1. Problem Statement

本文试图解决的研究问题是什么?这个问题重要吗?为什么?

当前的技术状态(即其他竞争方法/解决方案)是什么?

这篇论文的动机是什么?这篇论文的定位是什么?

如果有必要，你可以用图片来更好地说明研究问题。

本文试图解决的问题是使用简单的方法进行“do as I do” motion transfer，即将给定的视频中人的动作转移到另一个amateur人上面。同时给出方法来辨别视频是否是合成的。该方法可以用于影视特效制作和游戏等领域，同时用于辨别的方法也可以用来解决由视频真实性引发的法律和伦理问题。

早期的方法专注于利用现有的视频片段创建新的内容。有人使用optical flow或运动转移的经典计算机图形方法以在3d中达成目标。有一些方法依赖于校准的多摄像机系统扫描目标演员来建立3D模型，从而操纵他们的动作。最近的研究使用了深度学习，但依赖于更详细的输入表示。也有些方法将运动从外观中分离并synthesizing videos with novel motion，如采用无监督对抗训练来分离，但这些方法并不专注于合成详细的动作视频。现代方法能够生成新的人体姿态的单一图像，为之引入了新的结构和损失，并证明了姿势是未来预测和视频生成的有效监控信号，但是这些方法并不是专门为运动传递设计的。同时，也有方法通过学习视频之间的映射实现了动作转换，但是方法更复杂，消耗资源更多。

相比之下，本文提出的方法是合成新的动作并且使用2D视频来实现运动转移。该方法基于最近在两个独立方向上的研究：包括OpenPose和DensePose在内的现代姿势检测系统能够可靠和快速地提取姿势，最近出现的图像到图像转换模型如CoGAN，pix2pix等能够实现高质量的单幅图像生成。基于此，作者使用姿态检测作为中间表示，并考虑了时间相干性，从而获得逼真的视频。同时他们还给出了用以分辨合成视频的方法。

2. Summarise the paper’s main contributions

这篇论文声称有哪些贡献?

这篇论文有什么新内容吗?这篇论文有什么创新之处?

作者是否夸大了他们的贡献，如果是，又是如何夸大的?

作者声称他们设计了一个简单的方法来实现动作转移，基于现代姿势检测系统和image-to-image转换模型，使用姿势检查作为中间表示，生成时间相干的、逼真的视频，并给出了辨别视频真伪的方法。

虽然他们的工作大量基于前人的成果，但是他们创新性地有效地整合了两个独立方向的方法并加以改进，而且找到了合理的中间表示，从而使用较少的开销得到了较为逼真的结果。

3. Method and Experiment

他们提出了什么方法/想法/见解? 解释他们方法的关键(关键思想/洞察力/聪明才智)。

主要观点/论点是什么?

作者如何证实他们的主张?进行了哪些实验?这些实验的主要结果或发现是什么?

该方法分为三个阶段：姿态检测，global pose normalization和mapping from normalized pose stick figures to the target subject。

在姿态检测阶段是使用预先训练好的OpenPose姿态检测器估计2D的关节坐标x,y并创建原视频的pose stick figure。

全局姿态归一化阶段解释了源和目标身体形状和位置的差异。由于在不同视频中人的肢体比例可能不同，因而需要对源人的姿态关键点进行变换以使其和目标人一致。他们使用最近和最远的脚踝之间的线性映射找到了这种变换，并据此计算比例和平移。

在最后的阶段，他们使用生成对抗网络（GAN）来学习从pose stick figures到目标人物图像的映射。视频合成的方法是基于Wang等人[41]提出的对抗单帧生成过程。原始的GAN设置中生成器网络G合成图像以欺骗多尺度鉴别器D，同时D鉴别G产生的grand truth和虚假图像，从而同时训练G和D并相互推动以提高。基于单帧的方法在合成视频时会生成时间伪影，因此他们加入了一个learned model of temporal coherence以加强相邻帧之间的时间连贯性，预测帧xt-1的输出G（xt-1）和帧xt的输出G（xt），D也相应的判断假序列(xt-1, xt, G(xt-1), G(xt))和真实序列(xt-1, xt, yt-1, yt)的真实性。目标为：

文本, 信件

描述已自动生成

为了给面部增加更多细节和真实感，他们添加了一个专门的GAN Gf，输入面部周围的小部分图像G(x)F和xF即对应的pose stick figure，并生成一个残差r = Gf (xF , G(x)F )，最后与G生成的区域相加。鉴别器Df鉴别真脸和假脸，类似于pix2pix [16]目标：

文本, 信件

描述已自动生成

训练分阶段进行，且全图GAN和面部GAN分别训练。首先是针对全图, objective是

文本

描述已自动生成

LGAN(G, D) is the single image adversarial loss in the pix2pix paper [16]，LFM(G, D) is the discriminator feature-matching loss in pix2pixHD, and LP (G(x), y) is the perceptual reconstruction loss [17] which compares pretrained VGGNet [35] features at different layers of the network。之后是针对面部的GAN：

文本

中度可信度描述已自动生成

他们也比较了该方法与其他方法在视频和单帧方面的性能，并使用Mechanical Turk评价视频性能，使用SIMM和LPIPS方法评价单帧性能。在视频方面，调察显示与PoseWarp方法相比该方法无论使用Face GAN模块与否，其生成的视频都更加真实。而图片方面同样证实,与frame-by-frame(FBF)和temporal smoothing(FBF+TS)相比，该方法（FBF+TS+Face GAN）更为真实，虽然时间平滑设置的有效性并没有被强调。

他们还训练了一个fake-detector来分辨视频真伪。他们收集了一套62个主题的舞蹈视频数据集，使用48个作为训练对象，其余为测试对象，通过合成模型生成多个假视频来训练fake-detector。总的来说假检测器可以很好的分辨视频的真伪。

5. Critical Analysis

**5.1. Are the paper’s contributions** **significant?**

贡献/改进是微不足道的、增量的吗?

为什么之前的努力失败了?

论文的贡献是显著的。先前方法有些需要复杂的多摄像机系统，有些需要3D建模，有些需要更详细的输入，有些方法更复杂而且需要更多的资源。而该方法很好的整合了前人的工作，只需要使用易于获取的2D视频即可生成逼真的视频。同时他们还提出了辨别视频真伪的方法用于帮助解决可能产生的法律和伦理问题。

**5.2. Are the authors’ main claims valid?**

他们是否令人信服地证实了自己的主要观点?

他们的论点，推导，实验有漏洞吗?

作者的main claims是有效的。作者针对之前方法的局限提出了新的方法，给出了数学推导和网络模型，并通过实验对比了不同方法之间的差异。而且他们的调察也显示该方法确实表现更好。

**5.3. Limitation and weaknesses**

他们的方法有什么限制/弱点吗?可以做些什么来改进这项工作?

你会如何解决/克服他们的弱点?

从给出的结果图片和视频可以看出，有时候OpenPose对头发、衣服和身体的检测有时候也会出错，并且有时候会出现纹理伪影。这可以通过将目标视频与不同衣服或照明条件相结合来改善效果，改进姿态检测模块，和减轻高频纹理的伪影来改善结果。他们的姿势规范化阶段没有考虑不同的肢体长度或相机位置，使得生成的视频与源视频动作有差异。这可以通过优化规范化方法来改善。他们的模型有时候很难推断出训练数据集中完全没有的姿势。未来可以通过完善数据集来改善。

**5.4. Extension and future work**

你会建议作者做哪些额外的实验来强化结果?

你能想到论文中提出的方法/想法(假设有效)的其他可能应用吗?

未来可能的作品是什么?

作者可以尝试其他的姿势检测方法以寻求更好的姿势检测结果，并优化姿势规范化方法以减少动作的差异。同时他们还可以扩充数据集，尝试更多不同的衣物与光照条件的组合来减少纹理伪影，并尝试更多的动作来增强方法对不同动作的适应性。

未来该方法可以用在视频特效制作等娱乐性乃至专业性的领域，他们的视频真伪检测方法也能够用于解决法律和伦理问题。

**5.5. Is the paper stimulating or inspiring ?**

许多论文(甚至那些已经发表的论文)都是枯燥乏味的，而有些则是令人兴奋的。你对这篇论文有什么看法?为什么?

这篇论文是令人兴奋的。之前方法需要复杂的硬件设置、者详细的输入或3D建模，不容易实现。而该方法更容易实现，仅使用易于获取的视频即可产生逼真的视频。

**5.6. Conclusion and personal reflection**

首先，对本文进行总结。

那么，如果让你来解决这个研究问题，你会有什么不同的做法?还有别的解决方案吗?

最后，用一句话总结你从阅读本文中学到的东西。

In conclusion，这篇论文提出了一个简单的方法用于解决“do as I do”运动转移问题，整合了前人在两个独立方向上的成果，通过大量运用GAN，用2d视频作为输入源生成逼真的视频。他们还提出了用于辨别视频真伪的方法。

如果是我，我可能会尝试在此基础上结合光流法进行粗略的建模来辅助姿势检测算法，从而提高鲁棒性。

这篇文章告诉我找到合适的中间表达有时可以降低输入数据的详细度。同时合理使用GAN可以增强算法的效果。