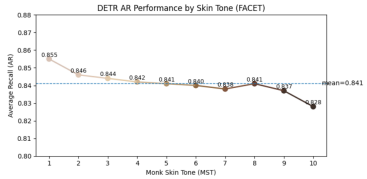
배포된 객체 탐지 모델에서 입력 섭동을 이용한 공정성 개선 기법

이도형[O,1*], 최서연[O,2*], 박성호[2†]
[1]인천대학교 정보통신공학과, [2]인천대학교 컴퓨터공학부
eeshape@inu.ac.kr, csy666666@inu.ac.kr, yunisomi@inu.ac.kr
* : equal contribution
† : corresponding author

Introduction

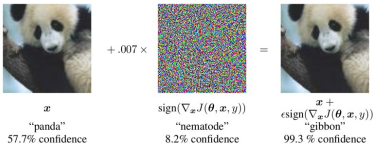
- **자율주행 시스템**에서 객체탐지는 안전 핵심 요소이며, 특히 **보행자 탐지**는 **사람의 안전**과 직결되는 중요한 과제이다.
- 하지만 기존의 보행자 탐지기는 평균 성능이 높더라도, **특정 인구통계학적 집단에 따라 탐지 성능이 달라지는 공정성(fairness) 문제**가 존재한다.



- 이러한 **집단 간 성능 격차**는 단순한 기술적 한계를 넘어 **사회적·윤리적 문제**로 확장될 수 있다.
- 본 연구는 이러한 배경을 바탕으로, **섭동을 통해 집단 간 탐지 성능 격차를 완화**하는 것을 목표로 한다.

Core Idea

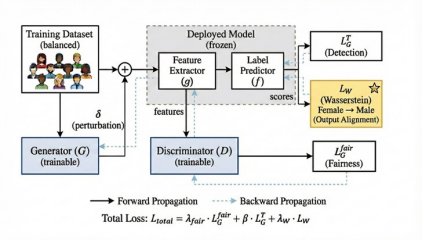
➢ 적대적 섭동(Adversarial Perturbation)



- **적대적 섭동**은 사람의 눈으로 구분하기 어려운 **작은 노이즈**로, 딥러닝 모델이 **잘못된 예측을 하도록 유도하는 공격 기법**이다.
- 본 연구에서는 적대적 섭동을 공격이 아닌 **공정성 개선을 위한 입력 수준 개입**으로 재정의하고, 입력 이미지에 가해지는 섭동을 최적화함으로써 (i) **전체 보행자 탐지 성능은 가능한 한 유지**하면서, (ii) **사전에 정의된 두 집단 간 탐지 성능 격차를 완화**하는 것을 목표로 한다.

Proposed Method

➢ GAN 기반 섭동 생성 방법



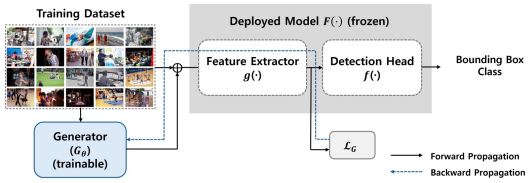
➢ 최종 목적함수

$$L_{total} = \lambda_{fair} L_G^{fair} + \beta(t) L_G^T + \lambda_W L_W$$

- L_G^{fair} : 판별기(D)가 교란된 이미지의 특징으로부터 성별을 구별하지 못하도록 생성기(G)를 학습시키는 적대적 손실
- L_G^T : 교란 적용 후에도 객체 탐지 성능이 유지되도록 보장하기 위한 손실 함수
- L_W : 탐지 결과의 공정성 확보를 위해 최종 출력 점수 분포를 직접 정렬하는 손실 함수

➢ MMD 기반 섭동 생성 방법

- 섭동을 통해 저성능 집단의 특징 분포를 고성능 집단의 특징 분포에 정렬한다.
- 두 분포 간 거리는 **Maximum Mean Discrepancy(MMD)**로 계산한다.



➢ Generator의 목적함수

$$\mathcal{L}_G = \lambda_{under}\,\mathrm{MMD}^2(\hat{h}^-, h^+) + \lambda_{over}\,\mathrm{MMD}^2(\hat{h}^+, h^+) + \lambda_{pert} \parallel \delta \parallel_2$$

- $\mathrm{MMD}^2(\hat{h}^-, h^+)$: 저성능 집단의 특징 분포를 고성능 집단에 정렬
- $\mathrm{MMD}^2(\hat{h}^+, h^+)$: 고성능 집단의 표현 보존
- $\parallel \delta \parallel_2$: 섭동의 크기를 제한하여 시각적 왜곡 방지

Experiments & Results

➢ GAN 기반 섭동 적용 결과

- **성별(gender)**을 보호 속성으로 설정

표 1. Baseline, Perturbation AP, AR 실험 결과

Gender	Base AP	Perturb AP	Δ_{AP}
Male	51.08	51.37	+0.29
Female	40.45	40.78	+0.34

Gender	Base AR	Perturb AR	Δ_{AR}
Male	83.39	83.59	+0.21
Female	82.58	83.28	+0.70

표 2. 성별 집단 간 AP, AR 성능 격차 실험 결과

Metric	Baseline	Perturbed	Δ_{gap}(%)
AP Gap	10.63	10.59	-0.42%
AR Gap	0.81	0.32	-60.60%

- AP 격차는 10.63pp에서 10.59pp로 거의 유지된 반면, AR 격차는 0.81pp에서 0.32pp로 60.6% 감소하여, 성별 성능 불균형을 유의미하게 완화함을 확인하였다.

➢ MMD 기반 섭동 적용 결과

- **피부톤(skin tone)**을 보호 속성으로 설정

표 3. Baseline, Perturbation AP, AR 실험 결과

Skin Tone	Base AP	Perturb AP	Δ_{AP}
Light	49.4	49.5	+0.1
Dark	42.3	42.6	+0.3

Skin Tone	Base AR	Perturb AR	Δ_{AR}
Light	84.3	84.4	+0.1
Dark	83.6	83.8	+0.2

표 4. 피부톤 집단 간 AP, AR 성능 격차 실험 결과

Metric	Baseline	Perturbed	Δ_{gap}(%)
AP Gap	7.1	6.9	-2.8%
AR Gap	0.7	0.6	-14.3%

- light 집단의 성능은 거의 유지된 반면 dark 집단의 성능은 상대적으로 더 개선되어, 피부톤 집단 간 격차가 AP 기준 2.8%, AR 기준 14.3% 감소하였다.

Conclusion

- 본 연구는 배포된 객체 탐지 모델을 수정하지 않고, 입력 단계에서의 **미세한 섭동**을 통해 **보호 속성(성별, 피부톤)에 따른 집단 간 탐지 성능 격차를 완화**하는 **GAN 기반 및 MMD 기반** 입력 보정 방법을 제안하였다.
- 제안한 방법은 **전체 탐지 성능을 유지**하면서, 보행자 누락 검출과 직결되는 **AR 격차를 효과적으로 감소**시켰다.
- 이러한 결과는 제안한 두 입력 보정 기법이 고성능 집단의 성능 저하를 최소화하면서 저성능 집단의 탐지 성능을 향상시켜, **집단 간 탐지 성능 격차 완화에 기여**할 수 있음을 시사한다.

INCHEON NATIONAL UNIVERSITY
TRUSTWORTHY AI LABORATORY