**GiantSteps 数据集**	**[114]**	**1325**	
	2002	GTZAN 数据集	[115]	500	
舞蹈动作	2016	MotionDance	[20]	73	754 分钟（12.6 小时）
	2021	Dance Revolution	[43]	790	
	2018	Dance with Melody	[24]	94	
	2020	Ahn 等人	[36]	94	
	2018	Listen to Dance	[32]	376	
	2019	Qi 等人	[33]	148	
	2020	ChoreoNet	[38]	94	
	2019	**Dancing2Music**	**[12]**	**4260**	
	2020	DanceIt	[39]	540	
	2020	Duan 等人	[40]	1980	
	2021	DeepDance	[45]	300	
	2020	Ren 等人	[35]	300	
乐器演奏	2018	Shlizerman 等人	[25]	513	734 分钟（12.2 小时）
	2019	URMP 数据集	[116]	78	
	2017	C4S 数据集	[117]	270	
	2015	Carrillo 等人	[118]	10	
	2006	ENST-Drums 数据集	[119]	225	
	2011	Abeber 等人.	[120]	72	
	2020	**Solo 数据集**	**[121]**	**3976**	
乐队指挥	**2014**	**Sarasúa 等人**	**[52]**	**250**	643 分钟（10.1 小时）
	2017	Karipidou 等人	[56]	36	
	2013	Sarasúa 等人	[50]	120	
	2019	Huang 等人	[59]	180	
	2019	IDEA 数据集	[64]	56	
	2013	Dansereau 等人	[9]	0.5	
	2021	***ConductorMotion100*（本文）**		**6000 分钟（100 小时）**	

4.2 指挥动作提取

在收集到的指挥视角演出录像视频中，很多情况下视频里除指挥外还有乐手、观众等人物出现。直接在视频中进行姿态检测会因模型无法判断哪个目标是指挥，而产生剧烈的抖动效果。因此，本文增加了一个指挥检测步骤，以将视频中指挥以外的区域遮住。具体地，首先标注了一个包含 300 张图像的指挥检测数据集，随后训练了一个经过预训练的 yolo-v3 目标检测网络。最后，根据指挥检测模型的输出，为每一帧生成一个带有 10%填充的遮罩，以在不遮挡指挥的情况下消除其他目标造成的影响。数据集中图像样本有两种来源，第一种是在收集到的指挥视频中随机采样，第二种是在互联网上收集音乐会相关的图片（用于提高泛化能力与鲁棒性）。由于 300 个样本的数量远不足以对基于深度学习的目标检测模型进行完整的训练，本文采用预训练[1]的 yolo-v3 模型[98]并进行迁移学习。在微调的过程中，yolo-v3 模型中的前 10 层的参数被冻结。微调过程中的训练、测试损失以及 precision、recall 的变化如图 4.3 所示。最终，模型达到了 0.861 的精确率与 0.991 的召回率，在测试集上的模型预测结果如

图 4.4 所示。

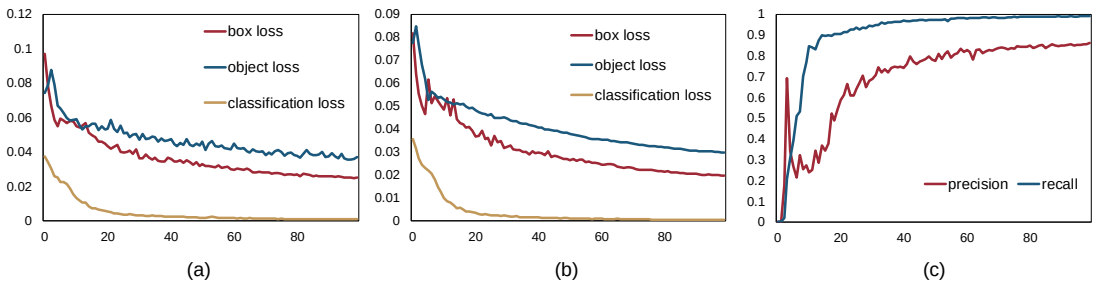


图 4.3 微调过程中的训练集损失(a)、验证集损失(b)、以及精确率、召回率(b)曲线

图 4.4 指挥检测效果图

[1] https://github.com/ultralytics/yolov3

由于在收集到的指挥视频中指挥朝向各不相同，为了更精准地提取指挥动作，本文采用基于 AlphaPose[2, [99][100][101]]的二维姿态估计与基于 VideoPose3D[3, [102]]的三维姿态估计。具体地，首先使用 AlphaPose 从图像中估计指挥关键点的二维坐标，再使用 VideoPose3D 将其投影至三维空间。然而，由于视角与遮挡原因，指挥的下半身往往不可见。这导致在 VideoPose3D 得到的 3 维人体姿态中，双膝、双足的关键点位置十分不准确。同时，在指挥动作中，下半身的姿态往往也不包含有用的信息。因此，如图 4.6 所示，本文仅保留了指挥上半身的 13 个关键点。此外，由于收集到的视频光照往往较差，AlphaPose 估计的 2 维关键点也包含着噪音。如图 4.5 所示，对比不同长度的卷积核，本文发现长为 3 的滑动卷积平滑（蓝色）可以在保留动作信息的前提下消除抖动现象。

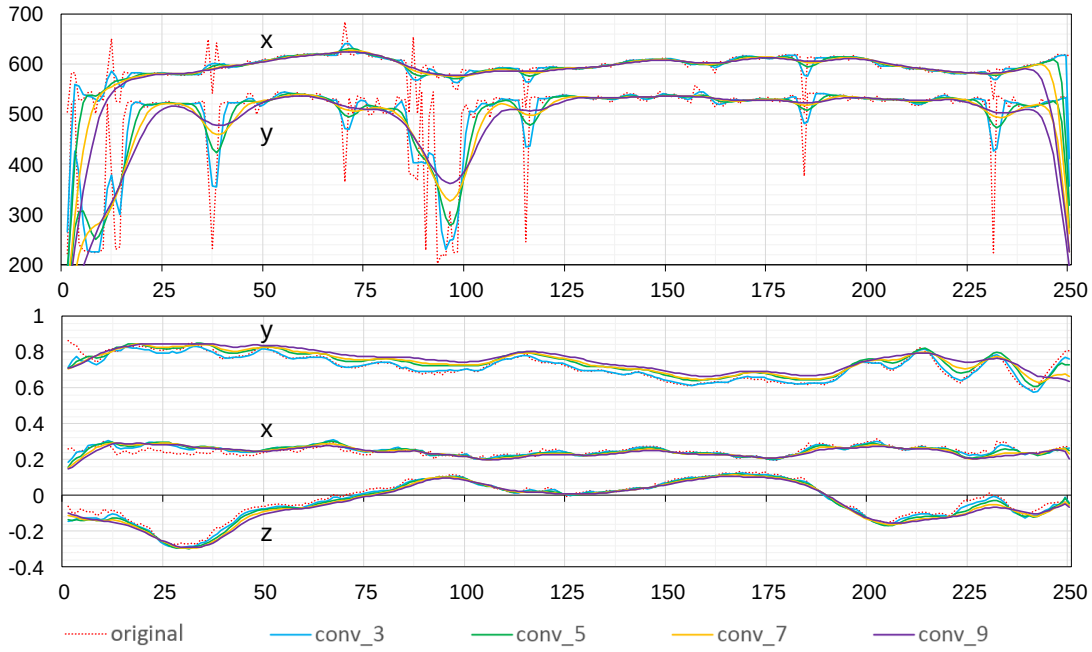


图 4. 5 不同卷积核下平滑效果的对比

随后，如图 4.6 所示，对经过平滑后得到的姿态数据进行朝向矫正，再从指挥的正面将姿态数据投影至二维空间。最后，对得到的二维数据进行归一化处理，将双肩平均宽度（第 5、6 号关键点之间的距离）缩放至 0.2，将上半身长度（5、6 号关键点中点与 11、12 号关键点中点之间的距离）缩放至 0.25。

[2] https://github.com/MVIG-SJTU/AlphaPose
[3] https://github.com/facebookresearch/VideoPose3D

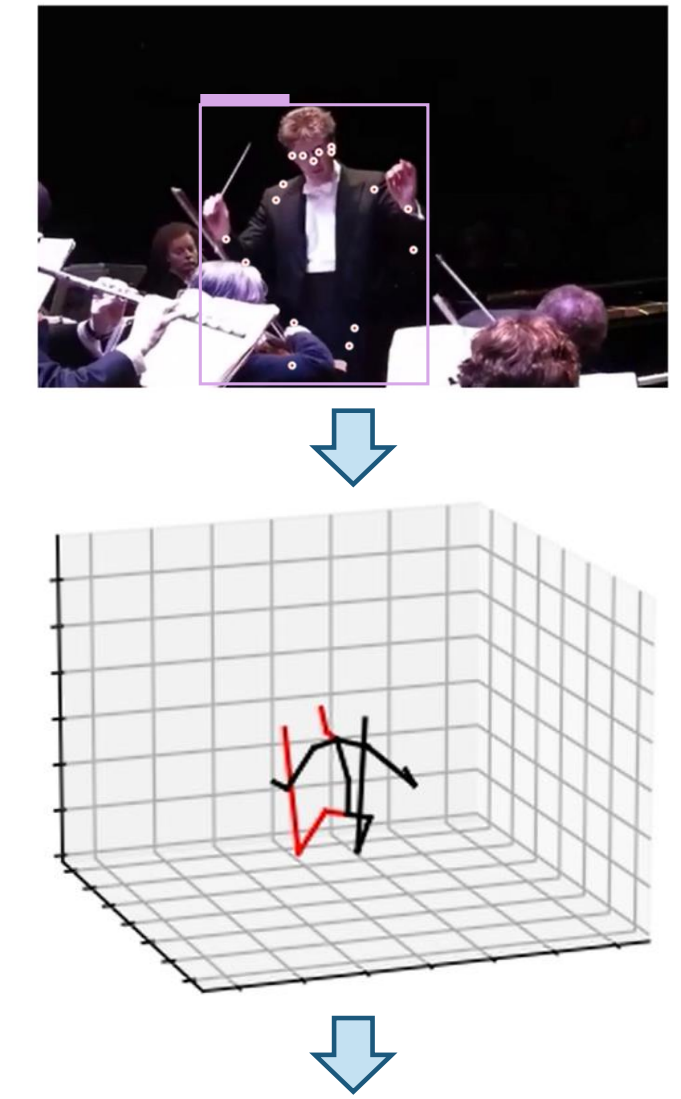


图 4.6 指挥姿态提取（上）、矫正（中）、归一化（下）示意图

4.3 音乐特征提取

本文采用 128 个频率窗口，hop-length=256 帧的梅尔频谱图（Mel Spectrogram）表示音乐数据。由 44100Hz 采样率的音频直接得到的梅尔频谱图的采样率是 86.52Hz，为了方便在模型中与动作数据（30Hz 采样率）对齐，本文将 86.52Hz 的梅尔频谱图重采样为 90Hz。最后，将梅尔频谱图转化为 dB 单位后归一化至(0,1)区间。图 4.7 展示了一个数据集中的梅尔频谱图样本，该样本对应于柴可夫斯基 1812 序曲的结尾乐段。

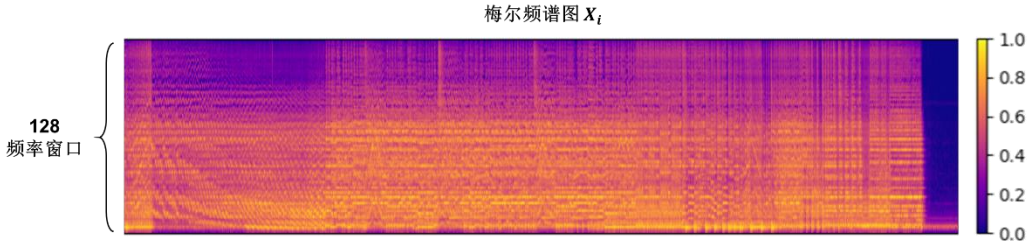


图 4.7 *ConductorMotion100* 数据集中梅尔频谱图的示例样本

第五章　实验与分析

5.1　实验设置

超参数：对比学习阶段中，M^{2}SNet由 AdamW[104]优化器优化，其学习率为0.001，β=(0.9, 0.999)，权重衰减率为 0.02。生成学习阶段中，生成器与判别器由RMSProp[105]优化器优化，其学习率为 0.0005。感知损失的权重ω_i都设为 1。判别器梯度惩罚的权重$\omega_{GP}=10$。batch size 为 30，每一个样本对为 30 秒的音频特征序列-指挥动作序列样本对。

实验细节：*ConductorMotion100* 数据集以 9:0.5:0.5 的比例划分为训练集、开发集与测试集，得到 90 小时的训练集、5 小时开发集与 5 小时测试集。对比学习阶段，由于 Hard Negatives 与 Super-hard Negatives 的难度较大，在第一个 epoch 时都使用 Easy Negatives 作为预训练。在生成学习阶段，遵循 WGAN 的一般操作，每步训练都训练 5 次判别器，再训练 1 次生成器。

实验环境：使用 Intel Core i5-9400 CPU 处理器(2.90 GHz)，32GB 内存，NVIDIA GeForce RTX 2080 Ti GPU (27GB)。算法基于 Python 3.6、Pytorch 1.6.0 实现。本方法的完整训练需要约 48 小时，对比学习阶段与生成学习阶段耗时相近。

5.2　评价指标

本文是第一个基于深度学习的音乐驱动的指挥动作生成方法，因此在实验中本文也无法沿用现存的评价指标。近年来，对于深度生成式模型的性能评价往往使用 Inception Score (IS)或 Frechet Inception Distance (FID)。然而，这些指标都需要一个经过分类任务预训练的特征提取器。然而，指挥动作上还没有通用的特征提取器，也不存在类似的分类标签，导致 IS 和 FID 不可用。因此，本文将提出一些新的评价指标。为了验证有效性，图 5.1 给出了这些指标在应对动作时空扰动时的变化情况。空间扰动指在(0%~200%)的范围内变化动作的幅值，时间扰动是在(80%~120%)的范围改变动作的速度。所得到的评价指标值是 100 次随机尝试的平均值。

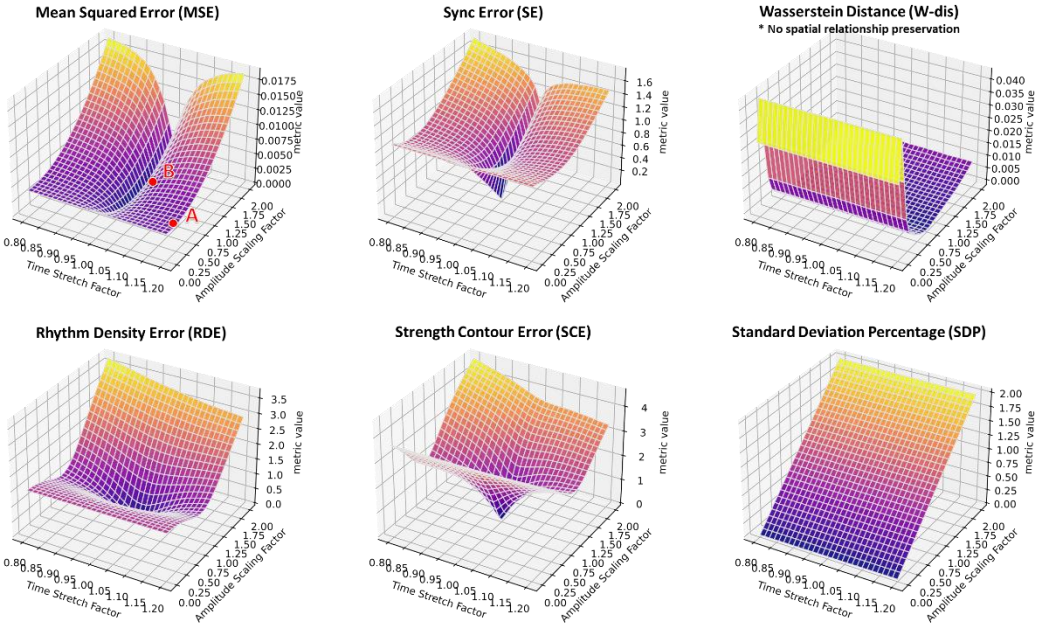


图 5.1 各个评价指标应对时空扰动的变化情况

（1） Mean Squared Error (MSE)

MSE 是衡量生成动作与真实动作相似程度最直接的一个方法。它在其他一些舞蹈生成或乐器演奏动作生成任务书也被采用。然而，MSE 有对于小幅值动作的偏好。如图 5.1 所示，时间扰动大但幅值较小的动作的 MSE（A 点）比时间扰动较小且幅值接近正常动作的 MSE（B 点）更小，这体现了 MSE 损失无法识别音乐与指挥动作之间一对多的对应关系。给定真实动作$Y_i = \{y_t\}_{t=1}^{T}$与生成动作$\hat{Y}_i = \{y_t\}_{t=1}^{T}$，MSE 的定义如下：

$$\mathrm{MSE}(Y_i, \hat{Y}_i) = \left\| Y_i - \hat{Y}_i \right\|_2^2 \tag{5.1}$$

（2） Sync Error (SE)

SE（同步误差）与同步损失十分相似。因为本文的模型直接优化同步损失，因此将其用作评价指标并不公平。相反，本文在测试集上进行M²S学习，并用得到的动作编码器已同步损失的方式计算同步误差。与 MSE 相比，SE 并没有对小幅值动作的偏好。SE 的定义如下：

$$\mathrm{SE}(Y_i, \hat{Y}_i) = \left\| E_{motion}(Y_i) - E_{motion}(\hat{Y}_i) \right\|_2^2 \tag{5.2}$$

（3） Wasserestein Distance (W-dis)

该指标度量真是动作与生成动作的推土机距离。对于本文中涉及的每一个对比模型，都在齐学习时并行地使用L_D训练一个判别器来计算 W-dis。图 5.1 对于 W-dis 的实验中，空间扰动同时也引入了空间关系的变化。而 W-dis 显示出它可以很好地识别空间关系的变化带来的不自然性。将并行训练的判别器仍记为$\boldsymbol{D}$，则 W-dis 的定义如下：

$$\text{W-dis}\left(Y_i, \hat{Y}_i\right) = D(Y_i) - D(\hat{Y}_i) \tag{5.3}$$

（4） Rhythm Density Error (RDE)

RDE 是本文提出的一个新的评价指标，它衡量生成动作与真是动作、频率分布的相似度。具体地，首先计算动作的功率谱密度（Power Spectral Density, PSD），再去除极低频的噪音（对应于身体转向、倾斜等低频大幅值动作成分）。本文假设 40BPM 是音乐节奏的一个大致的下界，因此使用$f = 40\text{BPM}/60\text{BPM} \approx 0.7\text{Hz}$作为频率下界，分离小于此界限的动作分量。最后，使用 log 和常数$k = 10^7$将指标值缩放至合适的区间。RDE 的定义如下：

$$\text{RDE}\left(Y_i, \hat{Y}_i\right) = \log\left[k\left\|\sum_j^{26} \text{PSD}_{f>0.7\text{HZ}}(Y_i[j]) - \sum_j^{26} \text{PSD}_{f>0.7\text{HZ}}\left(\hat{Y}_i[j]\right)\right\|_2^2 + 1\right] \tag{5.4}$$

（5） Strength Contour Error (SCE)

指挥动作的力度特已经被比较广泛地应用在了各种指挥动作感知方法中。本文定义 SCE 以比对生成动作与真是动作力度变化的相似程度。指挥动作的力度变化可以由各个关键点的一阶差分得到，但是直接比对这些一阶差分之和对于局部的错位不够鲁棒。因此，本文经验性地增加了一个池化降采样层，来提取更宽窗口内的力度变化趋势。具体地，对得到的一阶差分之和施加核为 60 帧（2 秒），步长为 30 帧（1 秒）的平均池化，称得到的曲线为力度轮廓（strength contour）。SCE 即是对比生成动作与真实动作力度轮廓之间的差异。类似地，最后使用 log 和常数$k = 10^7$将指标值缩放至合适的区间。SCE 的定义如下：

$$\mathrm{SCE}(Y_i,\hat{Y}_i)=\log\left[k\left\|\mathrm{pool}\left(\sum_j^{26}Y_i[j]\right)-\mathrm{pool}\left(\sum_j^{26}\hat{Y}_i[j]\right)\right\|_2^2+1\right] \tag{5.5}$$

（6） Standard Deviation Percentage (SDP)

本文使用生成指挥动作与真实指挥动作标准差之比来反映动作幅值的大小。例如，当有严重的过度平滑问题时，该动作的标准差将趋于 0%。而理想的生成器生成动作的标准差与真实动作的标准差之比应在 100%附近。令T^y表示动作Y_i的总帧数，$\bar{y}$表示平均姿态，则 SDP 可以定义为：

$$\mathrm{SDP}(Y_i,\hat{Y}_i)=\frac{\mathrm{SD}(\hat{Y}_i)}{\mathrm{SD}(Y_i)}\text{，其中 }\mathrm{SD(Y_i)}=\sqrt{\sum_{t=1}^{T^y}\frac{\|y_t-\bar{y}\|_2^2}{T^y-1}} \tag{5.6}$$

5.3 同步损失与对抗损失的平衡

本文首先使用开发集来寻找生成器损失函数L_G中λ_{sync}与λ_{adv}的最佳设定。具体地，固定$\lambda_{adv}=1$，使用$\lambda_{sync}=\{0.001,\ 0.01,\ 0.02,\ 0.05\ ,0.1,1\}$训练生成器，然后观察各个参数下模型的性能。其结果如图 5.2 与图 5.3 所示。本文发现，当$\lambda_{adv}=1,\lambda_{sync}=0.05$时模型取得了最佳的 RDE 与 SCE。根据同步损失与 W-dis 的变化情况，无论是更偏左（即更大的λ_{adv}，对真实性更多的强调）还是更偏右（更大的λ_{sync}，对同步性更多的强调）都会导致一项损失覆盖另一项损失，从而导致更差的 RDE 与 SCE。因此，在余下的实验中，都将采用$\lambda_{adv}=1,\lambda_{sync}=0.05$作为生成器损失函数$L_G$的超参数。

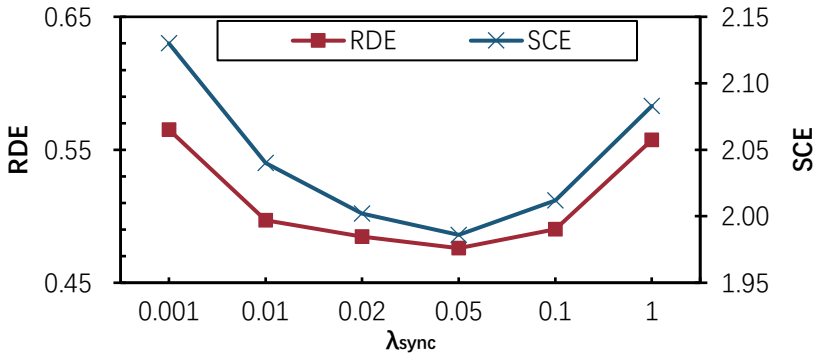


图 5.2 损失函数不同权重下 RDE 与 SCE 的变化

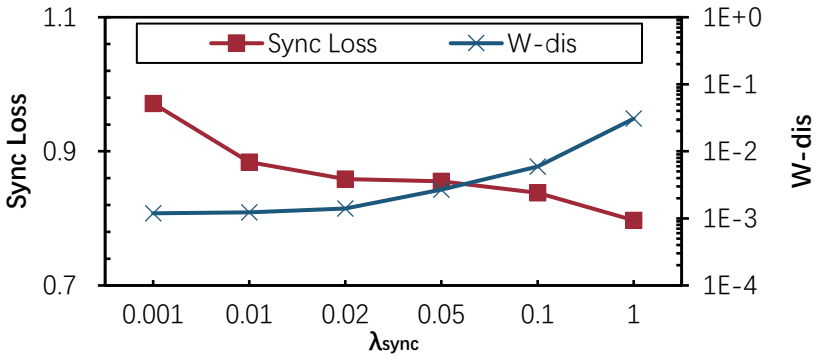


图 5.3 损失函数不同权重下同步损失（sync loss）与 W-dis 的变化

5.4 性能对比

在确定λ_{sync}与λ_{adv}的最佳设定后，本文在测试集上对比音乐驱动的指挥动作生成任务的性能。由于绝大部分的现有指挥动作生成方法都是基于规则的，很难为这些基于规则的方法与本文基于学习的方法设计公平的对比。此外，这些方法需要 MIDI 格式的音乐作为输入，但本文直接输入音乐音频数据。唯一以个现有的基于学习的指挥动作生成方法是 Wang 等人在 2003 年提出的 KHMM[4]方法，但该方法需要手动地为测试音乐选择合适的模型。同时，该模型原本是面向较小规模的数据集设计的，时间复杂度较高，无法在本文 *ConductorMotion100* 数据集上训练。由于现有的指挥动作生成方法都无法作为提出方法的对比组，本文选择了 3 个原本为其他音频-动作生成任务设计的模型作为对比。这些任务包括生成舞蹈动作[125]、说话动作[29]以及乐器演奏动作[25]。涉及的模型都不需要额外的先验知识，因此可以直接地用于学习生成指挥动作。

（1） Shlizerman 等人[25]，LSTM

该模型基于 LSTM，原本设计于预测乐器（钢琴或小提琴）演奏的动作。该模型包含一个带有 200 个隐藏神经元的单向 LSTM 层，以及若干全连接层。模型以梅尔倒谱系数（Mel-scale FrequencyCepstral Coefficients，MFCC）作为输入，通过优化 MSE 损失拟合真实的动作。

（2） Yalta 等人[125]，CNN-LSTM

该模型原本设计于生成舞蹈动作，使用一个 CNN 从梅尔频谱图中提取音乐特征，再使用 LSTM 的编码器-解码器进行动作生成。除了用于拟合真实动作的 MSE 损失外，该模型的损失函数还包括一项用于鼓励 LSTM 编码器输出的特征与真实动作标准差同步变化的对比损失。

（3） Ginosar 等人[29], GAN

该对比组结合了 KHMM[4]方法所采用的音乐特征与文献[29]中采用的损失函数。具体地，音乐特征是包含三个维度的音高，响度与节拍的特征。其中，节拍数据同按照原文献描述的方法[4]被转化为三角波的形式。损失函数是一个包含 L1 回归损失与对抗损失的联合损失。本文增加了一个额外的梯度惩罚项来提升对抗训练的稳定性。

以上三个对比模型与本文提出的M²SGAN模型的性能如表 2 所示。在评估指标 SE、W-dis、RDE、SCE、SDP 上，本文提出的M²SGAN都达到了最佳的性能。其中，在 SE、RDE、SCE 上的优势证明M²SGAN可以最准确地学习音乐与指挥动作之间的关联关系。在 W-dis 与 SDP 上的优势证明M²SGAN生成的动作真实度最高。值得注意的是，M²SGAN并没有达到最佳的 MSE。然而，由于具有对小幅值过度平滑动作的偏好，MSE 并不能准确地反映生成结果的准确性或真实性。可以看出， Shlizerman 等人-LSTM、Yalta 等人-CNN-LSTM 两个对照组上较低的 MSE 是由较低的 SDP 高度相关。

表 2 音乐驱动的指挥动作生成方法性能对比

对比模型	MSE ($\times 10^3$)	SE	W-dis ($\times 10^3$)	RDE	SCE	SDP
Shlizerman 等人[25], LSTM	3.50	1.301	87.47	0.9739	2.511	38.98%
Yalta 等人[125], CNN-LSTM	3.08	0.9105	50.85	0.9911	2.482	27.11%
Ginosar 等人[29], GAN	6.60	1.371	29.98	0.9437	2.864	97.93%
本文， M²SGAN	5.40	**0.8834**	**1.4264**	**0.049**	**2.046**	**99.62%**

为了更直观地展示生成动作的多样性，图 5.4 展示了不同方法生成动作的分布，以及真实动作的分布。这些指挥动作对应于贝多芬 C 小调第五交响曲 Op.67 的第一乐章。指挥动作动作以 0.1fps 的采样率采样并叠加。手部的轨迹对应 30 帧（即 1 秒）的长度。可以看出，基于 MSE 损失的两个对比方法（Shlizerman 等人-LSTM、Yalta 等人- CNN-LSTM）生成动作的幅值非常有限，它们存在着严重的过度平滑的问题。相比而言，基于生成对抗网络的方法（Ginosar 等人-GAN 与本文的M²SGAN）生成动作的分布更接近于真实动作。对比这两个方法，本文发现M²SGAN与音乐的吻合程度更好。然而，很难在本文中以图片的形式体现这一点。因此，本文以视频的形式对照了各种方法生成动作的效果，同时也设计了一个图灵测试。该视频位于 https://youtu.be/8lr5Q2qg58w 。

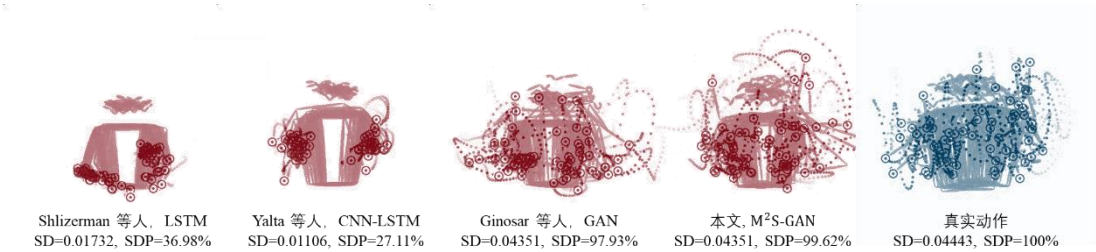
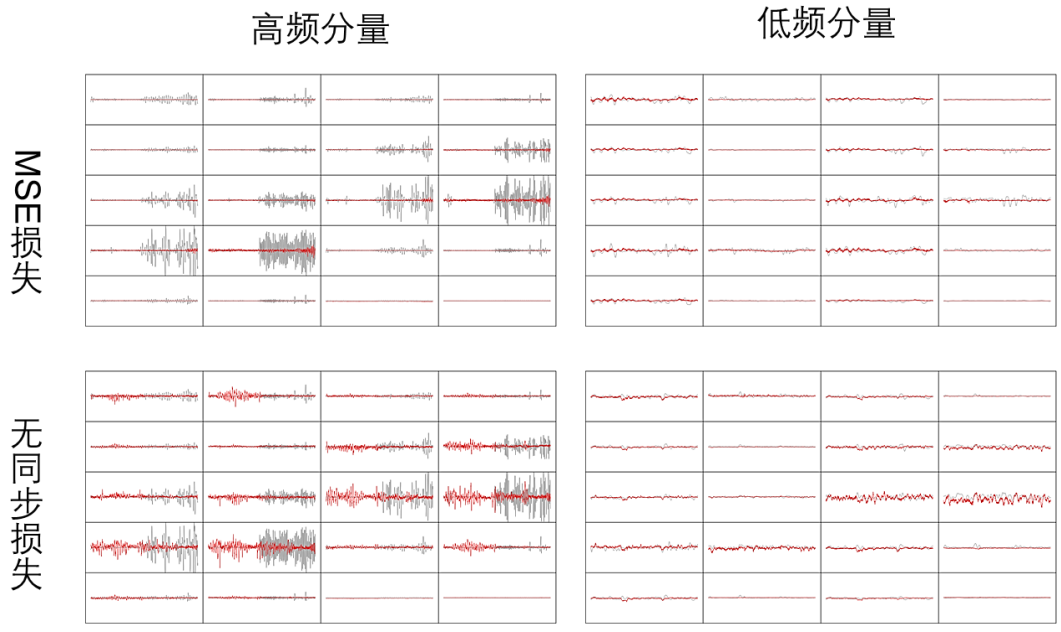


图 5.4 不同方法生成动作的分布

5.5 同步损失与对抗损失有效性的消融实验

本消融实验的目的在于证明本文提出的生成学习阶段中生成器损失函数L_G中，对抗损失与同步损失两者各自的有效性。为此，本文设置了 3 个对照组，分别是 MSE 损失、无同步损失、无对抗损失。MSE 损失只进行基于均方根误差损失的回归训练。无同步损失与无对抗损失指分别将L_G中的λ_{sync}与λ_{adv}设为 0 的对照组。

消融实验的结果如图 5.5 所示，其中，生成动作用红色表示，真实动作用灰色表示。可以看出 MSE 损失下的模型输出幅值很小，这一过度平滑的现象是由 MSE 本身的缺点造成的[28]。无同步损失下生成的动作与音乐不同步，无对抗损失下生成的动作缺乏真实性。这分别体现了模型在缺乏$P_G \to P_c$或$P_G \to P_{data}$限制时，都不能完成指挥动作条件生成的任务。相比之下，使用完整的生成器损失函数L_G（同步+对抗损失）训练的模型则能够生成真实且同步的指挥动作，且生成的动作不追求与真实动作一致，而实寻求拟合其概率分布P_{data}与P_c。



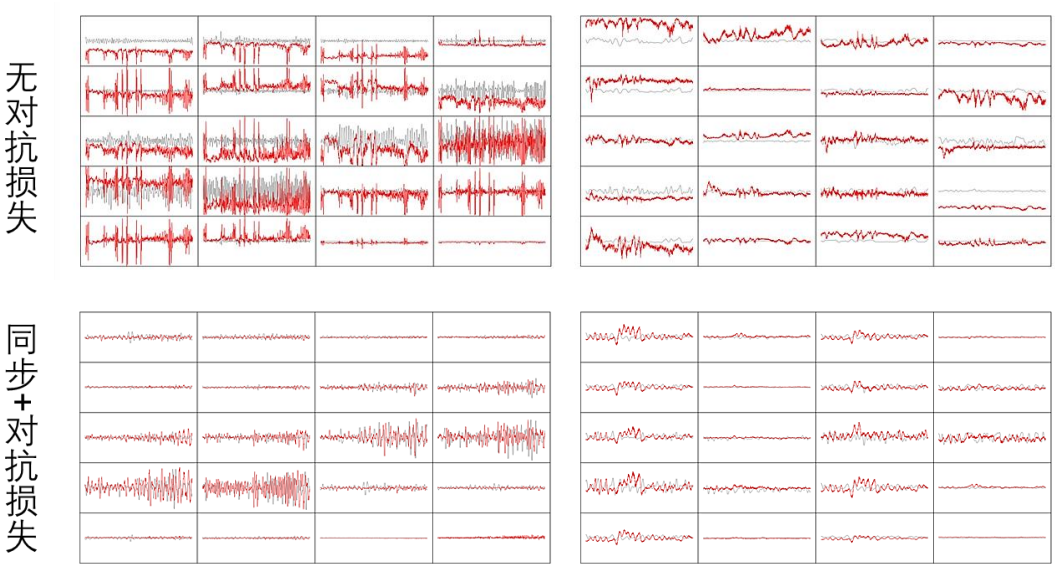


图 5.5 消融实验中不同对照组的生成效果

5.6 负样本采样策略的影响

本文认为，Hard Negatives 应是对于M²S学习而言的最佳负样本采样策略。在这一部分，本文将通过实验验证这一观点。本文分别使用 Easy Negatives、Hard Negatives 以及 Super-hard Negatives 训练M²SNet，然后观察训练好的模型在三种负样本上的测试正确率，其结果如表 3 所示。值得注意的是，使用 Hard Negatives 上训练的M²SNet在三种测试负样本上都达到了最佳的正确率。这证明了 Hard Negatives 同时包含语义的相关性（如 Easy Negatives）和时间的同步性（如 Super-hard Negatives）。

表 3 使用不同负样本采样策略下M²SNet的正确率

负样本采样策略	训练正确率	测试正确率		
		Easy	Hard	Super-hard
Easy	75.14%	67.56%	57.90%	53.01%
Hard	68.79%	**72.60%**	**67.83%**	**62.03%**
Super-hard	61.79%	65.71%	63.53%	61.27%

同时，由于较高的错误负采样机会，Easy Negatives 有时鼓励模型判别每个样本的身份，有时正常寻求学习语义相关性，如图 5.6 所示，这导致了在 Easy Negatives 下的M²SNet训练过程十分不稳定。相比之下，Hard Negatives 和 Super-hard Negatives 则十分稳定。

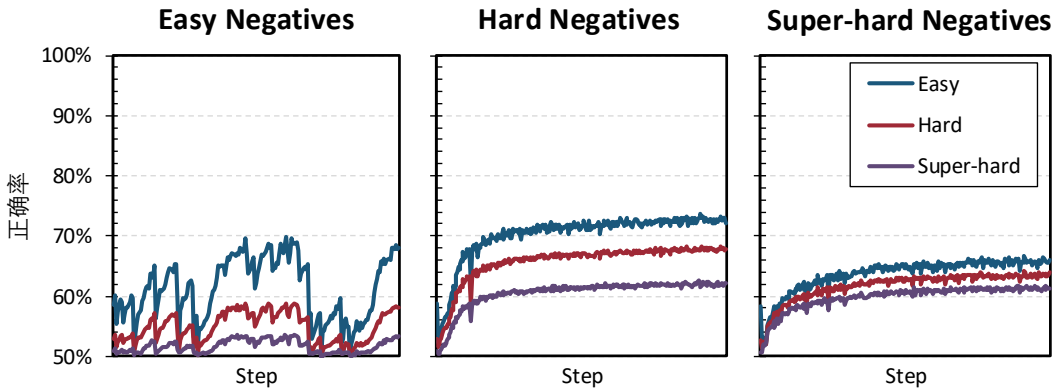


图 5.6 不同负样本采样策略下的训练曲线

5.7 训练集规模的影响

本文还进行了验证训练集规模影响的实验来证明本文提出的方法相较于传统基于回归损失方法的优越性。具体地，使用不同的训练集规模（1，4，8，16，32，64，90 小时）训练生成模型，并观察这些模型的性能变化。本实验对比了本文提出的M²SGAN与基于 MSE 训练的M²SGAN两种模型，其结果如图 5.7 所示。可以看出，对于小规模的训练集，MSE 方法可以很好地拟合训练集，达到较低的训练误差。但随着数据集规模逐渐超过模型的容量，其训练误差开始上升。此时模型逐渐发现降低输出动作的幅值可以达成更低的训练误差，这体现在 SDP 的变化中。最终，当训练集时长为 90 小时时，其 SDP 仅能达到约 50%。相反，本文提出的M²SGAN生成动作的 SDP 一直稳定在 100%附近。这是由于M²SGAN不寻求拟合真实动作，而是学习动作的概率分布，这也就避免了过度平滑问题的出现。

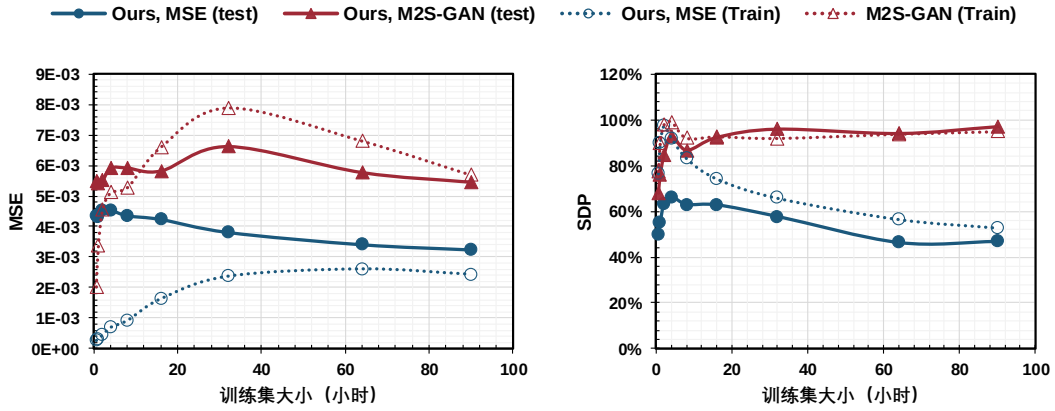


图 5.7 不同训练集规模下的模型性能

5.8 可视化M²SNet特征

根据网络结构，M²SNet需要为预测长约 3 秒的音频-动作样本对是否匹配，而这是一个相当困难的任务。为了完成这样的学习任务，M²SNet的 E_{music} 与 E_{motion} 需要将音频特征与指挥动作嵌入到一个共享的特征空间中。图 5.8 的左右两栏展示了经过训练后的两个M²SNet提取的特征。其中，两栏中的音频输入是相同的（贝多芬第五交响曲第一乐章的前 33.3 秒），而右侧一栏的动作序列分别为经过动态动作频域分解后的高频与低频分量。可以看出，随着逐层的特征提取 $X \rightarrow h_{x1} \rightarrow h_{x2} \rightarrow h_{x3} \rightarrow E_{music}(X)$、$Y \rightarrow h_{y1} \rightarrow h_{y2} \rightarrow h_{y3} \rightarrow E_{motion}(Y)$，来自两个模态的特征图逐渐趋于一致。更重要的是，分别对应于高频与低频动作分量的两个M²SNet从同一段音频输入中提取到的音频特征差异很大。这说明，M²SNet的音频分支面向不同的动作分量有效地学习了不同的音频语义表示，而M²SNet的动作分支以音频特征为指导，有效地学习了音乐相关的动作特征。

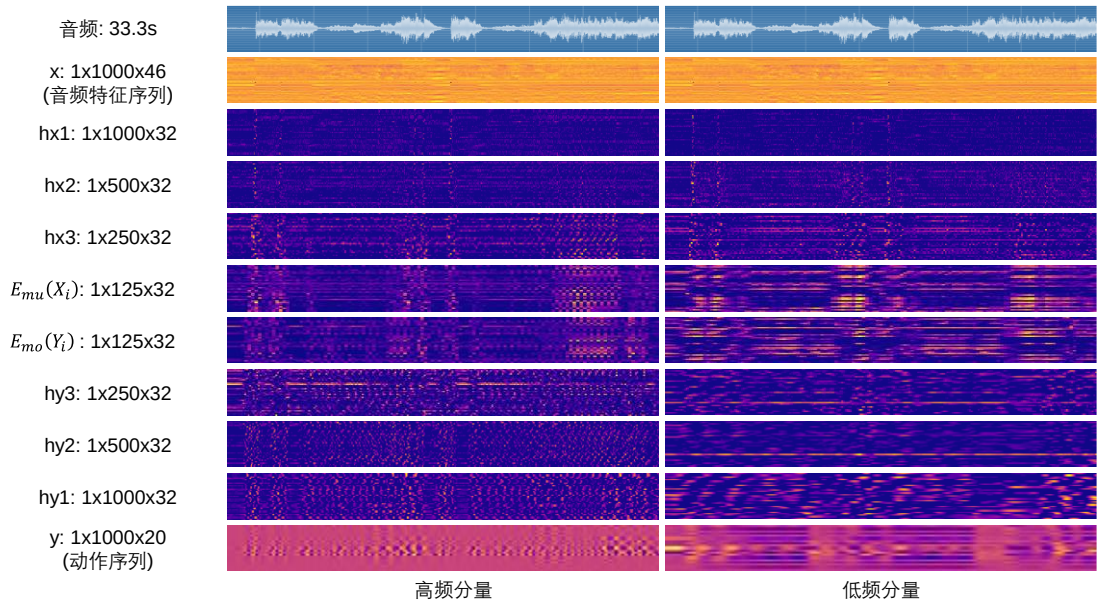


图 5.8 经过训练的两个M²SNet（分别对应高频与低频分量）提取的音频与动作特征

5.9 可视化指挥动作生成结果

为了更直观地展现本文提出方法的性能，本章节将对生成结果进行可视化展示。图 5.9 展示了对应于同一段音频的真实动作（蓝色）与生成动作（红色）。本文方法生成的动作灵活，真实，且并没有寻求拟合真实动作。此外，位于

https://www.bilibili.com/video/BV1Zy4y1W7Qq 的视频可以更加直观地展示了动态频域分解下高频分量与低频分量的生成效果。

图 5.10，图 5.11 与图 5.12 进一步展示了生成指挥动作基于 3 维动画渲染与姿态迁移的可视化。其中，图 5.10，图 5.11 为将生成的指挥动作填充为全身后使用 VideoPose3D [102]转化为 3 维坐标，再使用 MotionBuilder[4]将其绑定到 maxiamo[5]的 3 维人体模型上的渲染结果。图 5.12 展示的是使用基于 Liquid Warping GAN 的姿态迁移方法[106]将生成的指挥动作迁移至给定图像（上）上的效果。其对应的视频位于 https://www.bilibili.com/video/BV1aX4y1g7wh .

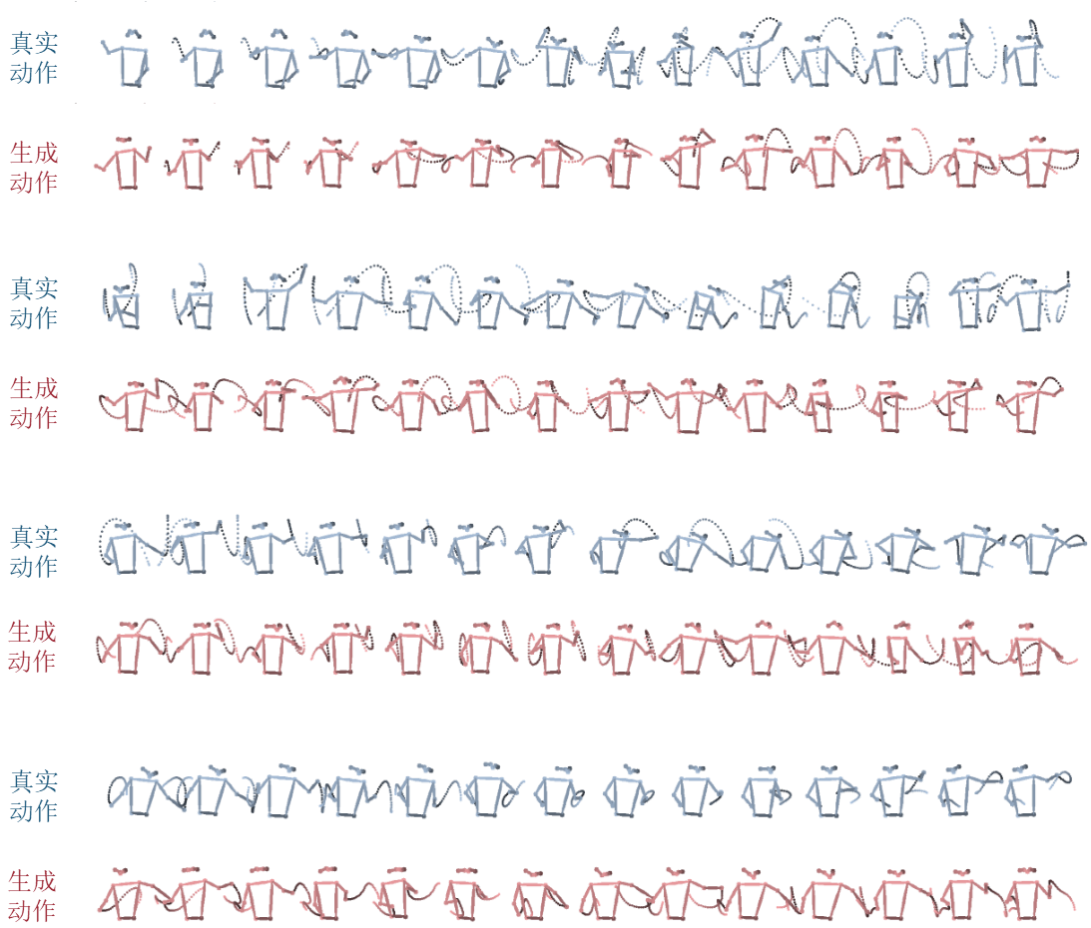


图 5.9 对应于同一段音乐的真实动作与生成动作

[4] https://www.autodesk.com/products/motionbuilder
[5] https://www.mixamo.com

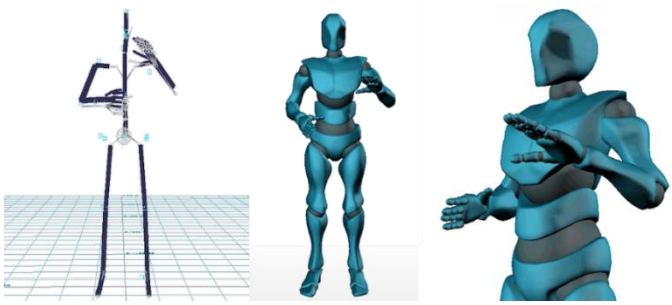


图 5.10 指挥动作的三维动画建模效果

图 5.11 指挥动作的 3 维动画建模效果

图 5.12 基于姿态迁移的指挥视频生成效果

第六章　总结与展望

本文首次将深度学习技术应用于音乐驱动的指挥动作生成任务。相较于现有的传统方法，本文的方法可以生成更加自然、美观、多样、且与音乐同步的指挥动作，其核心在于本文提出的动作的动态频域分解以及动作的跨模态感知。动作的动态频域分解以音乐节奏为依据将动作分解为高频分量与低频分量，显著降低了学习难度。跨模态感知是跨模态自监督学习与感知损失技术的融合，可以为生成其提供合理有效的音乐同步性监督信息，从而避免了传统回归损失的缺点。

本文还构建了一个大规模的指挥动作数据集，其总时长超过了很多现有的音乐-动作数据集以及音乐信息检索领域的数据集。根据本文所采用的跨模态自监督学习的核心思路，该数据集也有在音乐信息检索领域任务（如节拍跟踪）上潜在的应用价值。尽管提出的方法在实验中取得了较好的效果，但其仍有进一步改进和提升的空间。例如，在M^2S学习后音乐特征与指挥动作被嵌入到的共享特征空间中，指挥动作样本及对应音乐样本的潜在动作候选（即其他同样适合该音乐样本的指挥动作）构成一个以音乐样本位置为中心的分布。而本文所采用的跨模态感知损失相当于指引模型朝向分布中的特定一点，而非分布中心学习。

在未来的研究中，将从两个方向拓展本文的工作。首先是进一步挖掘 *ConductorMotion100* 数据集在音乐信息检索领域的应用前景，例如将其作为节拍跟踪（beat tracking）任务的预训练数据集，或者以在线学习的方式引入动态频域分解，循环地优化频域分解的准确率。第二个方向是进一步拓宽本文提出的结合自监督判别任务与跨模态生成任务的方法，在舞蹈生成、说话姿势生成等任务上验证本文提出方法的可行性，并尝试将该方法拓展为一个通用的面向跨模态条件生成的学习框架。

参考文献

[1] 李严君.乐队指挥及中西方指挥形式研究[J].北方音乐,2020(18):251-252。

[2] 张金鑫.多样化乐队指挥形式的探究[J].当代音乐,2018(06):108-109.

[3] 程酢培.试论指挥“图式”的重要性[J].当代音乐,2016(14):74-77.

[4] T. Wang, N. Zheng, Y. Li, Y.-Q. Xu, and H.-Y. Shum, “Learning kernel-based HMMs for dynamic sequence synthesis,” Graph. Model., vol. 65, no. 4, Art. no. 4, 2003, doi: 10.1016/S1524-0703(03)00040-7.

[5] Z. Ruttkay, Z. Huang, and A. Eliens, “The Conductor: Gestures for Embodied Agents with Logic Programming,” in Logic Programming, Joint Annual ERCIM/CoLogNet Workshop on Constraint and Logic Programming, 2003, pp. 9–16.

[6] D. Reidsma, A. Nijholt, and P. Bos, “Temporal interaction between an artificial orchestra conductor and human musicians,” Comput. Entertain., vol. 6, no. 4, Art. no. 4, 2008, doi: 10.1145/1461999.1462005.

[7] Á. Sarasúa, B. Caramiaux, and A. Tanaka, “Machine Learning of Personal Gesture Variation in Music Conducting,” in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016, 2016, pp. 3428–3432, doi: 10.1145/2858036.2858328.

[8] R. Takatsu, Y. Maki, T. Inoue, K.-i. Okada, and H. Shigeno, “Multiple virtual conductors allow amateur orchestra players to perform better and more easily,” in 20th IEEE International Conference on Computer Supported Cooperative Work in Design, CSCWD 2016, Nanchang, China, May 4-6, 2016, 2016, pp. 486–491, doi: 10.1109/CSCWD.2016.7566038.

[9] D. G. Dansereau, N. Brock, and J. R. Cooperstock, “Predicting an Orchestral Conductor’s Baton Movements Using Machine Learning,” Comput. Music. J., vol. 37, no. 2, Art. no. 2, 2013, doi: 10.1162/COMJ_a_00173.

[10] 唐郅,侯进.基于深度神经网络的语音驱动发音器官的运动合成[J].自动化学报,2016,42(06):923-930.

[11] 于灵云. 基于文本/语音驱动的高自然度人脸动画生成[D].中国科学技术大学,2020. doi: 10.27517/d.cnki.gzkju.2020.000412.

[12] H.-Y. Lee et al., “Dancing to Music,” in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing

Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 3581–359.

[13] H.-K. Kao and L. Su, “Temporally Guided Music-to-Body-Movement Generation,” in MM ’20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, 2020, pp. 147–155, doi: 10.1145/3394171.3413848.

[14] 殷瑞刚,魏帅,李晗,于洪.深度学习中的无监督学习方法综述[J].计算机系统应用,2016,25(08):1-7. doi: 10.15888/j.cnki.csa.005283.

[15] 王乐,杨晓敏.基于感知损失的遥感图像全色锐化反馈网络 [J/OL]. 计算机科学 :1-14[2021-06-01]. http://kns.cnki.net/kcms/detail/50.1075.TP.20210421.15 29. 049. html.

[16] 许春冬,凌贤鹏,应冬文,王晶.基于时频感知神经网络的语音频带扩展 [J/OL]. 信号处理: 1-12[2021-06-01]. http://kns.cnki.net/kcms/detail/11.2406.tn.20210517. 1238. 010.html.

[17] 孙浩,徐延杰,陈进,雷琳,计科峰,匡纲要.基于自监督对比学习的深度神经网络对抗鲁棒性提升[J/OL].信号处理: 1-11[2021-06-01]. http://kns.cnki.net/kcms/detail/11.2406.TN.20210422.1830.033.html.

[18] 王坤峰,苟超,段艳杰,林懿伦,郑心湖,王飞跃.生成式对抗网络 GAN 的研究进展与展望[J].自动化学报,2017,43(03):321-332.

[19] K. Haag and H. Shimodaira, “Bidirectional LSTM Networks Employing Stacked Bottleneck Features for Expressive Speech-Driven Head Motion Synthesis,” in Intelligent Virtual Agents - 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20-23, 2016, Proceedings, 2016, vol. 10011, pp. 198–207, doi: 10.1007/978-3-319-47665-0_18.

[20] N. Yalta, “Sequential Deep Learning for Dancing Motion Generation,” SIG-Challenge 2016. Nov. 2016.

[21] N. Sadoughi and C. Busso, “Joint Learning of Speech-Driven Facial Motion with Bidirectional Long-Short Term Memory,” in Intelligent Virtual Agents - 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings, 2017, vol. 10498, pp. 389–402, doi: 10.1007/978-3-319-67401-8_49.

[22] N. Sadoughi and C. Busso, “Novel Realizations of Speech-Driven Head Movements with Generative Adversarial Networks,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6169–6173, doi: 10.1109/ICASSP.2018.8461967.

[23] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, “Lip Movements Generation at a Glance,” in Computer Vision – ECCV 2018, Cham, 2018, pp. 538–553.

[24] T. Tang, J. Jia, and H. Mao, “Dance with Melody: An LSTM-autoencoder Approach to Music-oriented Dance Synthesis,” in 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018, 2018, pp. 1598–1606, doi: 10.1145/3240508.3240526.

[25] E. Shlizerman, L. M. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, “Audio to Body Dynamics,” in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 7574–7583, doi: 10.1109/CVPR.2018.00790.

[26] Y. Ferstl and R. McDonnell, “Investigating the use of recurrent motion modelling for speech gesture generation,” in Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018, Sydney, NSW, Australia, November 05-08, 2018, 2018, pp. 93–98, doi: 10.1145/3267851.3267898.

[27] B. Li, A. Maezawa, and Z. Duan, “Skeleton Plays Piano: Online Generation of Pianist Body Movements from MIDI Performance,” in Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018, 2018, pp. 218–224.

[28] Y. Ferstl, M. Neff, and R. McDonnell, “Multi-Objective Adversarial Gesture Generation,” Newcastle upon Tyne, United Kingdom, 2019, doi: 10.1145/3359566.3360053.

[29] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, “Learning Individual Styles of Conversational Gesture,” Jun. 2019, doi: 10.1109/cvpr.2019.00361.

[30] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, “Talking Face Generation by Adversarially Disentangled Audio-Visual Representation,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, Art. no. 01, Jul. 2019, doi: 10.1609/aaai.v33i01.33019299.

[31] N. Yalta, S. Watanabe, K. Nakadai, and T. Ogata, “Weakly-Supervised Deep Recurrent Neural Networks for Basic Dance Step Generation,” in International

Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019, 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8851872.

[32] J. Lee, S. Kim, and K. Lee, “Automatic Choreography Generation with Convolutional Encoder-decoder Network,” in Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019, 2019, pp. 894–899.

[33] Y. Qi, Y. Liu, and Q. Sun, “Music-Driven Dance Generation,” IEEE Access, vol. 7, pp. 166540–166550, 2019, doi: 10.1109/ACCESS.2019.2953698.

[34] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, “Talking Face Generation by Conditional Recurrent Adversarial Network,” in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, 2019, pp. 919–925, doi: 10.24963/ijcai.2019/129.

[35] X. Ren, H. Li, Z. Huang, and Q. Chen, “Self-supervised Dance Video Synthesis Conditioned on Music,” in MM ’20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, 2020, pp. 46–54, doi: 10.1145/3394171.3413932.

[36] H. Ahn, J. Kim, K. Kim, and S. Oh, “Generative Autoregressive Networks for 3D Dancing Move Synthesis From Music,” IEEE Robotics and Automation Letters, vol. 5, no. 2, Art. no. 2, 2020, doi: 10.1109/LRA.2020.2977333.

[37] A. Bogaers, Z. Yumak, and A. Volk, “Music-Driven Animation Generation of Expressive Musical Gestures,” in Companion Publication of the 2020 International Conference on Multimodal Interaction, ICMI Companion 2020, Virtual Event, The Netherlands, October, 2020, 2020, pp. 22–26, doi: 10.1145/3395035.3425244.

[38] Z. Ye et al., “ChoreoNet: Towards Music to Dance Synthesis with Choreographic Action Unit,” in MM ’20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, 2020, pp. 744–752, doi: 10.1145/3394171.3414005.

[39] X. Guo, J. Li, and Y. Zhao, “DanceIt: Music-inspired Dancing Video Synthesis,” CoRR, vol. abs/2009.08027, 2020.

[40] Y. Duan et al., “Semi-Supervised Learning for In-Game Expert-Level Music-to-Dance Translation,” CoRR, vol. abs/2009.12763, 2020.

[41] J.-W. Liu, H.-Y. Lin, Y.-F. Huang, H.-K. Kao, and L. Su, “Body Movement Generation for Expressive Violin Performance Applying Neural Networks,” in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, 2020, pp. 3787–3791, doi: 10.1109/ICASSP40776.2020.9054463.

[42] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, “End-To-End Generation of Talking Faces from Noisy Speech,” in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, 2020, pp. 1948–1952, doi: 10.1109/ICASSP40776.2020.9054103.

[43] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang, “Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning,” 2021.

[44] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, “Learn to Dance with AIST++: Music Conditioned 3D Dance Generation,” ArXiv, vol. abs/2101.08779, 2021.

[45] G. Sun, Y. Wong, Z. Cheng, M. S. Kankanhalli, W. Geng, and X. Li, “DeepDance: Music-to-Dance Motion Choreography With Adversarial Learning,” IEEE Trans. Multim., vol. 23, pp. 497–509, 2021, doi: 10.1109/TMM.2020.2981989.

[46] B. Li, Y. Zhao, and L. Sheng, “DanceNet3D: Music Based Dance Generation with Parametric Motion Transformer,” CoRR, vol. abs/2103.10206, 2021.

[47] M. Lee, G. Garnett, and D. Wessel, “An adaptive conductor follower,” in Proceedings of the International Computer Music Conference, 1992, p. 454.

[48] R. Typke, F. Wiering, and R. C. Veltkamp, “A Survey of Music Information Retrieval Systems,” in ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings, 2005, pp. 153–160.

[49] A. Brown and Y. Sasson, “Maestro: using Technology to Improve kinesthetic Skill Learning of Music conductors,” 2012.

[50] Á. Sarasúa, “Context-aware gesture recognition in classical music conducting,” in ACM Multimedia Conference, MM ’13, Barcelona, Spain, October 21-25, 2013, 2013, pp. 1059–1062, doi: 10.1145/2502081.2502216.

[51] S. Cosentino et al., “Natural human-robot musical interaction: understanding the music conductor gestures by using the WB-4 inertial measurement system,” Adv. Robotics, vol. 28, no. 11, Art. no. 11, 2014, doi: 10.1080/01691864.2014.889577.

[52] Á. Sarasúa and E. Guaus, “Beat Tracking from Conducting Gestural Data: a Multi-Subject Study,” in International Workshop on Movement and Computing,

MOCO ’14, Paris, France, June 16-17, 2014, 2014, p. 118, doi: 10.1145/2617995.2618016.

[53] R. Schramm, C. R. Jung, and E. R. Miranda, “Dynamic Time Warping for Music Conducting Gestures Evaluation,” IEEE Trans. Multim., vol. 17, no. 2, Art. no. 2, 2015, doi: 10.1109/TMM.2014.2377553.

[54] K. Lee, D. J. Cox, G. E. Garnett, and M. J. Junokas, “Express it!: An Interactive System for Visualizing Expressiveness of Conductor’s Gestures,” in Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition, C&C ’15, Glasgow, United Kingdom, June 22-25, 2015, 2015, pp. 141–150, doi: 10.1145/2757226.2757243.

[55] K. Lee, M. J. Junokas, M. Amanzadeh, and G. E. Garnett, “An analysis of basic expressive qualities in instrumental conducting,” in Proceedings of the 2nd International Workshop on Movement and Computing, MOCO 2015, Vancouver, British Columbia, Canada, August 14-15, 2015, 2015, pp. 148–155, doi: 10.1145/2790994.2791005.

[56] K. Karipidou, J. Ahnlund, A. Friberg, S. Alexanderson, and H. Kjellström, “Computer Analysis of Sentiment Interpretation in Musical Conducting,” in 12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017, 2017, pp. 400–405, doi: 10.1109/FG.2017.57.

[57] M. Lee, “Deep Neural Network Based Music Source Conducting System,” 2018.

[58] Y. V. S. Murthy and S. G. Koolagudi, “Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry: A Review,” ACM Comput. Surv., vol. 51, no. 3, Art. no. 3, Jun. 2018, doi: 10.1145/3177849.

[59] Y.-F. Huang, T.-P. Chen, N. Moran, S. Coleman, and L. Su, “Identifying Expressive Semantics in Orchestral Conducting Kinematics,” in Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019, 2019, pp. 115–122.

[60] F. Chin-Shyurng, S.-E. Lee, and M.-L. Wu, “Real-Time Musical Conducting Gesture Recognition Based on a Dynamic Time Warping Classifier Using a Single-Depth Camera,” Applied Sciences, vol. 9, no. 3, Art. no. 3, 2019, doi: 10.3390/app9030528.

[61] Y. Muraki, K. Kobayashi, K. Nishio, and K.-i. Kobori, “Generation of Brass Band Animation Synchronized with the Motion of Conductor’s Hand,” in HCI

International 2020 - Posters - 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part II, 2020, vol. 1225, pp. 204–211, doi: 10.1007/978-3-030-50729-9_29.

[62] A. Barmpoutis et al., “Assessing the Role of Virtual Reality with Passive Haptics in Music Conductor Education: A Pilot Study,” in Virtual, Augmented and Mixed Reality. Design and Interaction - 12th International Conference, VAMR 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part I, 2020, vol. 12190, pp. 275–285, doi: 10.1007/978-3-030-49695-1_18.

[63] F. Pedersoli and M. Goto, “Dance Beat Tracking from Visual Information Alone,” Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR, 2020.

[64] S. Lemouton, R. Borghesi, S. Haapamäki, F. Bevilacqua, and E. Fléty, “Following Orchestra Conductors: the IDEA Open Movement Dataset,” in Proceedings of the 6th International Conference on Movement and Computing, MOCO 2019, Tempe, AZ, USA, October 10-12, 2019, 2019, pp. 25:1–25:6, doi: 10.1145/3347122.3359599.

[65] G. Nan et al., “Generative Adversarial Networks for Spatio-Temporal Data: A Survey,” ACM Trans.Intell. Syst. Technol., vol. 37(4)7, no. 111, Art. no. 111, 2020, doi: 10.1145/1122445.1122456.

[66] M. Arjovsky and L. Bottou, “Towards Principled Methods for Training Generative Adversarial Networks,” 2017.

[67] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures,” IEEE Signal Processing Magazine, vol. 26, no. 1, Art. no. 1, 2009, doi: 10.1109/MSP.2008.930649.

[68] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, “ImageNet: A large-scale hierarchical image database,” in 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[69] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[70] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” in Computer Vision – ECCV 2016, Cham, 2016, pp. 694–711.

[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595, doi: 10.1109/CVPR.2018.00068.

[72] X. Wang et al., “ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks,” in Computer Vision – ECCV 2018 Workshops, Cham, 2019, pp. 63–79.

[73] T. Tariq, J. L. Gonzalez Bello, and M. Kim, “A HVS-Inspired Attention to Improve Loss Metrics for CNN-Based Perception-Oriented Super-Resolution,” in 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3904–3912, doi: 10.1109/ICCVW.2019.00484.

[74] M. S. Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H. K. Ekenel, and J.-P. Thiran, “SROBB: Targeted Perceptual Loss for Single Image Super-Resolution,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2710–2719, doi: 10.1109/ICCV.2019.00280.

[75] A. R. Tej, S. Sukanta Halder, A. P. Shandeelya, and V. Pankajakshan, “Enhancing Perceptual Loss with Adversarial Feature Matching for Super-Resolution,” in 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9207102.

[76] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, “SACNN: Self-Attention Convolutional Neural Network for Low-Dose CT Denoising With Self-Supervised Perceptual Loss Network,” IEEE Transactions on Medical Imaging, vol. 39, no. 7, Art. no. 7, 2020, doi: 10.1109/TMI.2020.2968472.

[77] R. Arandjelovic and A. Zisserman, “Look, Listen and Learn,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 609–617, doi: 10.1109/ICCV.2017.73.

[78] A. Owens and A. A. Efros, “Audio-Visual Scene Analysis with Self-Supervised Multisensory Features,” in Computer Vision – ECCV 2018, Cham, 2018, pp. 639–658.

[79] B. Korbar, D. Tran, and L. Torresani, “Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization,” in Proceedings of the 32nd

International Conference on Neural Information Processing Systems, Montréal, Canada, 2018, pp. 7774–7785.

[80] R. Arandjelović and A. Zisserman, “Objects that Sound,” in Computer Vision – ECCV 2018, Cham, 2018, pp. 451–466.

[81] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3852–3856, doi: 10.1109/ICASSP.2019.8682475.

[82] G. Verma, E. G. Dhekane, and T. Guha, “Learning Affective Correspondence between Music and Image,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3975–3979, doi: 10.1109/ICASSP.2019.8683133.

[83] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[84] X. Liu et al., “Self-supervised Learning: Generative or Contrastive,” CoRR, vol. abs/2006.08218, 2020.

[85] 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 Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 2020, pp. 3884–3892, doi: 10.1145/3394171.3413869.

[86] J.-B. Alayrac et al., “Self-Supervised MultiModal Versatile Networks,” in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 25–37.

[87] P. Morgado, N. Vasconcelos, and I. Misra, “Audio-Visual Instance Discrimination with Cross-Modal Agreement,” CoRR, vol. abs/2004.12943, 2020.

[88] M. Patrick, Y. M. Asano, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, “Multi-modal Self-Supervision from Generalized Data Transformations,” CoRR, vol. abs/2003.04298, 2020.

[89] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, “Self-Supervised Learning by Cross-Modal Audio-Video Clustering,” in

Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 9758–9770.

[90] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9726–9735, doi: 10.1109/CVPR42600.2020.00975.

[91] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, 2020, vol. 119, pp. 1597–1607.

[92] S. Bai, J. Z. Kolter, and V. Koltun, “Convolutional Sequence Modeling Revisited,” 2018.

[93] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[94] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” in The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016, 2016, p. 125.

[95] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” CoRR, vol. abs/1701.07875, 2017.

[96] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved Training of Wasserstein GANs,” in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5767–5777.

[97] I. Goodfellow et al., “Generative Adversarial Nets,” in Advances in Neural Information Processing Systems, 2014, vol. 27.

[98] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” CoRR, vol. abs/1804.02767, 2018.

[99] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional Multi-person Pose Estimation,” 2017.

[100] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, “CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark,” arXiv preprint arXiv:1812.00324, 2018.

[101] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, “Pose Flow: Efficient Online Pose Tracking,” 2018.

[102] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, “3D human pose estimation in video with temporal convolutions and semi-supervised training,” 2019.

[103] P. Grosche and M. Muller, “Extracting Predominant Local Pulse Information From Music Recordings,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 6, Art. no. 6, 2011, doi: 10.1109/TASL.2010.2096216.

[104] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” 2019.

[105] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio, “RMSProp and equilibrated adaptive learning rates for non-convex optimization,” CoRR, vol. abs/1502.04390, 2015.

[106] W. Liu and W. L. Zhixin Piao Min Jie, “Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis,” 2019.

[107] F. Gouyon et al., “An experimental comparison of audio tempo induction algorithms,” IEEE Trans. Speech Audio Process., vol. 14, no. 5, Art. no. 5, 2006, doi: 10.1109/TSA.2005.858509.

[108] M. E. P. Davies, N. Degara, and M. D. Plumbley, “Evaluation Methods for Musical Audio Beat Tracking Algorithms.” 2009.

[109] S. W. Hainsworth and M. D. Macleod, “Particle Filtering Applied to Musical Tempo Tracking,” EURASIP J. Adv. Signal Process., vol. 2004, no. 15, Art. no. 15, 2004, doi: 10.1155/S1110865704408099.

[110] F. Gouyon, “A computational approach to rhythm description - Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing,” 2005.

[111] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, “Selective Sampling for Beat Tracking Evaluation,” IEEE Trans. Speech Audio Process., vol. 20, no. 9, Art. no. 9, 2012, doi: 10.1109/TASL.2012.2205244.

[112] J. Hockman, M. E. P. Davies, and I. Fujinaga, “One in the Jungle: Downbeat Detection in Hardcore, Jungle, and Drum and Bass,” in Proceedings of the 13th

International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal, October 8-12, 2012, 2012, pp. 169–174.

[113] G. Peeters and J. Flocon-Cholet, “Perceptual tempo estimation using GMM-regression,” in Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies, MIRUM ’12, Nara, Japan, October 29 - November 02, 2012, 2012, pp. 45–50, doi: 10.1145/2390848.2390861.

[114] P. Knees et al., “Two Data Sets for Tempo Estimation and Key Detection in Electronic Dance Music Annotated from User Corrections,” in Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015, 2015, pp. 364–370.

[115] G. Tzanetakis and P. R. Cook, “Musical genre classification of audio signals,” IEEE Trans. Speech Audio Process., vol. 10, no. 5, Art. no. 5, 2002, doi: 10.1109/TSA.2002.800560.

[116] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, “Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications,” IEEE Transactions on Multimedia, vol. 21, no. 2, Art. no. 2, 2019, doi: 10.1109/TMM.2018.2856090.

[117] A. Bazzica, J. C. van Gemert, C. C. S. Liem, and A. Hanjalic, “Vision-based Detection of Acoustic Timed Events: a Case Study on Clarinet Note Onsets,” CoRR, vol. abs/1706.09556, 2017.

[118] A. P. Carrillo, J. L. Arcos, and M. M. Wanderley, “Estimation of Guitar Fingering and Plucking Controls Based on Multimodal Analysis of Motion, Audio and Musical Score,” in Music, Mind, and Embodiment - 11th International Symposium, CMMR 2015, Plymouth, UK, June 16-19, 2015, Revised Selected Papers, 2015, vol. 9617, pp. 71–87, doi: 10.1007/978-3-319-46282-0_5.

[119] O. Gillet and G. Richard, “ENST-Drums: an extensive audio-visual database for drum signals processing,” in ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October 2006, Proceedings, 2006, pp. 156–159.

[120] J. Abeßer, O. Lartillot, C. Dittmar, T. Eerola, and G. Schuller, “Modeling musical attributes to characterize ensemble recordings using rhythmic audio features,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and

Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, 2011, pp. 189–192, doi: 10.1109/ICASSP.2011.5946372.

[121] J. F. Montesinos, O. Slizovskaia, and G. Haro, “Solos: A Dataset for Audio-Visual Music Analysis,” in 22nd IEEE International Workshop on Multimedia Signal Processing, MMSP 2020, Tampere, Finland, September 21-24, 2020, 2020, pp. 1–6, doi: 10.1109/MMSP48831.2020.9287124.

[122] A. Vaswani et al., “Attention is All you Need,” in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[123] K. Cho et al., “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, 2014, pp. 1724–1734, doi: 10.3115/v1/d14-1179.

[124] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 2018, pp. 7444–7452.

[125] N. Yalta, S. Watanabe, K. Nakadai, and T. Ogata, “Weakly-Supervised Deep Recurrent Neural Networks for Basic Dance Step Generation,” in International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019, 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8851872.

附 录

A：个人简介

陈德龙，男，共青团员，2021 年 6 月毕业于河海大学信息学部计算机与信息学院计算机科学与技术专业。曾任中华学生联合会第二十七次代表大会代表、河海大学管弦乐团团长。曾获江苏省优秀共青团员、江苏省大学生年度人物提名奖、河海大学大学生年度人物、河海大学优秀毕业生等 10 余项省校荣誉荣誉。主持国家级创新训练项目 1 项，获国际科技竞赛、省级科技竞赛奖项各 1 项。本科期间共完成 8 篇学术论文，其中 7 篇为第一作者或通讯作者。共有 5 篇论文已发表，其中 3 篇一作论文分别获得 Best Demo，Best Presentation，Best Dataset Paper 奖项；3 篇期刊论文在投，包括 2 篇 SCI 一区的 IEEE Transection 汇刊论文，一篇 SCI 二区（影响因子 8.139）期刊的论文一审修改中。获国家发明专利 4 项受理，软件著作权 2 项授权。

B：本科期间撰写的学术论文

已发表或录用的论文

[C-1] MEP-3M: A Large-scale Multi-modal E-Commerce Products Dataset[C]. IJCAI 2021 Workshop on Long-Tailed Distribution Learning. （CCF-A 类，河海大学 A 类），第一作者，最佳数据集论文奖（Best Dataset Paper）

[C-2] VirtualConductor: Music-driven Conducting Video Generation System[C]. IEEE International Conference on Multimedia & Expo, ICME. 2021. （CCF-B 类,河海大学 A 类），第一作者，最佳演示奖（Best Demo），对应本毕业论文第六章 6.9 节

[C-3] Weakly Correlated Adversarial Learning for Cognitive Diagnosis System[C]. IEEE International Conference on Multimedia & Expo, ICME. 2021. （CCF-B 类，河海大学 A 类），第三作者

[C-4] A Review of Automated Diagnosis of COVID-19 Based on Scanning Images[C]. In the proceedings of The 6th International Conference on Robotics and Artificial Intelligence, ICRAI. 2020. （EI 检索），第一作者

[C-5] Significant Wave Height Prediction based on Wavelet Graph Neural Network[C]. The 4th International Conference on Big Data and Artificial Intelligence, BDAI. 2021. （EI 检索），第一作者，最佳报告奖（Best Presentation）

在投论文

[J-1] Deep Learning based Single Sample Per Person Face Recognition: A Riview[J]. Artificial Intelligence Review, AIRE.（SCI 二区，河海大学 A 类，IF=8.139），通讯作者，一审修改中

[J-2] M^2SGAN: Learning Music-driven Conducting Motion Generation from the Self-supervision of Music Motion Synchronization[J]. IEEE Transactions on Multimedia, TMM. （SCI 一区，CCF-B 类，河海大学 A 类，多媒体领域顶级期刊，IF=6.513），通讯作者，对应本毕业论文第三章

[J-3] A Review of Driver Fatigue Detection and Its Advances on the Use of RGB-D Camera and Deep Learning[J]. IEEE Transactions on Intelligent Transportation Systems, TITS.（SCI 一区，CCF-B 类，河海大学 A 类，IF=6.492），第二作者，导师一作

MEP-3M: A Large-scale Multi-modal E-Commerce Products Dataset

Delong Chen[1], **Fan Liu**[1*], **Xiaoyu Du**[2], **Ruizhuo Gao**[1] and **Feng Xu**[1]
[1]College of Computer and Information, Hohai University, China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
fanliu@hhu.edu.cn

Abstract

The product categories are vital for the e-commerce platforms due to the core applications on automatic product category assignment, personalized product recommendations, etc. Two key aspects of product classification are multi-modal information and fine-grained understanding. However, recent datasets could hardly support both sides. To address this issue, in this paper, we construct a large-scale **M**ulti-modal **E**-commerce **P**roducts classification dataset MEP-3M, which consists of over 3 million products and 599 fine-grained product categories. Each product is represented with an image-text pair and annotated with hierarchical labels. To our best knowledge, MEP-3M is the first e-commerce products dataset paying attention to the multi-modal and fine-grained aspects concurrently, and its scale achieves the largest in existing E-commerce datasets. We also present the performances of the several methods on this dataset as the baselines, where the best accuracy achieves 90.70%. This dataset is now available at https://github.com/ChenDelong1999/MEP-3M.

1 Introduction

The recent rise of deep learning can be traced back to the creation of ImageNet dataset [Deng *et al.*, 2009] and the revival of deep Convolutional Neural Network (CNN) [Krizhevsky *et al.*, 2012; Li *et al.*, 2021]. Since then, the combination of increasingly complex neural network architectures and increasingly large datasets fundamentally revolutionized the fields of Computer Vision (CV) and Natural Language Processing (NLP). In recent years, the research communities are gradually moving from these single-modal tasks to multi-modal tasks. Large-scale multi-modal datasets, especially vision-language datasets (e.g. Flickr30K [Young *et al.*, 2014], Multi30K [Elliott *et al.*, 2016], MS-COCO [Antol *et al.*, 2015], SBU Captions [Ordonez *et al.*, 2011], WIT [Srinivasan *et al.*, 2021]), have been constructed. These datasets enable us to develop multi-modal models, which learn to utilize the complementary information across different modalities and bring the opportunity to combine the advancements across different fields to further improve the model performance.

Figure 1: The comparison between our presented dataset and existing public e-commerce product dataset.

Recently, another hot topic in the deep learning field is fine-grained recognition, which aims to discover the subtle differences between different sub-categories, such as birds [Horn *et al.*, 2015], dogs [Sun *et al.*, 2018], cars [Yang *et al.*, 2015], and castles [Anderson *et al.*, 2021]. A lot of fine-grained datasets are created to promote the development of this domain, such as iNaturalist [Horn *et al.*, 2018], Products 10k [Bai *et al.*, 2020], and iMaterialist Fashion [Guo *et al.*, 2019]. Impressively, many e-commerce-related datasets emergence. A possible reason is the construction of this type of dataset can rely on the pre-defined hierarchical categorization information (e.g., Stock Keeping Unit, SKU).

However, recent e-commerce datasets only focus on one aspect from multi-modal or fine-grained without integrating them together. In this paper, we construct a large **M**ulti-modal **E**-commerce **P**roducts classification dataset named MEP-3M, which provides multi-modal and fine-grained data. It is collected from several Chinese large E-commerce platforms and consists of over 3 million image-text pairs of products and 599 classes. As demonstrated in Fig. 1, MEP-3M consists of the largest number of products, even compared with the single-modal E-commerce product datasets. Its scale is far better than the existing multi-modal dataset. The key characteristics of MEP-3M are summarized as follows:

*Contact Author

VIRTUALCONDUCTOR: MUSIC-DRIVEN CONDUCTING VIDEO GENERATION SYSTEM

Delong Chen, Fan Liu, Zewen Li, Feng Xu*

College of Computer and Information, Hohai University, China
fanliu@hhu.edu.cn

ABSTRACT

In this demo, we present the *VirtualConductor*, a system that can generate conducting video from a given piece of music and a single user's image. First, a large-scale conductor motion dataset is collected and constructed. Then, we propose an Audio Motion Correspondence Network (AMCNet) and adversarial-perceptual learning to learn the cross-modal relationship and generate diverse, plausible, music-synchronized motion. Finally, we combine 3D animation rendering and a pose transfer model to synthesize conducting video from a single given user's image. Therefore, any user can become a virtual conductor through the *VirtualConductor* system.

Index Terms— Adversarial learning, orchestral conductor, audio motion correspondence

1. INTRODUCTION

In recent years, deep learning has shown its advantages in learning discriminative feature representations [1] and learning high-quality generation [2] from massive data. As a notable research line in this field, learning the cross-modal mapping from sound to human motion has drawn a lot of attention. Various types of applications, including speech gesture generation and musical gesture generation (dancing and instrument playing), have been developed in recent years. But researchers pay little attention to the motion generation of an orchestral conductor. Moreover, there is not a large-scale conductor motion dataset currently available. Therefore, we build a system to make the first attempt towards music-driven conductor motion generation and realize a virtual conductor.

To build a large-scale conductor motion dataset, we first collect concert performance video recordings, then extract conductor motion by pose estimation [3]. Meanwhile, different types of audio features, including MFCC, spectral centroid, spectral bandwidth, onset envelope, estimated tempo, and predominant local pulse, are extracted. Finally, the constructed dataset consists of conductor motion data and aligned music features in a total of 40 hours.

However, modeling conductor motion still has several challenges. First, the conductor motion is highly complicated because it conveys various types of information, including tempo, strength, and emotion. Meanwhile, the generated motion should be closely synchronized with music. Moreover, because of different conducting styles, mapping music to conductor motion is a one-to-many task, which is difficult to learn by standard mean squared error (MSE) regression. In this demo, based on the constructed dataset, we propose the *VirtualConductor* system to tackle the above difficulties. We use a combination of MSE loss, pose perceptual loss, and adversarial loss to train the motion generator. In this way, the generated motion can be simultaneously diverse, plausible, and synchronized to music. Finally, by combining 3D animation rendering and pose transfer [4] module, the system can generate conducting video from given music and a single user's image. In the following sections, we will introduce our system in detail.

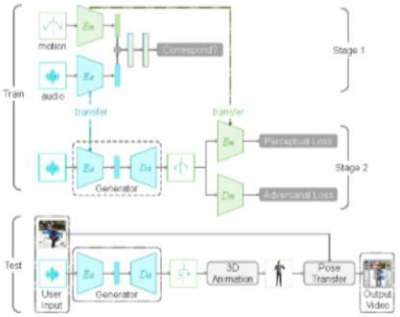


Fig. 1. The pipeline of presented demo *VirtualConductor*.

2. SYSTEM DESIGN AND IMPLEMENTATION

2.1. Audio Motion Correspondence Learning

We first design an AMCNet to learn the correspondence between audio and motion. As shown in Fig.1, the AMCNet consists of a music encoder E_a, a motion encoder E_m, and fuse layers. The features extracted by two encoders are concatenated and passed to fuse layers. The AMCNet output a

978-1-6654-4989-2/21/$31.00 ©2021 IEEE

WEAKLY CORRELATED ADVERSARIAL LEARNING FOR COGNITIVE DIAGNOSIS SYSTEM

Zhibin Chen, Fan Liu, Delong Chen, Jinyu He, Xiaohan Yan*

College of Computer and Information, Hohai University, China
fanliu@hhu.edu.cn

ABSTRACT

In traditional cognitive diagnosis models, the representations of students and questions tend to have a high correlation. It results in biases and poor performance in real-world applications. In order to weaken such correlation, we propose a **W**eakly **C**orrelated **A**dversarial **L**earning (WCAL) method. Based on WCAL, we design a cognitive system for both student knowledge state evaluation and exam results prediction which can help teachers select exams suitable for students. The experimental results show the proposed method can effectively model students' knowledge state and help teachers improve the teaching effect.

Index Terms— Adversarial learning, cognitive diagnosis, exam results prediction, knowledge state

1. INTRODUCTION

In real-world teaching scenarios, it is essential to let students participate in other schools' exams. But students' knowledge states in different schools are different. An exam suitable for one school's students may not be appropriate for other's. The selection of exam papers often relies on the teacher's subjective judgment, which lacks objectivity. In addition, the teacher can not estimate the difficulty of exam paper accurately. Recently, with the rise of artificial intelligence [1, 2], cognitive diagnosis has attracted more and more attention since it can be applied to predict student response. Base on the predicted results, teachers can select exams suitable for students. Therefore, it can help teachers improve the teaching effect.

However, in a cognitive diagnosis model, the representation of students and exams usually affect each other. As a result, the student latent attributes predicted by cognitive diagnosis will contain the characteristic of the question and vice versa. For example, a good student response may not be attributed to the easy exam but the student's ability. To solve this problem, in this demo, we propose **W**eakly **C**orrelated **A**dversarial **L**earning (WCAL) for cognitive diagnosis. As shown in Fig 1, by taking advantage of the adversarial learning scheme , WCAL can weaken the correlation between student and question representation and obtain precise student's

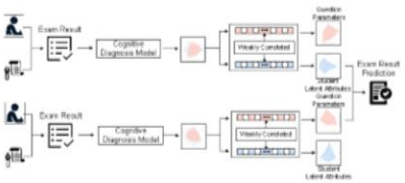


Fig. 1. The structure of the proposed WCAL model.

latent attributes and question parameters. The obtained representations enable the teacher to evaluate the student knowledge state and exam difficulty more precisely. Moreover, the system can also predict student response, i.e., predict the exam score given certain student and exam. Based on this, our proposed demo system can provide valuable information for planning exams and teaching strategies.

2. SYSTEM DESIGN

2.1. Cognitive Diagnosis

Suppose the student set is $V = \{v_1, v_2, ..., v_V\}$, the question set is $U = \{u_0, u_2, ..., u_U\}$, and the student response is Y, where $y_{i,j}$ is the result of student i answers question j. Then, as shown in Eq. 1, a cognitive diagnosis model can output a possibility $\hat{y}_{i,j}$ predicting whether the student v_i can correctly answer the question u_j,

$$\hat{y}_{i,j} = P(Y_{i,j} = 1|\theta_i, \xi_j) = f(\theta_{q_i}, \xi_{s_j}) \quad (1)$$

where θ_i represents the latent attributes of student i, ξ_j represents the parameters of question j.The cognitive diagnosis model are trained by following binary cross entropy loss function:

$$L_R = \sum_{i,j}^{U,V} y_{i,j} log\hat{y}_{i,j} + \sum_{i,j}^{U,V} (1 - y_{i,j}) log(1 - \hat{y}_{i,j}) \quad (2)$$

978-1-6654-4989-2/21/$31.00 ©2021 IEEE

A Review of Automated Diagnosis of COVID-19 Based on Scanning Images

Delong Chen
Key Laboratory of Ministry of Education for Coastal Disaster and Protection, Hohai University; College of Computer and Information, Hohai University, China
chendelong@hhu.edu.cn

Shunhui Ji*
College of Computer and Information, Hohai University
shunhuiji@hhu.edu.cn

Fan Liu
Key Laboratory of Ministry of Education for Coastal Disaster and Protection, Hohai University; College of Computer and Information, Hohai University, China
fanliu@hhu.edu.cn

Zewen Li
College of Computer and Information, Hohai University
servon@hhu.edu.cn

Xinyu Zhou
China Pharmaceutical University, Nanjing, China
magicme314@foxmail.com

ABSTRACT

The pandemic of COVID-19 has caused millions of infections, which has led to a great loss all over the world, socially and economically. Due to the false-negative rate and the time-consuming of the conventional Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests, diagnosing based on X-ray images and Computed Tomography (CT) images has been widely adopted. Therefore, researchers of the computer vision area have developed many automatic diagnosing models based on machine learning or deep learning to assist the radiologists and improve the diagnosing accuracy. In this paper, we present a review of these recently emerging automatic diagnosing models. 70 models proposed from February 14, 2020, to July 21, 2020, are involved. We analyzed the models from the perspective of preprocessing, feature extraction, classification, and evaluation. Based on the limitation of existing models, we pointed out that domain adaption in transfer learning and interpretability promotion would be the possible future directions.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**;

KEYWORDS

Deep learning, Machine learning, Biomedical Image Analysis, COVID-19

ACM Reference Format:
Delong Chen, Shunhui Ji, Fan Liu, Zewen Li, and Xinyu Zhou. 2020. A Review of Automated Diagnosis of COVID-19 Based on Scanning Images.

*Corresponding author.

ICRAI 2020, November 20–22, 2020, Singapore, Singapore

ACM ISBN 978-1-4503-8859-7/20/11…$15.00
https://doi.org/10.1145/3449301.3449778

In *2020 6th International Conference on Robotics and Artificial Intelligence (ICRAI 2020), November 20–22, 2020, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3449301.3449778

1 INTRODUCTION

It has been seven months since the first case of COVID-19 was confirmed. In the battle between human and the novel coronavirus, early diagnosing and early quarantine is of vital importance. However, testing based on Reverse Transcription Polymerase Chain Reaction (RT-PCR) is time-consuming and may cause certain false-negative reports. To solve this problem, diagnosing based on scanning images (CT or X-ray) has been proved to be practical and effective. In the virus-stricken area, radiologists have a heavy burden on analyzing scanning images. As shown in Figure 1, researchers therefore have started to pay more and more attention to the development of COVID-19 diagnosing models for reducing the diagnosing time and improve the accuracy of radiologists.

Due to the rapid development in this area, there have already been 9 reviews [79–87] existing on this topic, but they have various shortcomings. To our best knowledge, there are at least 70 deep learning based and machine learning based models that have been proposed, and many of them have not been covered by any of the existing surveys. Most reviews only covered about 10 different diagnosing models. Moreover, these reviews lack proper organization, comparison of performance and in-depth analysis of shortcomings of diagnosing models.

Therefore, in this paper we define a universal pipeline for diagnosing models, for both machine learning based models and deep learning based models. Then we organized the paper according to different stages of the model. The contributions of this paper are as follows:

- We systematically reviewed and analyzed 70 COVID-19 diagnosing models from the perspective of preprocessing, feature extraction, classification, and evaluation. These models are proposed from February 14 to July 21, 2020.
- Based on the discussion of the existing models' limitation, we pointed out that domain adaption in transfer learning and

2021 IEEE 4th International Conference on Big Data and Artificial Intelligence

Significant Wave Height Prediction based on Wavelet Graph Neural Network

Delong Chen [1], Tan Liu[1], Zheqi Zhang [2,*], Xiaomin Lu [1], Zewen Li [1]
[1]College of Computer and Information Hohai University, Nanjing, China
[2]Mihaylo College of Business and Economics, California State University, Fullerton, USA
email: tedzhang1988@gmail.com

2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI) | 978-1-6654-1270-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/BDAI52447.2021.9515293

***Abstract*—Computational intelligence-based ocean characteristics forecasting applications, such as Significant Wave Height (SWH) prediction, are crucial for avoiding social and economic loss in coastal cities. Compared to the traditional empirical-based or numerical-based forecasting models, "soft computing" approaches, including machine learning and deep learning models, have shown numerous success in recent years. In this paper, we focus on enabling the deep learning model to learn both short-term and long-term spatial-temporal dependencies for SWH prediction. A Wavelet Graph Neural Network (WGNN) approach is proposed to integrate the advantages of wavelet transform and graph neural network. Several parallel graph neural networks are separately trained on wavelet decomposed data, and the reconstruction of each model's prediction forms the final SWH prediction. Experimental results show that the proposed WGNN approach outperforms other models, including the numerical models, the machine learning models, and several deep learning models.**

***Index Terms*—Significant wave height, Wavelet transform, Deep learning**

I. Introduction

Ocean waves with high Significant Wave Height (SWH, or H_s) can overturn ships and destroy ocean or coastal engineering. It threatens human life, crop production, and the survival of aquaculture products. Therefore, the accurate prediction of SWH is vital since it can help reduce social and commercial losses. Moreover, SWH prediction can also bring several benefits. For example, optimizing ship routes according to the SWH prediction can avoid rough sea areas, thereby reducing the sailing time and fuel expenses. Furthermore, SWH prediction can provide valuable information for planning military and amphibious operations.

Due to its importance and valuable applications, SWH prediction approaches have been continuously developed for decades. The empirical-based and numerical-based SWH prediction approaches in the early years have high interpretability but low accuracy and limited generalization ability. As the rise of computational intelligence, machine learning-based SWH prediction models, such as the Support Vector Machine (SVM) and the Artificial Neural Network (ANN), have shown their advantages. Especially in recent years, deep learning-based models, which hold strong feature extraction ability, have also been applied to SWH prediction successfully [1], [2].

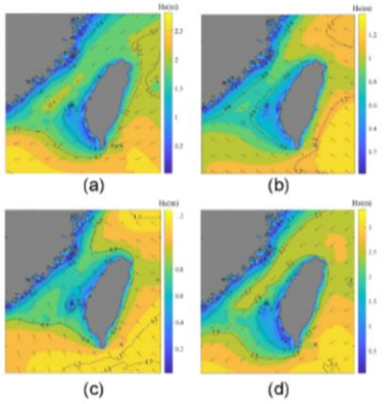


Fig. 1. The seasonal variation of SWH caused by the monsoon. (a)-(d) respectively show the average SWH value and wave direction of Q1-Q4. This figure clarified the motivation and necessity of wavelet decomposition for SWH prediction.

However, by reviewing the existing approaches, we find that there are the following two challenges that still remain and need to be solved for SWH prediction: **1)** effectively capture the relationships between different types of inputs and learn its complicated non-linear mapping and temporal dependencies with the SWH data, and **2)** distinguish occasional extreme sea conditions and seasonal SWH variation and learn both short-term and long-term SWH patterns (see Fig.1 for an example of long-term SWH variation).

In this paper, the above issues are addressed by the proposed **W**avelet **G**raph **N**eural **N**etwork (WGNN). The inputs and the target outputs are decomposed by the Debauches (Db)-type mother wavelet-based wavelet transform. For the derived components, several Graph Neural Networks (GNN) are separately

This work was partially funded by Natural Science Foundation of Jiangsu Province under Grant No. BK20191298, Fundamental Research Funds for the Central Universities under Grant No. B200202175.

978-1-6654-1270-4/21/$31.00 ©2021 IEEE

Artificial Intelligence Review manuscript No.
(will be inserted by the editor)

Deep Learning Based Single Sample Face Recognition: A Survey

Delong Chen · Fan Liu · Zewen Li

Received: date / Accepted: date

Abstract Face recognition has long been an active research area in the field of pattern recognition, particularly since the rise of deep learning in recent years. In some practical situations, however, each identity has only a single sample available for training. Face recognition under this situation is referred to as single sample face recognition and poses significant challenges to the effective training of deep models. In an attempt to unleash the full potential of deep learning as well as improving the model recognition performance in single sample situations, many deep learning based single sample face recognition methods have been proposed. While several comprehensive surveys have been conducted on traditional single sample face recognition approaches, emerging deep learning based methods are rarely involved in such reviews. Accordingly, we focus on these deep methods in this paper, classifying them into virtual sample methods and generic learning methods. In the former category, virtual images or virtual features are generated to benefit the training of the deep model. In the latter, additional multi-sample generic sets are used. There are three types of generic learning methods: combining traditional methods and deep features, improving the loss function, and improving network structure, all of which are covered in our analysis. Moreover, we review some datasets that have been commonly used for evaluating single sample face recognition models. We go on to compare the results of different types of models. We additionally discuss problems with existing SSPP face recognition methods, including identity information retention in virtual sample methods and domain adaption in generic learning methods. Furthermore, we regard the semantic gap as an important issue that needs to be considered in single sample face recognition.

Keywords Face recognition · Deep learning · Single Sample Per Person (SSPP) problem

D. Chen, F. Liu, Z. Li
College of Computer and Information, Hohai University Nanjing, China
E-mail: fanliu@hhu.edu.cn

Date:	19 Aug 2021
To:	"Fan Liu" fanliu@hhu.edu.cn
cc:	"Delong Chen" chendelong@hhu.edu.cn, "Zewen Li" servon@hhu.edu.cn
From:	"Artificial Intelligence Review (AIRE)" uthara.udayan.1@springernature.com
Subject:	Decision on your manuscript #AIRE-D-21-00005

Dear Dr. Liu:

We have received the reports from our advisors on your manuscript, "Deep Learning based Single Sample Face Recognition: A Survey", which you submitted to Artificial Intelligence Review.

Based on the advice received, the Editor has decided that your manuscript could be reconsidered for publication should you be prepared to incorporate major revisions.
When preparing your revised manuscript, you are asked to carefully consider the reviewer comments which are attached, and submit a list of responses to the comments.
Your list of responses should be uploaded as a file in addition to your revised manuscript.

M^{2}S-GAN: Learning Music-driven Conducting Motion Generation from the Self-Supervision of Music Motion Synchronization

Fan Liu, *Member, IEEE,* Delong Chen, Xiaoyu Du, and Feng Xu

***Abstract*—Music-motion correlation attracts much attention. Many recent works focus on the motion generations for dancers and musicians, but few works for the orchestral conductors. In this paper, we concentrate on music-driven conducting motion generation, which aims to generate the conducting motions according to a piece of music. We identify that two key problems in this task: 1) learning semantic music features and 2) learning music-related motion features, can be solved together by multi-modal self-supervised contrastive learning. Therefore, we put a contrastive learning stage ahead of the generative learning stage, resulting in a two-stage learning framework. In the first stage, we propose a Music Motion Synchronization (M^{2}S) learning task to train a two-branched Music Motion Synchronization Network (M^{2}S-Net) and obtain rich music and motion representations. In the second stage, the two branches of the M^{2}S-Net are integrated into a GAN-based framework and respectively provide semantic music features and calculate a proposed sync loss. These two branches, a generator and a discriminator form the Music Motion Synchronized Generative Adversarial Network (M^{2}S-GAN). The generator is trained by a joint loss composed of the sync loss and adversarial loss, which respectively poses constraints on motion consistency and realism. To train our model, we build a dataset *ConductorMotion100* that reaches an unprecedented 100 hours long. The extensive experiments demonstrate that our proposed approach achieves an impressive performance in generating the conducting motions. The dataset and code will be made public.**

***Index Terms*—Self-supervised learning, perceptual loss, orchestral conductor**

I. Introduction

MUSIC and human motion are closely related. When singing, playing musical instruments, or dancing with music, people's motion naturally follows the music's rhythm dynamics and emotion. While the music itself has been investigated for decades, the relationship between music and motion is a relatively emerging interdisciplinary research area [1]. With the development of generative techniques, the methods that can automatically generate musical motion from music are widely explored. Recently, researchers have successfully generated the dance motions [2], [3], instrument playing motions [4], [5], and singing motions [6] from music.

Conductors, the soul of an orchestra, always perform elegantly and charmingly in a concert. The conducting motions are well-designed before each performance, making it a great learning material. But in contrast to the generations of dancing, instrument playing, and singing motions, music-driven conducting motion generation received far less attention. In 2003, Wang *et al.* [7] design a Kernel-based Hidden Markov Model (KHMM) to predict conducting motion, but its rhythm adaptability and computational efficiency are poor. Several rule-based approaches were proposed in [8]–[14], but their diversity and realism are greatly limited. To our best knowledge, except from [7], these are no other learning-based conducting motion generation models so far.

The scarcity of learning-based conducting motion generation researches can be attributed to the great challenge of this task. The conducting motion not only conveys basic beat information, but also contains articulatory information (legato, staccato, etc.), hints towards different parts of the orchestra (string, winds, etc.), and music emotions. This task has the difficulty of generating instrument playing and dance motion at the same time because it has both low-level music texture dependencies and high-level music structure dependencies. Despite the complexity, the generation is also inherently ill-posed. Different conductors conduct with distinctive styles. For the same piece of music, the motions from different conductors may differentiate a lot. Using a standard L1 or L2 regression loss will fail to learn the one-to-many mapping and lead to over-smoothed results.

To address these challenges, we bring recent advances of multi-modal self-supervised learning [15], [16] into this task. As shown in Fig. 1, we integrate two types of self-supervised learning [17]: contrastive learning and generative learning, into a unified two-stage framework. In the contrastive stage, we propose a Music Motion Synchronization (M^{2}S) learning task, where a two-branched network, namely Music Motion Synchronization Network (M^{2}S-Net), learns rich music and motion features representation from the contrastive correlations between music and motions. In the subsequent generative stage, the music and motion features are respectively used to provide semantic control signal and calculate perceptual training metric. A Music Motion Synchronization-based Generative Adversarial Network (M^{2}S-GAN) is trained with a proposed sync loss and an Wasserstein distance-based adversarial loss [18], [19].

The key motivation here is to take advantage of the joint feature space constructed in the first contrastive stage. In this space, synchronized music and motion sequences are embedded into near points, while out-of-sync pairs are mapped

This work was partially funded by Natural Science Foundation of Jiangsu Province under Grant No. BK20191298, Fundamental Research Funds for the Central Universities under Grant No. B200202175. (Corresponding author: Delong Chen.)

Fan Liu, Delong Chen, and Feng Xu are with the College of Computer and Information, Hohai University, Nanjing, 210098, China (e-mail: fanliu@hhu.edu.cn, chendelong@hhu.edu.cn, xufeng@hhu.edu.cn).

A Review of Driver Fatigue Detection with Emphasis on the Use of RGB-D Camera and Deep Learning

Fan Liu, *Member, IEEE*, Delong Chen, Jun Zhou, Qiaolin Ye, Feng Xu

***Abstract*—Driver fatigue is an essential reason for traffic accidents, which poses a severe threat to people's lives and property. In this review, we summarize the latest research findings and analyze the developmental trends of driver fatigue detection technologies. Firstly, we analyze and discuss the four types of different fatigue detection technologies that are respectively based on driver physiological signals, behavior features, vehicle running features, and information fusion. Then, we study two state-of-the-art solutions in this field: RGB-D camera and deep learning. Finally, we present the idea of using RGB-D camera and deep learning technology simultaneously, where Generative Adversarial Networks (GAN) and multi-channel schemes have also been utilized to enhance the performance. The experimental results show that the fatigue features extracted by Convolutional Neural Networks (CNN) are superior to traditional handcrafted ones and single features cannot guarantee robustness. Moreover, the latent fatigue features extracted by deep learning technologies have been demonstrated to be effective for fatigue detection.**

***Index Terms*—Driver fatigue detection, review, information fusion, RGB-D, deep learning**

I. Introduction

IN today's society, due to the fast increase of vehicle and hectic lifestyle especially in proliferating economies, people are more and more sleep-deprived which can lead to driver fatigue. According to statistics, driver fatigue has been one of the major threats to life safety and the economy. The American Automobile Association also noted that 21% of fatal crashes result from fatigue driving. Therefore, driver fatigue detection has a vital practical significance for preventing traffic accidents.

In the last several decades, numerous fatigue detection methods and technologies have been developed. There are also some survey papers [1-16] reviewing and analyzing various fatigue detection methods from different perspectives. For example, Craig et al. [1] have presented the understanding of driver fatigue from the respect of psychology and found those elements that may influence driver fatigue. Sanjaya et al. [2] have reviewed and analyzed the measurements of physiological fatigue signals. Wang et al. [3] have mainly surveyed those fatigue detection approaches based on driver's behavior or performance. In [4] and [5], the above three kinds of fatigue detection methods have all been introduced and discussed. The literature [6] reviewed and discussed the sensors used by different measures for fatigue detection. Moreover, literatures [7-11] respectively listed representative systems, devices, tools, applications, and problems in driver fatigue detection. There were also some works specially designed for those professional drivers such as truck, taxi, and racing drivers, which have been introduced in [12-14]. In [14], some commercial devices were specially reviewed and evaluated to meet the requirements of the mining industry. Recently, [15] and [16] elaborated the advantages and disadvantages of those latest works from different aspects such as features, classifiers, accuracy, system parameters, and environment.

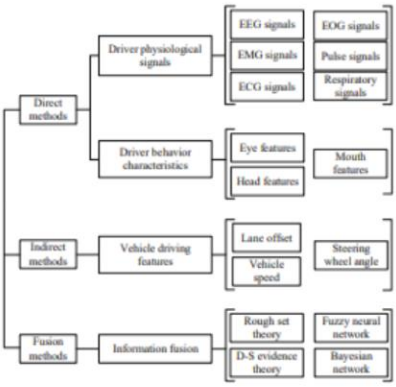


Fig. 1. The different categories of fatigue detection methods.

In recent years, fatigue detection has ushered in a new development period due to the continuous development and widespread use of RGB-D camera and deep learning technologies. However, all the above-mentioned survey papers have not reviewed or discussed the role of the two solutions in fatigue detection. Therefore, this review not only analyzes the previous driver fatigue detection methods in detail, but also

This work was partially funded by Natural Science Foundation of Jiangsu Province under Grant No. BK20191298, Fundamental Research Funds for the Central Universities under Grant No. B200202175. (Corresponding author: Fan Liu.)

Fan Liu, Delong Chen and Feng Xu are with the College of Computer and Information, Hohai University, Nanjing, 210098, China (e-mail: fanliu@hhu.edu.cn, chendelong@hhu.edu.cn, xufeng@hhu.edu.cn).

Qiaolin Ye is with the College of Information Science and Technology, Nanjing Forestry University, Nanjing, 210037, China (e-mail: yqlcom@njfu.edu.cn).

C：本科期间参加的科研项目

1. 2019 年河海大学大学生创新创业训练计划项目《基于跨年龄人脸识别的失踪人口匹配系统》（国家级），第一负责人
2. 国家自然科学基金，面上项目，61871444，鲁棒判别的多视角自适应子空间学习及其在异质图片识别上的应用研究
3. 江苏省自然科学基金，面上项目，BK20191298，基于语义的单样本人脸识别关键技术研究

河海大学大学生创新创业训练计划项目

结 题 证 书

项目名称：基于跨年龄人脸识别的失踪人口匹配系统

项目编号：201910294049

项目负责人：陈德龙　　指导教师：刘凡

项目组成员：徐胜捷、王文钦、何秋实、王越

项目等级：国家级　　验收结果：合格

河海大学大学生创新创业训练计划领导小组

（河海大学教务处代章）

2020年7月15日

D：本科期间申请的专利

[P-1] 基于动态频域分解的音乐驱动的指挥动作生成方法[P]. 中国发明专利. CN202111090067.5，2021-09-17.（受理），第二作者，导师一作

[P-2]基于自监督跨模态感知损失的乐队指挥动作生成方法[P]. 中国发明专利. CN202111090024.7，2021-09-17.（受理），第二作者，导师一作

[P-3]一种基于动机提取模型与卷积神经网络的自动作曲方法[P]. 中国发明专利. CN201910259941.X，2019-05-01.（受理），第一作者

[P-4] 一种基于机器学习的小微企业画像构建系统[P]. 中国发明专利. CN202010912809.7，2020-09-01.（受理），第一作者

[P-5] 基于 NNDT-Clus 算法的小微企业智能聚类分析系统. (小微企业智能聚类分析系统 V1.0) [Z]. 软件著作权. 2020SR1098487，2020-09-05.（授权），第一作者

[P-6] 智能布匹瑕疵识别系统. (智能布匹瑕疵识别 V1.0) [Z]. 软件著作权. 2020SR1094158，2020-09-14.（授权），第一作者


国家知识产权局

210008

2021 年 09 月 17 日

专利申请受理通知书


国家知识产权局

210008

2021 年 09 月 17 日

专利申请受理通知书

国家知识产权局

211100

江苏省南京市江宁区龙眠大道639号中国药科大学江宁校区
周欣宇(15956229989)

发文日：

2020年09月04日

申请号或专利号：202010912809.7　　发文序号：2020090400388490

专利申请受理通知书

根据专利法第28条及其实施细则第38条、第39条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日、申请人和发明创造名称通知如下：

申请号：202010912809.7
申请日：2020年09月03日
申请人：陈德龙,周欣宇
发明创造名称：一种基于机器学习的小微企业画像构建系统

经核实，国家知识产权局确认收到文件如下：
说明书 每份页数:3页 文件份数:1份
发明专利请求书 每份页数:4页 文件份数:1份
说明书附图 每份页数:1页 文件份数:1份
权利要求书 每份页数:1页 文件份数:1份 权利要求项数：2项
说明书摘要 每份页数:1页 文件份数:1份

提示：
1. 申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。
2. 申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。
3. 国家知识产权局收到向外国申请专利保密审查请求书后，依据专利法实施细则第9条予以审查。

审 查 员：自动受理　　审查部门：专利局初审及流程管理部

200101　纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路6号 国家知识产权局受理处收
2018.11　电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

国家知识产权局

210000

发文日：

2019年04月02日

申请号或专利号：201910259941.X　　发文序号：2019040201018060

专利申请受理通知书

中华人民共和国国家版权局
计算机软件著作权登记证书

证书号：软著登字第5977183号

软件名称：基于 NNDT-Clus算法的小微企业智能聚类分析系统
[简称：小微企业智能聚类分析系统]
V1.0

著作权人：陈德龙;徐胜捷;占鑫鹏;朱俯卿;陈海燕

开发完成日期：2020年06月01日
首次发表日期：未发表
权利取得方式：原始取得
权利范围：全部权利
登记号：2020SR1098487

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

2020年09月15日

中华人民共和国国家版权局
计算机软件著作权登记证书

证书号：软著登字第5972854号

软件名称：智能布匹瑕疵识别系统
[简称：智能布匹瑕疵识别]
V1.0

著作权人：陈德龙;徐胜捷;金刘宇

开发完成日期：2019年07月01日
首次发表日期：未发表
权利取得方式：原始取得
权利范围：全部权利
登记号：2020SR1094158

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

2020年09月14日

E：本科期间所获奖项

1. Best Dataset Paper Award at Long-Tailed Distribution Learning Workshop, IJCAI 2021.
2. Best Demo Award at IEEE International Conference on Multimedia and Expo (ICME) 2021.
3. Best Presentation Winner at 2021 4th International Conference on Big Data and Artificial Intelligence.
4. 第八届“中国软件杯”大学生软件设计大赛华东分赛区决赛三等奖，第一负责人（工业和信息化部、教育部、江苏省人民政府）
5. 2019 年 MCM/ICM 数学建模竞赛三等奖，第二负责人（美国数学及其应用联合会）
6. 河海大学计算机与信息学院 2017 年新生杯辩论赛“最佳辩手”称号（共青团河海大学计算机与信息学院委员会、计算机与信息学院科协）


IEEE International Conference on Multimedia and Expo 2021
July 5-9 2021
Shenzhen, CH || Virtual
ICME 2021 BEST DEMO PAPER AWARD
presented to
VirtualConductor: Music-driven Conducting Video Generation System
Delong Chen, Fan Liu and Feng Xu
for
Outstanding innovative contribution in the area of Multimedia Intelligence
Moncef Gabbouj
Houqiang Li
Guo-Jun Qi
Yonghong Tian
ICME 2021 General Chairs
Zhenzhen Hu
Zechao Li
ICME 2021 Demo Chairs


AWARD CERTIFICATE
THIS IS PRESENTED TO
Delong Chen
Hohai University, China
Your oral presentation has been selected as the best one of 2021 4th International Conference on Big Data and Artificial Intelligence (BDAI), Ocean University of China, Qingdao, China, July 02-04, 2021.
BDAI
Conference Committee

荣誉证书

同学：陈德龙

您所在的团队在由工业和信息化部、教育部、江苏省人民政府共同主办的第八届"中国软件杯"大学生软件设计大赛华东分赛区决赛中获三等奖。

特发此证，以资鼓励!

"中国软件杯"大学生软件设计大赛组委会

二〇一九年六月

2019

Mathematical Contest In Modeling

Certificate of Achievement

Be It Known That The Team Of

Yuxin Han

Delong Chen

Xu Tang

With Faculty Advisor

Yongzhong Zhu

Of

Hohai University

Was Designated As

Successful Participant

荣誉证书

陈德龙 同学：

在计算机与信息学院2017年"新生杯"科论赛中，表现优异，荣获"最佳辩手"称号

特发此证，以资鼓励！

计算机与信息学院科协

二〇一七年十月二十二日

F：本科期间所获荣誉

荣誉称号：

1. “江苏省优秀共青团员”称号（共青团江苏省委）
2. “2019 江苏省大学生年度人物”提名奖（中共江苏省教育工委、江苏省教育厅）
3. 河海大学“海韵风华大学生年度人物”称号（中共河海大学委员会）
4. 河海大学“海韵风华百佳学生”称号（中共河海大学委员会）
5. 河海大学 2021 届本科“优秀毕业生”荣誉称号（河海大学）
6. 河海大学大学生艺术团 2017 年度“优秀团员”荣誉称号（共青团河海大学委员会）
7. 2018-2019 学年河海大学“优秀学生干部”荣誉称号（河海大学）
8. 2019-2020 学年河海大学“优秀学生干部”荣誉称号（河海大学）
9. 2019 年河海大学大学生志愿者暑期文化“三下乡”社会实践活动“先进个人”（中共河海大学委员会）
10. 2019 年河海大学大学生志愿者暑期文化“三下乡”社会实践活动“优秀团队”（中共河海大学委员会）
11. 2018 年大学生暑期社会实践活动“优秀团队”荣誉称号（中共河海大学委员会）

奖学金：

1. 河海大学“校长奖学金”（河海大学）
2. 2017-2018 学年艺术体育优秀奖学金（河海大学）
3. 2018-2019 学年艺术体育优秀奖学金（河海大学）
4. 2018-2019 学年社会工作优秀奖学金（河海大学）
5. 2019-2020 学年学业进步奖学金（河海大学）
6. 2019-2020 学年艺术体育优秀奖学金（河海大学）
7. 2019-2020 学年科技创新奖学金（河海大学）
8. 2019-2020 学年社会工作优秀奖学金（河海大学）

授予 陈德龙 同志：

“江苏省优秀共青团员”称号。

共青团江苏省委

二〇二〇年五月

荣誉证书

河海大学 陈德龙：

荣获“2019江苏省大学生年度人物”提名奖。

特发此证，以资鼓励。

中共江苏省委教育工委

江苏省教育厅

2020年11月


河海大学
HOHAI UNIVERSITY

荣誉证书

陈德龙 同学：

在2020年“海韵风华十杰百佳”优秀学生评选中，荣获“海韵风华大学生年度人物”称号。

特发此证，以资鼓励。

中共河海大学委员会
2020年6月


河海大学
HOHAI UNIVERSITY

荣誉证书

陈德龙 同学：

荣获河海大学2021届本科“优秀毕业生”荣誉称号。

特发此证，以资鼓励。

河海大学
2021年6月


荣誉证书
Certificate Of Honors

陈德龙：

荣获2019年河海大学大学生志愿者暑期文化“三下乡”社会实践活动

先进个人

特发此证，以资鼓励。

中共河海大学委员会
2019年10月25日


荣誉证书
Certificate Of Honors

陈德龙 同学：

荣获河海大学大学生艺术团2017年度“优秀团员”荣誉称号。

特发此证，以资鼓励。

共青团河海大学委员会
2018年3月


荣誉证书

陈德龙 同学：

荣获2019-2020学年河海大学“优秀学生干部”荣誉称号，特发此证，以资鼓励。

河海大学
二〇二〇年十一月


荣誉证书

陈德龙 同学：

荣获2018-2019学年河海大学“优秀学生干部”荣誉称号，特发此证，以资鼓励。

河海大学
二〇一九年十一月


获 奖 证 书

陈德龙 同学：

荣获河海大学 2019-2020 学年艺术体育优秀奖学金，特发此证，以资鼓励。

河海大学

二〇二〇年十月


获 奖 证 书

陈德龙 同学：

荣获河海大学 2019-2020 学年学业进步奖学金，特发此证，以资鼓励。

河海大学

二〇二〇年十月


获 奖 证 书

陈德龙 同学：

荣获河海大学 2018-2019 学年社会工作优秀奖学金，特发此证，以资鼓励。

河海大学

二〇一九年十月


获 奖 证 书

陈德龙 同学：

荣获河海大学 2018-2019 学年艺术体育优秀奖学金，特发此证，以资鼓励。

河海大学

二〇一九年十月


获 奖 证 书

陈德龙 同学：

荣获河海大学 2017-2018 学年艺术体育优秀奖学金，特发此证，以资鼓励。

河海大学

二〇一八年十月


获 奖 证 书

陈德龙 同学：

荣获河海大学 2019-2020 学年社会工作优秀奖学金，特发此证，以资鼓励。

河海大学

二〇二〇年十月


获 奖 证 书

陈德龙 同学：

荣获河海大学 2019-2020 学年科技创新奖学金，特发此证，以资鼓励。

河海大学

二〇二〇年十月


荣誉证书

Certificate Of Honors

《马哲有点甜》陈德龙 初子清 唐 曦：

在河海大学第三十二届校园文化艺术节“汇聚同心 青春能量”创“艺”大赛中荣获二等奖。

特发此证，以资鼓励！

河海大学党委宣传统战部

二〇二〇年十月

G：本科期间社会工作

2020 年 8 月，中华全国学生联合会第二十七次代表大会代表；

2019 年 4 月-2020 年 9 月，河海大学大学生艺术团管弦乐团团长；

2018 年 3 月-2019 年 4 月，河海大学大学生艺术团管弦乐团声部长；

2017 年 9 月-2021 年 6 月，河海大学计算机与信息学院 2017 级计算机科学与技术 1 班文体委员.

2020 年 3 月至今，网络视频平台 bilibili.com（b 站）音乐区 up 主，粉丝量 2.3 万。本文算法生成效果视频发布后获得 35 万次播放，1.1 万人点赞，3000+次转发；

聘 书

Letter Of Appointment

兹聘任陈德龙同学为河海大学大学生艺术团管弦乐团声部长；聘期 2018 年 3 月——2019 年 4 月。

特此聘任。

共青团河海大学委员会

二〇一九年四月

聘 书

兹聘任陈德龙为河海大学信息学部计算机与信息学院2017级计算机科学与技术1班体育委员，聘期：2019年9月—2020年6月。

河海大学信息学部计算机与信息学院

2020年9月

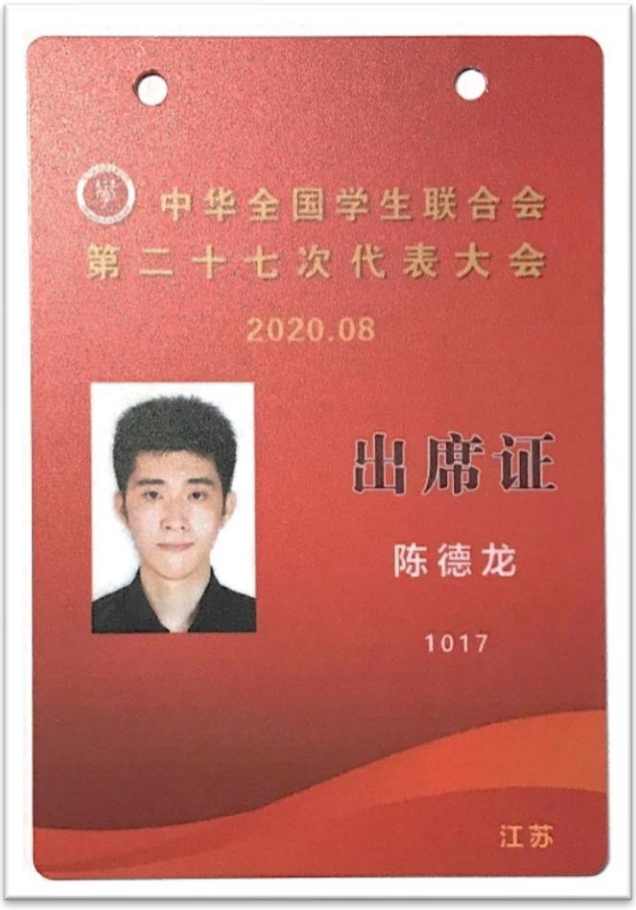
聘 书

兹聘任陈德龙为河海大学计算机与信息学院2017级计算机科学与技术1班体育委员，聘期：2018年9月—2019年6月。

特发此证，以资鼓励。

河海大学计算机与信息学院

二〇一九年六月


中华全国学生联合会
第二十七次代表大会
2020.08
出席证
陈德龙
1017
江苏

证明

兹证明计算机与信息学院计算机科学与技术专业陈德龙同学（学号：1706010120）担任河海大学大学生艺术团管弦乐团团长，任期从2019年4月起，任期一年。

特此证明。

共青团河海大学委员会

2019年9月24日