INCHEON NATIONAL UNIVERSITY
COMPUTER VISION LABORATORY

Balanced Datasets Are Not Enough- Estimating and Mitigating Gender Bias in Deep Image Representations (ICCV 2019)

Undergraduate Researcher at CVLab

Lee Dohyeong

2025.8.14

Contents

- Introduction
- Method
- Experiments
- Conclusion

Introduction

- Problem :
  - Blanced Data(예 : 남자,여자 1:1) -> Fair [직관적]
  - Blanced Data -> model-> gender bias [현실]
    - 라벨이 없는 (유추) 단서 : 아이동반여부, 헤어스타일 등
    - 예시 : Cooking과 Children은 자주 함께 등장하고, Children은 여성과 더 자주 등장

      => 모델이 여성에 대해 “Cooking” 을 예측할 가능성이 높아짐.

- Solution :
  - Adversarial Debiasing 접근 방식
  - 성별 단서를 직접적으로 제거하면서 가능한 한 많은

    작업 특정 정보를 보존

Introduction

- Leakage(**누설**) :

1. Dataset Leakage : **정답라벨 -> 성별 맞추는 정도**

$$\lambda_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{(Y_i, g_i) \in \mathcal{D}} \mathbb{1}[f(Y_i) == g_i],$$

   - 예시 : "공구", "차고", "정비"라는 라벨이 있으면, 성별이 '남성'으로 판단

2. Model Leakage : **모델 결과 -> 성별을 맞추는 정도**

$$\lambda_M(a) = \frac{1}{|\mathcal{D}|} \sum_{(\hat{Y}_i, g_i) \in \mathcal{D}} \mathbb{1}[f(\hat{Y}_i) == g_i)],$$

   - 예시 : 남자와 칼이 있는 사진

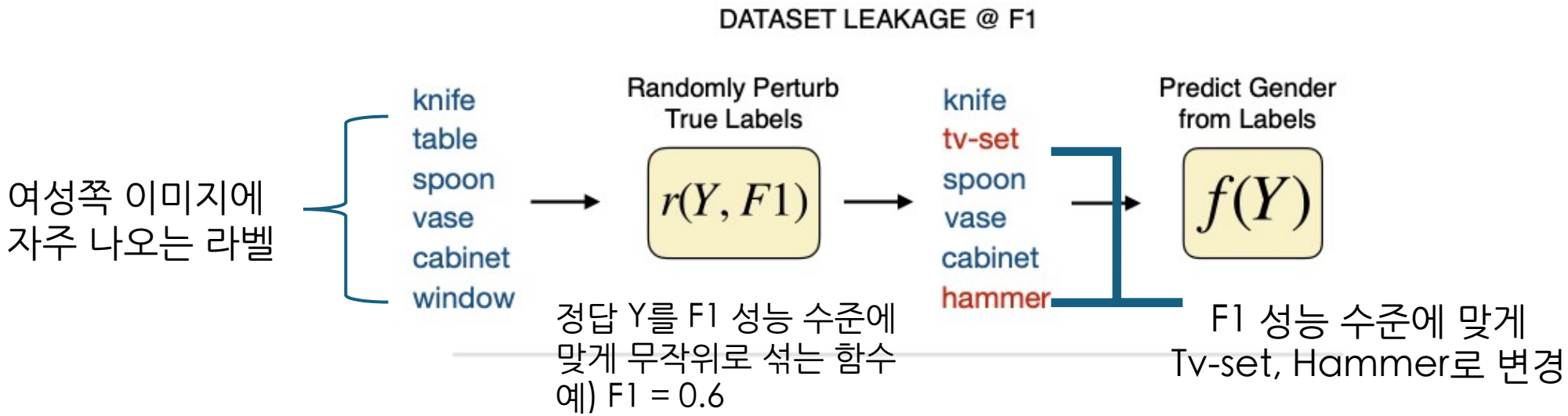
     => 칼만 보고 여자라고 판단

- f(Attacker) : Binary Classifier(male/female)
- Alternative Data Splits :
  - $\#(m,y)$: 라벨 y가 붙은 "남성" 이미지 수
  - $\#(w,y)$: 라벨 y가 붙은 "여성" 이미지 수
  - $\alpha \in \{3,2,1\}$: 비율 허용 폭

$$\forall y : \quad \frac{1}{\alpha} < \frac{\#(m,y)}{\#(w,y)} < \alpha$$

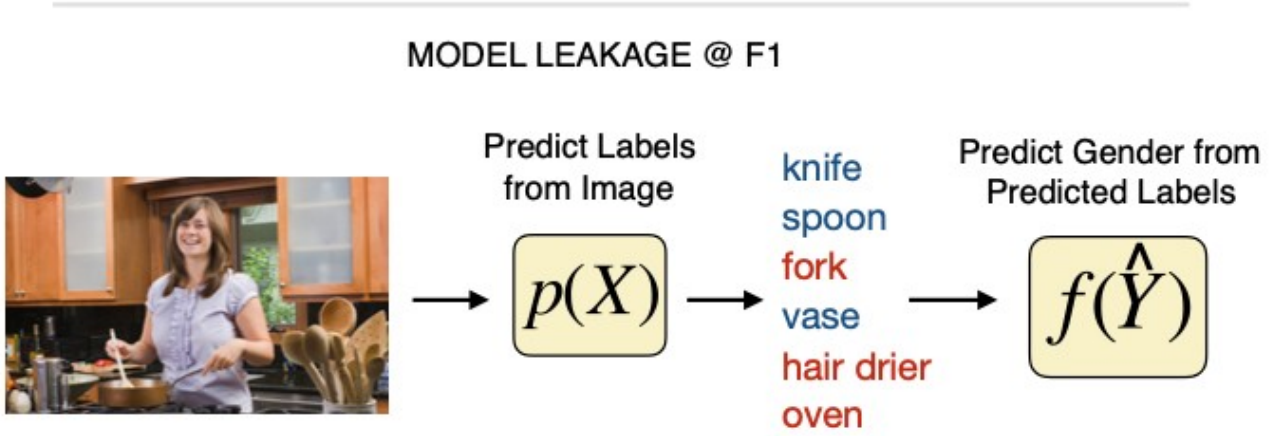
Introduction

F1= 정밀도(Precision)와 재현율(Recall)의 조화평균
Precision(**정밀도**) 과 Recall(**재현율**) 사이의 균형을 공정하게 반영하기 위해서
(**작은 값의 영향을 크게 받음**)



- 이미지의 라벨만 보고도 민감 속성(여기선 성별)을 예측할 수 있는 정도
- Task 목적 : 객체인식 -> 라벨만 보고 성별을 맞추면 문제 존재

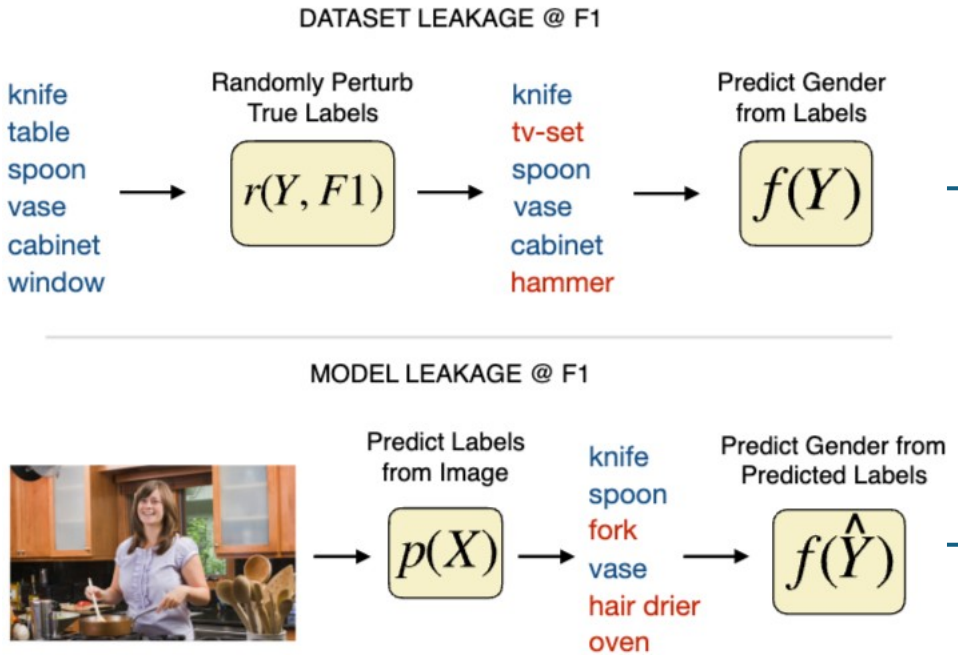
Introduction



1. image -> model -> label
2. 레이블을 보고 f가 남자인지 여자인지 판단함

원래 모델이 **성별을 직접 예측하지 않았더라도**, 출력된 라벨 패턴에서 성별 정보가 "누설(leakage)"될 수 있는지 측정

Introduction



Oven : 60%

Oven : 75%

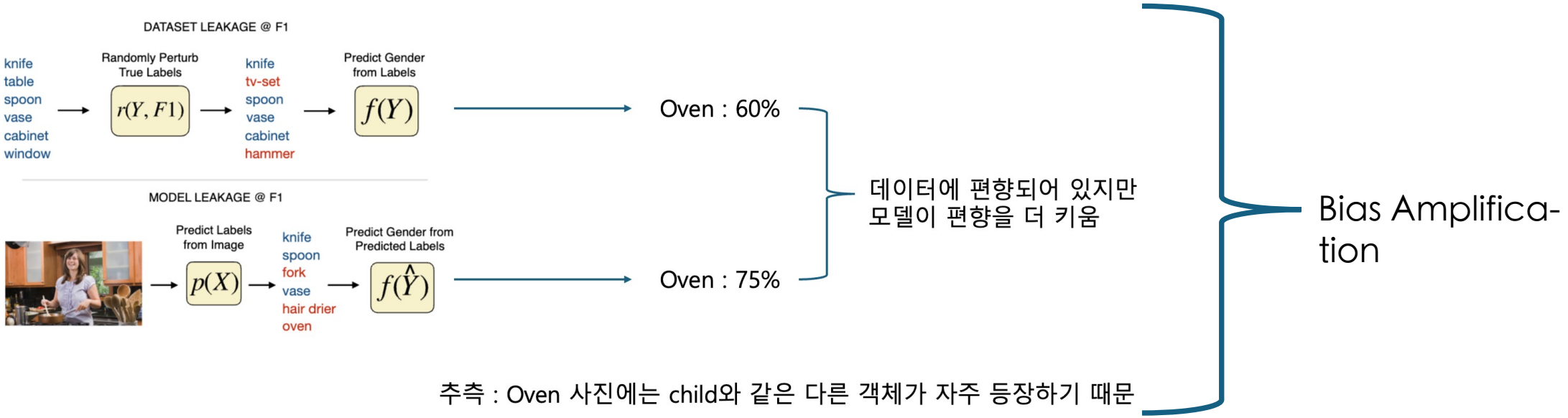
데이터에 편향되어 있지만 모델이 편향을 더 키움

추측 : Oven 사진에는 child와 같은 다른 객체가 자주 등장하기 때문

Introduction

- Amplification(증폭) : Model Leakage−Dataset Leakage
  - 모델이 데이터셋 편향보다 더 많은 성별정보를 증폭해서 노출하는 정도
  - Delta값이 높을 수록 성별과 관련 된 정보를 활용해서 예측한다는 뜻(0이면 비편향)

$$\Delta = \lambda_M(a) - \lambda_D(a)$$



Introduction

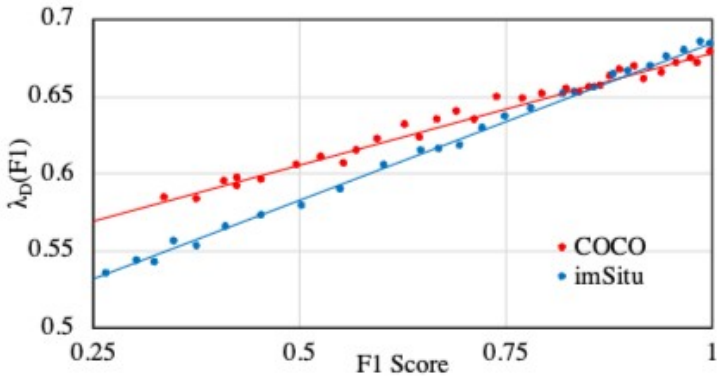
- Data Splits(Leakage, Performance)

CRF는 label 간 관계를 반영(객체와 성별 함께 예측, 결합확률을 높임)

dataset	split	Statistics #men	#women	Leakage λ_D	λ_M(F1)	λ_D(F1)	Performance Δ	mAP	F1
COCO [16]	original CRF	16,225	6,601	67.72 ± 0.31	73.20 ± 0.59	60.35	12.85	57.77	52.52
	no gender	16,225	6,601	67.72 ± 0.31	70.46 ± 0.36	60.53	9.93	58.23	53.75
	($\alpha = 3$)	10,876	6,598	62.00 ± 0.98	67.78 ± 0.29	57.50	10.28	57.04	52.60
	($\alpha = 2$)	8,885	6,588	56.77 ± 1.45	64.45 ± 0.56	54.72	9.73	56.21	51.95
	($\alpha = 1$)	3,078	3,078	53.15 ± 1.10	63.22 ± 1.11	52.85	10.37	48.23	42.89
imSitu [36]	original CRF	14,199	10,102	68.26 ± 0.31	78.43 ± 0.26	56.58	21.85	41.83	40.75
	no gender	14,199	10,102	68.26 ± 0.31	76.93 ± 0.20	56.46	20.47	41.02	40.11
	($\alpha = 3$)	11,613	9,530	68.11 ± 0.55	75.79 ± 0.49	55.98	19.81	39.20	37.64
	($\alpha = 2$)	10,265	8,884	68.15 ± 0.32	75.46 ± 0.32	55.74	19.72	37.53	36.41
	($\alpha = 1$)	7,342	7,342	53.99 ± 0.69	74.83 ± 0.34	53.20	21.63	34.63	33.94

객체 dataset

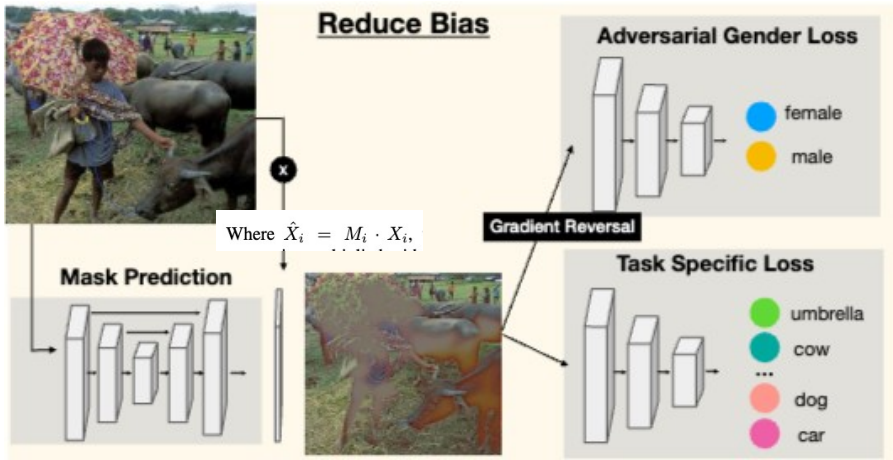
행동 dataset



Method

- Adversarial Debiasing Model:
  - Adv @ image : **이미지에서 직접 성별정보 제거.**

    U-Net **이용해서 마스크** M_i**를 예측하기위해** encoder-decoder**로 사용**
  - Gender Branch(ResNet-18) : **성별 예측**
    - **성별 예측자**(Adversary)
  - Main Branch(ResNet-50) : **타겟 분류(행동예측)**



Method

- Adversarial Debiasing Model:
  - Adv @ image
  - Adv @ conv 4
    - Resnet-50의 4번째 conv block에서 성별정보 제거
  - Adv @ conv 5
    - Resnet-50의 마지막 Conv Layer에서 성별정보를 제거

- 초기층은 필수 저수준 특징이 많아 성능 손실이 큽니다.
- 마지막층(conv5)은 고수준 성별 단서가 모여 정밀 제거가 유리

Method

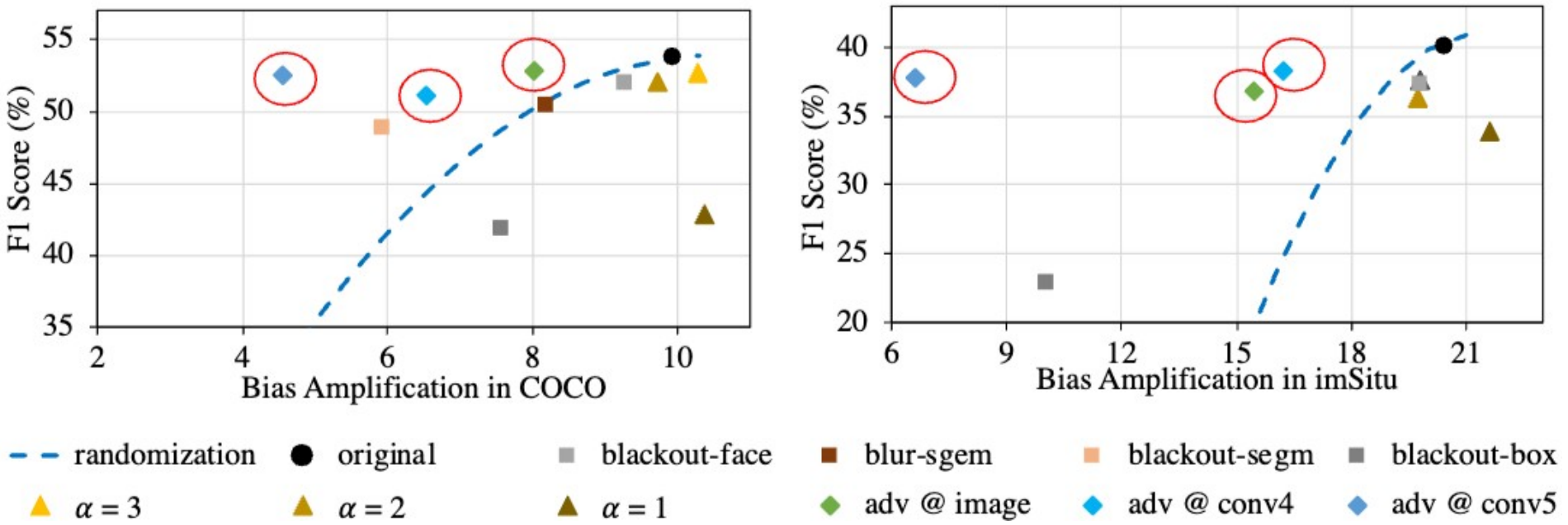
- Adversarial Debiasing Model:

$$L_p = \sum_i \Big[\underbrace{\beta \|X_i - \hat{X}_i\|_1}_{\text{① 원본 보존 항}} + \underbrace{L\big(p(\hat{X}_i), Y_i\big)}_{\text{② 주과제 손실}} - \underbrace{\lambda L_c\big(c(\hat{X}_i), g_i\big)}_{\text{③ 적대(성별 교란) 손실}} \Big]$$

1. **원본 보존항 :**
   - **마스크를 씌울 때 필요 이상으로 바꾸지 못하게**
   - L1 값이 크면 원본과 차이도 크다
   - B값이 크면 원본 보존을 더 강하게 요구
2. Main Branch
   - **가려진 이미지 X^도 정답을 잘 맞추도록**
3. Gender(Adversary) Branch
   - C(비평가)가 틀리게 유도
   - Gamma가 크면 성별 단서를 더 숨김
   - **성별을 못맞추게 하는 것이 목표**

Experiments

- Bias amplification vs F1
  - Dataset : COCO(객체), imSitu(행동)
  - Alpha : 성별간 라벨 동시 출현 비율
    - α = 1 : 균형, 동시 출현 비율 동일(α=n배 ; 여성n : 남성1)



Experiments

	Leakage		Performance		
	λ_M(F1)	λ_D(F1)	Δ	mAP	F1
original	70.46	60.53	9.93	**58.23**	**53.75**
blackout-face	69.53	60.24	9.29	55.93	51.81
blur-segm	68.19	59.99	8.20	55.06	50.26
blackout-segm	65.72	59.76	5.96	53.78	48.72
blackout-box	64.00	58.71	5.29	47.42	41.81
adv @ image	68.49	60.47	8.02	56.14	52.82
adv @ conv4	66.66	60.12	6.54	55.18	51.08
adv @ conv5	64.92	60.35	**4.57**	56.35	52.54
($\alpha = 1$)	63.22	52.85	10.37	48.23	42.89
adv @ conv5	**54.91**	52.40	**2.51**	43.71	38.98

Adv @ conv5방식이 편향을 가장 많이 줄임

α=1로 만든 데이터셋에서 성능이 떨어지는 이유는 **데이터 양과 다양성이 크게 줄어들었기 때문**입니다

Conclusion

- 데이터셋 누출(dataset leakage) · 모델누출(model leakage)을 정의하고, 둘의 차이(Δ)로 성별 편향증폭을 정량화
- 성별 라벨을 예측하지 않아도, 라벨-성별 동시출현을 '완전 균형'으로 맞춰도, 모델은 여전히 성별 편향을 증폭
- 원인: 어린이 · 의복 등 라벨되지 않은 성별-상관 단서가 중간표현에 남아 예측에 활용되기 때문입니다.
- 해결: 중간표현에 성별 판별자를 붙인 적대적 학습(gradient reversal)으로 단서를 제거
- "균형 데이터만으론 부족" → 누출 측정으로 진단하고, 표현 수준의 편향 완화(적대적 학습 등)를 결합해야 공정성과 정확도의 균형을 달성합니다.