Liquid Warping GAN: 模仿人体运动，外观转移和新颖视图合成的统一框架

Wen Liu[1] Zhixin Piao[1] Jie Min[1] Wenhan Luo[2] Lin Ma[2] Shenghua Gao[1]
[1]ShanghaiTech University [2]Tencent AI Lab
fliuwen,piaozhx,minjie,gaoshhg@shanghaitech.edu.cn
fwhluo.china,forest.linmag@gmail.com

摘 要

我们在一个统一的框架内处理人类的运动模仿，外观转移和新颖的视图合成，这意味着一旦受过训练的模型就可以用来处理所有这些任务。现有的特定于任务的方法主要使用 2D 关键点（姿势）来估计人体结构。但是，它们仅表达位置信息，无法表征个人的个性化形状并模拟四肢旋转。在本文中，我们建议使用 3D 身体网格恢复模块来解码姿势和形状，该模块不仅可以模拟关节的位置和旋转，而且可以表征个性化的身体形状。为了保留源信息，例如纹理，样式，颜色和脸部身份，我们提出了一种带有流动变形块（LWB）的流动变形 GAN，它可以在图像和功能空间中传播源信息，并合成依据于参考的图像。具体而言，通过降噪卷积自动编码器提取源特征，以很好地表征源身份。此外，我们提出的方法能够支持来自多个来源的更灵活的变形。此外，我们建立了一个新的数据集，即模仿者（iPER）数据集，用于评估人体运动模仿，外观转移和新颖的视图合成。广泛的实验从几个方面证明了我们方法的有效性，例如遮盖情况下的鲁棒性以及保持面部身份性，形状一致性和衣服细节。所有代码和数据集均可在下面访问：

https://svip-lab.github.io/ project/impersonator.html.

1. 介绍

人体图像合成，包括人体运动模仿[1，19，31]，外观转移[26，37]和新颖

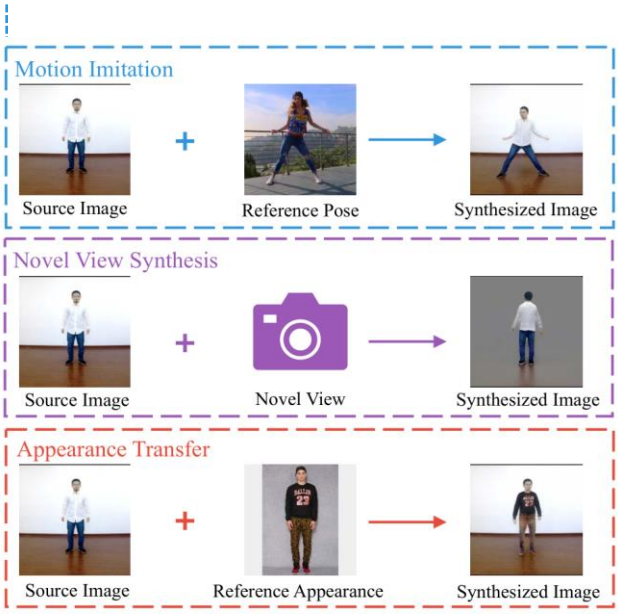


图 1. 人体运动模仿，外观转移和新颖视图合成的插图。 第一列是源图像，第二列是参考条件，例如图像或相机的新颖视图。 第三列是合成结果。

视图合成[40，42]在重演，角色动画，虚拟服装试穿，电影或游戏制作等方面具有巨大的潜在应用价值。具体来说，是给定源人物图像和参考人物图像，i）运动模仿任务的目标是生成具有源人物纹理和参考人物姿势的图像，如图 1 顶部所示； ii）人类新颖视图的合成旨在合成从不同角度捕获的人体新图像，如图 1 中部所示； iii）外观转移的目标是生成一个保留衣服参考身份的人像，如图 1 的底部所示，其中不同部分可能来自不同的人。

在人类图像合成领域，以前的工作分别针对特定任务渠道处理这些任务[19、26、42]，

Liu Wen 担任腾讯 AI Lab 的研究实习生时，贡献了均等的心血并完成了工作。

这似乎很难扩展到其他任务。最近，生成对抗网络（GAN）[6]在这些任务上取得了巨大的成功。以模仿人类动作为例，我们在图 2 中总结了最近的方法。在早期工作[19]中，如图 2（a）所示，源图像（及其姿势条件）和目标姿势的条件串联起来，然后通过对抗训练将其馈送到网络中，以生成具有所需姿势的图像。但是，直接级联而不考虑空间布局，并且生成器将来自源图像的像素企图直接放置到正确位置是不正确的。因此，它总是导致图像模糊并丢失源标识。后来，受空间变换器网络（STN）[10]的启发，提出了一种纹理变形方法[1]，如图 2（b）所示。它首先根据源姿势和参考姿势拟合一个粗糙的仿射变换矩阵，使用 STN 将源图像变形为参考姿势，并基于变形图像生成最终结果。但是，就颜色、样式或面部特征而言，纹理变形也无法保留源信息，因为生成器可能会在经过多个下采样操作（例如跨步卷积和合并）后丢弃源信息。同时，较新研究 [4，31]提出将源图像的深层特征变形到目标姿态中，而不是在图像空间中变形，如图 2（c）所示，这称为特征变形。然而，在特征变形中由编码器提取的特征不能保证准确地表征源身份，并因此以不可避免的方式产生模糊或低保真度的图像。

由于以下三个原因，上述现有方法在生成看起来不真实的图像时遇到了挑战：1）对于不同的衣物在质地，样式，颜色和高结构化面部特征等各个方面很难在网络中捕获和保存；2）铰接式和可变形的人体会导致较大的空间布局和过于复杂的姿势操纵的几何变化；3）所有这些方法都不能处理多个源输入，例如在外观转换中，不同部分可能来自不同源输入的人。

在本文中，为了保留源信息，包括衣服和脸部身份的详细信息，我们提出了流动变化块（LWB），以从三个方面解决源信息的丢失：1）去噪卷积自动编码器用于提取有用的特征，以保留源信息，包括纹理，颜色，样式和脸部身份；2）我们提议的 LWB 将每个局部部分的源要素合并到全局要素流中，以进一步保留源详细信息； 3）它支持多源变形，例如在外观转换中，从一个源变形头部特征，从另一个源变形人体特征，以及聚合为全局特征流。这将进一步增强每个源部分的局部身份标识。

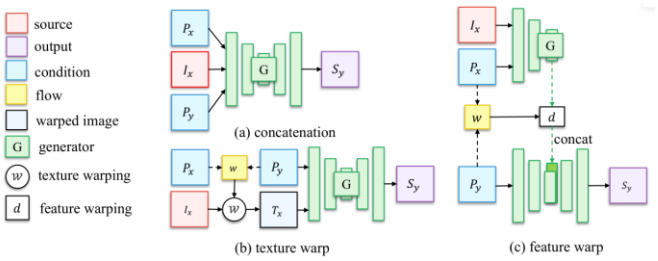


图 2.现有三种将源信息传播到目标条件的方法。（a）是早期连接，它将源图像和源条件以及目标条件连接到颜色通道中。（b）和（c）分别是纹理和特征变换，并且源图像或其特征在适合的转换流下传播到目标条件。

此外，现有方法主要依赖于 2D 姿势[1、19、31]，密集姿势[22]和身体解析[4]。这些方法仅考虑布局位置，却忽略了个性化的形状和肢体（关节）旋转，而这比人图像合成中的布局位置更为重要。例如，在一个极端的情况下，一个高个子男人模仿矮个子的动作，并且使用 2D 骨架，密集的姿势和身体解析条件将不可避免地改变高个子的身高和大小，如图 6 的底部所示。为了克服它们的缺点，我们使用参数统计人体模型 SMPL [2，18，12]将人体分解为姿势（关节旋转）和形状。它输出 3D 网格（不带衣服），而不是关节和零件的布局。此外，通过匹配两个 3D 三角网格之间的对应关系，可以很容易地计算出转换流，与先前从关键点拟合的仿射矩阵相比，该 3D 三角网格的精度更高，并且失准更少[1，31]。

基于 SMPL 模型和流动变形块（LWB），我们的方法可以进一步扩展到其他任务，包括人的外观转移和新颖的视图合成，并且一个模型可以处理这三个任务。我们总结如下：1）我们提出了一个 LWB 来传播和解决图像和特征空间中源信息（例如纹理，样式，颜色和面部识别）的丢失问题；2）通过利用 LWB 和 3D 参数模型的优势，我们的方法是一个适用于模仿人体运动，外观转换和新颖的视图合成的统一框架；3）我们为这些任务（尤其是视频中的人体运动模仿）建立了数据集，并发布了所有代码和数据集，以方便社区中的进一步研究。

2.相关工作

人体动作模仿。最近，大多数方法都是基于条件生成对抗网络（CGAN）[1、3、19、20、22、30]或变分自动编码器[5]。他们的关键技术思想是结合目标

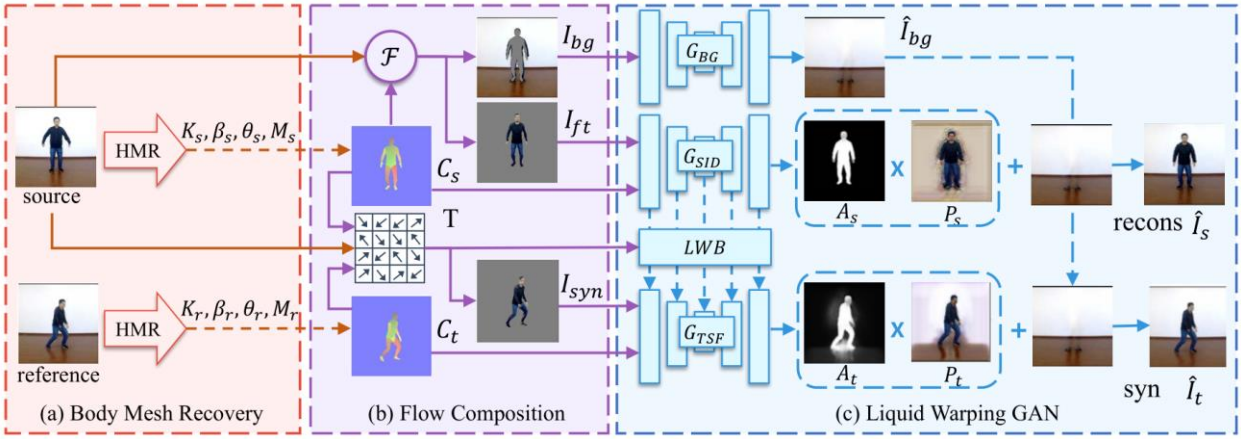


图 3.我们方法的训练流水线。我们从视频中随机采样一对图像，将其中一个称为源图像，命名为I_s，将另一个称为参考图像，命名为I_r。（a）身体网格恢复模块将估计每个图像的 3D 网格，并绘制其对应贴图C_s和C_t；（b）流合成模块将首先基于两个对应图及其在图像空间中的投影顶点来计算变换流 T。然后它将源图像I_s分为前景图像I_{ft}和蒙版背景I_{bg}。最后，它基于变换流 T 变换源图像，并生成变换图像I_{syn}；（c）在最后一个 GAN 模块中，生成器由三个流组成，分别由G_{BG}生成背景图像$\hat{I}_{bg}$，由G_{SID}重建源图像$\hat{I}_s$和在G_{TSF}的参考条件下合成目标图像$\hat{I}_t$。为了保留源图像的细节，我们提出了一种新颖的流动变形块（LWB，如图 4 所示），该块将G_{SID}的源特征传播到G_{TSF}的多层中，并在纹理，样式，颜色和样式方面保留了源信息。

图像以及源姿势（2D 关键点）作为输入，并运用 GAN 使用源姿势生成逼真的图像。这些方法的不同之处仅在于网络体系结构和对抗性损失。在[19]中，设计了一个 U-Net 生成器，并使用了一种从粗到精的策略来生成 256×256 图像。Si 等[1，30]提出了多阶段对抗损失，并分别生成了前景（或不同的身体部位）和背景。Neverova 等[22]用 DensePose [27]将稀疏的 2D 关键点替换为图像和人体表面之间的密集对应关系。Chan 等[3]使用 pix2pixHD [35]框架和专门的 Face GAN 来学习从 2D 骨架到图像的映射，并生成更真实的目标图像。此外，王等[34]将其扩展到视频生成，Liu 等[16]提出了人类演员视频的神经渲染。但是，他们的作品只是训练从 2D 姿势（或部位）到每个人的图像的映射-换句话说，每个人体都需要训练自己的模型。此缺点可能会限制其更广泛的应用。

人体外观转移。从计算机图形方法[24]到基于学习的方法[26，37]，实现人体外观建模或转换是一个广阔的课题，尤其是在虚拟试穿应用程序领域。基于图形的方法首先通过服装和 3d 扫描仪[38]或多相机阵列[15]估计衣服的详细 3D 人体网格，然后基于详细的 3D 网格可以将一个人的衣服呈现在另一个人身上。尽管这些方法可以产生高保真度的结果，但它们的成本，大小和受控制的环境对客户而言并不友好且不便。最近，根据深入的生成模型，SwapNet [26]首先学习一个姿势指导的服装分段合成网络，然后将具有纹理特征的服装解析结果从源图像馈送到编码器-解码器网络，以生成具有所需服装的图像。在[37]中，作者利用结合学习方法的几何 3D 形状模型，交换了三角网格的可见顶点的颜色，并训练了一个模型来推断不可见顶点的颜色。

人体新颖视图合成。新颖的视图合成旨在从任意角度合成同一对象以及人体的新图像。现有方法的核心步骤是通过卷积神经网络拟合从可观察视图到新颖视图的对应图。在[41]中，作者使用基于 CNN 预测外观流动并通过基于外观流动从源图像复制像素来合成同一物体的新图像的方法，并且他们已经获得了像车辆这样的刚性物体的不错的结果。接下来的工作[23]提出了基于外观流和生成对抗网络（GAN）推断不可见纹理的方法[6]，而 Zhu 等人则提出了[42]认为基于外观流的方法在诸如人体之类的关节和可变形物体上的性能较差，他们提出了一种外观形状流动策略来合成人体的新颖方法。此外，Zhao 等[40]设计了一种基于 GAN 的方法，以从粗到精的方式合成高分辨率视图。

3. 方法

我们的流动变形 GAN 包含三个阶段：身体网格恢复，流组成和 GAN 模块，其中 GAN 模块包括

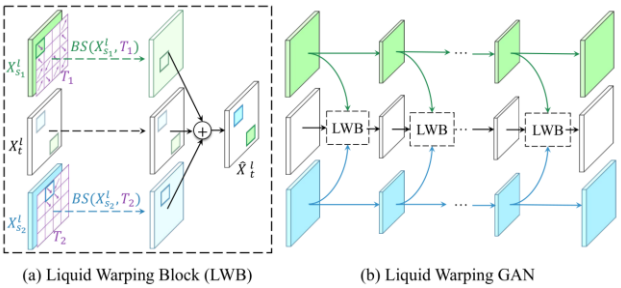
流动变形块（LWB）。对于不同的任务，训练通道是相同的。一旦在一项任务上训练了模型，它就可以处理其他任务。这里，我们以运动模仿为例，如图 3 所示。将源图像标记为I_s，将参考图像标记为I_r。第一身体网格恢复模块将估计I_s和I_r的 3D 网格，并渲染其对应贴图C_s和C_t。接下来，流合成模块将首先基于两个对应图及其在图像空间中的投影网格计算转换流 T。由此，源图像I_s被分解为前景图像I_{ft}和蒙版背景I_{bg}，并且基于变换流 T 而变形为I_{syn}。最后一个 GAN 模块具有一个带有三个流的生成器。它分别通过G_{BG}生成背景图像$\hat{I}_{bg}$，通过G_{SID}重构源图像$\hat{I}_s$，并在G_{TSF}的参考条件下合成图像$\hat{I}_t$。为了保留源图像的细节，我们提出了一种新颖的流动变形块（LWB），它在多层结构中将G_{SID}的源特征传播到G_{TSF}中。

3.1. 身体网格恢复模块

如图 3（a）所示，给定源图像I_s和参考图像I_r，该阶段的作用是预测运动姿势（四肢旋转）和形状参数以及每个图像的 3D 网格。在本文中，我们将 HMR [12]用作 3D 姿势和形状估计器，因为它在精度和效率之间取得了很好的折中。在 HMR 中，图像首先由 ResNet-50 [8]编码为具有$\mathbb{R}^{2048}$的特征，然后是预测 SMPL 的姿态$\theta \in \mathbb{R}^{72}$和形状$\beta \in \mathbb{R}^{10}$的迭代 3D 回归网络[18]，同时也包括弱视角摄像机$K \in \mathbb{R}^3$。 SMPL 是 3D 人体模型，可以定义为微分函数$M(\theta,\beta) \in \mathbb{R}^{N_v \times 3}$，它按照$N_v = 6,890$来设置三角网格的顶点参数，以及$N_f = 13,776$张脸，其具有姿态参数$\theta \in \mathbb{R}^{72}$和$\beta \in \mathbb{R}^{10}$。在此，形状参数β是从数千次已记录的扫描中获悉的低维形状空间的系数，而姿势参数θ是通过正向运动学使骨骼运动的关节旋转得到的。通过这样的过程，我们将获得源图像的身体重建参数信息$\{K_s,\theta_s,\beta_s,M_s\}$和参考图像的信息$\{K_r,\theta_r,\beta_r,M_r\}$。

3.2. 流组成模块

基于先前的估计，我们首先在相机视图K_s下绘制源网格M_s和参考网格M_r的对应图。在此，我们分别将源对应映射和参考对应映射表示为C_s和C_t。在本文中，我们使用了完全可区分的渲染器-神经网格渲染器（NMR）[13]。因此，我们通过弱透视相机将源V_s的顶点投影到 2D 图像空间中，$v_s = Proj(V_s,K_s)$。然后，我们计算每个网格面的重心坐标，并且



(a) Liquid Warping Block (LWB) (b) Liquid Warping GAN

图 4.流动变形块的图示。 （a）是 LWB 的结构。$X^l_{s_1}$和$X^l_{s_2}$是由第l^{th}层中不同源的G_{SID}提取的特征图。X^l_t是G_{TSF}在第l^{th}层的的特征图。 最终输出特征$\widehat{X}^l_t$将G_{TSF}的特征与双线性采样器（BS）相对于流T_1和T_2的变换源特征进行了总和。（b）是 LWB 的体系结构。

得到$f_s \in \mathbb{R}^{N_f \times 2}$。接下来，我们通过将源对应关系图与其网格面坐标$f_s$和参考对应关系图之间的对应关系进行匹配，来计算变换流$T \in \mathbb{R}^{H \times W \times 2}$，H×W 是图像的大小。因此，通过基于$C_s$对源图像$I_s$进行掩模来导出前景图像$I_{ft}$和蒙版背景$I_{bg}$。最后，我们通过变换流 T 变换源图像$I_s$，并获得变换图像$I_{syn}$，如图 3 所示。

3.3. 流动变形 GAN

该阶段在所需条件下合成高保真人像。更具体地说，它是 1）合成背景图像的大小； 2）根据可见部分预测不可见部分的颜色； 3）根据 SMPL 的重建生成衣服，头发和其他物体的像素。

生成器。我们的生成器以 3-流方式工作。如图 3（c）的顶部流中所示，一个名为G_{BG}的流将蒙版背景I_{bg}和蒙版（通过在彩色通道中对C_s进行二值化获得的蒙版（总共 4 个通道））连接起来，以生成真实的背景图像I_{bg}。其他两个流是源身份流，即G_{SID}和传输流，即G_{TSF}。 G_{SID}是一种去噪卷积自动编码器，旨在指导编码器提取能够保留源信息的特征。它与I_{bg}一起，将被掩盖的源前景I_{ft}和对应图C_s（总共 6 个通道）作为输入，并重建源正面图像I_s。G_{TSF}流合成最终结果，该结果通过双线性采样器接收变换的前景和对应图C_t（总共 6 个通道）作为输入。为了保留源信息，例如纹理，样式和颜色，我们提出了一种新颖的流动变形块（LWB），它将源与目标流链接在一起。它融合了G_{SID}的源特征并将其融合到传输流G_{TSF}中，如图 3（c）的底部所示。

我们提议的流动变形块（LWB）的优势之一

是针对多种来源的，例如在人的外观转换中，保留来源一的头部，并从来源二穿上外套，同时从来源三脱下外套。功能的不同部分通过自己的转换流独立地汇总到G_{TSF}中。 这里，我们以两个源为例，如图 4 所示。将$X^l_{s_1}$和$X^l_{s_2}$表示为第l^{th}层中不同源的G_{SID}提取的特征图。 X^l_t是G_{TSF}第l^{th}层的特征图。源要素的每个部分都由它们自己的转换流变形，并汇总到G_{TSF}的要素中。 我们使用双线性采样器（BS）分别针对转换流T_1和T_2变换源特征$X^l_{s_1}$和$X^l_{s_2}$。 最终的输出函数如下：

$$\widehat{X}^l_t = BS(X^l_{s_1}, T_1) + BS(X^l_{s_2}, T_2) + X^l_t$$

请注意，我们仅以两个来源为例，可以轻松地将其扩展到多个来源。

G_{BG}，G_{SID}和G_{TSF}具有类似的架构，称为 ResUnet，它是 ResNet [7]和 U-Net [28]的组合，但没有共享参数。对于G_{BG}，我们直接回归最终的背景图像，而对于G_{SID}和G_{TSF}，我们具体生成注意力图 A 和颜色图 P，如图 3（c）所示。 最终图像可以通过以下方式获得：

$$\hat{I}_s = P_s * A_s + \hat{I}_{bg} * (1 - A_s)$$
$$\hat{I}_t = P_t * A_t + \hat{I}_{bg} * (1 - A_t).$$

判别器。为了区分，我们遵循 Pix2Pix[9]的体系结构。补充材料中提供了有关我们的网络体系结构的更多详细信息。

3.4. 训练细节和损失函数

在这一部分中，我们将介绍损失函数以及如何训练整个系统。对于身体恢复模块，我们遵循 HMR 的网络架构和损失函数[12]。在这里，我们使用 HMR 的预训练模型。

对于流动变形 GAN，在训练阶段，我们从每个视频中随机采样一对图像，并将其中一个设置为源I_s，将另一个设置为参考I_r。请注意，我们提出的方法是用于模仿动作，外观转移和新颖视图合成的统一框架。因此，模型经过训练后，便可以应用于其他任务，而无需从头开始训练。在我们的实验中，我们训练了一个运动模仿模型，然后将其应用于其他任务，包括外观传递和新颖的视图合成。

整个损失函数包含四个项，分别是感知损失[11]，面部识别损失，注意力正则损失和对抗损失。

感知损失。它将重构的源图像$\hat{I}_s$和生成的目标图像$\hat{I}_t$正则化为更接近在 VGG [32]子空间中的完全真实值I_s和I_r。 其公式如下：

$$L_p = \|f(\hat{I}_s) - f(I_s)\|_1 + \|f(\hat{I}_t) - f(I_r)\|_1$$

在此，f 是预训练的 VGG-19 [32]。

面部识别损失。它根据合成的目标图像对裁剪后的面部进行正则化，使其与完全真实情况I_r的图像相似，从而推动生成器保留面部身份。显示如下：

$$L_f = \|g(\hat{I}_t) - g(I_r)\|_1$$

在这里，g 是预训练的 SphereFaceNet [17]。

对抗损失。它将合成大小图像的分布推向真实图像的分布。 如下所示，我们将$LSGAN_{-110}$ [21]损失以类似 PatchGAN 的方式用于生成的目标图像$\hat{I}_t$。 判别器 D 将其规范化，使其看起来更逼真。我们使用条件判别器，它将生成的图像和对应图C_t（6 个通道）作为输入。

$$L^G_{adv} = \sum D(\hat{I}_t, C_t)^2$$

注意力正则损失。它将注意力图 A 规范化为平滑的并防止它们饱和。考虑到没有注意力图 A 和颜色图 P 的完全真实图，它们是从上述损失的所得梯度中获悉的。但是，注意力遮罩很容易饱和到 1，这会阻止生成器工作。为了缓解这种情况，我们将蒙版调整为更接近 3D 身体网格渲染的轮廓 S。由于轮廓是粗糙的图像，并且包含不带衣服和头发的人体遮罩，因此我们还对[A]进行了总变差（TV）正则化，例如[25]，以弥补遮罩的缺点，并进一步增强平滑的空间色彩当组合来自预测背景$\hat{I}_{bg}$的像素和颜色图 P 时。 显示如下：

$$L_a = \|A_s - S_s\|_2^2 + \|A_t - S_t\|_2^2 + TV(A_s) + TV(A_t)$$
$$TV(A) = \sum_{i,j} [A(i,j) - A(i-1,j)]^2 + [A(i,j) - A(i,j-1)]^2$$

对于生成器，完整目标函数如下所示，其中λ_p，λ_f和λ_a分别是感知损失，面部识别损失和注意力损失的权重。

$$L^G = \lambda_p L_p + \lambda_f L_f + \lambda_a L_a + L^G_{adv}$$

对于判别器，完整的目标函数是

$$L^D = \sum [D(\hat{I}_t, C_t) + 1]^2 + \sum [D(I_r, C_t) - 1]^2$$

3.5. 推断

一旦对运动模仿任务进行了训练，就可以将其应用于推理的其他任务。对于各种任务的不同条件，差异在于转换流的计算。对于其余模块，如“身体网格恢复”和“流动变形 GAN”模块都相同。以下是测试阶段流程组成模块的每个任务的详细示意图。

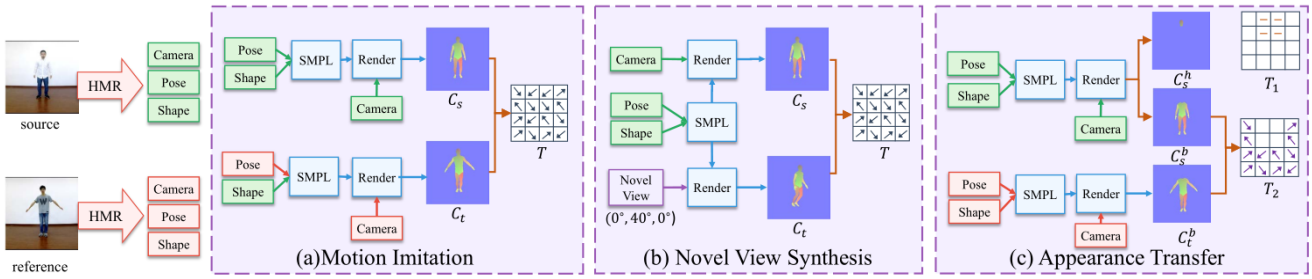


图 5.在测试阶段计算不同任务的转换流程的示意图。 左边是源和参考图像的“身体恢复”模块解开的身体参数。 右边是在不同任务中计算转换流的不同实现方式。

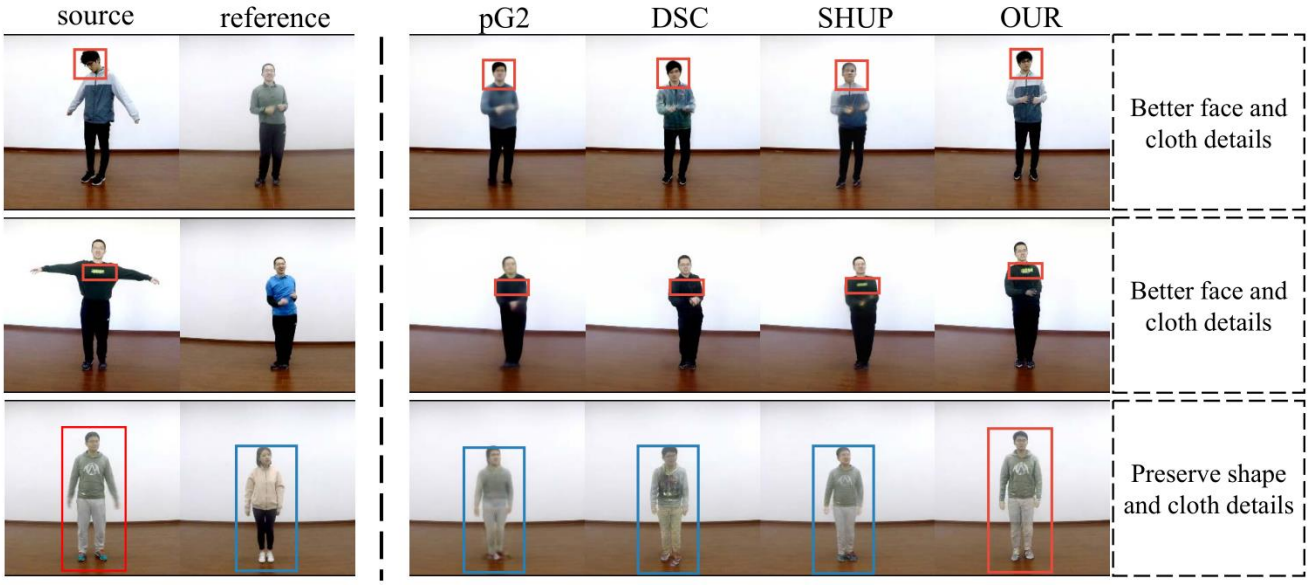


图 6.在 iPER 数据集上我们的方法与其他运动模仿方法的比较（放大以获得最佳视图）。 2D 姿势指导方法 pG2 [19]，DSC [31]和 SHUP [1]无法保留衣服细节，脸部身份和源图像的形状一致性。 我们用红色和蓝色矩形突出显示细节。

运动模仿。我们首先将参考的位姿参数值θ_r复制到源中，获得 SMPL 的综合参数以及 3D 网格$M_t = M(\theta_r, \beta_s)$。接下来，我们在相机视角$K_s$下绘制源网格$M_s$和合成网格$M_t$的对应图。在此，我们分别将源映射和合成对应图分别表示为C_s和C_t。然后，通过弱透视相机将源的顶点投影到 2D 图像空间中，$v_s = Proj(V_s, K_s)$。接下来，我们计算每个网格面的重心坐标，并具有$f_s \in \mathbb{R}^{N_f \times 2}$。最后，通过将源对应图与其网格面坐标$f_s$和的对应关系进行匹配，来计算变换流$T \in \mathbb{R}^{H \times W \times 2}$和生成对应图。该过程在图 5（a）中示出。

新颖的视图生成。在旋转 R 和平移 t 的情况下，给定新的摄影机视图。我们首先在新颖的视图下计算 3D 网格$M_t = M_s R + t$。流操作类似于模仿动作。我们提供源网格M_s和在弱视角相机K_s下建立新的网格M_t的关联图，并最终计算变换流量$T \in \mathbb{R}^{H \times W \times 2}$。这在图 5（b）中示出。

外观转移。它需要从参考图像中“复制”躯干或身体的衣服，同时保持源头（面部，眼睛，头发等）的身份。我们将转换流 T 分为两个子转换流，即源流T_1和参考流T_2。将头部网格表示为$M^h = (V^h, F^h)$，将身体网格表示为$M^b = (V^b, F^b)$。在此，$M = M^h \cup M^b$。对于T_1，我们首先将源头的网格M_s^h投影到图像空间中，从而获得剪影，S_s^h。然后，我们创建一个网格$G \in \mathbb{R}^{H \times W \times 2}$。然后，用$S^h$遮罩 G，并得出$T_1 = G * S^h$。在此，*表示逐元素乘法。对于$T_2$，它类似于运动模仿。我们绘制源主体$M_s^b$和参考主体$M_t^b$的对应图，分别表示为$C_s^b$和$C_t^b$。最后，我们根据$C_s^b$和$C_t^b$之间的对应关系计算转换流$T_2$。我们在图 5（c）中进行说明。

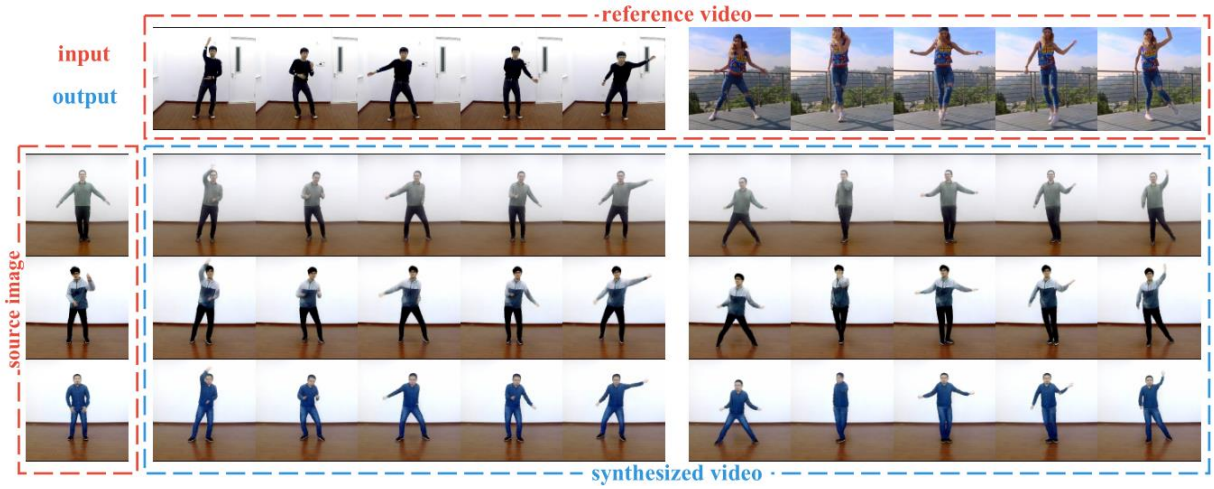


图 7. 在 iPER 数据集上从我们提出的方法中模仿运动的示例（放大以获得最佳视图）。我们的方法可以产生高保真度的图像，这些图像可以保留源的脸部身份，形状一致性和衣服细节，即使源图像（例如中排和底排）中也存在遮挡。 我们建议访问补充材料以获取更多视频结果。

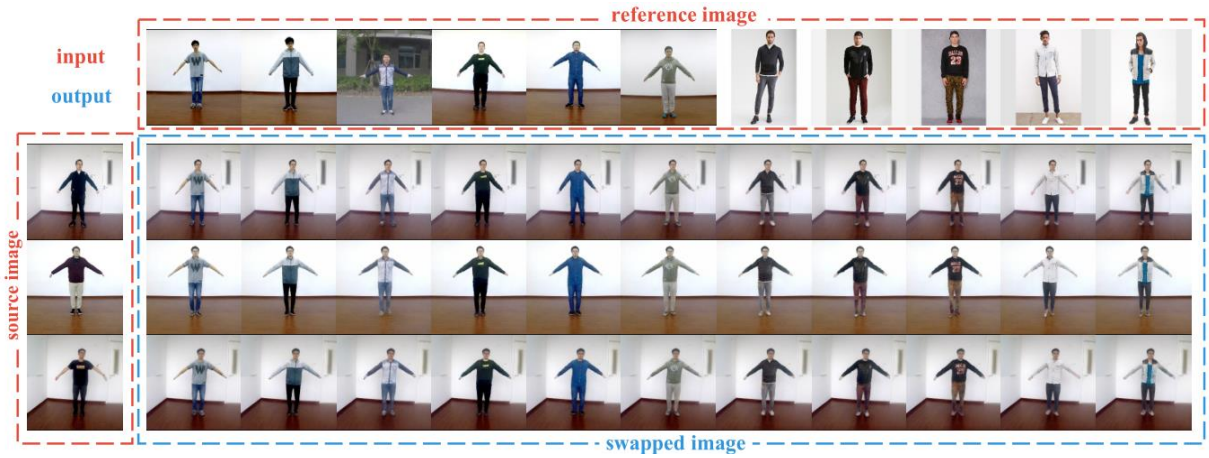


图 8. 在 iPER 测试集中的人类外观转移方法示例（放大以获得最佳视图）。 我们的方法可以生成高保真，体面的图像，该图像可以保留源图像的脸部身份和形状一致性，并保留参考图像的衣服细节。 我们建议您访问补充材料以获得更多结果。

4. 实验

数据集。为了评估我们提出的运动模仿，外观转移和新颖的视图合成方法的性能，我们建立了一个具有多种样式的衣服的新数据集，称为 Impersonator（iPER）数据集。 有 30 个不同形状，高度和性别条件的科目。每个对象穿着不同的衣服，并执行 A 姿势视频和随机动作的视频。一些对象可能会穿多件衣服，总共有 103 件衣服。整个数据集包含 241,564 帧的 206 个视频序列。 我们根据衣服的不同将其分为 8：2 的训练/测试集。

实施细节。为了训练网络，将所有图像标准化为[-1，1]，然后将其大小调整为 256×256。我们从每个视频中随机采样一对图像。在我们的实验中，mini-batch 大小为 4。λ_p，λ_f和λ_a分别设置为 10.0、5.0 和 1.0。 Adam[14]被用于生成器和判别器的参数优化。

4.1. 人体动作模仿的评估.

评估指标。我们提出了 iPER 数据集测试集的评估协议，它能够从不同方面指示不同方法的性能。详细信息如下：1）在每个视频中，我们选择三幅图像作为具有不同遮挡度的源图像（正面，侧面和遮挡）。正面图像包含的信息最多，而侧面图像会丢失一些信息，而封闭的图像会引起歧义。2）对于每个源图像，我们执行自我模仿，使演员模仿自己的动作。 SSIM [36]和学习知觉相似性（LPIPS）[39]是自我模仿设置中的评估指标。 3）此外，我们还进行交叉模仿，模仿演员模仿他人的行为。我们在称为 FReID 的预先训练的人员重识别模型[33]上使用初始分数（IS）[29]和 Frechet´距离来评估生成图像的质量。

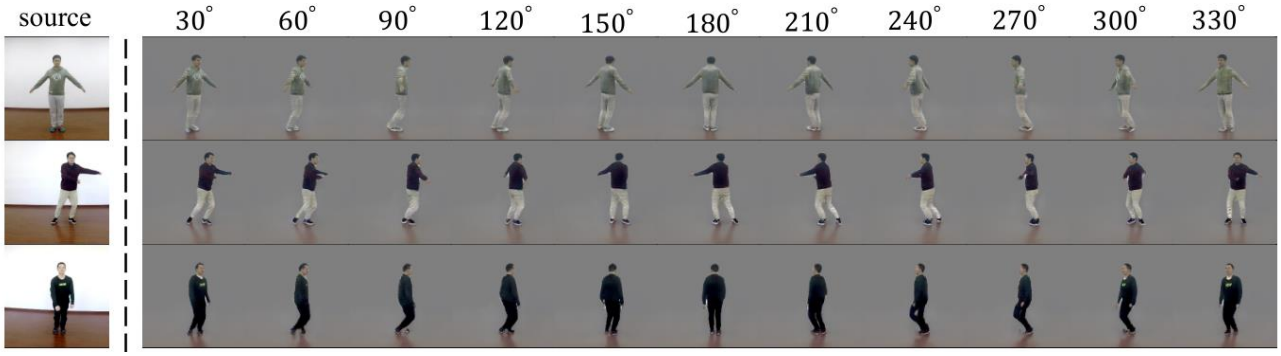


图 9.从我们的方法在 iPER 数据集上合成新颖视图的示例（放大以获得最佳视图）。我们的方法可以在不同的摄影机视角下生成逼真的结果，并且即使在自我遮挡的情况下（例如中间和底部），也能够保留源信息。

表 1. iPER 数据集上通过不同方法进行的运动模仿的结果。↑表示越大越好，而↓表示越小越好。SSIM 越高可能并不意味着图像质量更好[39]。

	Self-Imitation		Cross-Imitation	
	SSIM↑	LPIPS↑	IS↑	FReID↓
PG2 [19]	**0.854**	0.865	3.242	0.353
SHUP [1]	0.832	0.901	3.371	0.324
DSC [31]	0.829	0.871	3.321	0.342
W_C	0.821	0.872	3.213	0.341
W_T	0.822	0.887	3.353	0.347
W_F	0.830	0.897	3.358	0.325
Ours-W_{LWB}	0.840	**0.913**	**3.419**	**0.317**

与其他方法的比较。我们将我们的方法的性能与现有方法的性能进行比较，包括 PG2 [19]，SHUP [1]和 DSC [31]。我们在 iPER 数据集上训练所有这些方法，并将上述评估协议应用于这些方法。结果记录在表 1 中。可以看出我们的方法优于其他方法。此外，我们还分析了生成的图像，并与我们上面的方法进行了比较。从图 6 中，我们发现 1）二维姿态引导方法，包括 PG2 [19]，SHUP [1]和 DSC [31]，会改变源的体形。例如，在图 6 的第三行中，一个高个子模仿矮个子的运动，这些方法改变了源身的高度。但是，我们的方法能够保持身体形状不变。 2）当源图像表现出自我遮挡时，例如图 6 第 1 行中的不可见脸部，我们的方法可以生成模棱两可和不可见部分的更具真实感的内容。3）如图 6 的第二行和第三行所示，与其他方法相比，我们的方法在保留源身份（例如源的脸部身份和衣服细节）方面更加强大。4）我们的方法还具有很高的优势。交叉模仿设置中的逼真度图像（模仿他人的动作），我们在图 7 中进行了说明。

消融研究。为了验证我们提议的流动变形块（LWB）的影响，我们使用上述方法设计了三个基准来传播源信息，包括早期串联，纹理变换和特征变换。所有模块和损失函数都满足除了我们的方法和其他基准之间的传播策略不同外，其他均相同。在这里，我们表示早期串联，纹理变形，特征变形，以及我们建议的 LWB 分别为W_C，W_T，W_F和W_{LWB}。我们在 iPER 数据集上的相同设置下训练所有这些，然后评估它们在运动模仿上的性能。从表 1 中可以看出，我们提出的 LWB 优于其他基准。补充材料中提供了更多详细信息。

4.2. 人体外观转移的结果。

值得强调的是，一旦模型经过训练，就可以直接应用于三个任务，包括运动模仿，外观转移和新颖的视图合成。我们随机选择图 8 中显示的一些示例。通过我们的方法，可以很好地保留面部特征和衣服细节（在纹理，颜色和样式方面）。它证明了我们的方法即使在参考图像来自网络且不在 iPER 数据集范围之内的情况下，也可以在外观转换中取得不错的效果，例如图 8 中的最后五列。

4.3. 人体新颖视图生成的结果。

我们从 iPER 测试集中随机抽取源图像，并将视角从$30°$更改为 $330°$。结果如图 9 所示。当切换到其他视角时，我们的方法能够预测不可见部分的合理含量，并且即使在自我遮挡的情况下，也可以根据面部特征和衣服细节保留源信息，例如图 9 中的中排和下排。

5. 结论

我们提出了一个统一的框架来处理人类运动模仿，外观转换和新颖的视图合成。它采用身体恢复模块来估算 3D 身体网格，该网格比 2D 姿势更强大。此外，为了保留源信息，我们设计了一种新颖的变形策略，即流动变形块（LWB），它可以在图像和特征空间中传播源信息，并支持来自多个源的更灵活的变形。大量的实验表明，我们的框架胜过其他框架并产生了不错的结果。

参考文献

[1] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frdo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In European Confer-ence on Computer Vision, pages 561–578. Springer, 2016.

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. arXiv preprint arXiv:1808.07371, 2018.

[4] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In Advances in Neural Information Processing Systems 31: An-nual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montreal,´ Canada., pages 472–482, 2018.

[5] Patrick Esser, Ekaterina Sutter, and Bjorn¨ Ommer. A variational u-net for conditional appearance and shape generation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 8857–8866, 2018.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 2672–2680. Curran Associates, Inc., 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pat-tern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778, 2016.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Computer Vision - ECCV 2016 - 14th European Con-ference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV, pages 630–645, 2016.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with con-ditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5967–5976, 2017.

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer net-works. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Informa-tion Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 2017–2025, 2015.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Per-ceptual losses for real-time style transfer and super-resolution. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II, pages 694–711, 2016.

[12] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[13] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In 2018 IEEE Con-ference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 3907–3916, 2018.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Interna-tional Conference on Learning Representations, vol-ume abs/1412.6980, 2015.

[15] Vincent Leroy, Jean-Sebastien´ Franco, and Edmond Boyer. Multi-view dynamic shape refinement using local temporal integration. In IEEE International Con-ference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 3113–3122, 2003.

[16] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural ren-dering and reenactment of human actor videos. ACM Transactions on Graphics 2019 (TOG), 2019.

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hyper-sphere embedding for face recognition. In 2017 IEEE Conference on Computer Vision and Pattern Recog-nition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 6738–6746, 2017.

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia), 34(6):248:1– 248:16, oct 2015.

[19] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In Advances in Neural Information Processing Systems, pages 405–415, 2017.

[20] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentan-gled person image generation. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[21] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. On the effectiveness of least squares generative adversarial networks. CoRR, abs/1712.06391, 2017.

[22] Natalia Neverova, Rıza Alp Guler,¨ and Iasonas Kokki-nos. Dense pose transfer. In European Conference on Computer Vision (ECCV), 2018.

[23] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In The IEEE Conference on Computer Vi-sion and Pattern Recognition (CVPR), July 2017.

[24] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. Clothcap: seamless 4d clothing cap-ture and retargeting. ACM Trans. Graph., 36(4):73:1– 73:15, 2017.

[25] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Gan-imation: Anatomically-aware facial animation from a single image. In Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X, pages 835–851, 2018.

[26] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII, pages 679–695, 2018.

[27] Iasonas Kokkinos Rıza Alp Guler,¨ Natalia Neverova. Densepose: Dense human pose estimation in the wild. In The IEEE Conference on Computer Vision and Pat-tern Recognition (CVPR), 2018.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical im-age segmentation. In Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III, pages 234–241, 2015.

[29] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in Neural Information Processing Systems 29: Annual Confer-ence on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 2226– 2234, 2016.

[30] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. Multistage adversarial losses for pose-based human image synthesis. In The IEEE Conference on Com-puter Vision and Pattern Recognition (CVPR), June 2018.

[31] Aliaksandr Siarohin, Enver Sangineto, Stphane Lath-uilire, and Nicu Sebe. Deformable gans for pose-based human image generation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recog-nition. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 2015.

[33] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In Computer Vision - ECCV 2018 - 15th Eu-ropean Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV, pages 501–518, 2018.

[34] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Niko-lai Yakovenko, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In Advances in Neural Information Processing Systems 31: An-nual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montreal,´ Canada., pages 1152–1164, 20148.

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, An-drew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipula-tion with conditional gans. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[36] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing, 13(4):600–612, 2004.

[37] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In The IEEE Conference on Computer Vision and Pat-tern Recognition (CVPR), June 2018.

[38] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 5484–5493, 2017.

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable ef-fectiveness of deep features as a perceptual metric. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[40] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. In 2018 ACM Multimedia Con-ference on Multimedia Conference, MM 2018, Seoul,

Republic of Korea, October 22-26, 2018, pages 383–391, 2018.

[41] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jiten-dra Malik, and Alexei A. Efros. View synthesis by appearance flow. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Nether-lands, October 11-14, 2016, Proceedings, Part IV, pages 286–301, 2016.

[42] Hao Zhu, Hao Su, Peng Wang, Xun Cao, and Ruigang Yang. View extrapolation of human body from a sin-gle image. In The IEEE Conference on Computer Vi-sion and Pattern Recognition (CVPR), June 2018.