专题综述

Transformer 的视频背景音乐生成

北京航空航天大学 狄尚哲 姜泽仁 王肇凯 朱乐岩 何泽欣 刘偲

随着新媒体技术与相关产业的发展，人们对短视频的编辑和发布变得越来越便捷。通常，为了让视频更吸引人，视频制作者会为视频搭配背景音乐。然而，对于不具备音乐制作、视频剪辑等技能的人而言，这是一个困难工作。除此之外，这一过程还面临着版权等许多问题。

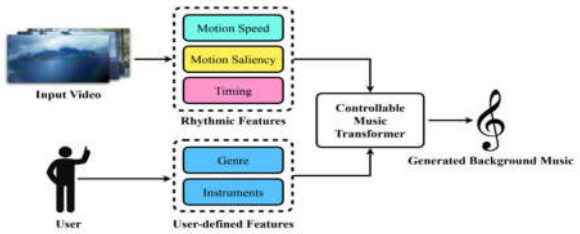

图 1 方法整体示意图

因此根据给定视频生成背景音乐成为一个十分重要的任务。然而，目前尚未有针对这一任务的有效研究。虽然在音乐生成领域已经有了一些研究工作[19][2][7]，但这些工作均未考虑视频信息。为了解决这个问题，我们提出一种新的音乐表示形式，并基于这种表示形式设计了音乐生成模型，实现了视频背景音乐生成的功能。

虽然先前尚未有与一般视频的背景音乐生成相关的工作，但已经有与音乐表示学习、音乐生成以及从静音演奏视频复原音乐的相关研究。大多数音乐生成研究工作是以类似 MIDI 的事件序列[7][12]作为输入。REMI[8]提出了一种表示音乐的结构，这种结构清晰地标注了小节、节拍、和弦、音高等信息。这种新的表示形式有助于维护音高局部变化的灵活性，提供了一种可以人为控制的节奏与和声结构。 Compound words[6] 将 REMI 的标记转换为一系列的复合词，大大缩短了序列的长度。

先前关于视频音乐生成的任务主要针对于乐器演奏的视频，例如小提琴、钢琴和吉他[4][15][16]。由于大部分生成结果，例如乐器类型、节奏、音调等都可以从人的手部动作判断，因此生成的音乐也基本是固定的，无法适应一般的视频背景音乐生成任务。

一、音乐与视频的关系

视听关系在心理学与认知科学等领域已经有了几个世纪的研究，特别是音乐与视频，它们在许多方面存在着联系。例如，一个人会希望在看浪漫电影时听到抒情的音乐，或者在观看战斗场景时听到激昂的音乐。

为了更好地使生成的背景音乐匹配视频，我们分析并建立了若干音乐-视频关系。首先建立音视频之间的时间与节奏对应关系；基于此，我们进一步建立视频光流强度与音符密度的联系；最后我们将视频的运动显著性与音乐的音符强度进行对应。以上三种关系将用于指导如何从视频中提取信息监督音乐的生成。

1.1. 视频帧与音乐节拍

理想情况下，生成的背景音乐随着视频的开始与结束应当平滑地出现和减弱。我们考虑将需要生成的音乐分成固定数量的片段(音乐节拍)，并给每一个片段设置位置相关的独特编码，使得模型能够学习到相关的开始和结束等位置信息。通过这种设计，我们能方便地对生成的音乐长度进行控制，使得与相应的视频匹配。

1.2. 动作速度和 Simu-note 密度

我们发现视频的运动速度与音乐的 Simu-note 密度间存在正相关的关系，即快速的画面应当对应激烈的背景音乐，而缓慢的画面应当对应舒缓的背景音乐。如图 2 所示，音乐的 Simu-note 密度定义为一个小节内

Simu-note 的数量。

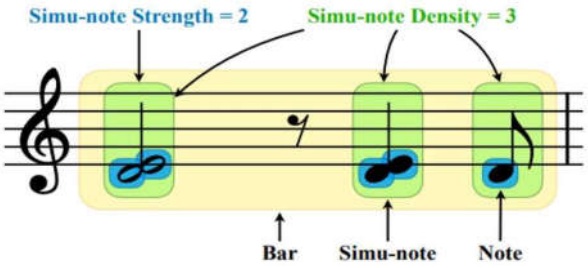

图 2 Simu-note 强度与密度示意图

1.3. 动作显著性和 Simu-note 强度

视觉节拍可以类比为音乐节拍在视觉上的表示形式，其代表视觉动作在时间上的分布。当画面中出现比较明显的转场时，视觉节拍强度应当是一个比较大的值；相反，当画面没有比较大的变化时，视觉节拍强度应当是一个比较小的值。我们对 Simu-note 的强度与视觉节拍强度建立起了一个正相关的对应关系，通过这种方式，视频中的转场点在生成出的音乐中将表现为强拍。如图2所示，Simu-note 强度定义为单个 Simu-note 发生时刻音符的个数。

二、可控音乐生成模型

在上述建立的视频-音乐节奏关系的基础上，我们提出了一种基于 Transformer 的方法以生成视频背景音乐，称为 Controllable Music Transformer (CMT)。整体架构如图 3 所示。我们从视频和 MIDI 文件中分别提取节奏特征(参见前一部分)。在训练时，我们只使用了音乐数据集及其中的节奏特征。在生成时，我们将节奏特征替换为视频中提取出的节奏特征，用来对音乐的生成过程进行控制。

2.1. 音乐表示形式

我们为可控的多轨音乐生成模型设计了一种结构化的表示形式。启发于 PopMAG[13] 与 CWT[6]，我们将一组相互关联的属性合并为同一个 token，来缩短序列的长度。如图 3 所示，每个 token 共有 7 种属性：类别、乐器、时值、音高、强度、密度、拍数。上述属性被分为两组：其一是节奏相关的属性(图 3 中标识为 R)，包括强度、密度和拍数；另一组是音符本身的属性(图 3 中标识为 N)，包括乐器、时值和音高。我们使用类别属性(图 3 中 R/N 的一行)来区分这两组属性。

我们将节奏相关 token 中的乐器、时值和音高三个属性的值设为 None，将音符相关 token 中的强度、密度和拍数三个属性设为 None。图 3 中的空位即对应于该位置的值为 None 的情况。每个节奏相关 token 包含强度属性，表示后续音符 token 的数量。此外，密度属性在每个小节内单调递减，表示该小节内剩余 simu-note 的数量。每个 token 的不同属性分别进行 embedding 并连接到一起，作为 token 的 embedding。此外，我们提取流派和乐器种类作为每段音乐的初始 token，对其使用独立的 embedding 层。

2.2. 对生成过程的控制

在训练结束后，CMT 已经理解了强度和密度两个属性的含义。由此，我们在生成过程中只需将这两个属性替换为我们需要的值，从而生成出与指定视频更和谐匹配的音乐。

2.2.1. 密度属性替换

为了使生成的音乐各处音符密度与视频的光流强度相符，我们将每个小节的密度属性替换为从视频中提取出的光流强度信息，按照特定比例位数进行替换。由于 CMT 已经理解了小节 token 上密度属性的含义，模型将在这个小节中自动生成出对应数量的音符，从而对音符的密度进行控制。

2.2.2. 强度属性替换

类似地，我们利用视频的视觉节拍信息来控制每个 simu note 的强度。如果 CMT 在一个视觉节拍处生成出了一个 simu note，这个 simu note 的强度会被替换为这个视觉节拍的强度(根据比例位数)。然后 CMT 会在这一拍中预测出指定数量的音符，从而控制音符强度。

2.2.3. 调节控制程度

我们使用了一个超参数 C 来调节对模型生成音乐过程的控制程度。在这个过程中，我们需要权衡两个因素。其一是视频与旋律的匹配程度，其二是音乐的质量。也就是说，在生成的过程中，加入的限制条件越多，生成的音乐听起来也越不自然。为了解决这个问题，我们设计了一个超参数 C，用来表示对生成过程的控制程度。C 的值越大，生成过程中加入的限制条件也就越多。当 C 为 0 时，我们生成音乐的过程是完全不受控制的；当

C 为 1 时，我们将得到与视频节奏完全匹配的音乐。 C 的值可以由用户根据需求来指定。

2.2.4. 节拍时间编码

为了利用视频的时间(长度)特征，我们在训练和生成过程的 embedding 中加入了节拍时间编码。也就是说，它指导 CMT 在合适的时间开始和结束生成过程。节拍时间编码的 embedding 表示了当前节拍在整个视频中的位置比例。我们将该比例等分为 M 个区间，并使用一个可学习的 embedding 层来将其映射到和其他 token 相同的维度，并将其一同作为 CMT 的输入。

2.2.5. 流派和乐器种类

我们的方法中包含六种音乐流派(乡村，舞蹈，电子，金属，流行和摇滚)以及五种乐器(鼓，钢琴，吉他，贝斯和弦乐)，将它们各作为 CMT 模型的初始 token。用户可以改变不同的初始 token 以选择不同的流派和乐器，从而使得音乐与视频的情感相互对应。

2.3. 序列建模

音乐 token 序列(如 2.1 小节所述)被输入进 Transformer[17]模型以建模元素之间的依赖关系。我们使用 Linear Transformer[9] 作为模型主干结构，考虑到其轻量级的架构和注意力机制的线性复杂度等优势。

Multi-head 输出模块(依照[6]的设计)，按照两阶段的方式预测每个 token 的 7 种属性。第一阶段中，模型将 transformer 的输出进行线性投影，预测出类别属性；第二阶段中，使用类别属性通过六个前馈 head 同时预测剩余的 6 种属性。

在生成阶段，上述提到的控制策略被结合在一起。我们使用了随机温度控制采样策略[5]以提升生成序列的多样性。

三、实验评估

我们针对提出的音乐生成模型进行了一系列消融实验。在实验中，我们兼用了客观性评价指标与主观性评价指标。 我们使用 Lakh Pianoroll Dataset (LPD) 来训练 CMT 模型。LPD 是从 Lakh MIDI Dataset (LMD) 提取出的 174154 首多轨音乐的集合。我们使用的是 LPD-5-Cleansed 版本的 LPD 数据集，这个版本是由 LMD 经过一系列数据清洗，并将所有音轨合并为鼓、钢琴、吉他、贝斯和弦乐五个音轨后得到的。

3.1. 客观评价性实验

我们使用了 3 个统计性的评价指标来客观地评价生成出的音乐。 Pitch Histogram Entropy，评估音乐的音调质量； Grooving Pattern Similarity，衡量音乐的节奏，和 Structureness Indicator，衡量音乐的结构重复性。如表 1 所示，我们使用消融实验来评价提出的可控性属性在音乐生成过程中发挥的作用。这里列出的所有指标越接近数据集（Data）中的指标代表效果越好。在进行客观指标评估时我们不对音乐施加控制，只增加 3 种可控属性。从结果上看，增加了一些额外的属性后，对音乐生成的节奏和结构性帮助较大。

表 1 客观指标实验结果

Model	Data	Baseline	Ours
Pitch Histogram Entorpy	4.452	3.634	3.617
Grooving Pattern Similarity	0.968	0.677	0.810
Structure Indicator	0.488	0.219	0.241

3.2. 主观评价性实验

对于主观性的用户评价方法，我们设计了一个调查问卷，并邀请了 36 人来对我们提出的可控性指标进行评价。我们主要从两个方面设计主观性评价指标来评价生成出的音乐，首先是音乐本身的音乐性，这里我们主要考虑以下几点：(1) 丰富性：生成出音乐的多样性和趣味性；(2) 正确性：作曲技巧与演奏错误；(3) 结构性：音乐是否具有某种风格或音乐模式。另一个方面是生成的音乐和视频的匹配程度，主要考虑以下几点：(1) 节奏性：生成音乐与视频运动的匹配程度。举例来说，一个激烈体育运动 vlog 会与一首快节奏的音乐相匹配，而一个平缓的旅行 vlog 会与一首轻柔的慢节奏音乐相匹配。(2) 同步性：音乐的重音或边界是否与视频的视觉节拍相匹配。举例来说，对于一个比较有节奏感的视频，例如舞蹈视频，音乐的重音应当落在主要的舞步上。(3) 视频结构性：音乐的起点和终点是否与视频的起点和终点相匹配。类似地，音乐与视频都有序章、插曲和终章，音乐与视频的相应部分应当相互匹配。问卷

需要大约十分钟来完成。结果如表 2 所示，我们生成的音乐虽然在音乐性上不如匹配的人工创作的音乐，但是在音视频匹配程度上和最后的排名上，我们提出的算法均领先。

表 2 主观指标实验结果

Model	Baseline	Matched	Ours
Melodiousness↑	3.4	4.0	3.8
Compatibility↑	3.4	3.7	3.9
Overall Rank↓	2.3	1.9	1.8

四、总结

本文中我们针对未被探索过的视频背景音乐生成任务，提出了关联信息可控的生成模型 CMT，使用基于复合词的音乐表示形式，并加入了“组合音符”的密度与强度两个属性来与视频中的运动强度和视觉节拍相关联。我们的模型无需配对的视频-音乐训练数据，在训练时只使用音乐，而在生成时直接利用视频提取出的关联信息。经过实验，我们的模型生成出了能够与视频的节拍、感情风格相匹配的音乐，生成音乐的质量也比较可观。该项研究对于短视频制作、直播与电商等场景具有实际应用价值。

责任编委 储珺

参考文献

[1] Chris Donahue, Huanru Henry Mao, Yiting Ethan Li, Garrison W Cottrell, and Julian McAuley. 2019. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. arXiv preprint arXiv:1907.04868 (2019).

[2] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

[3] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2018. GANSynth: Adversarial Neural Audio Synthesis. In International Conference on Learning Representations.

[4] Chuang Gan, Deng Huang, Peihao Chen, and Joshua B Tenenbaum. [n.d.]. Foley music: Learning to generate music from videos. ([n. d.]).

[5] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019).

[6] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. arXiv preprint arXiv:2101.02402 (2021).

[7] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer: Generating Music with Long-Term Structure. In International Conference on Learning Representations.

[8] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop Music Transformer: Beat-based modeling and generation of expressive Pop piano compositions. In Proceedings of the 28th ACM International Conference on Multimedia. 1180–1188.

[9] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. 2020. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In Proceedings of the International Conference on Machine Learning (ICML).

[10] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2016. SampleRNN: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837 (2016).

[11] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016).

[12] Christine Payne. 2019. MuseNet. OpenAI Blog 3 (2019).

[13] Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Popmag: Pop music accompaniment generation. In Proceedings of the 28th ACM International Conference on Multimedia. 1198–1206.

[14] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2018. A hierarchical latent vector model for learning long-term structure in music. In International Conference on Machine Learning. PMLR, 4364–4373.

[15] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Audeo: Audio generation for a silent performance video. arXiv preprint arXiv:2006.14348 (2020).

[16] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Multi-Instrumentalist Net: Unsupervised Generation of Music from Body Movements. arXiv preprint arXiv:2012.03478 (2020).

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 6000–6010.

[18] Ziyu Wang, Yiyi Zhang, Yixiao Zhang, Junyan Jiang, Ruihan Yang, Junbo Zhao, and Gus Xia. 2020. Pianotree vae: Structured representation learning for polyphonic music. arXiv preprint arXiv:2008.07118 (2020).

[19] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. arXiv preprint arXiv:1703.10847 (2017).

[20] Andrea Valenti, Antonio Carta, and Davide Bacciu. 2020. Learning Style-Aware Symbolic Music Representations by Adversarial Autoencoders. arXiv preprint arXiv:2001.05494 (2020).

刘偲

北航副教授，博导。研究方向是跨模态多媒体智能分析(跨模态包含自然语言，计算机视觉以及语音等)以及经典计算机视觉任务(目标检测、跟踪和分割)。
个人主页：http://colalab.org/
Email: liusi@buaa.edu.cn

专题综述

动物姿态估计研究与展望

悉尼大学 张敬

一、引言

近年来，对动物的保护和动物行为的分析受到越来越多的关注。准确的动物姿态估计是动物行为分析的重要基础。动物姿态估计旨在对动物运动的关键点进行准确的描述和估计，从而为后续对动物行为的分析创造条件。此外，准确的动物姿态估计在动画制作，仿真设计，动物保护等应用场景中发挥重要作用。

现有的姿态估计方法主要聚焦于人体姿态估计，即识别检测人体上具有语义信息的关键点[8][9][10][12][15][16]，如图 1 所示。自 2014 年以来，人体姿态估计算法经历了快速的发展并在一系列应用中展现出极佳的识别准确度和应用前景。

图 1 人体姿态估计示意图

相较于人体姿态估计，动物姿态估计任务存在更多挑战。其中，最主要的挑战是如何应对自然界中动物种类的多样性。自然界中存在大量不同种类的动物。由于物种、生存环境等不同，不同种类的动物在皮毛，行为，姿态等方面存在巨大差异。先进的人体姿态估计算法是否可以应对这样巨大的差异性是未知且需要探索的。

目前用于动物姿态估计的数据集大多聚焦于特定的动物种类，如老虎，马，猫，狗等，而忽略了自然界中动物种类的多样性。使用这些数据集进行训练和测试只能对算法在某一特定动物种类上的表现进行评估。考虑到动物种类的多样性，这样的数据集是远远不够的。这也让动物姿态估计领域的一个重要问题悬而未决，即现有姿态估计算法是否可以很好的泛化到自然界中各种各样的动物上？

为了探索这个问题的答案，迫切需要构建一个包含多种动物和其姿态标注的大规模数据集。此外，这样一个的数据集也有利于推动动物姿态估计算法的研究从聚焦于单一动物种类的动物姿态估计到一个普适的具有良好泛化性的通用动物姿态估计的发展。

二、姿态估计算法

现有人体姿态估计算法方面的研究可以粗略划分为两类，即自下向上的方法和自上向下的方法。前者指直接根据输入图片回归出人体的关键点信息，并对其进行分组，以得到不同个体的人体关键点检测结果；自上向下的方法指根据人体在图片中所在的位置，将人从图片中分割出来，并针对各个人体进行独立的人体姿态估计。虽然自下向上的方法具有较快的推理速度，尤其是在图像中具有较多个体的情况下，自上向下的方法往往能得到更好的人体姿态估计效果。这些方法往往在通用的人体姿态估计数据集上进行训练和效果评估，包括 MS COCO[6]和 MPII[1]。其中，MS COCO 包含大概 250000 多个人体实例以用于训练和测试，MPII 数据集中包含超过 40000 张有标注的人体实例。为了评估人

体姿态估计算法在更复杂场景中的表现能力，一些包含复杂场景的人体姿态估计数据集也被提出，比方说包含拥挤人群场景的数据集 CrowdPose[4]和包含遮挡场景的数据集 OCHuman[13]。这些数据集极大的帮助了人体姿态估计算法的快速发展，并帮助这些人体姿态估计算法应对各种挑战的场景，例如不同光照条件的变化，人体姿态的变化，人体尺度的变化，以及人群拥挤和遮挡等场景。这些丰富的数据集提供的测试基准为各个先进的人体姿态估计算法在各种应用场景中发挥出良好的效果提供了重要参考。

动物姿态估计和人体姿态估计算法本质上是相似的，即都是将包含动物或者人体的图片作为输入，通过网络预测对应的关键点信息。两者的区别更多的在于由于目标动物或人体姿态、纹理、生物结构等的不同而导致的关键点定义不同。然而，相较于人体姿态估计数据集的多样性和丰富程度，目前只有少数几个关于动物姿态估计的数据集提供给动物姿态估计识别算法进行训练和测试。然而，由于深度学习算法是数据驱动的，这样有限的数据集会影响和限制目前的动物姿态估计模型的性能和泛化性。此外，由于动物数据整理和标注的难度，这些数据集往往只关注特定的动物种类，例如，马，斑马，蚊子，老虎等。这样有限的动物种类使这些数据集中只包含有限的动物姿态信息、纹理信息和动物栖息地的背景信息等。Animal Pose Dataset[2] 尝试包含更多种类的动物，以帮助模型学到更具备泛化性的特征。然而，Animal Pose 数据集也仅仅包含 5 类动物，

这还是难以让网络学到足够具有泛化性的特征表示。此外，尽管不同种类的动物有不同的外观、行为模式和骨骼分布情况，他们往往会遵循一定的生物学规律，例如生物进化过程中自然产生的科，目，种等分类方式。属于同一属的动物往往会比属于不同属的动物有更为接近的生活习性，行为模式和姿态分布等。具体来说，相较于牛和黑猩猩，牛和马有更为相近的关键点分布。这是因为牛和马同属于偶蹄目，而黑猩猩则并不属于偶蹄目。利用不同动物生物学上的相似特性可以帮助在有限动物种类上训练好的动物姿态估计网络更好的泛化到未知的动物种类上。因此，一个大规模的，按照生物学规律进行整理的动物姿态估计数据集是有必要的。

表 1 动物关键点定义

Keypoint	Definition	Keypoint	Definition
1	Left Eye	10	Right Elbow
2	Right Eye	11	Right Front Paw
3	Nose	12	Left Hip
4	Neck	13	Left Knee
5	Root of Tail	14	Left Back Paw
6	Left Shoulder	15	Right Hip
7	Left Elbow	16	Right Knee
8	Left Front Paw	17	Right Back Paw
9	Right Shoulder		

三、 动物姿态估计数据集

3.1 数据集收集

本文构建了一个大规模动物姿态估计数据集 AP-10K[11]。为了得到高质量动物数据，AP-10K 以 9 个公开发布的用于动物分类的数据集为基础，经过仔细清洗、鉴别、再组织和标记，构建了一个包含 59658 张图片的动物数据集。在这个数据集中，不同动物按照科和物种的生物学概念进行了准确划分，物种之间的生物学关系得到了清晰的体现。在此基础上，经过仔细分析和挑选，本着“每个物种选取 200 张作为基础，稀有物种充分标记”的原则，我们对其中 54 类动物进行标记，最终得到了 10015 张包含姿态信息的图片。表 1 展示了 17 个关键点的定义，图 2 展示了一幅黑猩猩图片其对应的标记。

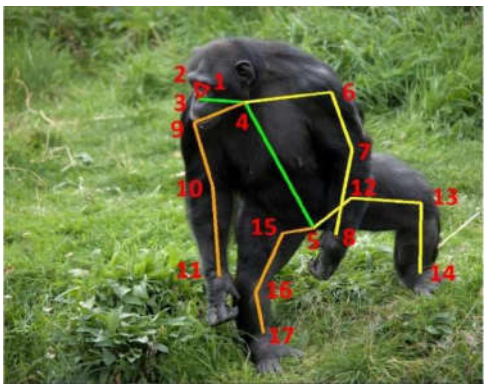

图 2 黑猩猩图片及其对应标记点

表 2 不同动物数据集对比

Dataset	Species	Family	Labeled image	Unlabeled image	Keypoint	Instance
Animal-Pose Dataset [2]	5	N/A	4,666	0	20	6,117
Horses-10 [7]	1	N/A	8,110	0	22	8,110
ATRW [5]	1	N/A	8,076	0	15	9,496
AP-10K	54	23	10,015	50k	17	13,028

表 3 全监督学习结果比较

	HRNet-w32 [9]	HRNet-w48 [9]	ResNet50 [3]	ResNet101 [3]	Hourglass [8]
w/o pretraining	$0.703_{\pm 0.002}$	$0.713_{\pm 0.002}$	$0.646_{\pm 0.001}$	$0.667_{\pm 0.002}$	$0.686_{\pm 0.006}$
w/ pretraining	$0.738_{\pm 0.006}$	$0.744_{\pm 0.004}$	$0.699_{\pm 0.004}$	$0.698_{\pm 0.002}$	$0.729_{\pm 0.001}$

3.2 数据集整理

为了利用好生物进化规律以帮助使用有限动物种类进行训练的动物姿态估计模型更好的泛化到自然界中各种动物上，我们在构建 AP-10K 的过程中对收集到的数据按照分类阶元进行了重新整理和标注。为了简化分类层级，我们没有按照科-属-种关系对动物进行整理，而是按照科-种关系对动物进行整理，即对属和种不作进一步区分。此外，按照生物学进化规律对动物进行划分可以对网络泛化能力进行更为公平的评估，并为提升网络在特定动物类别关键点估计能力提供依据和指导。

3.3 数据集标注

为了获得高质量的标记效果，我们招募了 13 名经过训练的志愿者对数据进行标注。此外，我们提供了详尽的文档对于标记者可能遇到的标记状况进行了详细的解说，其中包括对于多个体、遮挡情况等情形的处理情况等。这些举措保证了多个体、遮挡等有难度的少见样本的准确标记效果。为了更进一步保证标注信息的质量，我们采取了自动化和人工两种校验手段。其中自动化校验是指根据预设规则对于标记好的坐标信息进行自动化检查，去除一些低质量标记和错误标记。例如标记点落在检测框外侧，同一个实例出现重复的标记名称等。人工校验是指组织者和标记者进行了三轮检查，这确保了高质量的标注信息。三轮检查过程如下：首先，标记者在分配的标记工作完成后，将标记结果提交组织者进行检查，组织者将检查出的错误信息反馈给标记者，这是一轮检查；标记者根据反馈的勘误表对标记进行修改，并将二次修改结果反馈给组织者，这是二轮检查；最后组织者拿到二次标记结果，对于标记进行最后的检查，如果发现错误就进行本地修改，这是三轮检查。三轮检查的过程如同 TCP 协议的三次握手一般，加强了标注可靠性。

图 3 AP-10K 动物种类分布情况

3.4 数据集统计指标

AP-10K 数据集中包含 10015 张完全标记的动物图片及其对应标注，共 13028 个不同的动物个体，涵盖了 23 科，54 种不同的动物种类。这使得 AP-10K 有更为复杂的动物姿态分布和纹理信息。这样的特点让使用 AP-10K 进行训练的模型展现出更好的泛化性能。下图展示了 AP-10K 数据集的特点(表 2)和动物种类分布(图 3)。由图中可以看出，AP-10K 数据集不论是在动物种类还是在标记数量均具有显著优势。值得一提的，AP-

10K 数据集的标记图片具有长尾分布的特点，比如对于猫科(Felidae)来说，一共有 10 个标记物种，1913 张标记图片，而河狸科只包含 1 个物种，178 张标记图片。这些特性对于小样本学习、零样本学习或者元学习等研究方向是很有意义的。此外，AP-10K 数据集中额外包含 50K 张含有类别标注但是缺少关键点标注的动物图片。这些图片和对应的生物学标注可以为研究跨物种动物姿态估计的自监督[15]和半监督学习等课题提供条件。

表 4 人体姿态估计模型迁移学习效果比较

epoch	AP	$AP_{.5}$	$AP_{.75}$	AP_M	AP_L
20	$0.606_{\pm 0.004}$	$0.906_{\pm 0.005}$	$0.635_{\pm 0.006}$	$0.501_{\pm 0.037}$	$0.610_{\pm 0.003}$
30	$0.642_{\pm 0.002}$	$0.921_{\pm 0.010}$	$0.680_{\pm 0.002}$	$0.521_{\pm 0.044}$	$0.645_{\pm 0.002}$
40	$0.667_{\pm 0.003}$	$0.934_{\pm 0.004}$	$0.714_{\pm 0.007}$	$0.547_{\pm 0.059}$	$0.671_{\pm 0.003}$
210	$0.753_{\pm 0.005}$	$0.962_{\pm 0.002}$	$0.827_{\pm 0.003}$	$0.616_{\pm 0.031}$	$0.756_{\pm 0.004}$

表 5 牛科科内泛化实验

Train / Test	Bov./Ant	Bov./A.S.	Bov./Bis	Bov./Buf	Bov./Cow	Bov./She.	Average
Antelope	$0.607_{\pm 0.010}$	$0.742_{\pm 0.013}$	$0.775_{\pm 0.004}$	$0.842_{\pm 0.005}$	$0.729_{\pm 0.002}$	$0.853_{\pm 0.002}$	$0.788_{\pm 0.051}$
A.S.	$0.836_{\pm 0.016}$	$0.655_{\pm 0.015}$	$0.805_{\pm 0.022}$	$0.840_{\pm 0.007}$	$0.725_{\pm 0.027}$	$0.697_{\pm 0.008}$	$0.781_{\pm 0.059}$
Bison	$0.731_{\pm 0.017}$	$0.646_{\pm 0.009}$	$0.530_{\pm 0.006}$	$0.605_{\pm 0.006}$	$0.616_{\pm 0.009}$	$0.693_{\pm 0.014}$	$0.658_{\pm 0.047}$
Buffalo	$0.783_{\pm 0.010}$	$0.748_{\pm 0.031}$	$0.726_{\pm 0.017}$	$0.658_{\pm 0.004}$	$0.794_{\pm 0.022}$	$0.750_{\pm 0.008}$	$0.760_{\pm 0.025}$
Cow	$0.597_{\pm 0.011}$	$0.691_{\pm 0.004}$	$0.740_{\pm 0.007}$	$0.732_{\pm 0.009}$	$0.586_{\pm 0.006}$	$0.683_{\pm 0.002}$	$0.689_{\pm 0.051}$
Sheep	$0.707_{\pm 0.012}$	$0.607_{\pm 0.006}$	$0.681_{\pm 0.004}$	$0.676_{\pm 0.007}$	$0.645_{\pm 0.005}$	$0.520_{\pm 0.001}$	$0.663_{\pm 0.034}$

表 6 狗科科内泛化实验

Train / Test	Can./Dog	Can./Fox	Can./Wolf	Average
Dog	$0.224_{\pm 0.011}$	$0.699_{\pm 0.009}$	$0.699_{\pm 0.003}$	$0.699_{\pm 0.000}$
Fox	$0.614_{\pm 0.013}$	$0.627_{\pm 0.005}$	$0.732_{\pm 0.013}$	$0.673_{\pm 0.059}$
Wolf	$0.663_{\pm 0.024}$	$0.694_{\pm 0.013}$	$0.633_{\pm 0.006}$	$0.679_{\pm 0.016}$

四、应用

4.1 全监督学习

AP-10K 评估了五种主流的人体姿态估计模型在动物姿态估计任务上的表现，它们分别是 HRNet-w32[9]，HRNet-w48[9]，SimpleBaseline[10] (ResNet50[3]骨干网络)[3]，SimpleBaseline[10](ResNet101[3]骨干网络)和 Hourglass[8]，然后又对比了使用流行的 ImageNet 预训练模型和随机初始化网络进行训练的效果。实验结果(表 3)表明：使用 ImageNet[14] 预训练比随机初始化的效果要更好，ImageNet 预训练能够提升上述 5 种模型的性能。随着网络规模的增大，HRNet[9] 和 SimpleBaseline[10] 的训练指标也逐渐提升，这展现了大网络具有更好的表征能力这一特性。值得注意的是，增加训练代数可以提升随机初始化模型的最终性能，降低其与采用 ImageNet 预训练权重初始化模型的性能差距，详细结果可以参考 AP-10k 论文[11]。

4.2 人体姿态估计模型的迁移学习

因为人和四足动物的相似性，评估人体姿态估计模型到动物姿态估计模型的泛化能力是一个很有必要的事情。AP-10K 使用 HRNet-w32 模型，加载基于 COCO 的人体姿态估计任务预训练模型的权重，然后在 AP-10K 数据集上进行微调并测试。实验结果(表 4)表明当训练 epoch 较少时，人体姿态估计算法迁移到动物姿态估计的结果不够好，这是因为动物和人在外形

表 7 猫科科内泛化实验

Train / Test	Fel./Bob.	Fel./Cat	Fel./Che.	Fel./Jag.	Fel./K.C.	Fel./Leo.	Fel./Lio.	Fel./Pan.	Fel./S.L.	Fel./Tig.	Average
Bob.	0.631 ±0.005	0.714 ±0.016	0.664 ±0.004	0.674 ±0.013	0.673 ±0.013	0.663 ±0.006	0.691 ±0.016	0.623 ±0.004	0.669 ±0.005	0.713 ±0.008	0.676 ±0.026
Cat	0.638 ±0.002	0.332 ±0.004	0.625 ±0.018	0.552 ±0.010	0.629 ±0.007	0.641 ±0.009	0.601 ±0.004	0.609 ±0.010	0.582 ±0.014	0.608 ±0.007	0.609 ±0.027
Che.	0.715 ±0.002	0.716 ±0.012	0.660 ±0.003	0.762 ±0.013	0.731 ±0.014	0.747 ±0.010	0.734 ±0.021	0.790 ±0.008	0.713 ±0.008	0.662 ±0.008	0.730 ±0.034
Jag.	0.757 ±0.005	0.770 ±0.017	0.754 ±0.006	0.704 ±0.008	0.750 ±0.004	0.759 ±0.012	0.798 ±0.013	0.724 ±0.008	0.756 ±0.011	0.734 ±0.005	0.756 ±0.020
K.C.	0.961 ±0.008	0.804 ±0.035	0.692 ±0.042	0.771 ±0.028	0.779 ±0.010	0.958 ±0.008	0.713 ±0.017	0.924 ±0.026	0.864 ±0.033	0.838 ±0.016	0.836 ±0.094
Leo.	0.730 ±0.005	0.697 ±0.007	0.766 ±0.014	0.741 ±0.006	0.682 ±0.005	0.686 ±0.009	0.700 ±0.012	0.705 ±0.012	0.775 ±0.010	0.744 ±0.004	0.727 ±0.031
Lio.	0.623 ±0.016	0.582 ±0.023	0.639 ±0.012	0.694 ±0.010	0.688 ±0.002	0.690 ±0.018	0.528 ±0.002	0.638 ±0.007	0.630 ±0.011	0.625 ±0.024	0.645 ±0.036
Pan.	0.705 ±0.020	0.722 ±0.011	0.718 ±0.020	0.720 ±0.023	0.727 ±0.013	0.785 ±0.014	0.763 ±0.026	0.511 ±0.014	0.719 ±0.004	0.684 ±0.018	0.727 ±0.028
S.L.	0.792 ±0.011	0.776 ±0.008	0.810 ±0.018	0.779 ±0.019	0.790 ±0.024	0.818 ±0.004	0.821 ±0.009	0.760 ±0.015	0.724 ±0.010	0.855 ±0.012	0.800 ±0.027
Tig.	0.754 ±0.008	0.741 ±0.018	0.751 ±0.012	0.715 ±0.015	0.768 ±0.021	0.753 ±0.015	0.797 ±0.005	0.848 ±0.023	0.744 ±0.011	0.675 ±0.007	0.763 ±0.036

表 8 科间泛化实验

train	Bov.	$0.782_{\pm 0.002}$	Bov.	$0.782_{\pm 0.002}$	Bov.	$0.782_{\pm 0.002}$	Cerc.	$0.695_{\pm 0.007}$
	Ant.	$0.856_{\pm 0.001}$	Ant.	$0.856_{\pm 0.001}$	Ant.	$0.856_{\pm 0.001}$	Alo.	$0.697_{\pm 0.020}$
	A.S.	$0.887_{\pm 0.006}$	A.S.	$0.887_{\pm 0.006}$	A.S.	$0.887_{\pm 0.006}$	Mon.	$0.725_{\pm 0.013}$
	Bis.	$0.643_{\pm 0.005}$	Bis.	$0.643_{\pm 0.005}$	Bis.	$0.643_{\pm 0.005}$	N.N.M.	$0.750_{\pm 0.027}$
	Buf.	$0.815_{\pm 0.004}$	Buf.	$0.815_{\pm 0.004}$	Buf.	$0.815_{\pm 0.004}$	S.M.	$0.581_{\pm 0.008}$
	Cow.	$0.737_{\pm 0.004}$	Cow.	$0.737_{\pm 0.004}$	Cow.	$0.737_{\pm 0.004}$	Uak.	$0.720_{\pm 0.009}$
	She.	$0.754_{\pm 0.005}$	She.	$0.754_{\pm 0.005}$	She.	$0.754_{\pm 0.002}$		
test	Cer.	$0.641_{\pm 0.007}$	Equ.	$0.468_{\pm 0.019}$	Hom.	$0.015_{\pm 0.001}$	Hom.	$0.446_{\pm 0.007}$
	Der.	$0.724_{\pm 0.004}$	Hor.	$0.618_{\pm 0.005}$	Chi.	$0.005_{\pm 0.000}$	Chi.	$0.446_{\pm 0.011}$
	Moo.	$0.558_{\pm 0.010}$	Zeb.	$0.319_{\pm 0.035}$	Gor.	$0.026_{\pm 0.003}$	Gor.	$0.445_{\pm 0.011}$

和纹理上有较大的差异性。随着训练时间的增加，微调的效果也逐渐增加，并显著优于采用 ImageNet 预训练模型进行训练的结果。该结果表明，人体姿态估计和动物姿态估计任务之间域间隔(Domain Gap)相比姿态估计任务和图像分类任务之间域间隔更小。

4.3 动物姿态估计模型在科内和科间的泛化性能

为了验证动物姿态估计模型在同一科内和相似动物科之间的泛化性能，我们选择了 AP-10K 中三个数量最多的科(牛，狗和猫)进行实验。在每科中，一个物种被用作测试集而剩下的物种构成训练集。科内实验结果(表 5-7)表明，在三个不同科中，测试物种的分数虽然不如在第一部分中使用大量物种进行训练的效果好，但是也能达到一个不错的结果。这是因为同科物种在生物学关系和外形上具有高度相似性。实验结果中狗(Dog)的分数偏低，这是因为相比狐狸(Fox)和狼(Wolf)，狗(Dog)包含了更多的图片，将其排除之后训练集图片数量较少。其次，狗(Dog)中包含了许多人工培育的宠物类型，它们的外形差异较大，类似现象也存在于猫(Cat)中。

在科间实验中，牛科被用作为训练集，鹿科(Cervidae)、马科(Equidae)和人科(Hominidae)被分别用作测试集。科间实验结果(表 8)表明，使用牛科作为训练集的模型在鹿科和马科的泛化结果很好，但是在人科上泛化效果较差。因为牛科和鹿科、马科的生物学关系相近，外形差异也较小。而人科物种和牛科生物学关系

表 9　科间迁移学习和少样本学习效果

Species	Setting	Performance	Species	Setting	Performance
Deer	Generalization	$0.723_{\pm 0.036}$	Moose	Generalization	$0.587_{\pm 0.025}$
	Few-Shot	$0.742_{\pm 0.034}$		Few-Shot	$0.648_{\pm 0.025}$
	Transfer	$0.751_{\pm 0.024}$		Transfer	$0.726_{\pm 0.011}$
Horse	Generalization	$0.592_{\pm 0.047}$	Zebra	Generalization	$0.324_{\pm 0.021}$
	Few-Shot	$0.635_{\pm 0.034}$		Few-Shot	$0.480_{\pm 0.029}$
	Transfer	$0.718_{\pm 0.023}$		Transfer	$0.708_{\pm 0.024}$
Chimpanzee	Generalization	$0.009_{\pm 0.006}$	Gorilla	Generalization	$0.017_{\pm 0.006}$
	Few-Shot	$0.022_{\pm 0.010}$		Few-Shot	$0.144_{\pm 0.121}$
	Transfer	$0.550_{\pm 0.032}$		Transfer	$0.662_{\pm 0.039}$

表 10 跨数据集泛化效果比较

	Direct Test(mAP)	Finetune&Test(mAP)	Train&Test(mAP)
Animal-Pose Dataset[2] → AP-10K	0.424	0.722	0.727
AP-10K → Animal-Pose Dataset[2]	0.913	0.935	0.932

较远，外形和生存环境也差异较大，所以泛化效果不好。作为对照，表格最后一列使用了猴科(Cercopithecidae)作为训练集来测试人科的物种，性能得到大幅提升，这再次证明了 AP-10K 在构建过程中采用生物学进化规律的必要性：生物学关系和外形相似的物种，彼此之间域差异也越小，更利于姿态估计模型的泛化。

4.4 科间的迁移学习和少样本学习

在科间泛化实验的基础上，AP-10K 进一步探究了少样本学习和迁移学习带来的性能提升。与科间迁移实验相同，牛科图片被作为训练集，然后鹿科(Deer 和 Moose)、马科(Horse 和 Zebra)和人科(Chimpanzee 和 Gorilla)图片被用于微调和测试，其中少样本学习对每个物种抽样 20 张进行微调，而迁移学习采用该物种全部训练集图片进行微调。实验结果(表 9)表明少样本学习和迁移学习效果均相对于直接泛化测试有了不同程度的提升。即便是对人科这样和训练集差距较大的测试集，采用更多的图片进行迁移也能得到性能的提升。

4.5 跨数据集泛化能力比较

如表 10 所示，我们使用 Animal Pose Dataset (包含 5 类动物)和 AP-10K 数据集分别训练姿态估计模型并对比了它们的双向泛化效果。结果表明，采用包含更多物种的 AP-10K 数据集进行(预)训练的模型的泛化性能优于使用少量动物数据进行训练的模型。

五、展望

AP-10K 是第一个大规模的哺乳动物姿态数据集。它的物种数量、姿态多样性，以及按照生物学关系组织上的优势可以极大的促进相关领域的研究，例如动物保护和动物行为研究等。我们基于 AP-10K 训练了 5 种经典的姿态估计模型并测试了它们的在不同物种上的表现能力，初步探究了动物和人体姿态估计之间的联系以及不同物种之间的泛化效果。总得来说，AP-10K 数据集为动物姿态估计领域提供新的可能性和发展发向。

致谢： 本文由博士生徐宇飞(悉尼大学)、喻航(西安电子科技大学)撰写初稿，指导老师张敬(悉尼大学)进行修改。本文对应发表在 NeurIPS2021 的学术论文作者还包括赵伟教授(西安电子科技大学)、管子玉教授(西安电子科技大学)、陶大程教授(京东探索研究院)。

论文链接：
https://openreview.net/forum?id=rH8yliN6C83

数据集和代码链接：
https://github.com/AlexTheBad/AP-10K

责任编委　崔海楠

参考文献

[1] Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In CVPR.

[2] Cao, J., Tang, H., Fang, H. S., Shen, X., Lu, C., & Tai, Y. W. (2019). Cross-domain adaptation for animal pose estimation. In CVPR.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In CVPR.

[4] Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H. S., & Lu, C. (2019). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In CVPR

[5] Li, S., Li, J., Tang, H., Qian, R., & Lin, W. (2020). ATRW: A Benchmark for Amur Tiger Re-identification in the Wild. In Proceedings of the 28th ACM International Conference on Multimedia.

[6] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In ECCV.

[7] Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M., & Mathis, M. W. (2021). Pretraining boosts out-of-domain robustness for pose estimation. IEEE/CVF Winter Conference on Applications of Computer Vision.

[8] Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In ECCV.

[9] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[10] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In ECCV.

[11] Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D. (2021). AP-10K: A Benchmark for Animal Pose Estimation in the Wild. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

[12] Zhang, J., Chen, Z., & Tao, D. (2021). Towards high performance human keypoint detection. International Journal of Computer Vision, 129(9), 2639-2662.

[13] Zhang, S. H., Li, R., Dong, X., Rosin, P., Cai, Z., Xi, H., ... & Hu, S. M. (2019). Pose2seg: Detection free human instance segmentation. In CVPR.

[14] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In CVPR.

[15] Xu, Y., Zhang, Q., Zhang, J., & Tao, D. (2021). RegionCL: Can Simple Region Swapping Contribute to Contrastive Learning? arXiv preprint arXiv:2111.12309.

[16] Xu, Y., Zhang, Q., Zhang, J., & Tao, D. (2021). ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias. In Thirty-fifth Conference on Neural Information Processing Systems.

张敬

悉尼大学工程学院计算机系，博士后研究员。主要研究方向为计算机视觉和深度学习，已经在计算机视觉及人工智能相关领域的国内外著名学术期刊和会议发表论文 40 余篇，担任多个国际学术期刊和会议的审稿人，以及 IJCAI、AAAI 的 Senior Program Committee Member。
Email: jing.zhang1@sydney.edu.au