Hiding in Plain Sight

Rodrigo Alfonso Mansilla Dubón
Departmento de Ingeniería en Ciencias de la Computación y Tecnologías de la Información
Universidad del Valle de Guatemala
man22611@uvg.edu.gt

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1. Introducción

El aumento en el uso de mecanismos de reconocimiento facial ha convertido la identidad biométrica en un activo transaccional, integrandose en la infraestructura de seguridad y en la vida cotidiana de millones de personas. Sin embargo, dicha integración plantea desafíos críticos que limitan su implementación de mánera ética y segura. En consecuencia, la identidad visual de una persona deja de ser propiedad privada del individuo y pasa a ser un input rastreable para sistemas de vigilancia másiva.

Agregado a los riesgos inherentes a la privacidad , el enterenamiento empleado en estos sistemas incluyen sesgos que no reflejan un panorama demográfico general, favoreciendo la inequidad algorítmica. Ante esto, las soluciones tradicionales de anonimización enfrentan un dilema que se divide en dos ramas, o se destruye la utilidad visual de la imagen, o se preserva la imagen a costa de exponer la identidad.

Este trabajo presenta una arquitectura generativa adversaria residual diseñada para inyectar perturbaciones como mascaras de ruido que permiten disociar la identidad biométrica de la apariencia física. A diferencia de tecnologías de DeeoFakes, que buscan la suplantación, este enfoque busca la evasión de la identidad, lo que envenena las características semánticas que las redes neuronales utilizan para el reconocimiento,sin degradar la experiencia visual para el observador humano.

2. Antecedentes

La búsqueda de la privacidad facial sin sacrificar la utilidad visual ha impulsado una evolución desde técnicas de ofuscación simples hacia modelos generativos complejos. El objetivo fundamental de disociar la identidad biométrica de la apariencia física forma la base de enfoques recientes como CIAGAN (Lange and et al., 2020), DeepPrivacy2 (Hukkelås and Lindseth, 2023) y Mask-Invariant Face Recognition (MIFR) Tran and et al. (2023).

CIAGAN Lange and et al. (2020) y DeepPrivacy2 Hukkelås and Lindseth (2023) emplean redes generativas adversarias (GANs) para sintetizar rostros completos, reemplazando la identidad original mientras intentan preservar atributos no identificativos (como la pose o el fondo) mediante máscaras de segmentación o control de atributos. De manera similar, MIFR Tran and et al. (2023) se enfoca en el reconocimiento robusto bajo oclusiones, utilizando técnicas de in-painting para reconstruir áreas faciales. Aunque estos modelos logran cierto grado de anonimización, el número de operaciones y la complejidad requerida para sintetizar una identidad facial completamente nueva a menudo resulta en artefactos visuales perceptibles y una pérdida de la .esencia"de la imagen original, especialmente en

regiones de alta frecuencia como los ojos y la boca. Además, la dependencia de módulos externos de detección o segmentación limita su aplicabilidad en tiempo real.

El uso de perturbaciones adversarias ofrece una alternativa prometedora, enfocándose en modificar la percepción de la máquina en lugar de la humana. Adv-Makeup Yin and et al. (2021) y PGT-GAN Guo and et al. (2022) han demostrado éxito en generar "maquillaje"digital imperceptible que engaña a los sistemas de reconocimiento facial. Sin embargo, estos enfoques a menudo requieren procesos de optimización iterativos para cada imagen o dependen de la transferencia de estilo, lo que puede ser computacionalmente costoso y difícil de generalizar a través de diferentes dominios o arquitecturas de reconocimiento no vistas durante el entrenamiento.

En este trabajo presenta ChameleonNet, un enfoque de anonimización que se distingue por integrar el aprendizaje de una perturbación residual estructurada en una arquitectura de inferencia de un solo paso. Este diseño busca reducir la complejidad del pipeline al prescindir de módulos externos como la segmentación semántica previa, el in-painting o redes de control de atributos explícitas. A diferencia de los enfoques que priorizan el reemplazo total de la identidad visual Lange and et al. (2020); Hukkelås and Lindseth (2023), el modelo presentado se enfoca en preservar la coherencia estructural de la imagen original. Para ello, aprovechamos las conexiones de salto propias de la arquitectura U-Net e introducimos una función de pérdida de atención diseñada específicamente para guiar la perturbación hacia regiones biométricamente relevantes, buscando un balance óptimo entre evasión y eficiencia computacional. En las siguientes secciones, se describe la arquitectura de ChameleonNet y motivaremos su ventaja sobre los modelos generativos tradicionales.

3. Arquitectura del Modelo

El objetivo de ChameleonNet es aprender una función de mapeo generativo G que transforme una imagen facial de entrada x en una versión anonimizada x_{adv}, de tal manera que la identidad biométrica sea suprimida mientras se preserva la apariencia visual. El marco de trabajo se compone de dos módulos principales: un **generador entrenable de perturbaciones** (G) y un **extractor de características biométricas fijo** (F).

El generador G opera bajo un esquema de aprendizaje residual, donde no sintetiza la imagen final directamente, sino una "máscara de ruido" aditiva $\delta = G(x)$. La imagen anonimizada se obtiene mediante $x_{adv} = clamp(x + \alpha \cdot \delta)$, donde α es un factor de escala que controla la magnitud máxima de la perturbación.

El extractor de características F es una red neuronal profunda pre-entrenada en reconocimiento facial , cuyos pesos permanecen congelados durante todo el proceso. Este módulo actúa como un "juez biométrico", proporcionando la retroalimentación necesaria para guiar el entrenamiento de G hacia la evasión de identidad. El sistema completo se optimiza de extremo a extremo mediante una función de pérdida híbrida diseñada para balancear tres objetivos competitivos: fidelidad visual, divergencia de embeddings biométricos y disrupción de mapas de atención intermedios.

3.1. Generador de Perturbaciones Residuales (G)

La arquitectura del generador G se basa en una U-Net adaptada, una red completamente convolucional tipo encoder–decoder.

El codificador consta de tres bloques de reducción de resolución, cada uno con una convolución, normalización de instancia y LeakyReLU. Esta ruta extrae características jerárquicas, reduciendo la dimensión espacial y aumentando la profundidad hasta un cuello de botella semántico.

El decodificador reconstruye la resolución espacial mediante tres bloques de aumento con convoluciones transpuestas, normalización de instancia y ReLU. Incorpora conexiones de salto que concatenan características de alta resolución del codificador con el decodificador, mitigando el desvanecimiento del gradiente y recuperando detalles finos esenciales para perturbaciones imperceptibles.

La capa final utiliza una activación $\tanh$ para generar una máscara de ruido cruda, $\delta_{raw} \in [-1, 1]$. Esta se escala por un factor α para controlar la magnitud: $\delta = \alpha \cdot \delta_{raw}$. La imagen adversaria se obtiene mediante una **conexión residual aditiva**: $x_{adv} = clamp(x + \delta, -1, 1)$. Esta formulación residual simplifica la optimización al permitir que la red aprenda únicamente la sutil diferencia (δ) necesaria para la evasión.

3.2. Funciones Objetivo Híbridas

El entrenamiento del generador G no sigue un esquema adversarial tradicional. En su lugar, se optimiza una función de pérdida compuesta diseñada para ajustar de manera dinámica la realización de la tarea en 3 frentes, donde G debe satisfacer simultáneamente objetivos competitivos de fidelidad visual, evasión biométrica y disrupción de la atención. Esta formulación híbrida se inspira en enfoques exitosos de síntesis de imágenes que combinan pérdidas a nivel de píxel con pérdidas semánticas en espacios latentes (Isola et al., 2017; Johnson et al., 2016). La función de pérdida total se define como la suma ponderada de tres componentes:

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \lambda_{attn}\mathcal{L}_{attn} \tag{1}$$

donde λ_{rec}, λ_{id} y λ_{attn} son hiperparámetros que controlan la importancia relativa de cada objetivo. A continuación, se detalla y justifica cada componente.

3.2.1. Pérdida de Preservación Visual ($\mathcal{L}_{rec}$)

Para asegurar que la perturbación δ permanezca imperceptible y que la imagen adversaria x_{adv} retenga la máxima fidelidad respecto a la entrada original x, se emplea una pérdida de reconstrucción basada en la norma L_1. Siguiendo la evidencia presentada en trabajos de traducción de imagen a imagen (Isola et al., 2017), se selecciona la norma L_1 sobre la norma L_2, ya que esta última tiende a penalizar desproporcionadamente los grandes errores, lo que a menudo conduce a artefactos borrosos en la imagen generada. Por el contrario, la norma L_1 fomenta la dispersión de la perturbación en el dominio del píxel, preservando mejor los bordes y detalles de alta frecuencia esenciales para la fidelidad perceptual:

$$\mathcal{L}_{rec} = \mathbb{E}_{x\sim p_{data}(x)}[||G(x)||_1] = \mathbb{E}_{x\sim p_{data}(x)}[||x_{adv} - x||_1] \tag{2}$$

Minimizando $\mathcal{L}_{rec}$, se fuerza al generador a encontrar la perturbación más pequeña posible para lograr sus objetivos de evasión.

3.2.2. Pérdida de Evasión Biométrica ($\mathcal{L}_{id}$)

El objetivo central de este componente es suprimir la identidad biométrica presente en la imagen, manipulando su representación en el espacio latente de un reconocedor facial. Para ello, se emplea el extractor de características fijo F, basado en la arquitectura InceptionResnetV1 entrenada bajo el paradigma de FaceNet (Schroff et al., 2015), con pesos preentrenados en VGGFace2. Se obtienen los embeddings biométricos $e_x = F(x)$ y $e_{adv} = F(x_{adv})$.

Como se establece en (Schroff et al., 2015), estos sistemas están diseñados para que la similitud de coseno o distancia euclidiana en el espacio de embeddings corresponda directamente a la similitud facial. Por lo tanto, la pérdida de evasión se diseña para minimizar dicha similitud, desplazando el vector de identidad de la imagen adversaria hacia la ortogonalidad con respecto al vector original:

$$\mathcal{L}_{id} = \mathbb{E}_{x\sim p_{data}(x)}\left[CosineSimilarity(F(x), F(x_{adv}))\right] = \mathbb{E}_{x\sim p_{data}(x)}\left[\frac{F(x)\cdot F(x_{adv})}{\|F(x)\|_2\,\|F(x_{adv})\|_2}\right] \tag{3}$$

Al minimizar esta magnitud, el generador G aprende a producir perturbaciones que desplazan x_{adv} fuera de su clúster de identidad original en el espacio latente definido por F.

3.2.3. Pérdida de Disrupción de Atención ($\mathcal{L}_{attn}$)

Con el fin de aumentar la robustez y la transferibilidad del ataque, se introduce una función de pérdida diseñada para **interferir** en los mecanismos de atención intermedios de la red de reconocimiento. Esta técnica se basa en el concepto de pérdidas perceptuales o coincidencia de características (feature matching) (Johnson et al., 2016), donde se comparan las activaciones en capas profundas de una red preentrenada para capturar la semántica de la imagen en lugar de sus valores de píxel exactos.

Adaptando este enfoque para la disrupción, se extraen los mapas de características de una capa convolucional intermedia l del extractor F, denotados como $F^{(l)}(x) \in \mathbb{R}^{C\times H\times W}$. A partir de estas activaciones, se construye un mapa de atención espacial $A^{(l)}(x) \in \mathbb{R}^{H\times W}$ mediante el promedio del cuadrado de las activaciones a lo largo del canal, una técnica común para visualizar la relevancia espacial:

$$A^{(l)}(x) = \frac{1}{C}\sum_{c=1}^{C}\left(F_c^{(l)}(x)\right)^2.$$

La pérdida de disrupción de atención se define como el error cuadrático medio entre el mapa de atención de la imagen original y el correspondiente mapa generado por la imagen adversaria. Mientras que en tareas de superresolución se busca minimizar esta diferencia para preservar la semántica (Johnson et al., 2016), aquí se busca maximizar la divergenciapara alterar la percepción semántica de la red:

$$\mathcal{L}_{attn} = \mathbb{E}_{x\sim p_{data}(x)}\left[\left\|A^{(l)}(x) - A^{(l)}(x_{adv})\right\|_2^2\right]. \tag{4}$$

Al maximizar esta divergencia, el generador G se ve obligado a alterar las regiones espaciales que la red F considera más relevantes para la identificación, provocando una **fragmentación de gradiente** que dificulta la extracción coherente de características biométricas, incluso cuando la geometría facial permanece intacta.

4. Resultados

Se evalua la eficacia de ChameleonNet utilizando el conjunto de validación de CelebA-HQ (64×64). Los experimentos se realizaron tras 30 épocas de entrenamiento. Los resultados demuestran que nuestra arquitectura logra una disociación entre la identidad biométrica y la apariencia visual, validando la hipótesis de que es posible evadir el reconocimiento facial mediante perturbaciones residuales imperceptibles para el ojo humano.

4.1. Evaluación Cuantitativa y Convergencia

La Tabla 1 resume el rendimiento promedio del modelo en el conjunto de validación. Se utiliza la Similitud de Coseno de embeddings (FaceNet) como métrica de privacidad , y PSNR, SSIM y LPIPS como métricas de fidelidad visual.

Tabla 1: Métricas de rendimiento promedio en el conjunto de validación de CelebA-HQ. La identidad se reduce drásticamente mientras se mantiene una alta fidelidad estructural y perceptual.

Estado	ID Sim. ↓	PSNR ↑	SSIM ↑	LPIPS ↓	AUC ↓
Original	1.00	∞	1.00	0.00	1.00
Camuflado	**-0.0271**	**27.32 dB**	**0.8156**	**0.0450**	**0.3853**

Los resultados muestran una reducción drástica en la similitud de identidad, pasando de 1.0 a un promedio de -0.0271. Dado que los umbrales de verificación típicos en benchmarks estándar como LFW rondan el rango de 0.4-0.6 para mantener bajas tasas de falsa aceptación (Schroff et al., 2015), un valor negativo indica que la red biométrica percibe la imagen camuflada como diametralmente opuesta a la identidad original.

Simultáneamente, las métricas de fidelidad visual establecidas en la literatura de evaluación de calidad de imagen confirman que la degradación perceptual es mínima. Específicamente, se obtienen valores promedio de Índice de Similitud Estructural (SSIM) superiores a 0.8 (Wang et al., 2004) y distancias de Similitud Perceptual Aprendida (LPIPS) inferiores a 0.05 (Zhang et al., 2018), rangos que corresponden a diferencias difícilmente perceptibles para el observador humano promedio.

Al observar la dinámica del entrenamiento, se observa una rápida convergencia de la pérdida de identidad ($\mathcal{L}_{id}$) en las primeras épocas, estabilizándose junto con la pérdida de reconstrucción ($\mathcal{L}_{rec}$),

lo que demuestra la eficacia del esquema de optimización simultánea sin necesidad de fases de calentamiento.

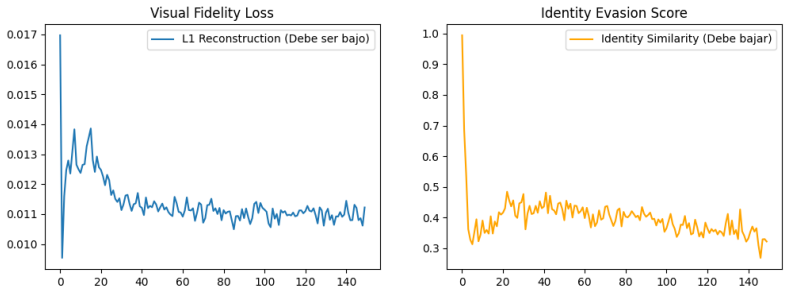

Figura 1: Dinámica de entrenamiento durante 30 épocas

4.2. Efectividad del Ataque en el Espacio Latente

Para visualizar el impacto del ataque en la distribución de características biométricas, se analiza el espacio latente del extractor FaceNet.

La Figura 2 presenta la curva ROC. Conforme a la metodología estándar de evaluación (Fawcett, 2006), un clasificador perfecto tendría un Área Bajo la Curva (AUC) de 1.0, mientras que la línea de base aleatoria corresponde a un AUC de 0.5. Nuestro ataque reduce el AUC a 0.3853, situándolo por debajo del azar. Esto indica que el sistema de reconocimiento facial se desempeña peor que una predicción aleatoria al intentar vincular las imágenes camufladas con sus identidades originales, confirmando la ruptura efectiva del enlace biométrico.

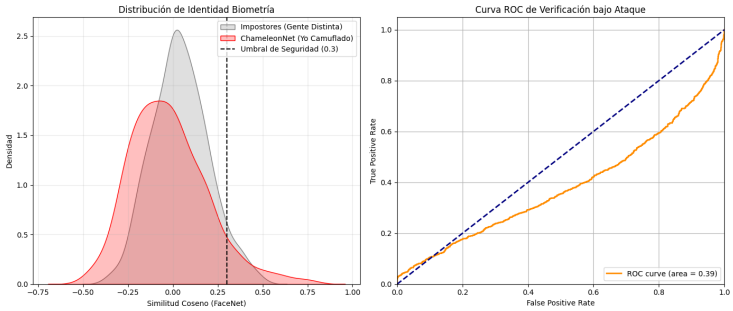

Figura 2: La curva ROC cae por debajo del azar (AUC=0.38), indicando que el clasificador falla sistemáticamente.

Este fenómeno se visualiza espacialmente en la proyección t-SNE (van der Maaten and Hinton, 2008). Mientras que las imágenes originales se agrupan densamente según su identidad, las versiones camufladas son expulsadas de sus clústeres originales y dispersadas en regiones distantes del espacio de embeddings, haciendo imposible su reidentificación mediante proximidad.

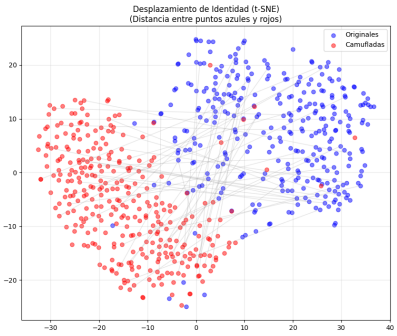

Figura 3: t-SNE muestra cómo las imágenes camufladas son dispersadas lejos de los clústeres de identidad original.

4.3. Análisis Forense y Mecanismos de Evasión

Figura 4: Análisis forense de casos de estudio (Best, Average, Worst, Edge).

Se utilizan técnicas de Inteligencia Artificial Explicable (XAI) para diseccionar el mecanismo de ataque en casos representativos (Figura 4).

4.3.1. Envenenamiento de Atención y Fragmentación de Gradiente

Los mapas de Grad-CAM (Selvaraju et al., 2017) (columnas 3 y 4) revelan un fenómeno de .^{en}venenamiento de atención". En casos exitosos, la atención de FaceNet se desplaza de regiones clave como los ojos hacia áreas menos informativas o se dispersa difusamente.

El análisis de Integrated Gradients (Sundararajan et al., 2017) presenta hallazgos aún más fundamentales. En una imagen normal, el IG delinearía contornos faciales suaves. En las imágenes camufladas, se observa "fragmentación de gradiente": los mapas de relevancia se vuelven ruidosos y granulados.

Esto exhibe que la red víctima ya no percibe una estructura facial coherente, sino un cúmulo de píxeles de alta frecuencia inyectados por el generador, lo que impide la extracción de características biométricas que revelen estructuras faciales. La columna 6 confirma que el ruido inyectado (δ) no es aleatorio, sino que posee una estructura semántica alineada con los rasgos faciales, posible gracias a las conexiones de salto de la U-Net.

4.4. Validación de Integridad del Modelo

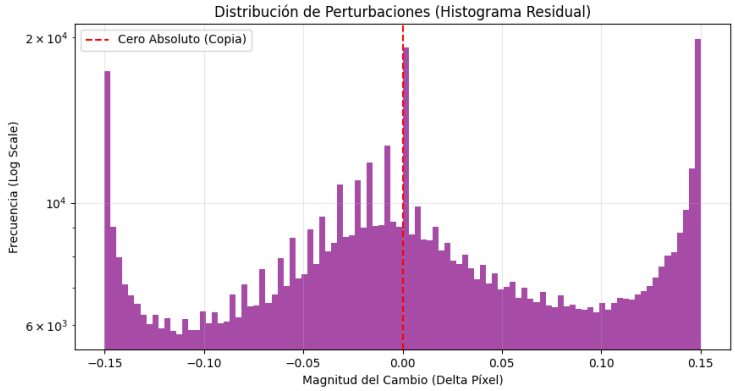

Figura 5: Histograma de residuos de píxel (δ). La distribución bimodal en los extremos (± 0.15) y la casi nula concentración en cero confirman una perturbación activa generalizada.

Un riesgo inherente al entrenar generadores bajo estrictas restricciones de fidelidad visual es el colapso de identidad, un modo de fallo donde el generador converge hacia la función trivial $G(x) \approx x$, minimizando la pérdida de reconstrucción pero fallando en el ataque.

Para descartar este fenómeno, se analiza la distribución estadística de las perturbaciones residuales a nivel de píxel, definidas como $\delta = x_{adv} - x$. La Figura 5 muestra el histograma de estos residuos para el conjunto de validación. Lejos de una concentración gaussiana en torno al cero (que indicaría inactividad), observamos una clara distribución bimodal saturada en los límites de la escala de ruido permitida (e.g., ± 0.15).

El diagnóstico cuantitativo corrobora esta observación visual: la diferencia media absoluta de píxel es de 0.0724, y apenas un $1.33\,\%$ del total de píxeles permanecen matemáticamente idénticos a la entrada. Esta evidencia confirma que ChameleonNet no colapsa hacia la identidad; por el contrario, el modelo optimiza agresivamente el "presupuesto"de perturbación disponible en toda la extensión espacial de la imagen para maximizar la distancia biométrica.

5. Discusión

Los resultados obtenidos con ChameleonNet validan empíricamente la hipótesis central de este trabajo: la identidad biométrica percibida por una red neuronal profunda no está inextricablemente ligada a la apariencia visual macroscópica de un rostro. Se ha demostrado que es posible alterar la materia matemática de la imagen como sus características profundas y gradientes, hasta el punto de la irreconocibilidad algorítmica, mientras se preserva intacta su forma perceptual para el observador humano.

El éxito de esta disociación no reside en la potencia bruta del generador U-Net, sino en la naturaleza adversaria híbrida de la función de objetivo. Los experimentos sugieren que la optimización simultánea en múltiples frentes es crucial. Si solo se optimizara la evasión biométrica ($\mathcal{L}_{id}$), el modelo colapsaría rápidamente hacia la destrucción de la imagen. Si solo se priorizara la reconstrucción ($\mathcal{L}_{rec}$), colapsaría hacia la función identidad.

El factor diferenciador clave es la introducción de la Pérdida de Disrupción de Atención. Como evidencian los análisis forenses de Gradient Shattering, esta pérdida actúa como un veneno cognitivo que impide que el extractor de características fijo se enfoque coherentemente en rasgos faciales salientes. Esto fuerza al generador a encontrar soluciones en un espacio latente estrecho: aquel donde las perturbaciones son visualmente invisibles pero semánticamente catastróficas .

Esto tiene implicaciones profundas para la seguridad y la privacidad. Demuestra que los sistemas de reconocimiento facial actuales, a pesar de su precisión sobrehumana, son frágiles ante ataques estructurados de alta frecuencia. ChameleonNet ofrece una prueba de concepto de que la privacidad en la era digital no requiere necesariamente la ofuscación visual, sino que puede lograrse mediante la alteración selectiva de la información que las máquinas consumen, permitiendo a los usuarios .[es]conderse a plena vista".

6. Conclusiones

El objetivo principal de disociar la identidad biométrica de la apariencia física sin degradación perceptual se alcanzó exitosamente mediante el desarrollo de ChameleonNet. Primero, se demostró la viabilidad técnica de una arquitectura generativa adversaria residual capaz de anonimizar rostros en un solo paso de inferencia, prescindiendo de módulos externos de segmentación o control.

Los resultados validan la hipótesis de evasión, logrando reducir la similitud de identidad a niveles de desconocido mientras se mantiene una alta fidelidad visual en escenarios de caja blanca. Finalmente, aunque la evasión es robusta contra el modelo objetivo, se identificó que la transferibilidad a arquitecturas de caja negra (como HOG) es limitada, lo que define la necesidad de investigar el entrenamiento contra ensambles de modelos como trabajo futuro.

Referencias

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Hui Guo and et al. Pgt-gan: Perpetual generative tensor-based adversarial networks for face anonymization. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2022.

Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic and controllable face anonymization. *arXiv preprint arXiv:2211.09454*, 2023.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*, 2016.

Maximilian Lange and et al. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 618–626, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328. PMLR, 2017.

L. Tran and et al. Mask-invariant face recognition through template-level fusion. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Bangjie Yin and et al. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 586–595, 2018.