Sistema de Reconocimiento Facial: Identificación de Personas a partir de Imágenes

Nombre del Autor
Afiliación / Dirección 1
Afiliación / Dirección 2
email@dominio

Segundo Autor
Afiliación / Dirección 1
Afiliación / Dirección 2
email@dominio

problema fundamental en el que nos enfocamos es la capacidad de los modelos para operar con conjuntos de datos

Abstract

Como proyecto final de la asignatura de Machine Learning, se nos planteó el reto de desarrollar un sistema de reconocimiento facial capaz de identificar a nuestros compañeros de clase en una imagen, es decir, reconocer quién aparece en una fotografía.

El abordaje de este desafío se descompone en dos fases: la detección de rostros en imágenes y su posterior reconocimiento. En este proyecto, nos enfocaremos principalmente en la segunda fase, analizando el uso de Redes Neuronales Convolucionales (CNNs) en el reconocimiento facial, con especial atención a las redes siamesas. Estas redes han demostrado un rendimiento efectivo al comparar características faciales y en la verificación de identidad.

A través de la implementación de un modelo siamés, se explorará la capacidad de esta metodología para mejorar la precisión y robustez en el reconocimiento de personas no vistas en el conjunto de entrenamiento. Este enfoque innovador podría abrir nuevas vías para la identificación en entornos educativos, donde la diversidad de rostros y poses puede presentar desafíos únicos.

Los resultados de este proyecto se espera que contribuyan a la comprensión de la efectividad de las redes siamesas en el reconocimiento facial, proporcionando así un análisis crítico que podría ser útil para futuras investigaciones y aplicaciones en el ámbito de la inteligencia artificial y la visión por computadora.

1 Introducción

El reconocimiento facial representa un desafío técnico complejo, especialmente en entornos con datos limitados. En este proyecto, nos enfrentamos a la problemática de contar con un conjunto de imágenes reducido, en particular, la ausencia de imágenes de varios compañeros de aula. Esto planteó una dificultad central: construir un modelo capaz de identificar a personas que no habían sido incluidas previamente en el entrenamiento, exigiendo que el sistema generalizara de manera efectiva a nuevos individuos.

Para abordar este desafío, optamos por mantener el enfoque más sencillo posible para la detección de rostros, utilizando el algoritmo Viola-Jones como base para detectar caras en las imágenes. Este algoritmo permitió detectar rostros con precisión, lo que estableció una sólida base para la fase de reconocimiento, donde se pondrían a prueba las capacidades del sistema para identificar a compañeros nuevos sin haber sido entrenado específicamente para ello.

A medida que avanzamos, diseñamos y ajustamos nuevos modelos con el fin de mejorar continuamente los resultados, centrándonos en la simplicidad y en maximizar la capacidad de generalización del sistema. Las mejoras fueron evaluadas en función de los siguientes criterios clave:

- Precisión en la identificación de nuevos individuos: Evaluamos si el sistema era capaz de reconocer a compañeros que no formaban parte del conjunto de entrenamiento inicial. - Eficiencia en la detección y reconocimiento: Medimos la velocidad y la precisión en la detección de rostros mediante el algoritmo Viola-Jones, garantizando que el proceso fuese eficiente y sencillo. - Capacidad de generalización: Analizamos hasta qué punto los nuevos modelos mejoraban la capacidad del sistema para manejar imágenes de personas no vistas previamente, priorizando la generalización

sin necesidad de grandes volúmenes de datos. - Comparación iterativa: En cada fase del proyecto, comparamos el desempeño de los nuevos modelos con las versiones anteriores, midiendo las mejoras en precisión, velocidad y generalización.

Con este enfoque iterativo y basado en soluciones sencillas, logramos refinar progresivamente el sistema de reconocimiento facial, superando las limitaciones iniciales impuestas por la falta de datos. Esto nos permitió mejorar la capacidad del modelo para reconocer personas no incluidas en el conjunto de entrenamiento, mientras manteníamos la simplicidad y la eficiencia en el proceso.

2 Estado del Arte en Verificación de Identidad Facial

El campo del reconocimiento facial ha avanzado de manera significativa en la última década, impulsado por la evolución de las redes neuronales profundas y la disponibilidad de grandes conjuntos de datos. La verificación de identidad facial, que busca determinar si un rostro nuevo pertenece a una persona previamente registrada en una base de datos, se beneficia de estos avances. En esta sección, exploraremos las metodologías más recientes y efectivas, describiendo sus características técnicas y arquitectónicas, así como su precisión en diversas condiciones.

2.1 FaceNet

Uno de los modelos más influyentes en este campo es **FaceNet**, desarrollado por Google en 2015. FaceNet emplea una arquitectura de red neuronal siamesa que genera embeddings faciales. Esta técnica se basa en la idea de que las imágenes de la misma persona deben estar cercanas entre sí en el espacio de características, mientras que las imágenes de diferentes personas deben estar alejadas. FaceNet utiliza una función de pérdida triplete, que se centra en optimizar la distancia entre un ancla (una imagen de referencia), una imagen positiva (de la misma persona) y una imagen negativa (de otra persona). Este enfoque permite una comparación eficiente utilizando la distancia euclidiana entre embeddings. En el conjunto de datos **LFW** (Labeled Faces in the Wild), FaceNet logró una precisión del 97.3%, destacándose en condiciones de iluminación controladas y variaciones de poses.

2.2 ArcFace

Un avance posterior significativo es **ArcFace**, que se basa en una arquitectura de red residual profunda, específicamente **ResNet-100**. La principal innovación de ArcFace es su función de pérdida conocida como **Additive Angular Margin Loss (AM-Softmax)**, que maximiza la separación angular entre los embeddings de diferentes identidades. Este enfoque no solo mejora la discriminación entre clases, sino que también optimiza el rendimiento en condiciones donde los rostros pueden compartir características similares. En pruebas realizadas en **LFW**, ArcFace alcanzó una impresionante precisión superior al 99.8%, así como un 99.63% en el conjunto de datos **MegaFace**, evidenciando su efectividad tanto en escenarios controlados como desafiantes.

2.3 CosFace

Otra alternativa destacada es **CosFace**, que introduce el uso de una distancia cosenoidal para mejorar la generalización en la verificación facial. También basado en **ResNet-50**, CosFace implementa una variación de la función de pérdida AM-Softmax, centrada en la distancia entre embeddings faciales. Los resultados obtenidos en el conjunto de datos LFW revelaron una precisión del 99.73%, mientras que en MegaFace logró un 95.1%. Esto destaca la capacidad de CosFace para adaptarse a la variabilidad de los datos y su robustez en condiciones cambiantes.

2.4 VGGFace2

Por su parte, **VGGFace2** se centra en la robustez frente a variaciones faciales, como diferencias en pose, iluminación y expresión. Este modelo utiliza una arquitectura que combina **ResNet-50** y **SENet-50** (Squeeze-and-Excitation Networks). SENet introduce un mecanismo de atención que ajusta el peso de cada canal de características, permitiendo que el modelo se concentre en las características más relevantes para la verificación facial. VGGFace2 ha alcanzado una precisión del 99.2% en LFW y ha mostrado resultados competitivos en entornos no controlados, lo que subraya su aplicabilidad en escenarios del mundo real.

2.5 CurricularFace

Más recientemente, **CurricularFace** ha introducido un enfoque de aprendizaje jerárquico que optimiza el rendimiento en la clasificación de

imágenes difíciles. Utilizando una arquitectura basada en **ResNet-100**, este modelo presenta imágenes más simples en las primeras etapas del entrenamiento y gradualmente introduce imágenes más complejas. Este método permite que la red aprenda a resolver primero los casos fáciles, ajustándose posteriormente a los más desafiantes. CurricularFace ha mostrado precisiones de hasta 99.8% en LFW, siendo especialmente útil en situaciones donde la calidad de las imágenes puede variar significativamente.

2.6 Vision Transformers y Reconocimiento Facial enmascarado

En los últimos años, hemos visto la creciente popularidad de los **Transformers** en el ámbito de la visión por computadora. Modelos como **Vision Transformers (ViT)** y **Swin Transformer** han comenzado a ser explorados para tareas de reconocimiento facial. Estas arquitecturas permiten procesar las imágenes como secuencias de parches, mejorando la capacidad de capturar relaciones espaciales de largo alcance, crucial en el análisis facial. Este enfoque se complementa con el reconocimiento de rostros enmascarados, impulsado por la pandemia de COVID-19, que ha llevado al desarrollo de modelos como **Masked ArcFace** y **CurricularFace** adaptados para reconocer rostros parcialmente cubiertos. **Masked ArcFace** ha demostrado mantener una precisión del 91.5% en el reconocimiento facial enmascarado, evidenciando su eficacia en situaciones desafiantes donde los rostros están parcialmente ocultos.

3 Objetivos

El objetivo principal de este proyecto es desarrollar un sistema de reconocimiento facial eficiente, capaz de identificar correctamente a personas no incluidas en el conjunto de entrenamiento, utilizando un enfoque sencillo y con datos limitados. Los objetivos específicos son:

- **Desarrollar un modelo de detección y reconocimiento facial**: Utilizar el algoritmo Viola-Jones para la detección de rostros y anlizar difernetes técnicas para el reconocimiento.
- **Evaluar la capacidad del modelo para generalizar**: Medir la capacidad del sistema para identificar correctamente individuos no vistos durante el entrenamiento.
- **Mejorar la precisión y eficiencia del sistema**: Optimizar el balance entre precisión en la identificación y eficiencia computacional. experimentando con estas

4 Metodologías

En este proyecto de reconocimiento facial, implementaremos y evaluaremos diversos enfoques que van desde los más sencillos hasta los más complejos. A lo largo del proceso, experimentaremos con estas técnicas con el objetivo de mejorar la precisión y la eficiencia del sistema.

4.1 Detección de rostros con Viola-Jones

En las primeras etapas del proyecto, utilizamos el algoritmo **Viola-Jones** para la detección de rostros. Este algoritmo transforma las imágenes a escala de grises para reducir la cantidad de datos a procesar, y emplea características *Haar* para identificar rostros dentro de una ventana que se desplaza por la imagen.

Para mejorar la precisión en el reconocimiento, aplicamos *data augmentation*,Esta técnica ayudó a aumentar la diversidad del conjunto de entrenamiento sin requerir más imágenes reales, lo que permitió obtener mejores resultados en la predicción posterior.

4.2 Reconocimiento con LBPH (Local Binary Patterns Histogram)

El algoritmo **LBPH (Local Binary Patterns Histogram)** fue utilizado para el reconocimiento facial. Este enfoque genera una representación intermedia de la imagen en la que se resaltan las características del rostro. Posteriormente, se divide la imagen en una cuadrícula y se calcula un histograma para cada celda, el cual describe la distribución de los valores de intensidad de los píxeles. Finalmente, los histogramas se concatenan para formar un descriptor único que se compara con los histogramas previamente entrenados.

El proceso de reconocimiento consistía en los siguientes pasos:

1. Captura de la imagen o video a través de una cámara.
2. Aplicación de Viola-Jones para detectar el rostro en la imagen.
3. Uso de LBPH para extraer el histograma de la cara detectada.

4. Comparación del histograma con los modelos previamente entrenados.

5. Retorno del ID correspondiente al rostro más parecido.

Aunque el sistema era relativamente rápido, la precisión no superaba el 60%, incluso con técnicas de *data augmentation*. Esto nos llevó a explorar otros enfoques más avanzados.

4.3 Redes Convolucionales Siamesas

En un intento por mejorar la precisión, probamos redes convolucionales siamesas. A diferencia de los modelos tradicionales, las redes siamesas no clasifican imágenes en etiquetas específicas, sino que miden la distancia entre los *embeddings* de dos imágenes. El flujo del proceso es el siguiente:

1. Dos imágenes son procesadas por redes con la misma arquitectura.

2. Se obtienen los *embeddings* de ambas imágenes.

3. Se calcula la distancia euclidiana entre los *embeddings*; cuanto menor sea la distancia, más similares son las imágenes.

4. Se aplica una función sigmoide para normalizar los resultados y una función de pérdida para ajustar los pesos de la red.

A pesar de que este enfoque parecía prometedor, en la práctica presentó varios inconvenientes: requería un gran número de imágenes para alcanzar una precisión aceptable, lo que lo hizo inviable en nuestro contexto de datos limitados. Además, su demanda computacional lo hacía poco práctico para equipos con recursos limitados.

4.4 Transición hacia FaceNet

Finalmente, optamos por implementar **FaceNet**, un modelo de reconocimiento facial basado en *embeddings* euclidianos, que requiere solo unas pocas imágenes para reconocer a una persona con alta precisión. FaceNet fue entrenado previamente con más de 140 mil imágenes, permitiendo una precisión del 99.2% con solo 2 o 3 fotos por persona.

El modelo FaceNet funciona de la siguiente manera:

1. Selección de una imagen de referencia o "ancla".

2. Comparación de la imagen del ancla con imágenes positivas (de la misma persona) y negativas (de personas diferentes).

3. Ajuste de los parámetros del modelo para reducir la distancia entre imágenes positivas y aumentar la distancia entre las negativas.

5 Experimentación

En el proceso de implementación de un modelo de reconocimiento facial utilizando una red neuronal siamesa, se llevaron a cabo diversas experimentaciones con el objetivo de mejorar la precisión y la capacidad de generalización del modelo. Estas experimentaciones estuvieron principalmente enfocadas en la creación y estructuración adecuada del conjunto de datos, así como en la metodología de entrenamiento, con el fin de obtener resultados consistentes y robustos.

Una red siamesa requiere pares de imágenes que representen relaciones entre una imagen de referencia, conocida como **ancla** (anchor), y otra imagen que puede ser de la misma persona (**par positivo**) o de una persona diferente (**par negativo**). Estos pares están etiquetados como 1 (si el par es positivo) o 0 (si el par es negativo), formando las siguientes tuplas:

- $\langle anchor, positive, 1 \rangle$: Ambas imágenes pertenecen a la misma persona.

- $\langle anchor, negative, 0 \rangle$: Las imágenes pertenecen a personas diferentes.

Las decisiones sobre cómo estructurar y seleccionar estos pares tienen un impacto crítico en el rendimiento del modelo, como lo evidencian las siguientes iteraciones de experimentación.

5.1 Primera Iteración: Entrenamiento con Datos Limitados

En la primera fase, diseñamos un conjunto de datos limitado con el objetivo de que el modelo aprendiera a identificar una única persona. Para ello, creamos pares de imágenes donde las anclas y las imágenes positivas provenían exclusivamente de una sola persona, y las imágenes negativas fueron seleccionadas del conjunto de datos LFW (Labeled Faces in the Wild). Recopilamos un total de 3000 pares, con aproximadamente la mitad de ellos siendo positivos y la otra mitad negativos.

Este enfoque inicial resultó ser ineficaz. El modelo falló en generalizar adecuadamente, mostrando una tasa de precisión extremadamente baja. Los pares de entrenamiento no eran lo suficientemente diversos, y al centrarse en una sola persona, el modelo no logró aprender representaciones discriminativas útiles para otras identidades. Los resultados mostraban que el modelo confundía con frecuencia a la persona objetivo con otras, revelando una falta de robustez.

Este fallo resaltó la necesidad de introducir una mayor diversidad en los datos y nos llevó a la conclusión de que el enfoque inicial era inadecuado para entrenar un modelo siamesa de reconocimiento facial.

5.2 Segunda Iteración: Expansión y Selección Aleatoria del Conjunto de Datos

Reconociendo las limitaciones de la primera iteración, decidimos ampliar considerablemente el tamaño del conjunto de datos. Aumentamos la cantidad de pares a 14,000, seleccionados de un conjunto de aproximadamente 100,000 imágenes de la base de datos LFW. Estos pares incluían 7000 pares positivos ($\langle anchor, positive, 1\rangle$) y 7000 pares negativos ($\langle anchor, negative, 0\rangle$).

Sin embargo, cometimos un error metodológico clave: la selección de los pares fue completamente aleatoria, lo que condujo a un desequilibrio en la distribución de las imágenes. Específicamente, no se garantizó que cada imagen de ancla tuviera tanto un par positivo como un par negativo. Esto provocó que algunas imágenes de ancla solo estuvieran emparejadas con positivos o con negativos, lo que resultó en un conjunto de datos incompleto y desbalanceado.

Como resultado, el modelo se entrenó en un conjunto de datos inconsistente, lo que produjo un rendimiento subóptimo. La tasa de precisión mejoró ligeramente respecto a la primera iteración, pero aún no alcanzaba los estándares esperados. Este fallo nos enseñó la importancia de mantener un equilibrio adecuado entre las tuplas positivas y negativas para cada ancla.

5.3 Tercera Iteración: Balance y Reducción del Dataset

En la tercera iteración, corregimos los errores cometidos en la fase anterior. Decidimos reducir el tamaño del conjunto de datos a 6000 pares, asegurando que para cada imagen de ancla existiera al menos un par positivo y uno negativo. Esta selección fue diseñada cuidadosamente para mantener un equilibrio estricto entre los pares, evitando que algunas imágenes de ancla fueran emparejadas solo con positivos o solo con negativos.

Este enfoque balanceado mejoró significativamente la capacidad de generalización del modelo. A través de una evaluación rigurosa, observamos que el modelo comenzó a distinguir de manera más efectiva entre imágenes de la misma persona y de personas diferentes, lo que se tradujo en una mayor precisión en las pruebas de validación.

5.4 Lecciones Aprendidas sobre la selección del dataset

Las iteraciones anteriores revelaron varias lecciones fundamentales que son de vital importancia en la implementación de redes siamesas para tareas de verificación de identidad facial:

1. **Balance en la Estructura del Dataset**: Uno de los errores más críticos fue la falta de equilibrio en la distribución de pares positivos y negativos. Es esencial que cada imagen de ancla tenga tanto un par positivo como un par negativo para asegurar que el modelo aprenda de manera efectiva a discriminar entre las dos clases.

2. **Diversidad de Datos**: La primera iteración mostró que entrenar el modelo con imágenes de una sola persona limita severamente su capacidad de generalización. Un conjunto de datos más diverso permite que el modelo aprenda a representar diferentes identidades y reduce el riesgo de sobreajuste (*overfitting*).

3. **Selección Estratégica de Imágenes**: La selección aleatoria sin planificación puede generar conjuntos de datos desequilibrados y poco representativos. Es fundamental utilizar criterios claros y estratégicos para seleccionar los pares de imágenes que alimentarán al modelo.

4. **Tamaño del Dataset vs. Calidad de los Pares**: Aunque un conjunto de datos más grande puede parecer una solución intuitiva, nuestra experiencia mostró que la calidad y balance de los pares es mucho más importante que simplemente aumentar el tamaño. Un conjunto de datos más pequeño pero bien equilibrado y estructurado puede generar

mejores resultados que uno grande y desbalanceado.

5.5 Experimentación con Tasa de Aprendizaje y Funciones de Pérdida

Durante la fase de experimentación, realizamos pruebas exhaustivas variando tanto las tasas de aprendizaje como las funciones de pérdida para evaluar su impacto en el rendimiento de la red neuronal siamesa. Nuestro objetivo era identificar los valores óptimos que permitieran una convergencia eficiente y estable del modelo, así como maximizar la precisión en la tarea de verificación facial.

5.5.1 Tasa de Aprendizaje

Inicialmente, probamos una tasa de aprendizaje (*learning rate*) de 10^{-4}, ya que es un valor comúnmente recomendado para redes profundas. Sin embargo, observamos que el modelo tardaba mucho en converger y no alcanzaba un nivel de precisión aceptable en un tiempo razonable.

A continuación, experimentamos con tasas de aprendizaje de 10^{-3} y 10^{-5}, pero sin mejoras significativas. Finalmente, encontramos que la tasa de aprendizaje de 10^{-2} era la más adecuada. Este valor permitió que el modelo convergiera rápidamente, manteniendo la estabilidad del proceso de entrenamiento y alcanzando una precisión considerablemente mayor. Optamos por esta tasa de aprendizaje de 10^{-2} debido a su buen equilibrio entre velocidad de entrenamiento y rendimiento.

5.5.2 Comparativa de Funciones de Pérdida

La función de pérdida es un elemento clave en el entrenamiento de una red siamesa, ya que mide la distancia entre las representaciones generadas por el modelo para los pares de imágenes y permite optimizar el proceso de aprendizaje. A continuación, presentamos las funciones de pérdida probadas, junto con una evaluación comparativa de su rendimiento en nuestro escenario de verificación facial.

- **Triplet Loss**: Esta función de pérdida optimiza la distancia entre una imagen de ancla (*anchor*) y un positivo, mientras maximiza la distancia entre el ancla y un negativo. La idea es mantener los ejemplos de la misma clase cercanos en el espacio de *embeddings*, y los de clases diferentes, alejados. **Ventajas**: Ayuda a generar representaciones altamente discriminativas. **Desventajas**: La selección del margen puede ser complicada y afecta considerablemente el rendimiento.

- **Contrastive Loss**: Contrastive Loss optimiza la distancia euclidiana entre las imágenes del mismo par, minimizando la distancia para pares positivos y maximizándola para pares negativos. Es un enfoque más sencillo que Triplet Loss, ya que solo requiere trabajar con pares de imágenes, no con tríos. **Ventajas**: Simplicidad en su implementación. **Desventajas**: Menos eficaz que Triplet Loss para datos donde es difícil separar las clases.

- **Binary Cross-Entropy (BCE) Loss**: Esta función de pérdida mide la probabilidad de que un par de imágenes pertenezca a la misma clase. Se compara la predicción del modelo con la etiqueta binaria (1 para pares positivos y 0 para negativos). **Ventajas**: Es una función de pérdida estándar y simple. **Desventajas**: No optimiza explícitamente las distancias entre las representaciones, lo que puede limitar su eficacia en tareas siamesas.

5.5.3 Comparativa de Resultados

A lo largo de las pruebas, se evaluaron las funciones de pérdida en términos de su capacidad para optimizar la verificación de identidad facial, utilizando como métrica principal la precisión obtenida en un conjunto de validación. A continuación, se muestra una tabla comparativa de las funciones de pérdida probadas:

Función de Pérdida	Precisión (Conjunto de Validación)
Triplet Loss	94.5%
Contrastive Loss	91.8%
Binary Cross-Entropy	89.7%

Table 1: Comparativa de funciones de pérdida.

5.5.4 Conclusión sobre la Mejor Función de Pérdida

De las funciones de pérdida evaluadas, **Triplet Loss** se destacó como la más eficaz para la tarea

de verificación facial, debido a su capacidad para generar representaciones discriminativas y separadas de manera efectiva. Aunque **Contrastive Loss** es más sencilla de implementar, no ofreció el mismo nivel de discriminación, y **Binary Cross-Entropy** resultó menos adecuada al no aprovechar la estructura de las distancias entre *embeddings* de manera explícita.

Por lo tanto, recomendamos **Triplet Loss** como la mejor opción para tareas de verificación facial, especialmente en combinación con técnicas como el *hard negative mining* para mejorar la eficiencia en el aprendizaje.

6 Conclusiones y Futuras Mejoras

Este proyecto demuestra la viabilidad del reconocimiento facial como una herramienta eficaz para la identificación de personas. En el futuro, se explorará la integración de técnicas de aprendizaje profundo más avanzadas y se buscará mejorar la diversidad y el tamaño de la base de datos utilizada.

7 Referencias

1. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *CVPR 2015*. https://arxiv.org/abs/1503.03832.

2. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *CVPR 2019*. https://arxiv.org/abs/1801.07698.

3. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., ... & Liu, W. (2018). CosFace: Large Margin Cosine Loss for Deep Face Recognition. *CVPR 2018*. https://arxiv.org/abs/1801.09414.

4. Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A Dataset for Recognising Faces Across Pose and Age. *FG 2018*. https://arxiv.org/abs/1710.08092.

5. Huang, Y., Wang, W., & Wang, F. (2020). CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. *CVPR 2020*. https://arxiv.org/abs/2004.00288.

6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021*. https://arxiv.org/abs/2010.11929.

7. Wang, Y., Chen, Y., & Gu, Y. (2021). Masked ArcFace: A Loss for Robust Face Recognition under Masks. *ICCV 2021*. https://arxiv.org/abs/2107.09603.