THÔNG TIN CHUNG

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://youtu.be/R-hkp3pw6pk
- Link slides (dạng .pdf đặt trên Github):
  https://github.com/taileuit/ppnckh/blob/main/Presentation.pdf

- Họ và Tên: Lê Minh Tài
- MSHV: 240202026

- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 8/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 9
- Link Github: https://github.com/taileuit/ppnckh

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ỨNG DỤNG MÔ HÌNH ĐỐI KHÁNG ĐỂ BẢO VỆ DANH TÍNH TRÊN ẢNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

APPLICATION OF ADVERSARIAL MODELS FOR IDENTITY PROTECTION IN IMAGES

TÓM TẮT *(Tối đa 400 từ)*

Trong bối cảnh công nghệ nhận dạng khuôn mặt đang ngày càng được triển khai rộng rãi trong các hệ thống giám sát, an ninh và nền tảng trực tuyến, những lo ngại về nguy cơ rò rỉ danh tính cá nhân trở nên ngày một rõ rệt. Thực tế cho thấy chỉ một bức ảnh khuôn mặt được công khai cũng bị khai thác để truy vết hoặc phục vụ cho các mục đích không mong muốn. Điều này đặt ra yêu cầu cấp thiết đối với các giải pháp bảo vệ quyền riêng tư trên ảnh số, đặc biệt là ảnh chân dung.

Đồ án này tiếp cận vấn đề thông qua việc nghiên cứu và ứng dụng CLIP2Protect – một phương pháp tạo ảnh đối kháng có định hướng, kết hợp mô hình CLIP với StyleGAN2. Cốt lõi của phương pháp nằm ở việc sử dụng mô tả văn bản như một tín hiệu điều khiển để sinh ra các biến thể khuôn mặt vừa mang yếu tố thẩm mỹ theo yêu cầu (ví dụ: “mắt khói”, “son đỏ”), vừa làm suy giảm đáng kể khả năng nhận dạng của các thuật toán học máy. Đây là hướng tiếp cận mới mẻ, khi kết hợp giữa bảo vệ danh tính và yếu tố cá nhân hóa thông qua tương tác ngôn ngữ.

Trong quá trình triển khai, tác giả xây dựng hệ thống từ mã nguồn công khai, thử nghiệm trên tập dữ liệu khuôn mặt phổ biến, và tiến hành đánh giá dựa trên ba tiêu chí chính: tỷ lệ giảm độ chính xác nhận diện, mức độ giữ nguyên thị giác tự nhiên, và tính khả dụng trong môi trường thực tế.

Đồ án hướng đến việc cung cấp một góc nhìn thực nghiệm rõ ràng cùng công cụ trực quan, góp phần nâng cao nhận thức và hỗ trợ kỹ thuật cho bài toán bảo vệ danh tính khuôn mặt trong kỷ nguyên số.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong thời đại AI phát triển, hình ảnh khuôn mặt cá nhân khi chia sẻ trên mạng xã hội có khả năng bị thu thập và nhận dạng trái phép. Điều này đặt ra nhu cầu bảo vệ quyền riêng tư trong ảnh kỹ thuật số, đặc biệt là các giải pháp không làm giảm chất lượng hình ảnh và vẫn đảm bảo tính cá nhân hóa.

Một giải pháp mới đầy tiềm năng là CLIP2Protect, được giới thiệu tại CVPR 2023. Phương pháp này tạo ra ảnh đối kháng – hình ảnh được chỉnh sửa để làm giảm khả năng nhận diện của AI nhưng vẫn giữ được vẻ tự nhiên với mắt người. CLIP2Protect kết hợp mô hình sinh ảnh StyleGAN2 và mô hình ngôn ngữ–hình ảnh CLIP (mô tả văn bản như "son đỏ", "mắt khói", v.v.) từ đó tạo ra ảnh chân dung mới được "trang điểm".

Không giống các phương pháp khác, CLIP2Protect không cần can thiệp vào hệ thống AI nhận dạng, mà cho phép người dùng chủ động bảo vệ khuôn mặt của mình. Ảnh đầu ra vừa đáp ứng tính thẩm mỹ, vừa có khả năng đánh lừa các hệ thống nhận dạng như Face++ hay Aliyun.

Phạm vi áp dụng của phương pháp tập trung vào ảnh chân dung đơn lẻ, đặc biệt là ảnh selfie hoặc ảnh đại diện được chia sẻ công khai. Đây là các trường hợp phổ biến nhất mà người dùng cá nhân có nguy cơ bị thu thập dữ liệu khuôn mặt.

Đề tài này nhằm cài đặt, kiểm thử và đánh giá CLIP2Protect trong thực tế, đặc biệt trên mạng xã hội. Quá trình xử lý gồm:

- Input: Ảnh chân dung + mô tả phong cách/trang điểm khuôn mặt.
- Output: Ảnh đã chỉnh sửa theo mô tả, có tính đối kháng, khó bị AI nhận diện nhưng vẫn đẹp và tự nhiên.

Thử nghiệm ban đầu cho thấy CLIP2Protect chưa tối ưu về độ ẩn danh trong mọi trường hợp, tuy nhiên vẫn đảm bảo tính thẩm mỹ và đúng mô tả văn bản. Đây là bước tiến quan trọng trong cân bằng giữa bảo mật và trải nghiệm người dùng.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

Đề tài hướng đến việc khám phá và đánh giá hiệu quả thực tiễn của mô hình CLIP2Protect trong việc bảo vệ danh tính trên ảnh khuôn mặt. Các mục tiêu cụ thể bao gồm:

1. Tìm hiểu và phân tích mô hình CLIP2Protect, bao gồm cách kết hợp giữa CLIP (hiểu ngôn ngữ–hình ảnh) và StyleGAN2 (sinh ảnh), để tạo ra các ảnh khuôn mặt đối kháng được điều hướng bằng mô tả văn bản. Mục tiêu là làm rõ cơ chế kỹ thuật giúp mô hình vừa làm suy giảm khả năng nhận diện của AI, vừa giữ được tính thẩm mỹ của ảnh đầu ra.
2. Xây dựng một hệ thống thử nghiệm mô hình CLIP2Protect hoạt động ổn định, sử dụng mã nguồn công khai:
    - Thiết lập môi trường lập trình (Google Colab hoặc máy cục bộ có GPU).
    - Xử lý các ảnh khuôn mặt đầu vào từ tập dữ liệu công khai (như CelebA-HQ) kết hợp với mô tả văn bản ("son đỏ", "mắt khói",...).
    - Sinh ảnh đầu ra và đánh giá hiệu quả bằng các chỉ số định lượng (tỉ lệ nhận diện đúng, FID Score, mức độ phù hợp với mô tả).
3. Đánh giá tính khả dụng và tiềm năng ứng dụng của hệ thống này trong đời sống thực tế, đặc biệt là với người dùng cá nhân trên mạng xã hội. Mục tiêu cuối cùng là hướng tới phát triển một công cụ thân thiện, dễ sử dụng giúp người dùng:
    - Chủ động bảo vệ danh tính cá nhân trên ảnh đại diện/selfie.
    - Cá nhân hóa phong cách hình ảnh mà vẫn đảm bảo an toàn dữ liệu khuôn mặt.
    - Giảm rủi ro bị khai thác trái phép bởi các hệ thống nhận dạng khuôn mặt AI.

NỘI DUNG VÀ PHƯƠNG PHÁP

Mô hình CLIP2Protect có thể làm suy giảm hiệu quả nhận dạng khuôn mặt của các hệ thống AI như thế nào mà vẫn duy trì tính thẩm mỹ tự nhiên của ảnh? Mức độ chính xác của ảnh đầu ra so với mô tả văn bản đầu vào là bao nhiêu, và ảnh hưởng thế nào đến mức độ nhận diện của AI? Hay CLIP2Protect có tiềm năng được ứng dụng thực tế như một công cụ bảo vệ quyền riêng tư cho người dùng phổ thông không? Để trả lời các câu hỏi này thì đồ án tập trung vào việc nghiên cứu phương pháp bảo vệ quyền riêng tư khuôn mặt hiện đại – CLIP2Protect – thông qua quá trình cài đặt, chạy thử và phân tích hiệu quả của mô hình. Nội dung chính của đề tài được triển khai theo các bước sau, kèm theo tiêu chí đánh giá thành công cho từng giai đoạn:

1. **Nghiên cứu lý thuyết và phân tích mô hình CLIP2Protect**

- Đọc và tổng hợp nội dung từ các bài báo gốc liên quan: CLIP2Protect (CVPR 2023), StyleGAN2 (CVPR 2020), CLIP (ICML 2021).
- Phân tích pipeline kỹ thuật của mô hình: cách sinh latent code ban đầu, tối ưu hóa theo mô tả văn bản, sử dụng loss function từ CLIP và từ mô hình nhận dạng.
- So sánh với các phương pháp bảo vệ khuôn mặt khác như AdvMakeup, TIP-IM, AMT-GAN.

2. **Cài đặt và chạy thử mô hình trên tập dữ liệu thực nghiệm**

- Tải mã nguồn mô hình từ GitHub chính thức của tác giả.
- Thiết lập môi trường thực thi trên Google Colab hoặc máy tính cục bộ có hỗ trợ GPU.
- Chuẩn bị tập dữ liệu ảnh chân dung công khai (ví dụ: CelebA-HQ).
- Chạy thử mô hình với các mô tả văn bản như “mắt khói”, “tóc highlight”, “son đỏ” để sinh ảnh đối kháng.

3. **Đánh giá định lượng hiệu quả mô hình**

So sánh ảnh gốc và ảnh đã xử lý bằng

- Tỉ lệ nhận dạng đúng trước và sau (nếu truy cập được API như Face++, Aliyun).
- Chỉ số FID Score để đánh giá mức độ khác biệt về thị giác.
- Độ chính xác giữa ảnh đầu ra và mô tả văn bản (có thể dùng CLIP cosine similarity hoặc đánh giá chủ quan).

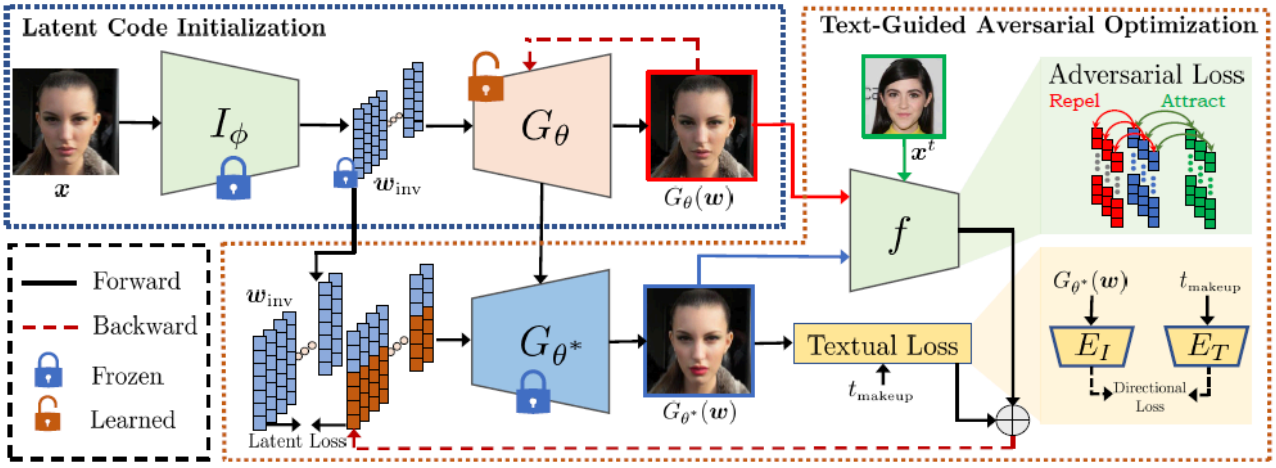
4. **Phân tích kết quả và đề xuất cải tiến**

- Tổng hợp dữ liệu đầu ra (ảnh, số liệu) để phân tích hiệu quả mô hình.
- Viết báo cáo đánh giá: ưu điểm, hạn chế, khả năng triển khai thực tế.
- Đề xuất mở rộng như: xây dựng giao diện người dùng, tích hợp vào nền tảng mạng xã hội, khảo sát trải nghiệm người dùng.

KẾT QUẢ MONG ĐỢI

1. **Hiểu rõ kiến trúc và quy trình hoạt động của mô hình.**

Sơ đồ dưới đây mô tả tổng quan quy trình của CLIP2Protect, gồm hai giai đoạn chính: khởi tạo mã ẩn từ ảnh gốc dựa trên mô tả văn bản, và tối ưu hóa ảnh đầu ra để vừa đảm bảo tính thẩm mỹ, vừa làm giảm khả năng nhận diện của AI.



Hình 1. Pipeline mô hình CLIP2Protect gồm hai bước: (1) khởi tạo mã ẩn (latent code initialization), và (2) tối ưu hóa ảnh có điều hướng bằng văn bản (text-guided adversarial optimization). Mô hình sử dụng CLIP để đo độ tương thích giữa ảnh và

văn bản, đồng thời duy trì tính đối kháng với hệ thống nhận diện.

Nguồn: Shamshad et al., "CLIP2Protect: Text-Guided Visual Privacy Protection via Adversarial Makeup", CVPR 2023.

	Adv-Makeup [71]	TIP-IM [70]	AMT-GAN [22]	Ours
Natural outputs	Yes	Partially	Partially	Yes
Black box	Yes	Yes	Yes	Yes
Verification	Yes	No	Yes	Yes
Identification	No	Yes	No	Yes
Unrestricted	Yes	No	Yes	Yes
Text guided	No	No	No	Yes

Bảng 1. So sánh một số phương pháp bảo vệ khuôn mặt phổ biến, bao gồm Adv-Makeup, TIP-IM, AMT-GAN và CLIP2Protect (Ours). CLIP2Protect là phương pháp duy nhất hỗ trợ điều khiển bằng văn bản, hoạt động tốt trong môi trường black-box, đồng thời cho phép người dùng cá nhân hóa ảnh đầu ra theo ý muốn.

Nguồn: Shamshad et al., "CLIP2Protect: Text-Guided Visual Privacy Protection via Adversarial Makeup", CVPR 2023.

2. **Xây dựng được hệ thống thử nghiệm mô hình CLIP2Protect**

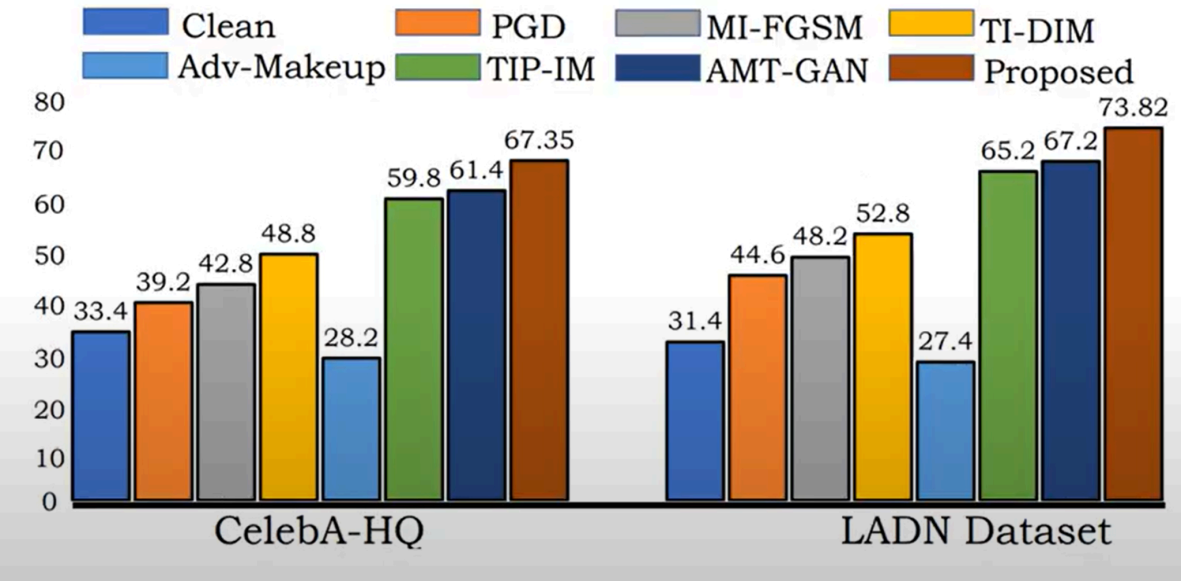
Đảm bảo tính hoạt động ổn định của thử nghiệm, từ đó mở rộng các phiên bản chính thức cho phép người dùng nhập vào ảnh khuôn mặt và mô tả văn bản để tạo ra ảnh đối kháng có tính chất bảo vệ danh tính, đồng thời vẫn giữ được tính thẩm mỹ tự nhiên của ảnh.

3. **Chứng minh được khả năng bảo vệ danh tính của phương pháp**

Phương pháp hiệu quả khi đo lường thông qua tỷ lệ nhận diện đúng giảm từ 30–50% so với ảnh gốc khi thử nghiệm trên các API nhận diện như Face++. FID Score dưới 50, thể hiện ảnh đầu ra vẫn giữ được mức độ tự nhiên cao. Và mức độ tương thích giữa ảnh và mô tả văn bản đầu vào cao (có thể đo bằng cosine similarity từ CLIP).

Hình minh họa dưới đây thể hiện kết quả đánh giá của CLIP2Protect so với các phương

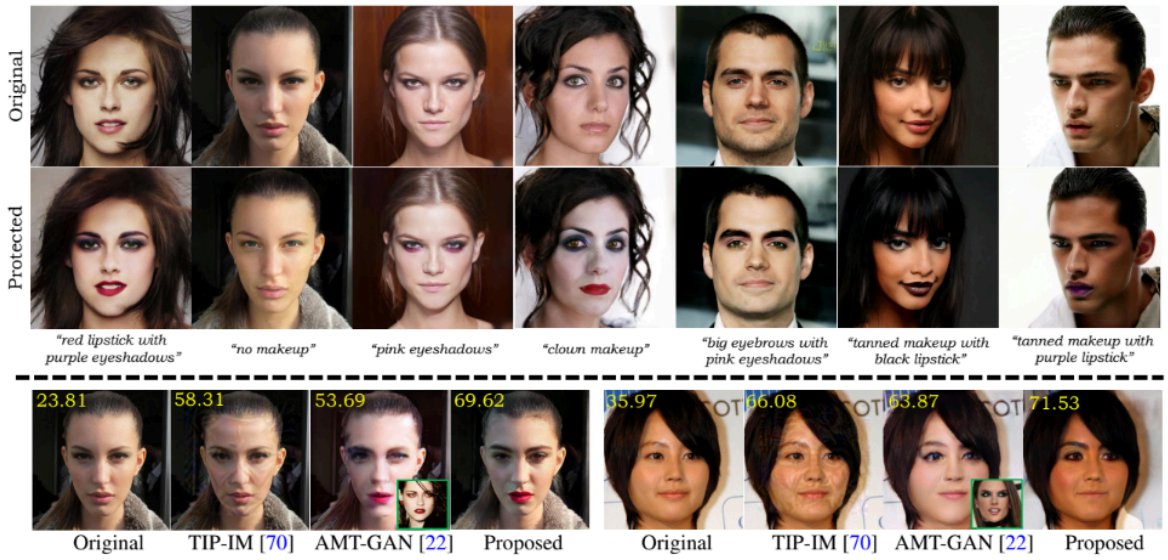
pháp bảo vệ danh tính khác (TIP-IM, AMT-GAN, v.v.) trên hai tập dữ liệu phổ biến. Biểu đồ cho thấy CLIP2Protect đạt điểm tự tin cao nhất trong tấn công giả mạo, cho thấy hiệu quả thuyết phục trong môi trường đánh giá thực tế:



Hình 2. Kết quả đánh giá trung bình từ API Face++ đối với các phương pháp khác nhau trong bối cảnh tấn công giả mạo. "Proposed" là mô hình CLIP2Protect. Cao hơn nghĩa là hình ảnh gây nhầm lẫn tốt hơn, và mô hình đề xuất vượt trội so với các phương pháp trước đó cả trên tập CelebA-HQ và LADN.

4. **Xác định được tiềm năng ứng dụng thực tế, hạn chế và khả năng mở rộng của mô hình CLIP2Protect**

Tiềm năng có CLIP2Protect là rất lớn trong việc bảo vệ quyền riêng tư hình ảnh, đặc biệt là trong môi trường mạng xã hội hoặc các nền tảng chia sẻ công khai. Kết quả của đồ án sẽ góp phần minh họa một giải pháp đơn giản, thân thiện với người dùng nhưng có hiệu quả rõ rệt trong việc hạn chế nguy cơ nhận dạng trái phép từ các hệ thống AI hiện đại.



Hình 3. *So sánh ảnh gốc (Original), ảnh đã được xử lý bằng các phương pháp bảo vệ danh tính gồm TIP-IM, AMT-GAN, và CLIP2Protect (Proposed). Chú thích màu vàng hiển thị tỷ lệ nhận diện chính xác của hệ thống Face++ (càng thấp càng tốt). Phương pháp CLIP2Protect đôi khi không đạt mức nhận diện thấp nhất, tuy nhiên ảnh đầu ra đạt tính thẩm mỹ cao hơn và thể hiện rõ đặc điểm trang điểm theo văn bản mô tả, chẳng hạn như "pink eyeshadows" hoặc "tanned makeup with purple lipstick". Điều này phản ánh định hướng của CLIP2Protect trong việc cân bằng giữa khả năng bảo vệ danh tính và yếu tố cá nhân hóa, thân thiện với người dùng.*
Nguồn: Shamshad et al., "CLIP2Protect: Text-Guided Visual Privacy Protection via Adversarial Makeup", CVPR 2023.

Mặc dù CLIP2Protect thể hiện hiệu quả trong việc giảm khả năng nhận diện khuôn mặt, tuy nhiên hệ thống hiện tại vẫn còn một số hạn chế về khả năng tiếp cận và mức độ ứng dụng thực tế. Một số hướng cải tiến có thể xem xét:

- Tích hợp giao diện người dùng trực quan: Phát triển một ứng dụng web đơn giản (dựa trên Streamlit hoặc Flask) cho phép người dùng tải ảnh, nhập mô tả, và nhận ảnh đã được xử lý.
- Tăng cường đánh giá từ người dùng: Thực hiện khảo sát trực tuyến để thu thập phản hồi về mức độ thẩm mỹ và mức độ hài lòng với ảnh đầu ra, từ đó đánh giá

tính chấp nhận của người dùng cuối.

- Khảo sát khả năng tích hợp thực tế: Đề xuất một kịch bản thử nghiệm tích hợp CLIP2Protect như một lớp xử lý ảnh trước khi người dùng đăng lên nền tảng mạng xã hội như Facebook, Twitter.

Những mở rộng này có thể nâng cao tính ứng dụng của mô hình, từ môi trường nghiên cứu sang công cụ hỗ trợ thực tế cho người dùng đại chúng trong việc bảo vệ danh tính trên ảnh số.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar.
Clip2Protect: Protecting Facial Privacy Using Text-Guided Makeup via Adversarial Latent Search.
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20595–20605.

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila.
Analyzing and Improving the Image Quality of StyleGAN.
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever.
Learning Transferable Visual Models From Natural Language Supervision.
arXiv preprint arXiv:2103.00020, 2021.

[4] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, Libing Wu.
Protecting Facial Privacy: Generating Adversarial Identity Masks via Style-Robust Makeup Transfer.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15014–15023.