Generación de gestos de lengua de señas con redes neuronales generativas basadas en poses y etiquetas

Gaston Gustavo Rios[1,3] [0000-0003-0252-7036], Pedro Dal Bianco[1,3] [0000-0001-7197-8602], Franco Ronchetti[1,2][0000-0003-3173-1327], Facundo Quiroga[1] [0000-0003-4495-4327], Oscar Stanchi[1,3] [0000-0003-0294-2053], and Waldo Hasperué[1,2] [0000-0002-9950-1563]

[1] Instituto de Investigación en Informática LIDI - Universidad Nacional de La Plata., La Plata, Argentina
{grios,pdalbianco,fronchetti,fquiroga, ostanchi,whasperue}@lidi.info.unlp.edu.ar
https://weblidi.info.unlp.edu.ar

[2] Comisión de Investigaciones Científicas de la Pcia. de Bs. As. (CIC-PBA). Argentina
https://www.cic.gba.gob.ar

[3] Becario de postgrado UNLP
https://unlp.edu.ar

Abstract. Obtener datos etiquetados para el entrenamiento de redes neuronales en tareas de reconocimiento de lengua de señas es un desafío difícil y costoso. En este artículo investigamos la factibilidad de generar datos utilizando Generative Adversarial Networks (GAN), para mejorar el entrenamiento de redes neuronales. Específicamente, generamos imágenes de manos condicionando los modelos GAN con información semántica de poses y etiquetas. Comparamos los modelos ReACGAN y SPADE en la generación de nuevas imágenes de alta calidad. Evaluamos la generación de señas en dos conjuntos de datos: RWTH y HaGRID. Se entrenaron modelos generativos utilizando subconjuntos de tamaño reducido para probar el efecto de la reducción de datos de entrenamiento. Medimos la calidad de los modelos resultantes utilizando métricas cuantitativas (FID, IS, cobertura y densidad) y cualitativas (encuestas). Como resultado obtuvimos modelos GAN capaces de generar señas con un buen nivel de realismo que luego podrán ser utilizados para aumentar conjuntos de datos de lengua de señas.

Keywords: Reconocimiento de señas · Lengua de señas · Red generativa antagónica · SPADE · ACGAN · Estimación de pose.

1. Introducción

En los últimos años, el rendimiento de los modelos de aprendizaje profundo ha aumentado considerablemente. Sin embargo, el rendimiento de los modelos está estrechamente relacionado con la calidad del conjunto de datos utilizado para entrenarlos, lo que requiere conjuntos de datos grandes con suficientes muestras para aproximar la verdadera distribución del dominio. Además, la creación de nuevos conjuntos de datos etiquetados es una tarea difícil y costosa, que a menudo requiere la participación de expertos durante la recopilación y etiquetado de los datos [3]. Este es el caso de los conjuntos de datos de lenguas de señas, donde se necesita un intérprete para etiquetar los datos existentes o grabar nuevas muestras con precisión[4].

Dado que las lenguas de señas no son mutuamente inteligibles, cada una requiere su propio conjunto de datos [3]. Debido a la dificultad de la tarea en la creación de

conjuntos de datos de lenguas de señas, las comunidades con menos recursos se ven limitadas en el tamaño y la calidad de sus conjuntos de datos. Incluso las comunidades con recursos abundantes tienen conjuntos de datos de lengua de señas limitados que no son adecuados para entrenar modelos. La falta de datos dificulta el entrenamiento de modelos que se puedan aplicar a un escenario real en esta tarea. Para nuestros experimentos utilizamos los conjuntos de datos RWTH y HaGRID. RWTH es un conjunto de lengua de señas alemanas tomado de un canal de noticias del clima. Presenta poca variación de señantes y señas, imágenes borrosas y una cantidad limitada de datos etiquetados. En contraste, HaGRID es un conjunto de datos de gestos de gran tamaño creado para entrenar modelos detectores de gestos. HaGRID por lo tanto es un buen punto de comparación de la capacidad de los modelos de generar imágenes de alta calidad.

En este artículo evaluamos diferentes métodos para entrenar modelos generativos condicionados por poses y etiquetas. Entrenamos las arquitecturas de modelos GAN: ReACGAN y SPADE. Adicionalmente evaluamos diferentes pesos para la parte condicional de la función de pérdida, lo que nos llevó a la decisión de entrenar nuestros modelos con un alto peso condicional, aumentando el condicionamiento de las imágenes para que no se alejen de la distribución de las clases sin perder calidad. Comparamos el uso de las poses de las manos extraídas con OpenPose como líneas 2D en las uniones de las articulaciones o puntos en cada articulación. Realizamos experimentos utilizando los 2 métodos en RWTH llegando a resultados similares, por lo que decidimos utilizar las uniones para entrenar los modelos en HaGRID. Como resultado obtuvimos modelos generadores capaces de generar, a partir de etiquetas o poses, nuevas imágenes de gestos de manos con un alto nivel de realismo.

2. Marco Teórico

2.1 Modelos generativos

Recientemente, los modelos generativos han mostrado grandes mejoras en la calidad de las imágenes sintéticas. Los modelos más exitosos, como las Redes Generativas Antagónicas (GAN) [7], los Autoencoders Variacionales (VAE) [12] y los Modelos de Difusión[8], pueden generar nuevas imágenes realistas sin memorizar los datos del conjunto de entrenamiento. Con estos modelos, podemos generar un número arbitrario de nuevas muestras de datos. Sin embargo, estas imágenes pueden presentar artefactos que agregan ruido al entrenar nuevos modelos con ellas. Además, el colapso de moda puede afectar la variación de las imágenes generadas, lo que resulta en una cantidad limitada de imágenes únicas[1].

Redes Generativas Antagónicas (GAN) Las GAN funciona entrenando de manera conjunta un discriminador y un generador, donde el generador minimiza la distancia entre las distribuciones de los datos generados y reales con el objetivo de que el discriminador no pueda distinguirlos. Por otro lado, el discriminador es entrenado para indicar si los datos son reales o falsos [6]. De esta forma, ambos modelos son entrenados hasta alcanzar un equilibrio de Nash. Esto se dá cuando, dados los parámetros actuales de ambos modelos, la función de costo de cada modelo llega a un mínimo local.

2.2 Generación condicionada

Las GAN no condicionales no tienen forma de controlar la moda de los datos generados, ya que estos son generados a partir de un vector de ruido *z*. Las Conditional Generative Adversarial Nets (CGAN) [15] introducen la idea de utilizar datos adicionales *y* para de esta forma condicionar los resultados del generador. Estos datos adicionales serán utilizados tanto en el generador como en el discriminador para condicionar el entrenamiento. Este tipo de modelo permite la utilización de etiquetas o de otros tipos de datos (por ejemplo, texto utilizando *embeddings* para codificar la entrada de texto generando vectores similares para conceptos relacionados) como condicionador. ReACGAN [9] se propuso como una mejora de los métodos utilizados por Auxiliary Classifier GAN (ACGAN). ACGAN utiliza información condicional durante el entrenamiento, incluyendo en este una pérdida de clasificación adicional a la pérdida del modelo, lo que aumentó su rendimiento en comparación con GAN regular. Sin embargo, ACGAN ha mostrado tener un entrenamiento inestable cuando el número de clases aumenta y colapsa a una cantidad pequeña de datos generados fácilmente clasificables. Estos problemas se abordaron en ReACGAN al proyectar los vectores de entrada sobre una hiperesfera unitaria y al utilizar comparaciones de dato a dato en cada minibatch. Al hacer esto, ReACGAN logró resultados estado del arte en modelos GAN y obtuvo un mejor rendimiento que algunos modelos de difusión. SPADE [17] es una GAN condicional que originalmente se diseñó para usar una capa de segmentación para condicionar la generación de nuevos datos sintéticos. Introduce un nuevo método de normalización similar al módulo de Batch Normalization que permite el uso de datos 2D mediante convoluciones.

2.3 Generación de lengua de señas

La generación de lengua de señas es una tarea que ha realizado avances recientemente con la maduración de los modelos de redes neuronales generativas. La generación de señas realistas provee una nueva forma de agilizar la interacción entre personas señantes y no señantes, y el aprendizaje de las diversas lenguas. También este tipo de modelos permitirán utilizar los nuevos datos creados para entrenar otro tipo de modelos con conjuntos de datos de lengua de seña con datos etiquetados limitados. La tarea de la generación de señas de la lengua de señas ha sido enfrentada anteriormente mediante la generación de poses [24,20], basada en avatares[11] y de imágenes foto-realistas [21,23,16]. En particular, la generación de imágenes foto-realistas ha mostrado que los modelos GAN son un método efectivo para lograr crear imágenes de personas comunicándose mediante lengua de señas con un buen nivel de realismo.

3. Métodos Propuestos

Elegimos GAN en lugar de los Modelos de Difusión debido a su velocidad de inferencia más rápida y similar rendimiento, lo que permite la generación de grandes conjuntos de datos de imágenes sintéticas sobre la marcha. Esta capacidad puede ser beneficiosa para reducir dinámicamente el sobreajuste, similar a los beneficios del aprendizaje activo [19]. Sin embargo, los modelos GAN convencionales son propensos

al colapso de moda, lo que disminuye la diversidad de las imágenes generadas. Además, el entrenamiento de estos modelos es inestable y requiere muchas soluciones alternativas para entrenarlos de manera efectiva. Un método conocido para estabilizar el entrenamiento y reducir el colapso de moda es agregar información condicional a los datos de entrenamiento. Esto incluye a los modelos ReACGAN condicionados por etiquetas y SPADE, que condiciona la salida a través de un mapa semántico.

En los experimentos realizados utilizamos Generative Adversarial Networks (GAN) condicionados por etiquetas, puntos clave de la mano y las uniones entre los puntos clave. Para condicionar a partir de la etiqueta 1 utilizamos un clasificador auxiliar [9], mientras que para condicionar en los puntos clave y las uniones entre ellos 2 se utilizaron las capas de normalización SPatially-Adaptive (DE) normalization (SPADE) [17].

4. Experimentos realizados

En esta sección detallamos los conjuntos de datos utilizados, la metodología utilizada para realizar los entrenamientos de los modelos generativos y los resultados obtenidos. Cada modelo fue evaluado utilizando las métricas Fréchet Inception Distance (FID), Inception Score (IS), Cobertura y Densidad.

4.1 Conjuntos de Datos

RWTH [13] está compuesto por 3359 imágenes etiquetadas y 1 millón débilmente etiquetadas de señales capturadas del canal de televisión alemán PHOENIX. Las señales pertenecen a la lengua de señas alemana. Todas las imágenes están recortadas centradas en los señantes. El conjunto de datos contiene un total de 45 señas diferentes. El conjunto de datos está altamente desequilibrado y presenta una gran variación intraclase y similitud entre diferentes clases. Los intérpretes siempre visten ropa negra sobre un fondo blanco.

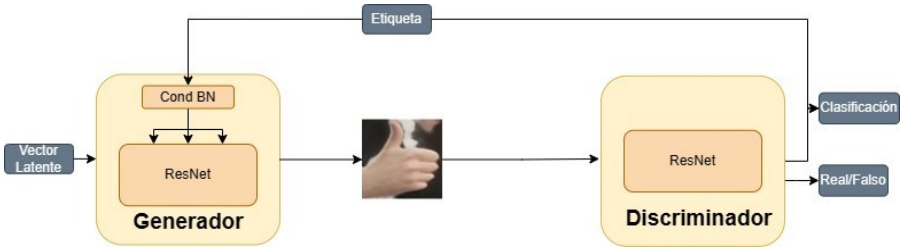


Fig.1: Diagrama del modelo ReACGAN. El generador toma como entrada un vector latente muestreado de una distribución gauseana y una etiqueta. El discriminador toma como entrada una imagen generada o real. El discriminador tiene dos salidas que son utilizadas para calcular la perdida *Data-to-Data CrossEntropy* (D2D-CE) y antagónica.

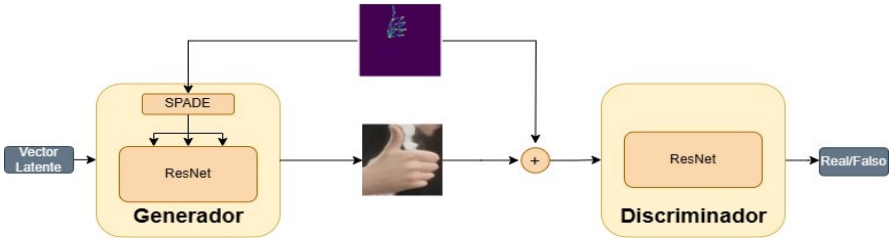
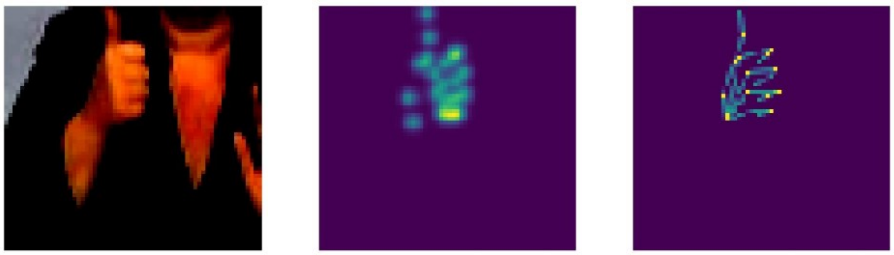


Fig.2: Diagrama del modelo SPADE. El generador toma como entrada un vector latente muestreado de una distribución gauseana y una pose codificada en *c* canales con un tamaño idéntico a la imagen de salida. El discriminador tomara como entrada la unión a nivel de los canales entre una imagen falsa o verdadera y su respectiva pose. Posteriormente utilizara la salida para calcular la perdida antagónica.

HaGRID - HAnd Gesture Recognition Image Dataset [10] fue creado para la clasificación y detección de gestos estáticos de mano. HaGRID consta de 552,992 imágenes RGB FullHD de cuerpo completo con 18 clases de gestos de mano y una clase "sin gesto". Hay un total de 34,730 personas únicas con al menos la misma cantidad de escenas. HaGRID muestra una alta diversidad entre cada persona, iluminación y fondo. El conjunto de datos también proporciona las poses para cada mano.



(a) Original (b) Puntos clave (c) Uniones

Fig.3: Visualización de la imagen original, sus puntos clave y uniones. Cada punto clave es representado como un canal con una distribución normal multivariada centrada en la coordenada del punto clave. Las uniones consisten en una línea por canal que une dos puntos anatómicamente adyacentes.

4.2 Preprocesamiento

Para obtener las poses de las señas de RWTH utilizamos OpenPose [5,22]. Esto resultó en un total de 21 puntos clave por mano, pero dado que trabajamos con señas de una sola mano, solo utilizamos la mano con mayor confianza en cada imagen. Eliminamos las muestras para las cuales no pudimos extraer ninguna pose, reduciendo el total de imágenes de 3359 a 2098.

En el caso de HaGRID, decidimos recortar las manos en las imágenes para obtener una mejor precisión en la generación especifica de estas, por lo que solo trabajamos con las manos de las imágenes ignorando el resto del cuerpo.

Para ambos conjuntos de datos modificamos el tamaño de cada imagen a 64x64 y normalizamos los valores de los píxeles. Para el aumento de datos incluimos volteo, translación, *cutout* y cambios en el brillo, saturación y contraste de las imágenes. Aislamos algunas muestras de cada conjunto de datos para usarlas como conjunto de prueba para nuestro clasificador. El resto de los datos se utilizó para el entrenamiento y la validación.

Las poses son representadas como un mapa de características $P \in \mathrm{R}^{cx64x64}$ donde c es la cantidad de canales. En el condicionamiento por pose utilizando puntos clave, cada canal posee una distribución normal multivariada centrada en la coordenada de un punto. Por lo tanto, este tipo de entrada posee un total de 21 canales. En el caso del condicionamiento utilizando las líneas de las uniones entre puntos se dibuja una línea por canal, creando un total de 20 canales.

4.3 Configuración de las evaluaciones

Empleamos la misma arquitectura base para todos los modelos GAN y aplicamos Normalización espectral para estabilizar el entrenamiento. Para condicionar los modelos con las etiquetas, utilizamos Conditional Batch Normalization en ReACGAN, y, alternativamente, utilizamos módulos SPADE para condicionar el modelo con los puntos. En todas las configuraciones utilizamos perdidas bisagra [14] y asignamos una pérdida condicional alta para mejorar la capacidad del modelo de generar imágenes con la etiqueta o pose correcta, ya que una pérdida condicional alta demostró aumentar la calidad de las imágenes generadas.

(a) RWTH real

(b) HaGRID real

(c) RWTH etiqueta

(d) HaGRID etiqueta

(e) RWTH pose

(f) HaGRID pose

Fig.4: Ejemplos reales y generados de RWTH y HaGRID con los modelos ReACGAN y SPADE condicionados por etiqueta y datos de pose respectivamente.

Table 1: Comparación de la performance de los modelos GAN en el conjunto de datos RWTH. La tabla muestra los resultados de la evaluación de los modelos utilizando las métricas Fréchet Inception Distance (FID), Inception Score (IS), Cobertura y Densidad.

RWTH GAN	**FID(↓)**	**IS(↑)**	**Cobertura(↑)**	**Densidad(↑)**
ReACGAN λ_{cond}=0.5	45.45	2.21	0.52	0.33
ReACGAN λ_{cond}=1	**45.19**	2.11	**0.60**	**0.48**
SPADE-puntos-clave	51.96	2.24	0.43	0.30
SPADE-uniones	51.05	**2.25**	0.38	0.23

Table 2: Comparación de la performance de los modelos GAN en el conjunto de datos HaGRID. La tabla muestra los resultados de la evaluación de los modelos utilizando las métricas cuantitativas Fréchet Inception Distance (FID), Inception Score (IS), Cobertura y Densidad. También se muestra como métrica cualitativa con participantes humanos la media del nivel de realismo basada en encuestas con una puntuación del 1 al 5.

HaGRID GAN	**FID(↓)**	**IS(↑)**	**Cobertura(↑)**	**Densidad(↑)**	**Humano**
ReACGAN	13.9	3.62	0.88	0.80	3.97
SPADE-uniones	33.21	3.88	0.65	0.70	-

4.4 Resultados

De cada modelo medimos su FID e IS, dos métricas estándar para medir la calidad de las imágenes generadas de los modelos generativos [2]. Ambas métricas están basadas en la utilización del modelo Inception, de ahí proviene sus nombres, y proveen un método cuantitativo para medir el realismo de las imágenes basado en la distancia entre la distribución de datos reales y generados. Por otro lado, estas métricas tienen sus limitaciones, por lo que se agregaron las métricas de cobertura y densidad, las cuales permiten medir por separado la diversidad y fidelidad de las imágenes respectivamente. La fidelidad indica el grado de similaridad entre las imágenes generadas y las reales. La diversidad, por otro lado, mide si los elementos generados

cubren la totalidad de la variabilidad de los elementos reales. También se realizaron pruebas cualitativas para el modelo ReACGAN entrenado con HaGRID por medio de cuestionarios donde los participantes fueron dados la tarea de puntuar las imágenes generadas del 1 al 5 en 3 áreas: realismo de la mano, correcta anatomía de los dedos y correcta posición de los dedos. Como resultado de estas pruebas cualitativas se obtuvo una media de 3.97 en cuanto al realismo de las imágenes, 3.70 respecto a la anatomía de los dedos y 4.03 respecto a la posición de los dedos.

Las Tablas 1 y 2 muestran el rendimiento de los diferentes modelos GAN entrenados en los conjuntos de datos RWTH y HaGRID. Aumentar el peso de la pérdida condicional genera una ligera mejora en las métricas del modelo generador. Sin embargo, no hubo una diferencia clara en el rendimiento de los modelos condicionados en puntos clave o en las líneas de las manos. Los modelos condicionados por etiqueta mostraron una mejor calidad, obteniendo mejores métricas que los modelos condicionados por pose en todos los casos.

5. Conclusiones

En este articulo realizamos la generación de manos utilizando modelos con arquitecturas y métodos de condicionamiento variados. Como resultado, se obtuvieron modelos capaces de generar imágenes con un buen nivel de realismo, lo cual fue validado mediante métricas cualitativas y cuantitativas. Los modelos condicionados por etiquetas obtuvieron un mejor resultado que los condicionados en pose, esto puede deberse a la utilización de D2D-CE para calcular la perdida en ReACGAN.

Los modelos generadores creados son capaces de generar nuevos conjuntos de datos sintéticos de lengua de seña que luego podrían ser utilizados para entrenar otro tipo de modelos. Esto permitiría el entrenamiento de modelos con menor cantidad o variabilidad de datos, un problema común en conjuntos de datos de la lengua de señas. A futuro, sería interesante la búsqueda de una mejor pérdida condicional para el modelo condicionado en poses. De esta forma, dada la semejanza de los dominios de las señas y los gestos, sería posible la realización de una generación *zero-shot*[18] de RWTH con un modelo entrenado condicionado con poses en HaGRID.

Referencias

1. Arora, S., Zhang, Y.: Do gans actually learn the distribution? an empirical study. CoRR abs/1706.08224 (2017)
2. Borji, A.: Pros and cons of gan evaluation measures (2018)
3. Bragg, D., Caselli, N., Hochgesang, J.A., et al.: The fate landscape of sign language ai datasets: An interdisciplinary perspective. ACM Trans. Access. Comput. 14(2) (2021). https://doi.org/10.1145/3436996
4. Bragg, D., Koller, O., Bellard, M., et al.: Sign language recognition, generation, and translation: An interdisciplinary perspective (2019)
5. Cao, Z., Hidalgo Martinez, G., Simon, T., et al.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)

6. Goodfellow, I.J.: NIPS 2016 tutorial: Generative adversarial networks. CoRR abs/1701.00160 (2017)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: Generative Adversarial Networks. arXiv e-prints p. arXiv:1406.2661 (2014).
8. https://doi.org/10.48550/arXiv.1406.2661
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. CoRR abs/2006.11239 (2020)
10. Kang, M., Shim, W., Cho, M., et al.: Rebooting ACGAN: auxiliary classifier gans with stable training. CoRR abs/2111.01118 (2021)
11. Kapitanov, A., Makhlyarchuk, A., Kvanchiani, K.: HaGRID - HAnd Gesture Recognition Image Dataset. arXiv e-prints p. arXiv:2206.08219 (2022)
12. Kim, J.H., Hwang, E.J., Cho, S., et al.: Sign language production with avatar layering: A critical use case over rare words. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1519–1528. European Language Resources Association, Marseille, France (2022)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022)
14. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3793–3802
15. (2016). https://doi.org/10.1109/cvpr.2016.412
16. Lim, J.H., Ye, J.C.: Geometric gan (2017)
17. Mirza, M., Osindero, S.: Conditional generative adversarial nets. CoRR abs/1411.1784 (2014)
18. Natarajan, B.S., R., E.: Dynamic gan for high-quality sign language video generation from skeletal poses using generative adversarial networks. Soft Computing 26, 13153–13175 (2021)
19. Park, T., Liu, M., Wang, T., et al.: Semantic image synthesis with spatiallyadaptive normalization. CoRR abs/1903.07291 (2019)
20. Ramesh, A., Pavlov, M., Goh, G., et al.: Zero-shot text-to-image generation (2021)
21. Ren, P., Xiao, Y., Chang, X., et al.: A survey of deep active learning. CoRR abs/2009.00236 (2020)
22. Saunders, B., Camgoz, N.C., Bowden, R.: Adversarial training for multi-channel sign language production (2020)
23. Saunders, B., Camgoz, N.C., Bowden, R.: Everybody sign now: Translating spoken language to photo realistic sign language video (2020)
24. Simon, T., Joo, H., Matthews, I., et al.: Hand keypoint detection in single images using multiview bootstrapping. In: Cvpr (2017)
25. Stoll, S., Camgoz, N.C., Hadfield, S., et al.: Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. Int. J. Comput. Vision 128(4), 891–908 (2020). https://doi.org/10.1007/s11263019-01281-2
26. Xiao, Q., Qin, M., Yin, Y.: Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. Neural Networks 125, 41–55 (2020). https://doi.org/https://doi.org/10.1016/j.neunet.2020.01.030