프라이버시 보호를 위한 얼굴 부위 조합 딥페이크 생성기법

팀 교수님제주도에서는차긁으시면안됩니다

1. 개요

- 최근 딥페이크 기술의 발전은 영상 합성의 현실감을 극대화하며 다양한 콘텐츠 창작에 활용되고 있다[1]. 그러나 동시에 개인의 얼굴 정보가 무단으로 사용되거나, 원치 않는 영상 생성에 악용될 우려가 커지고 있다. 이에 따라, 프라이버시 보호를 위한 얼굴 변형 및 익명화 기술의 중요성이 부각되고 있다[2-3].
- **실제 인물의 얼굴을 직접적으로 활용하지 않으면서도, 현실적이며 자연스러운 얼굴 합성 비디오를 생성**하는 새로운 방식의 딥페이크 생성 기법이 필요하다.
- 본 보고서에서는 프라이버시 침해에 악용되던 딥페이크 기술을 역으로 **프라이버시 보호에 활용하고자, 여러 사람의 얼굴 부위를 조합하여 새로운 얼굴을 합성하고, 이를 기반으로 고화질 비디오를 생성하는 프라이버시 보호형 딥페이크 생성 기법**을 제안한다.
- 전체 얼굴을 하나의 인물로부터 복제하는 기존의 방식과 달리, 본 연구는 얼굴의 주요 부위(눈, 코, 입 등)를 다수의 서로 다른 인물로부터 선택적으로 조합함으로써, 합성된 얼굴이 특정 개인에 귀속되지 않도록 설계되었다[4-8].
- 그림 1은 본 팀에서 제안한 기법의 개념도이다.

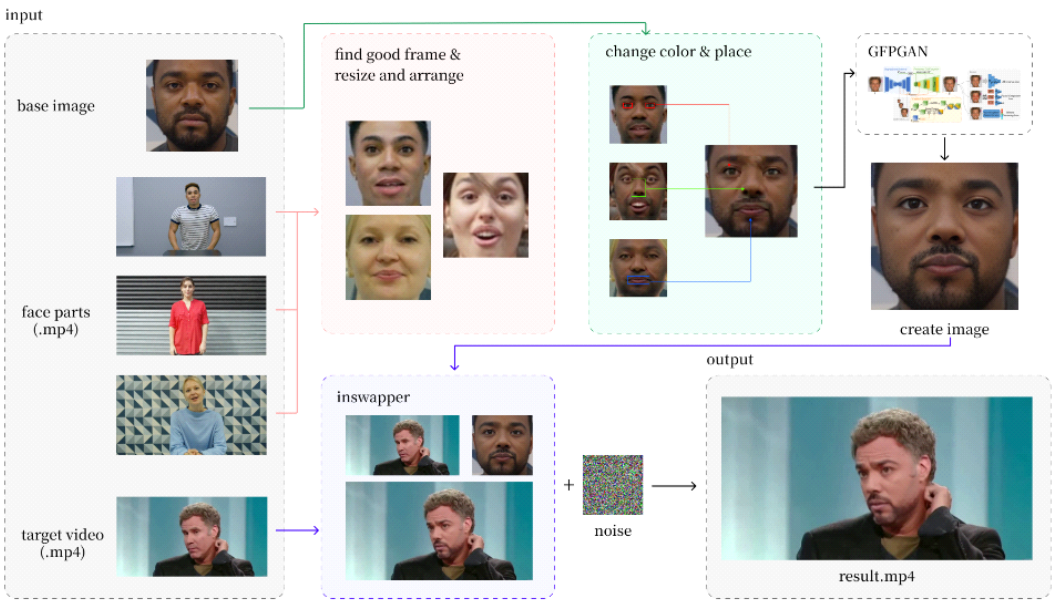

그림 1 제안한 기법의 개념도

2. 기존 기법의 한계와 본 연구의 필요성

- 2024년 NeurIPS에서 발표된 Stable Diffusion 기반 얼굴 부위 조합 기법인 FuseAnyPart는, 다수의 참조 이미지로부터 특정 얼굴 부위를 선택적으로 추출하여 하나의 타겟 이미지에 조합하는 방식의 얼굴 부위 교체(facial part swapping) 기술이다[9].
- 하지만 얼굴 파트를 자연스럽게 결합하려면, 각 부위를 정밀하게 추출하고 정렬하는 전처리 과정이 필수적이다[10]. 그럼에도 불구하고 FuseAnyPart는 얼굴 정렬(face

alignment)이나 피부톤 보정(color harmonization)과 같은 핵심적인 처리 과정을 포함하지 않아, 합성 결과의 품질에 한계가 있다.

- **실제로 FuseAnyPart를 통해 조합된 얼굴은 파트 간의 위치, 크기, 색조 등이 서로 맞지 않아 시각적으로 이질적인 결과를 발생시킨다.** 또한, FuseAnyPart는 **고화질의 정적 이미지 데이터를 기반으로 학습되었기 때문에, 실제 환경에서 사용되는 데이터의 다양성과 품질 저하 요소들을 충분히 반영하지 못하고 있다.** 특히 서로 다른 해상도와 촬영 조건을 가진 비디오 간 얼굴 파트를 조합할 경우의 품질이 현저히 낮다.
- 실제 실험에서도 눈, 코, 입과 같은 주요 얼굴 요소들이 상대적으로 어긋나거나 비율이 맞지 않아 얼굴의 일부가 과도하게 부각되는 현상이 관찰되었다.
- 더욱이, 현재까지 공개된 얼굴 부위 교체 기법 중 비디오를 대상으로 하는 연구는 거의 전무하다. 대부분의 연구는 정적 이미지 기반으로 설계되어 있으며, 이러한 점은 facial part swapping 기술의 실제 활용 가능성을 크게 제한하는 요소로 작용한다.
- 이에 따라 **본 연구는 보다 정교하고 자연스러운 얼굴 부위 조합을 가능하게 하기 위해 구조적인 전처리 절차와 생성 파이프라인을 새롭게 설계하였다**. 이 파이프라인은 얼굴 정렬, 피부색 정규화, 그리고 선명도 향상과 같은 후처리 기법을 포함하여, 고품질의 합성 결과를 생성할 수 있도록 설계되었다.
- 그림 2는 기존 FuseAnyPart 기법과 본 연구에서 제안한 기법을 동일한 입력 조건 하에서 비교한 이미지로, **제안한 방법이 기존의 FuseAnyPart 기법에 비해 조합된 얼굴 부위 간의 정렬 정확도 및 시각적 자연스러움 측면에서 우수한 성능을 나타냄을 명확히 확인할 수 있다.**

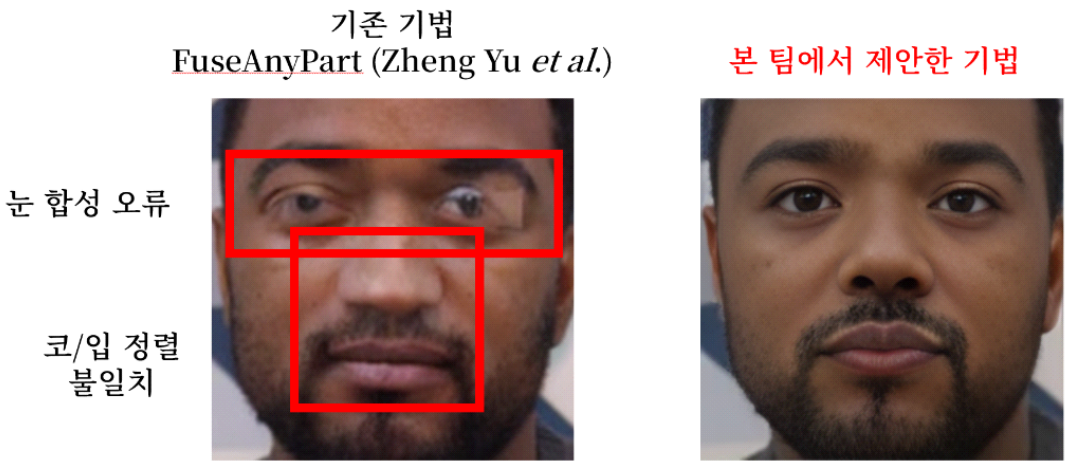

그림 2 기존 기법과 제안한 기법의 비교

3. 제안 기법 설명

- 제안하는 기법은 총 7단계로 구성되어 있으며, 그림 3부터 그림 9까지는 각 단계의 구성 요소와 처리 과정을 시각적으로 설명한다.
- 전체 파이프라인은 초기 얼굴 감지 및 얼굴 부위 추출 단계부터 시작하여, 개별 파트의 정렬 및 조합, 합성 이미지 생성, 최종적으로 딥페이크 비디오를 생성하는 과정을 포함한다.
- 본 연구에서는 해당 모듈의 요구사항과 호환성을 고려하여, 그에 가장 적합한 모델과 기법을 적용하였다. 이러한 방식을 통해 각 모듈의 안정성과 성능을 극대화하는 데 중점을

두었다.

- **1단계: 원본 비디오에서 프레임 추출**
  - 입력된 비디오로부터 정면을 바라보고, 눈을 뜨고 있으며 입을 다문 상태가 동시에 나타나는 첫 번째 프레임을 선택하여 분석 대상 프레임으로 지정하였다. 이는 표정의 왜곡이 적고, 눈·코·입의 위치가 안정적이기 때문에 정확한 얼굴 정렬 및 파트 분리 기준점 확보에 유리하다.
  - 얼굴 감지에는 작은 얼굴까지 식별이 가능한 Adrian Bulat의 S3FD 모델, 랜드마크 추출에는 얼굴의 깊이 정보를 고려한 3DFAN 모델을 사용하였다.
  - 이 단계는 **후속 얼굴 정렬 및 부위 분리 작업의 기준 프레임 확보**를 목적으로 한다.

그림 3 (1단계) 원본 비디오에서 프레임 추출

- **2단계: 추출한 이미지에서 얼굴 감지 및 정렬**
  - 입력 영상 프레임에서 얼굴을 감지한 후 해당 영역을 크롭하고, 눈, 코, 입 등의 주요 랜드마크를 기반으로 얼굴 부위를 정렬하여, **다양한 얼굴 각도나 위치 변화에도 일관된 기준으로 얼굴 부위를 추출한다**.
  - 얼굴 탐지에는 YOLOv8 또는 MTCNN을 사용하며, 정밀한 랜드마크 추출 및 정렬을 위해 dlib를 이용한다. 먼저 dlib+yolo 조합으로 얼굴 추출 및 얼굴 랜드마크를 탐지한다. 이때 dlib+yolo로 얼굴 인식이 불가능할 경우 MTCNN의 facenet_pytorch 기반의 landmark detector를 활용해 얼굴 추출 및 랜드마크를 탐지하고, 탐지 결과를 바탕으로 얼굴 정렬을 진행한다.

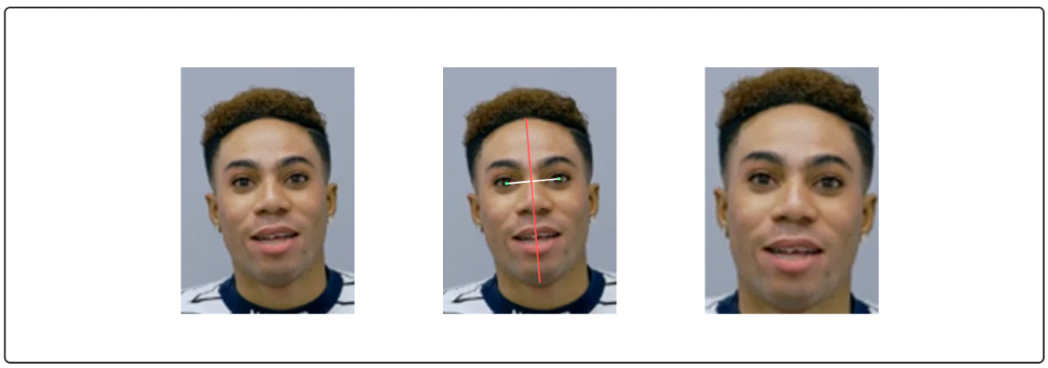

그림 4 (2단계) 추출한 이미지에서 얼굴 감지 및 정렬

- **3단계: 이미지 전처리**
  - 얼굴 부위 배치 전, 입력 이미지에서 해상도가 낮거나 크기가 작은 얼굴은 그대로 조합할 경우 형태가 깨지거나 인식 오류가 발생할 수 있다.
  - 이를 방지하기 위해, GFPGAN 기반의 얼굴 복원 모델을 적용하여 저해상도 얼굴의 해상도를 보간하고, 윤곽 및 질감 등의 디테일을 복원한다[11].
  - 이 과정은 작은 얼굴 부위를 고해상도로 전처리함으로써, 이후 합성 단계에서 보다 자연스럽고 일관된 결과를 도출할 수 있도록 돕는다.

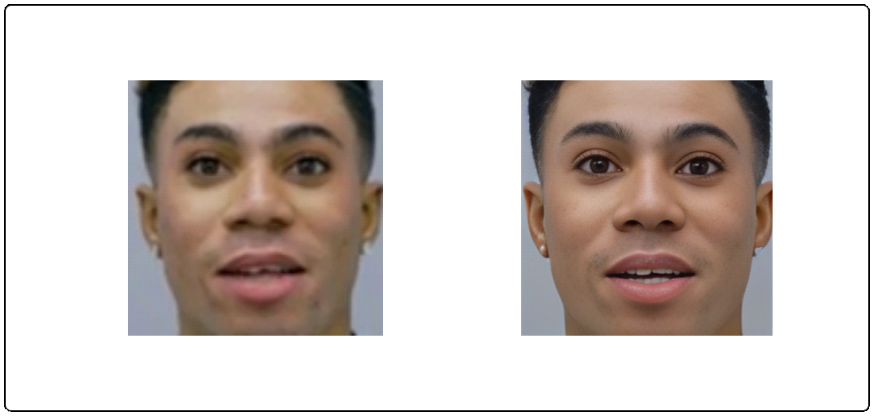

그림 5 (3단계) 이미지 전처리

- **4단계: 피부톤 변경 및 얼굴 부위 조합**
  - YOLOv8 기반 얼굴 부위 탐지를 통해 **얼굴 부위를 추출하고 원본 비디오의 얼굴에 조합함으로써, 특정 개인의 얼굴이 아닌 합성된 얼굴 이미지를 생성**한다.
  - 얼굴 부위를 정밀하게 분리하기 위해 BiSeNet기반의 얼굴 파싱(parsing)을 수행한다.
  - 눈, 코, 입 등의 얼굴 파츠는 각기 다른 영상에서 랜덤으로 선택되며, 부위 간 연결 부위에서 발생할 수 있는 피부색 이질감을 최소화하기 위해 CSD-MT모델을 사용하여 피부톤을 통일한다.

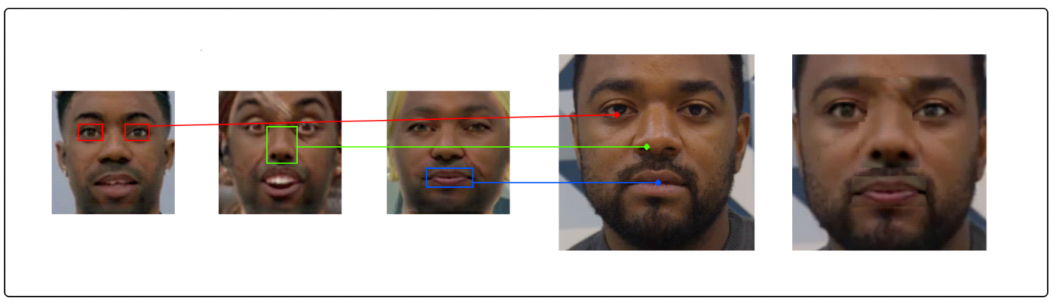

그림 6 (4단계) 원본 비디오에서 프레임 추출

- **5단계: 이미지 정제**
  - 본 단계에서는 이전 단계에서 생성된 합성 얼굴 이미지의 선명도 향상 및 시각적 자연스러움 개선을 위해 GFPGAN을 적용한다[11].

- **생성된 얼굴 이미지의 경계 영역을 자연스럽게 연결하고, 피부 톤과 세부 질감을 자연스럽게 보정하여 고해상도/시각적으로 안정된 얼굴 합성 결과**를 얻을 수 있다.

그림 7 (5단계) 이미지 정제

- **6단계: 비디오 속 사람의 얼굴 교체**
  - YOLOv8-nano와 ArcFace 기반 ResNet-50 얼굴 임베딩 모델을 이용하여, 참조 이미지로부터 얼굴 임베딩 벡터를 추출한다. 해당 벡터는 ArcFace 방식의 고차원 특징 표현으로, 참조 인물의 정체성을 효과적으로 내포하고 있다.
  - 추출된 임베딩 벡터를 InsightFace에서 제공하는 얼굴 교체 모델(inswapper)에 입력함으로써, 비디오의 얼굴을 합성한 얼굴로 자연스럽게 교체한다.

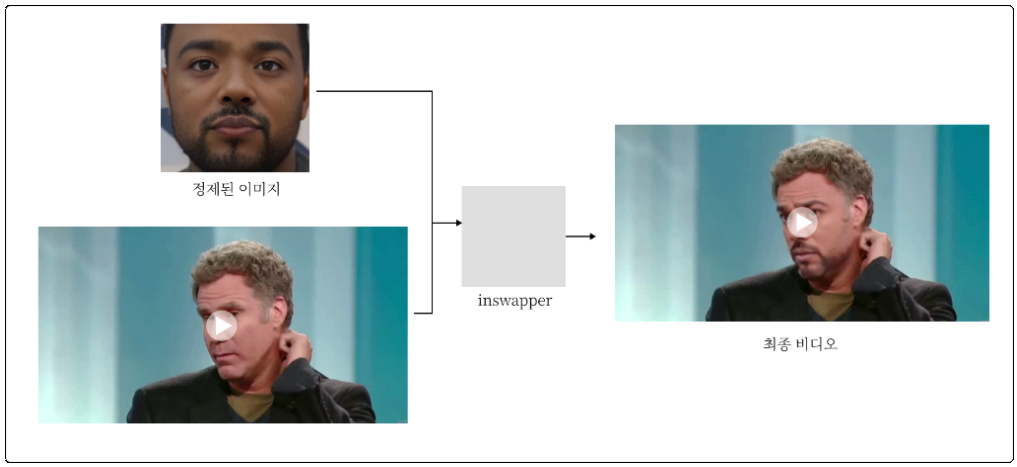

그림 8 (6단계) 비디오 속 사람의 얼굴 교체

- **7단계: 노이즈 추가**
  - 생성한 비디오의 각 프레임에 노이즈를 추가하여 딥페이크 탐지기가 활용하는 주파수 기반 통계적 특징을 교란시킨다.
  - 동시에, 노이즈는 시각적으로 거의 인식되지 않기 때문에 영상 품질 저하 없이 탐지를 회피하는 효과를 기대할 수 있다.
  - 향후 특정 딥페이크 탐지 모델에 의존하지 않고 작동하는 모델 비종속적 공격인 StatAttack을 이용하여 탐지기의 분포 기반 분류 기준을 무력화시킬 수 있다[12].

그림 9 (7단계) 노이즈 추가

4. 수행 결과

- 제안한 얼굴 부위 조합 기법을 통해, **원본 인물의 식별 정보를 제거한 합성 얼굴 비디오를 생성한다**.
- 그림 10은 생성된 영상들의 예시이다.

5. 결론 및 향후 연구

- 본 연구에서는 개인 프라이버시 보호를 목적으로, 얼굴 부위 조합 기반 딥페이크 생성 기법을 설계 및 구현하였다.
- **입력 비디오에서 눈·코·입 등의 주요 얼굴 부위를 서로 다른 인물의 프레임에서 무작위로 추출·조합함으로써, 원본 인물의 식별 정보를 효과적으로 제거하면서도 자연스러운 외관을 유지하는 얼굴 합성이 가능함을 확인하였다.**
- 이러한 비식별화 기반 합성 방식은 개별 인물의 정체성을 유지하지 않는다는 점에서, 향후 프라이버시 보호형 콘텐츠 생성에 사용할 수 있는 가능성을 제시한다.
- 그러나 얼굴 부위 조합은 정면 얼굴에 최적화되어 있어, 숙이거나 측면을 향한 얼굴에서는 감지 실패로 인해 일부 프레임에서 얼굴 교체가 누락되거나 비자연스러운 결과가 발생할 수 있다. **이를 해결하기 위해서는 얼굴 부위 조합 과정에서 시선 방향이나 포즈 정보를 반영하는 정렬 및 보정 전략이 필요하다.**

- 본 연구의 기법은 개별 프레임 단위에서 합성이 이루어지므로, 시간적 연속성을 고려한 RNN 기반 시계열 탐지기에서는 여전히 이질적이라고 평가될 가능성이 있다[13-14]. 이는 프레임 간 불일치가 탐지의 단서로 작용될 수 있다는 점에서, 완전한 탐지 회피에는 구조적인 한계로 볼 수 있다.
- **향후 연구에서는 얼굴 부위 조합 과정에서, 원본 인물의 시선 방향이나 포즈 정보를 분석할 필요가 있다. 또한 프레임 간 시간적 일관성을 유지할 수 있도록 확장하고, 프라이버시 보호 기법으로서의 실효성 및 안정성 검증 체계를 구축하여, 보다 정교하고 지속 가능한 익명화 컨텐츠 생성 프레임워크로 확장할 예정이다.**

6. 참조 문헌

[1] R. Yan et al., "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2024, pp. 8984-8994.

[2] F. Rosberg et al., "FIVA: Facial image and video anonymization and anonymization defense," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Paris, France, Oct. 2023, pp. 427-436, doi: 10.1109/ICCVW60793.2023.00043.

[3] U. A. Ciftci et al., "My Face My Choice: Privacy enhancing deepfakes for social media anonymization," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 1369-1379.

[4] Y. Nirkin et al., "FSGAN: Subject Agnostic Face Swapping and ReenactMENT," in Proc. IEEE/CVPR Workshops, 2019.

[5] I. Perov et al., "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," arXiv preprint, 2020.

[6] C. Xu et al., "Region-Aware Face Swapping," arXiv preprint, 2022.

[7] R. Natsume et al., "FSNet: An identity-aware generative model for image-based face swapping," arXiv preprint, 2018.

[8] Z. Zhao et al., "DiffSwap: High-fidelity and controllable face swapping via 3D-aware masked diffusion," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023.

[9] Z. Yu et al., "FuseAnyPart: Diffusion-driven facial parts swapping via multiple reference images," in Advances in Neural Inf. Process. Syst. (NeurIPS), 2024.

[10] Z.Sun et al., "DiffSwap: High-fidelity andcontrollable face swapping via 3D-aware masked diffusion," in Proc. IEEE/CVFConf. Comput. Vis. Pattern Recognit. (CVPR), 2023.

[11] X. Wang et al., "Towards real-world blind face restoration with generative facial prior," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 9168-9178.

[12] Y. Hou et al., "Evading DeepFake Detectors via Adversarial Statistical Consistency," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 12271-12280.

[13] B. Liu et al., "TI2Net: Temporal Identity Inconsistency Network for Deepfake Detection," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 4680-4689,

[14] Z. Gu et al., "Spatiotemporal inconsistency learning for deepfake video detection," in Proc. ACM Int. Conf. Multimedia (ACM MM), Ottawa, ON, Canada, Oct. 2023, pp. 6504-6512, doi: 10.1145/3664654.