基于人体姿势的图像合成论文报告

董合德
21821187
21821187@zju.edu.cn

摘要（Abstract）

基于人体姿势的图像合成[39]是指给定一个人的图像及其期望姿势骨架图，合成出满足期望姿势的图像（即除了与原图姿势不同以外保持人和背景的外观）。该论文提出了一个模块化的生成神经网络，它使用人类动作视频中的训练对图像和姿势来合成看不见的姿势。该网络将场景分为不同的身体部位和背景层，将身体部位移动到新的位置并重新定义它们的外观，并将新的前景与空洞填充的背景相结合。并使用单个目标图像作为监督标签来联合训练这些子任务的模块。

1. 引言

该论文的主要工作是给定一个人的图像以及目标姿势，我们自动合成一个逼真的图像，描绘人物在该姿势中的样子。我们在变换中保留了人物和背景的外观，如图 1 所示。

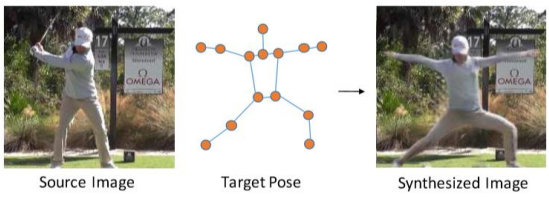

图 1：基于人体姿势的图像合成示例图

对于该图像合成任务，需要保留人物和背景的外观，并捕捉与新姿势一致的身体部位细节。姿势的差异可能导致图像空间的复杂变化。阴影和边缘等微妙的细节应该在感知上与身体的动作变化一致。

该论文通过训练以图像及其姿势为输入的监督学习模型来解决这些问题。主要思想是将这个复杂的问题分解为更简单的模块化子任务，作为一个生成神经网络共同训练。该网络首先将源图像分成背景层和与不同身体部分相对应的多个前景层，使其能够将身体部位空间移动到目标位置。然后对移动的身体部位进行修改和融合以合成新的前景图像，同时背景分别填充适当的纹理以解决由于解除紊乱而导致的间隙。最后，网络合成前景和背景以产生输出图像。所有这些操作都作为一个网络联合执行，并且仅使用目标图像作为监督标签一起训练。

2. 相关工作

图像合成是计算机视觉的任务之一，即给定之前的图像情况合成新的图像。大多数的图像合成工作都集中在简单的固定对象，例如车和家具[8,10,16,20,30,34]。最近的研究开始考虑合成人类所想象不到的场景[11,13,32]。这些方法使用编码器神经网络来捕获输入图像和期望变换之间的复杂关系，以及使用解码器来合成输出图像。相比之下，我们将场景表示为可单独操作的层，允许我们将任务构建为更简单的模块化子问题的组合。

计算机视觉中的许多相关问题可以作为图像翻译的实例，或者将场景的一个表示转换为另一个。例如场景分割[18,21]，表面法线预测[2]，色彩[31]，样式转移[12,23]，边缘检测[29]和草图反转[5]。在这些任务中，像素被修改而不是从输入图像移动到输出图像。最近的一项研究表明，U-Net（带跳过连接的编码器 - 解码器神经网络）能够处理各种各样的翻译任务[6]。该论文在该问题上结合了U-Net架构用于多个图像转换子任务，例如图像分割和空洞填充。

3. 基于人体姿势的图像合成

与之前的工作不同的是，该论文采用了模块化的框架，即将复杂的操作分解为若干个子任务，利用不同的模块去训练学习这些子任务，最后将其结合。

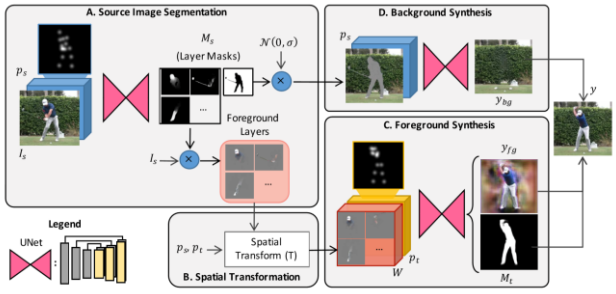

图 2：网络结构示例图

如图 2,整个模型分为 4 个部分（1）图片背景分割（Source Image Segmentation）（2）空间转换（Spatial Transformation）（3）前景合成（Foreground Synthesis）（4）背景合成（Background Synthesis）。整体的任务难点是在（1）前景合成：人物合成的细节（边缘、阴影）（2）背景修补：因人物动作改变，填补空缺的部分。

3.1. 模型

该论文提出了一种神经网络模型，它可以学习如何对图像空间进行变换。该模型的输入是一个形（example，label）的元组 $((I_s, p_s, p_t), I_t)$ ，其中 I_s, p_s, p_t 分别表示源图像，源图像 2D姿势和目标 2D姿势， I_t 表示目标图像（如图 3）。另外假设源图像和目标图像所描绘的人的穿着和背景相同。

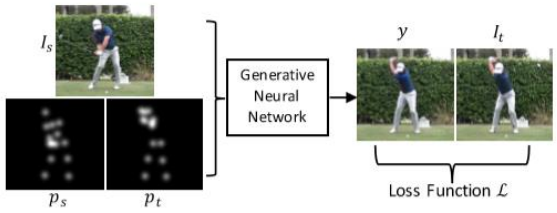

图 3：模型输入输出示例图

3.2. 姿势表征

和之前表征姿势的工作一样[15,28],该论文将 2D 图像中的 p_s, p_t 表示 3D图像中的体积 $R^{H\times W\times J}$ ，其中 H, W 表示输入图像的高度和宽度，并且每个 J 通道在不同的关节的位置加入高斯噪声，有利于网络更快学习到这个特征。在该论文中 J=14，分别表示头、脖子、肩膀、手肘、手腕、臀部、膝盖和脚踝这些位置。

3.3. 图片背景分割

该模块是将图片的前景和背景进行分割，并且将前景分层 $L(L=10)$ 个部分，即将身体分成 10 个部位：头部、上臂、下臂、上腿、下腿和躯干。身体部位与关节不同，前 9 个部分由 2 个关节组成，躯干包含 4 个关节，具体如图 4。

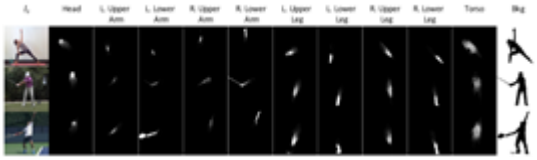

图 4：前景分割示例图

故该模块根据源图及其姿势，利用U-Net学习每个姿势属于图片中的哪个部分，最终输出身体姿势中的L个位置和 1 个背景的Mask（如图 4）。根据这个Mask与原本的图片相乘，就可以得到前景的各部分（Foregrounds Layers）。

3.4. 空间转换

该部分是通过简单的转换，将Foregrounds Layers做位移、旋转、缩放等，得到 W ，主要目的是将分割后的部分与目标姿势一一对应。

3.5. 前景合成

利用 3.4 空间转换得到了目标姿势对应L个位置部分，然后对其进行组合。将 $[W, p_t]$ 作为U-Net的输入，以目标前景及其Mask M_t 作为输出，进行训练。

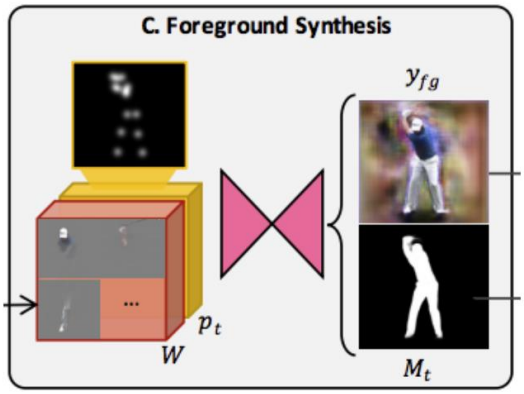

图 5 前景合成结构示例图

3.6. 背景合成

该部分将图片分割中的 M_s 的最后一层背景取出，并将背景层填补孔洞后与 I_s 源图像进行结合得到

I_{st} ，并与 p_s 源姿势作为U-Net的输入，从而训练得到背景图 y_{bg} 。

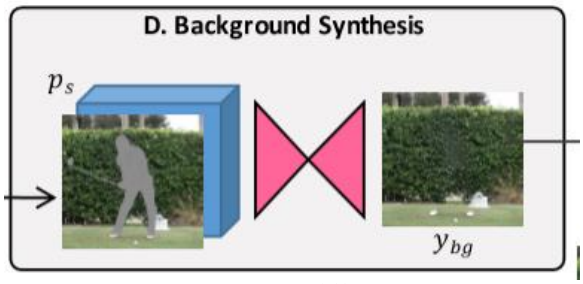

图 6 背景合成结构示例图

最终将前景合成和背景合成的输出进行结合，得到最终的指定人体姿势的目标图。

$$y = M_t \otimes y_{fg} + (1 - M_t) \otimes y_{bg}$$

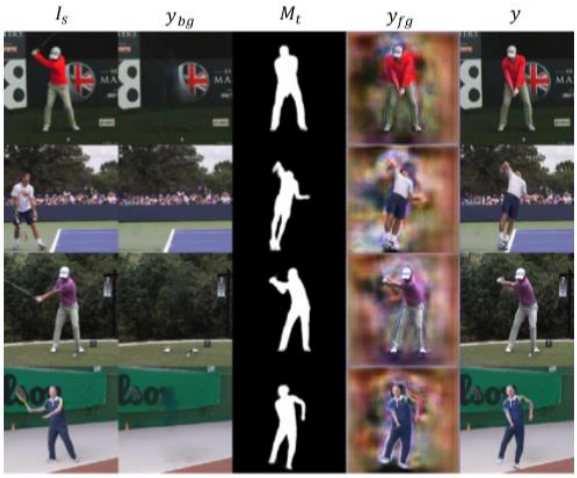

图 7 前景合成与背景合成模块输出及最终结合示例图

3.7. 损失函数

以 $G: y = G(I_s, p_s, p_t)$ 表征该模型，该模型的损失函数分为两个部分一个是VGG LOSS，将VGG19的前 16 层输出连接并计算L1 距离，以及传统的GAN LOSS。

$$L_{VGG+GAN}(G,D) = L_{VGG}(G) + \lambda L_{GAN}(G,D)$$

$$L_{GAN}(G,D) = E_{I_s, I_t, p_s, p_t}[\log D(I_t, p_t) + \log(1 - D(y, p_t))]$$

4. 实验结果与分析

该论文从YouTube收集的操作人员的视频来评估他们的方法。每个训练示例是同视频的一对图像及其相应的姿势。所选择的视频主要是静态背景。通过使用来自同一视频的成对图像，保证其图中人的外观和背景不变，同时允许其姿势可以改变。共收集了三类视频：高尔夫挥杆、瑜伽/运动训练和网球动作。数据集大小分别为 136, 60 和 70 个视频。我们将所有动作类组合成一个数据集。我们将随机数据增强应用于每个示例：缩放，平移，旋转，水平移动和图像饱和度。我们随机抽出了 10%的视频用于测试，保证其姿势不会出现在训练集和验证集。

表 1 Errors and SSIM score

Model	L1 Error	VGG Error	SSIM Score
UNet	0.038(0.018)	0.215(0.091)	0.847(0.103)
Ours	0.034(0.018)	0.200(0.092)	0.863(0.105)

4.1. 结果对比

图 8 展示了与U-Net在合成图片（同类别但不同姿势）的效果，从图 8 可知该模型在一些细节处理的比较好例如服装和阴影。

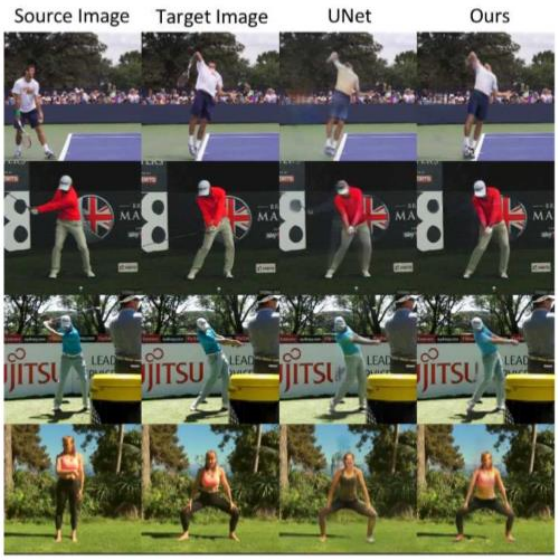

图 8 与U-Net效果对比图

5. 结论

该论文提出了一种将姿势合成的任务拆分成多个子任务的思想，即使用模块化的神经网络来合成新姿势的人类图像。该模型分层执行合成，将前景与背景和不同身体部位相互分离。它将身体部位移动到目标位置，这样可以捕捉大的姿势变化，同时保持正确的前景外观。 通过将前景与背景分离，它还能够比典型的UNet架构合成更逼真的背景。

参考文献

[1] J. Andreas et al. Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705, 2016. 2.

[2] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) , pages 2650–2658, 2015. 2

[3] L. A. Gatys et al. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. arXiv preprint arXiv:1505.07376, 12, 2015. 5

[4] I. Goodfellow et al. Generative adversarial nets. In Advances in neural information processing systems (NIPS), pages 2672–2680, 2014. 2

[5] Y. G¨uc¸l¨ut¨urk et al. Convolutional sketch inversion. In European Conference on Computer Vision (ECCV), pages 810– 824. Springer, 2016. 2

[6] P. Isola et al. Image-to-image translation with conditional adversarial networks. arXiv preprint , 2017. 2, 3

[7] M. Jaderberg et al. Spatial transformer networks. In Advances in neural information processing systems (NIPS), pages 2017–2025, 2015. 4

[8] D. Ji et al. Deep view morphing. Computer Vision and Pattern Recognition (CVPR) , 2017. 2

[9] J. Johnson et al. Inferring and executing programs for visual reasoning. arXiv preprint arXiv:1705.03633, 2017. 2

[10] T. D. Kulkarni et al. Deep convolutional inverse graphics network. In Advances in Neural Information Processing Systems (NIPS), pages 2539–2547, 2015. 2

[11] C. Lassner et al. A generative model of people in clothing. arXiv preprint arXiv:1705.04098, 2017. 2, 6

[12] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In European Conference on Computer Vision (ECCV), pages 702–716. Springer, 2016. 2

[13] L. Ma et al. Pose guided person image generation. arXiv preprint arXiv:1705.09368, 2017. 2, 6

[14] M. Mathieu et al. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440, 2015. 2

[15] A. Newell et al. Stacked hourglass networks for human pose estimation. In European Conference on Computer Vision (ECCV), pages 483–499. Springer, 2016. 3, 6

[16] [16] E. Park et al. Transformation-grounded image generation network for novel 3d view synthesis. arXiv preprint arXiv:1703.02921, 2017. 2

[17] D. Pathak et al. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pages 2536–2544, 2016. 2

[18] T. M. Quan et al. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. arXiv preprint arXiv:1612.05360, 2016. 2

[19] A. Radford et al. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015. 2

[20] [K. Rematas et al. Novel views of objects from a single image. TPAMI , 2016. 2

[21] O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (MICCAI) , pages 234–241. Springer, 2015. 2, 3

[22] T. Salimans et al. Improved techniques for training gans. In Advances in Neural Information Processing Systems (NIPS), pages 2234–2242, 2016. 2

[23] Y. Shih et al. Data-driven hallucination of different times of day from a single outdoor photo. ACM Trans. Graph. , 32(6):200:1 –200:11, 2013. 2

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 5

[25] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. IEEE Trans. Image Processing, 3(5):625 – 638, 1994. 3

[26] X. Wang and A. Gupta. Generative image modeling us- ing style and structure adversarial networks. In European Conference on Computer Vision (ECCV), pages 318–335. Springer, 2016. 2

[27] Z. Wang et al. Image quality assessment: From error visibility to structural similarity. Trans. Img. Proc., 13:600 –612, 2004. 6

[28] S.-E. Wei et al. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pages 4724–4732, 2016. 3

[29] S. Xie and Z. Tu. Holistically-nested edge detection. In Proceedings of the IEEE international conference on computer vision (ICCV) , pages 1395–1403, 2015. 2

[30] J. Yang et al. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In Advances in Neural Information Processing Systems (NIPS), pages 1099–1107, 2015. 2

[31] R. Zhang et al. Colorful image colorization. In European Conference on Computer Vision (ECCV), pages 649–666. Springer, 2016. 2

[32] B. Zhao et al. Multi-view image generation from a single-view. arXiv preprint arXiv:1704.04886, 2017. 2

[33] J. J. Zhao et al. Energy-based generative adversarial network. CoRR, abs/1609.03126, 2016. 2

[34] T. Zhou et al. View synthesis by appearance flow. In European Conference on Computer Vision (ECCV), pages 286–301. Springer, 2016. 2

[35] J.-Y. Zhu et al. Generative visual manipulation on the natural image manifold. In Proceedings of European Conference on Computer Vision (ECCV), 2016. 2

[35] A. Alpher, , J. P. N. Fotheringham-Smythe, and G. Gamow. Can a machine frobnicate? Journal of Foo, 14(1):234–778, 2004.

[36] A. Alpher. Frobnication. Journal of Foo, 12(1):234–778, 2002.

[37] Actual Author Name. The frobnicatable foo filter, 2014. Face and Gesture (to appear ID 324).

[38] Actual Author Name. Frobnication tutorial, 2014. Some URL al tr.pdf.

[39] Balakrishnan, Guha, et al. "Synthesizing images of humans in unseen poses." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.