基于姿势的手语外观转换

Amit Moryossef[1,2*], **Gerard Sant**[1*], **Zifan Jiang**[1]
[1]University of Zurich, [2] 签名.mt
amit@sign.mt

Abstract

我们介绍了一种在手语骨骼姿势中转移签名者外观的方法，同时保留手势内容。利用估计的姿态，我们将一个签名者的外观转移到另一个身上，保持自然的动作和过渡。这种方法改善了基于姿态的渲染和手势拼接，同时模糊了身份识别。我们的实验表明，尽管该方法降低了签名者识别的准确性，但它略微损害了手势识别性能，突显了隐私与效用之间的权衡。我们的代码可在 https://github.com/sign-language-processing/pose-anonymization 获取。

1 介绍

个人数据，特别是能够识别个人身份的信息，在许多国家的数据保护法律中居于核心地位，包括欧盟通用数据保护条例（GDPR；European Parliament and Council of the European Union (2016)）。在有声语言中，标识信息通过外貌、语调、动作模式和手势选择嵌入到每一个表达中 (Bragg et al., 2020; Battisti et al., 2024)。因此，从信息论的角度来看，移除所有标识信息需要移除所有信息。然而，可以通过有选择地移除一些信息来实现隐私与效用之间的权衡。

我们提出了一种简单而有效的方法，用于改变手语姿势中签署者的外观（图 1），同时保留其潜在的手语内容（§3）。具体而言，给定签署者 α 的手语视频和人物 β 的图像，我们的方法生成人物 β 执行与签署者 α 相同手语的内容。

定性地说，该方法有效地平滑了骨架姿态拼接 (Moryossef et al., 2023b)，并改进了基于姿态的视频渲染 (Saunders et al., 2021)。然而，作为数据增强的方法定量评估显示，虽然它可以帮助混淆手势识别模型，但它会损害手语识别 (§5)。

2 相关工作

关于手语姿势外观变化的研究。如 Isard (2020) 所强调的，视频匿名化主要分为两类：隐藏视频的部分内容 (Hanke et al., 2020; Rust et al., 2024) 或者在不包含某些信息的情况下重新生成视频。本研究侧重于后者。

例如，Saunders et al. (2021) 替换签名者的视觉外观，针对人类消费。他们从原始视频中估计姿态，并使用生成对抗网络（GAN；Goodfel-

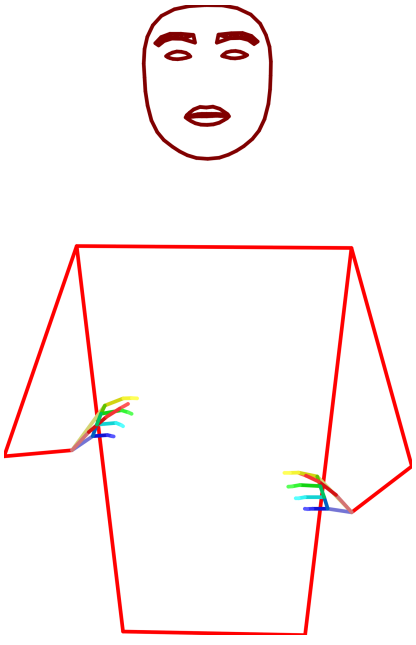


图 1: 从一个大型手语数据集（≈ 50 百万帧）中提取的平均 MediaPipe 整体框架（为视觉清晰度减少的关键点）。

low et al. (2014)) 生成外表不同的人类。这个过程如果正常工作，会像单独的姿态估计一样有效地匿名化签名视频，因为所有的原始姿态信息都被捕捉和重现。同样地，基于卡通的匿名化方法用动画角色复制签名，但常常遗漏关键细节如面部表情和手势配置 (Tze et al., 2022)。

Battisti et al. (2024) 发现仅姿势估计并不能隐藏手语者的身份。他们指出可以从姿势数据中识别出手语者，强调了需要先进的匿名化技术来更好地保护隐私。我们的工作通过提出外观转移来解决这一问题，以帮助模糊手语姿势。

3 方法

我们的外观转换方法专注于在姿态序列中改变签署者的外观，同时保留底层的符号信息。该方法假设视频从放松的姿态开始，而不是从中途签署开始。

给定一个手势序列由签署者 $\alpha(P_\alpha)$ 给出和一个单一姿势帧由签名人 $\beta(P_\beta)$ 提供，这两个姿态都基于肩膀宽度被规范化为相同的尺度，使用了姿势格式 (Moryossef et al., 2021a) 库。假设两个手势者的外观存在于每个姿态的第一帧中。

忽略手势，将签名人 β 的外观转移到签名人 α 的视频中，我们通过移除 α 的外观并添加 β 的外观来修改姿势序列（方程 1）。

$$\hat{P_\alpha} = P_\alpha - P^0_\alpha + P^0_\beta \tag{1}$$

为了进行标准化匿名处理，我们选择人员 β 作为大型手语数据集中的平均框架（图 1）。这导致了一个比例平均的人体形象，并不特别类似于任何特定的个人。我们注意到，从信息论的角度来看，这种方法并不能保证匿名性。用法在算法 1 中有所展示。

4 定性评估

这种简单的方法产生了出色的结果。首先，我们展示了一些来自不同姿态的姿态帧，当它们被转移到平均外观（匿名化）时，以及当它们被转移到不同人的外观时（表 1）。

Algorithm 1 对姿态序列进行匿名处理

```
from pose_format import Pose
from pose_anonymization.appearance \
    import remove_appearance

with open("example.pose", "rb") as f:
    pose = Pose.read(f.read())

pose = remove_appearance(pose)
```

我们考虑一篇关于手语拼接和渲染的近期论文 (Moryossef et al., 2023b)。该论文通过从词汇表中识别相关手势，以智能方式将它们拼接在一起（裁剪中立位置并平滑过渡），然后使用训练于单个翻译员上的渲染模型生成视频来实现口语文本到手语视频的转换。我们引入了一个单一干预措施—在找到相关的词汇条目后，我们将姿态的外观转移到渲染器训练所用的翻译员的姿态上。

渲染 渲染模型是一个使用 ControlNet(Zhang and Agrawala, 2023) 进行可控性微调的 Stable Diffusion 模型 (Rombach et al., 2021)。由于该模型是在单个人的外观上训练的，因此对于各种输入外观不够鲁棒。总的来说，这不是一个很好的模型，我们希望最大化从其中获得的结果。图 2 展示了原始姿势与新姿势的脸部渲染对比。我们可以看到，当转换为模型训练所基于的解说员的外观时，结果更“真实”。

(a) 无转移　　(b) 带转移

图 2: 来自 ControlNet 渲染的面部

签名缝合 给定统一的外观，缝合的姿态序列现在更加连贯且不那么跳跃。不同身体部位的大小在句子中不会发生变化，并且缝合点看起

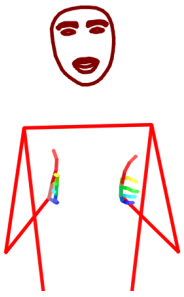
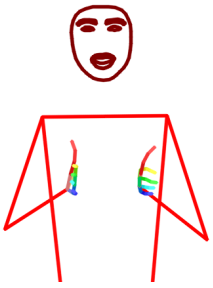
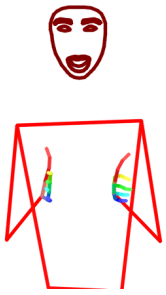
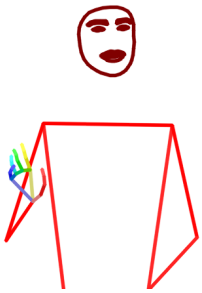
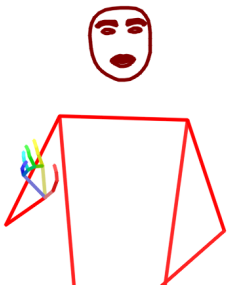
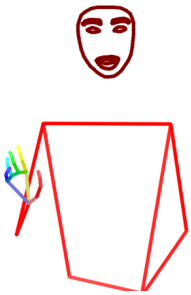
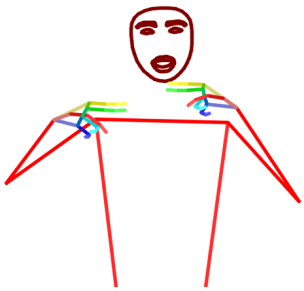
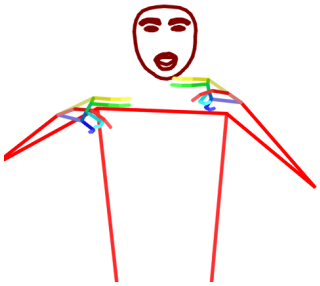
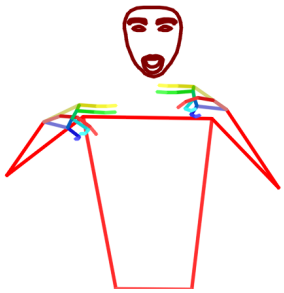
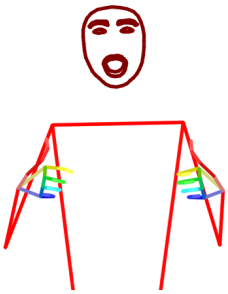
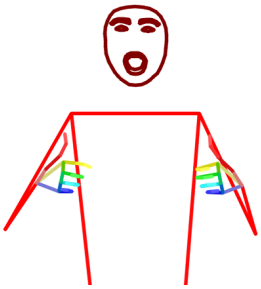
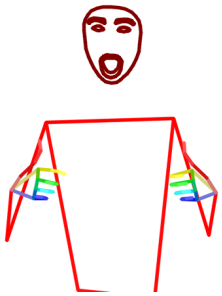
Sign	Original	Anonymized	Transferred
Kleine (‘small’)			
Kinder (‘children’)			
essen (‘eat’)			
Pizza (‘pizza’)			

表 1: 示例的四个手语符号。左边显示了原始手势的中间帧。中间是一个使用大型手语数据集中的平均姿势进行匿名化的版本。右边将外观转换为特定口译员的手势。要观看视频比较，请查看 https://github.com/sign-language-processing/pose-anonymization。

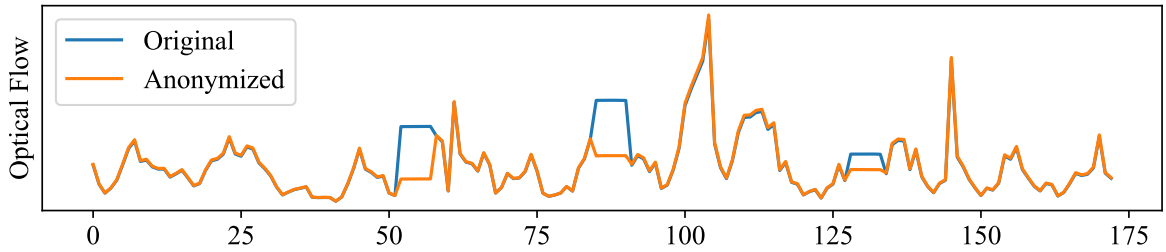


图 3: 光流（两帧之间的变化幅度）对于由四段原始视频和匿名化视频拼接而成的视频。较高值表示较大的局部变化，曲线下方的较大面积表示整体上更大的变化。除了拼接区域外，所有帧的流动完全相同。

来更平滑。当跨姿态序列跟踪光流（图 3）时，使用匿名化和原始姿势进行比较时，符号转换更平滑且不太明显。

5 实验与结果

为了量化我们的外观转换方法对手语识别的影响，我们使用了由 Moryossef et al. (2021b) 提供的代码进行手势和手语者识别任务。我们假设转换后的姿态可以作为一种有效的数据增强技术，使我们在训练和测试阶段都能隐藏手语者的身份同时达到相似的质量水平。

对于我们的实验，我们使用了包含 226 个不同词汇手势类别的 AUTSL 数据集 (Sincan and Keles, 2020)。重要的是，外观转换过程没有修改手部姿态特征，而是专注于身体和面部。

我们在四种条件下训练了模型：(1) 使用原始姿势序列；(2) 将单一外观转移到图 1 中显示的平均姿势上；(3) 每个样本转移多种外观；(4) 结合所有这些数据源，包含 10% 个原始姿势、10% 个平均姿势和 80% 个转移的外观。在测试过程中，每个模型都在原始姿势序列上进行了评估，转移到了平均姿势，并且转移到了 10 种不同的外观上，后者的评估使用多数投票法，称为*传输*方法。

如表 2 所示，没有任何配置能超越使用原始姿态序列进行训练和测试的模型（左上）。然而，在原始姿态和转移姿态的组合上进行训练使模型在对增强外观数据进行推理时更具鲁棒性（右下）。

为了评估我们的外观转换方法在多大程度上混淆了手势者身份，我们使用原始的姿态序列重新训练了模型，但用手势者分类层替换了最终的手势分类层，其余网络部分保持冻结，如同 Sant and Escolano (2023) 所示。

训练	测试		
	原始	匿名化	传输
(1) Original Poses	**80.97%**	65.82%	71.46%
(2) Anonymized Poses	63.26%	64.48%	51.50%
(3) Transferred Poses	67.08%	66.54%	57.32%
(4) Combined	79.96%	60.88%	76.78%

表 2: 手势识别准确率在 AUTSL 测试集上。‘迁移’是从随机选择的相同 10 种不同外观中集成的预测结果。

当在原始姿势上训练和测试时，模型在识别签名者方面达到了 80.18% 的准确性，表明存在可识别的特征。当在匿名化姿势上进行训练和测试时，准确率下降到 65.34%，并且使用转移姿势时，进一步降低到 52.20%。这些结果表明，虽然我们的方法显著减少了可识别信息，但并未完全消除它，因为随机机会只能达到 3.23% 的准确性。

6 结论

我们提出了一种在手语姿势中进行外观转移的方法，允许在保持关键手语信息的同时改变手势者在外貌序列中的外观。通过规范化姿势并选择性地从另一个个体那里转移外观（排除手部几何形状以保持自然运动），我们在手语渲染和拼接任务中实现了平滑且一致的结果。

我们的定性评估显示，外观迁移有效平滑

了姿势过渡并增强了拼接手势序列的视觉一致性。然而，定量结果表明，虽然该方法有助于匿名化手语者身份，但它可能对手语识别性能产生负面影响。

限制

我们相信隐私与实用性的平衡在于移除所有信息，仅保留符号的选择。这类似于口语文本通过选择词汇程度上的匿名化来实现语音的匿名性。实际上，在对手语视频进行匿名化处理时，我们建议将手语分割 (Moryossef et al., 2023a) 与音系学手语文本转写相结合。转写后引入的手势片段瓶颈保证了移除识别信息如外观、韵律提示和动作模式。然后，一个手语合成组件应将转写的符号序列重新合成为视频。

我们研究的一个主要限制是没有进行人工评估。虽然该方法旨在保留重要的手势信息，但关键是要评估改变手势者外观是否会影响人类观众对手势的自然性和可理解性，尤其是在现实世界的情境中。评估匿名化或转移后的外观是否仍能让观众识别或辨认个别手势者是确保该方法在模糊身份方面成功的关键。这项评估将提供关于技术如何平衡隐私与符号内容的实用性和可理解性的见解。

致谢

该项工作由苏黎世大学数字社会倡议（DSI）的 SIGMA 项目（G-95017-01-07）资助，并由 sign.mtltd 资助。

References

Alessia Battisti, Emma van den Bold, Anne Göhring, Franz Holzknecht, and Sarah Ebling. 2024. Person identification from pose estimates in sign language. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 13–25, Torino, Italia. ELRA and ICCL.

Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA. Association for Computing Machinery.

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Amy Isard. 2020. Approaches to the anonymisation of sign language corpora. In *SIGNLANG*.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023a. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2021a. pose-format: Library for viewing, augmenting, and handling .pose files. https://github.com/sign-language-processing/pose.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023b. An open-source gloss-based baseline for spoken to signed language translation. In *2nd International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Available at: https://arxiv.org/abs/2305.17714.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgöz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling. 2021b. Evaluating the immediate applicability of pose estimation for sign language recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3429–3435.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Annual Meeting of the Association for Computational Linguistics*.

Gerard Sant and Carlos Escolano. 2023. Analysis of acoustic information in end-to-end spoken language translation. In *INTERSPEECH 2023*, pages 52–56.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2021. Anonysign: Novel human appearance synthesis for sign language video anonymisation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Christina O. Tze, Panagiotis P. Filntisis, Anastasios Roussos, and Petros Maragos. 2022. Cartoonized anonymization of sign language videos. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models.