INCHEON NATIONAL UNIVERSITY
COMPUTER VISION LABORATORY

Fairness in Autonomous Driving: Towards Understanding Confounding Factors in Object Detection under Challenging Weather (2024 CVPR Workshop)

Undergraduate Researcher at CVLab

Lee Dohyeong

2025.6.26

Contents

- Introduction
- Related Work
- Method
- Experiments
- Conclusion

Introduction

- **Autonomous Driving :**
  - Object Detection → 경로예측 → 경로계획 → 제어
  - 작은 오류 ⇒ 충돌(collision)
- **객체 탐지 실패 두가지 유형 :**
  1. **물체 탐지 못함(FN : False Negative)**
  2. **존재하지 않은 물체 탐지(FP : False Positive)**
- **오류의 원인 의심 :**
  1. 특정 인구 집단에 주로 발생하는가?
  2. 다양한 작동조건에서 이러한 오류가 해당 집단에 어떤 영향을 미치는지

Introduction

용어	정의	객체 탐지에서의 의미
TP (True Positive)	예측이 정답이고 실제로도 정답인 경우	실제 보행자가 있는 위치에 정확하게 박스를 쳤고, 정답과 일치
FP (False Positive)	예측은 정답이나 실제는 오답인 경우	보행자가 없는 곳에 박스를 쳐서 잘못 탐지(헛것을 봄)
FN (False Negative)	예측은 오답이나 실제는 정답인 경우	실제 보행자가 있었는데 탐지하지 못한 경우 (미탐지)
TN (True Negative)	예측과 실제 모두 오답인 경우	많은 탐지하지 않은 영역이 존재하므로 유의미하지 않음

Introduction

Q1. Object detection model exhibit performance biases based on protected attributes?
Protected attributes(보호 속성): Skin tone, Gender, Body size.

Q2. External environmental conditions such as adverse weather impact these biases?

Related Work

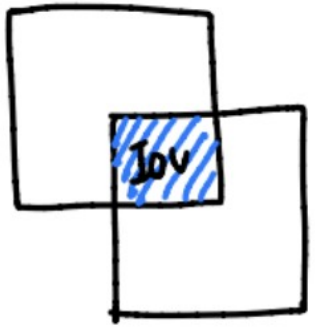
- CNN-based object detectors limitations
  - Confounding Factors(혼란요인) :
    - COCO and Pascal VOC are biased.
    - When evaluating whether "skin tone affects detection accuracy," lighting conditions may also affect the result.
  - Local Feature Bias :
    - CNNs rely on local receptive fields.
    - Overly focus on superficial features like skin tone or clothing color

=> Transformer models like DETR make object detection fairer through self-attention?

Method

- Object Detector Prediction Method
  1. Predicted Bounding Box, Confidence Score, Predicted Class
  2. Compare Ground Truth Box and Prediction Box
  3. Find **IoU(Intersection over Union) and classified as TP or FP.**
     - If the IoU is over 0.5 or 0.75

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



Method

- **Key Fairness Metrics(existing method) - AR & AP**

1. AR (Average Recall, 평균 재현율)

- Low AR means there are many false negatives (FN).
- Means missing people on the road.(Most important safety metric)
- Ensuring high AR for all groups is the minimum for equal opportunity(기회균등).
  - Equal opportunity : Equal chance for all groups.

$$AR = \frac{TP}{TP + FN}$$

Method

- **Key Fairness Metrics(existing method) - AR & AP**

  2. AP (Average Precision, 평균 정밀도)

  - Measures how accurate the “person” predictions
  - Low AP means there are many false positives (FP).
  - Cause unnecessary sudden stops

$$AP = \frac{TP}{TP + FP}$$

Method

- **Key Fairness Metrics(New Metrics) - ATPC & AFPC**

1. ATPC (Average True Positive Confidence, TP에 대한 신뢰도) :
   - Average confidence score of all TP predictions.
   - How confident was the model when it was correct?
   - Even if AR is high, a low ATPC means the model barely recognized the object
   - Weak predictions can easily turn into FN

$$ATPC = \frac{1}{N_{\text{TP}}} \sum p(\hat{y} = 1 | x, \text{class} = \text{person})$$

Method

- **Key Fairness Metrics(New Metrics) - ATPC & AFPC**

  2. AFPC (Average False Positive Confidence, FP에 대한 신뢰도) :

     - How confident was the model when it was wrong?
     - Average confidence score of all FP predictions.
     - High AFPC means the model is strongly confident in its wrong guesses.
     - Less able to filter out its own errors, increasing risk.

$$\text{AFPC} = \frac{1}{N_{\text{FP}}} \sum p(\hat{y} = 1 | x, \text{class} = \phi)$$

Method

- **Dataset Construction(Real World, Simulated)**
  - FACET(Real World) :
    - Dataset for fairness evaluation in computer vision.
    - 50,000 human annotations
    - Skin tone (MST scale 1-10)
    - Lighting labels (dimly-lit, well-lit)
    - Reflects the complexity and bias of the real world.
    - Demographic imbalance(Fig.5)

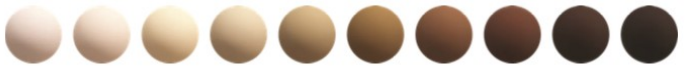


Figure 3: Monk Skin Tone (MST) [38] scale where MST=1 is the lightest skin tone and MST=10 is the darkest skin tone

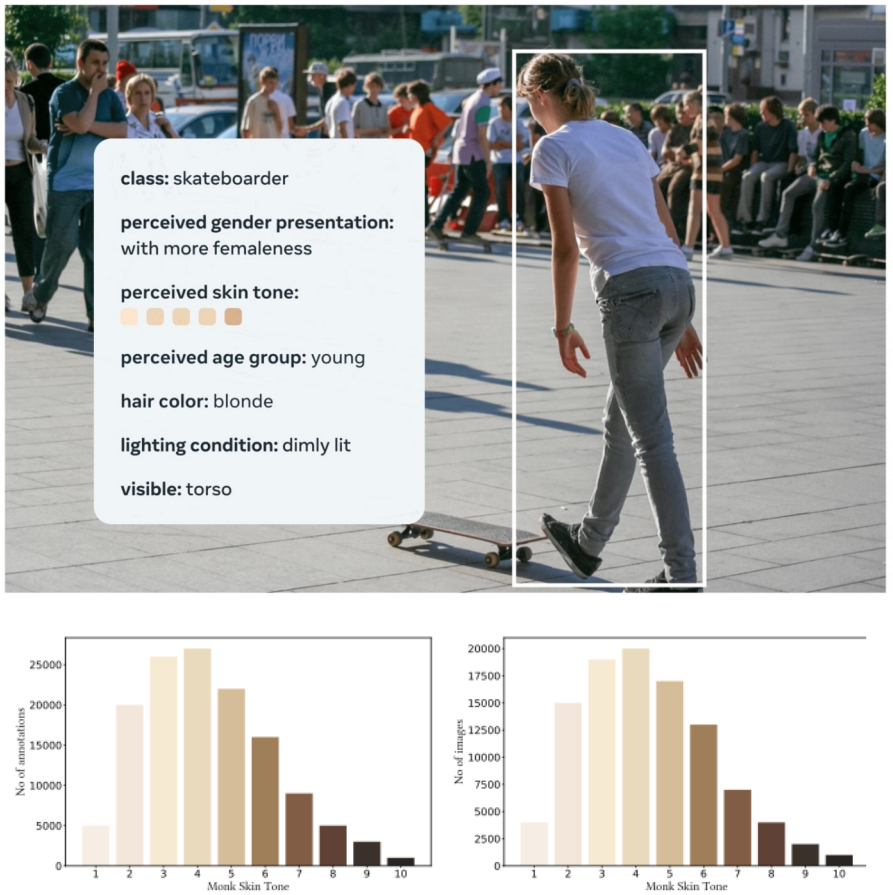


Figure 5: Histogram of the images and annotation with Monk Skin Tone scale for FACET dataset. Lighter skin tone annotations are more prominent in the dataset compared to the darker skin tone annotations.

Method

- Dataset Construction(Real World, Simulated)
  - FACET(Real World) :

FACET Mask Statistics	person	clothing	hair
perceived gender presentation			
with stereotypical maleness	6608	32103	3788
with stereotypical femaleness	4127	18136	3346
non-binary presentation	50	223	36
cannot be determined	72	193	13
perceived skin tone			
MST 1	2198	10687	1389
MST 2	5154	24328	3496
MST 3	6121	28825	4263
MST 4	5651	26583	3889
MST 5	4849	22738	3349
MST 6	3816	17931	2452
MST 7	2542	11845	1544
MST 8	1619	7564	922
MST 9	1216	5727	666
MST 10	521	2481	293
cannot be determined	2839	11844	1611
perceived age group			
younger	4145	19440	3107
middle	5443	25458	3319
older	1134	5352	733
cannot be determined	135	405	24
Hair color			
black	4053	18137	3323
brown	2726	12205	2267
blonde	1024	4633	952
red/orange	148	674	136
colored	84	340	96
grey	747	3519	559
cannot be determined	2885	14863	485

Hair type			
wavy	2090	9526	1897
curly	241	1141	253
straight	5141	22109	4395
coily	178	750	158
dreadlocks	113	522	109
bald	265	1167	81
Unknown	3626	19129	905
Additional attribute			
eyeware	1509	6993	957
headscarf	665	3634	256
tattoo	184	926	143
cap	3305	18209	797
facial hair	1511	7382	963
mask	591	3271	377

Method

- **Dataset Construction(Real World, Simulated)**
  - Carla(Simulated) :
    - High-fidelity open-source driving simulator.
    - Control gender, body size, skin tone, fog, rain, and pedestrian distance.
    - Enables pure causal analysis.
    - Used semantic segmentation to extract accurate bounding boxes efficiently.

Figure 4: Carla simulation sample image across fog intensities of 0%, 25%, 50%, 75%, and 100%. The visibility of the road incrementally reduces as the fog intensity increases.

Experiments

- **Effect of Artificial Darkness**
- Hypothesis : Darkness reduces detection, differently by skin tone.
- Result(Fig.7) :
  - AR decreased for all skin tones as darkness increased.
  - Lighter skin tones have consistently higher ATPC values, indicating higher model confidence.

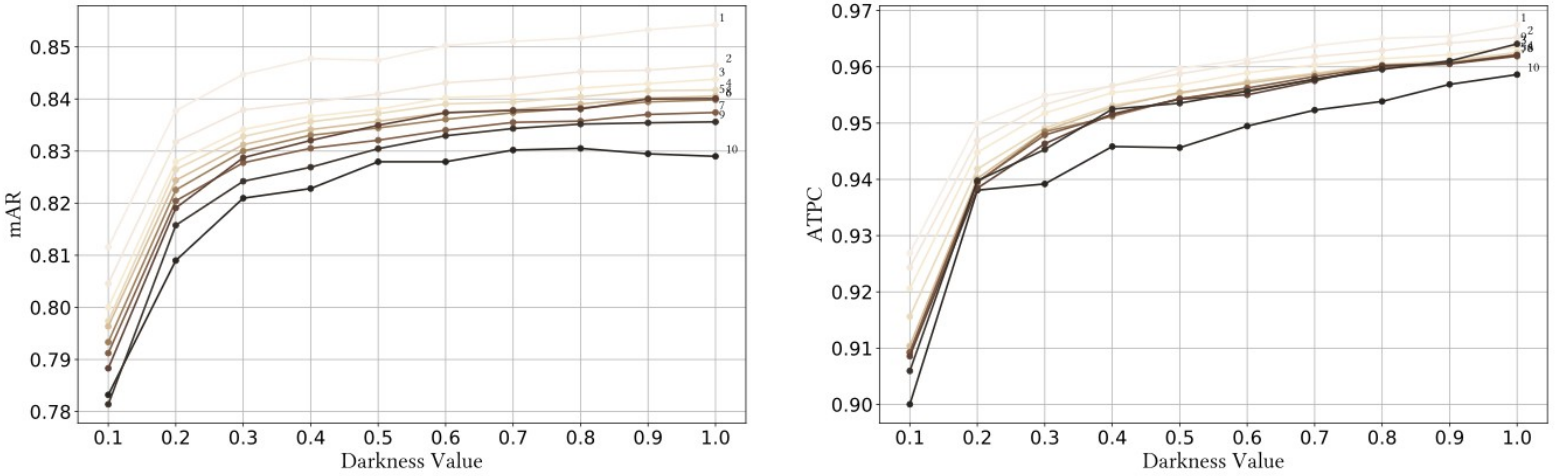


Figure 7: Performance disparities of ResNet50-backbone DETR model on Monk Skin Tone scale on FACET dataset. The metrics mAR and ATPC shows the model is more capable of identifying lighter skin tone people more confidently than darker skin tone people. For any skin tone, the model's performance drops with the ambient darkness (0=dark).

Experiments

- **Effect of Artificial Darkness**
- Hypothesis : Darkness reduces detection, differently by skin tone.
- Result(Fig.8) : (Randomly sampled 1000 images for each skin tone group)
  - Lighter skin tones (MST 1-4) showed higher mAR and ATPC than darker tones (MST 7-10).
  - Model is biased to detect lighter skin tones
  - Variance in ATPC values increased

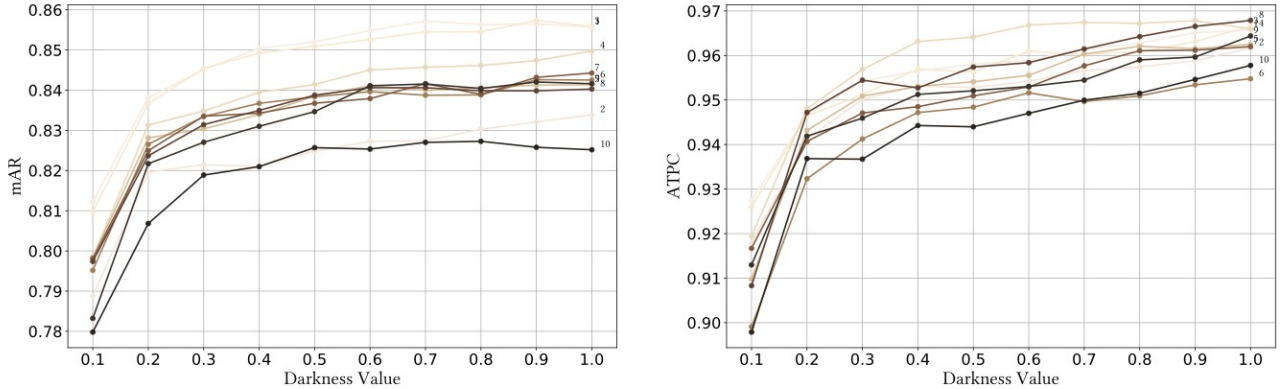


Figure 8: Evaluation of skin color and darkness on FACET dataset with Monk Skin Tone scale using same sample size. The results indicate a general advantage for lighter skin tones in the mAR metric but no significant disparity in ATPC metric. The disparity trends can vary even for randomly selected subsets and the full dataset.

Experiments

- **Effect of Real Lighting Conditions**
- Hypothesis : Lighting causes skin tone gap.
- Result(Fig.9) :
  - Dimly-lit conditions, the skin tone gap almost disappears.
  - Well-lit conditions, lighter skin tones show clearly better AR and ATPC.

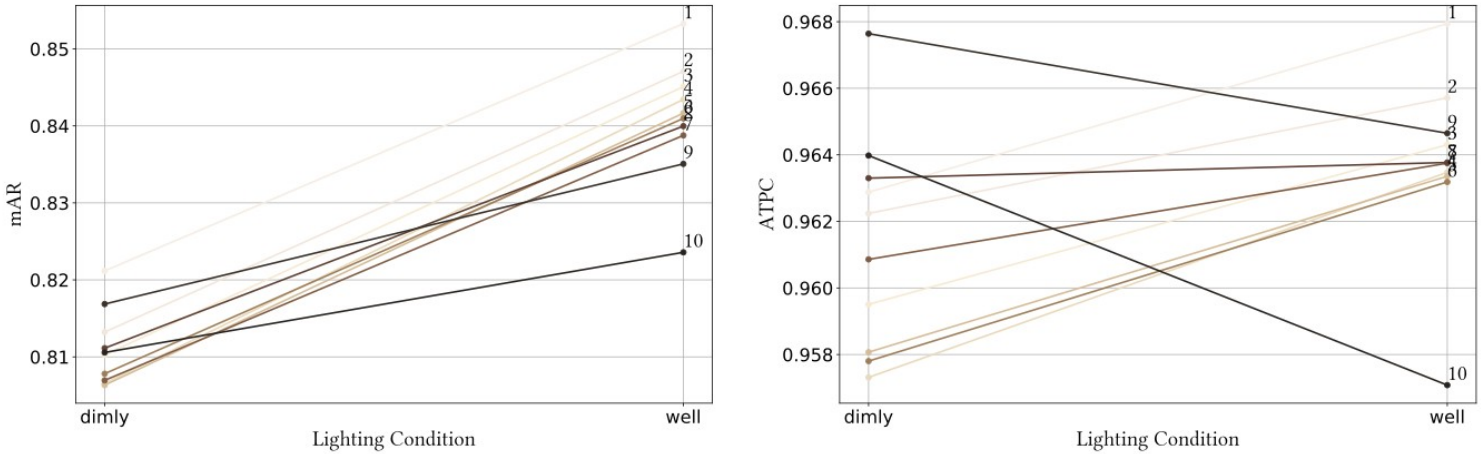
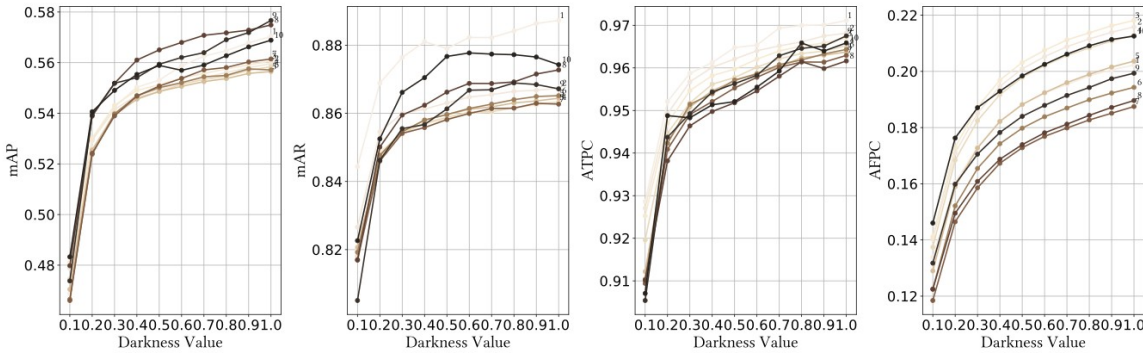


Figure 9: Analysis on the annotated lighting conditions, "well-lit" and "dimly-lit", and the skin tone. While the disparity for skin tones in the dimly-lit is not significant, lighter skin tones stands a better way of getting identified in well-lit conditions.

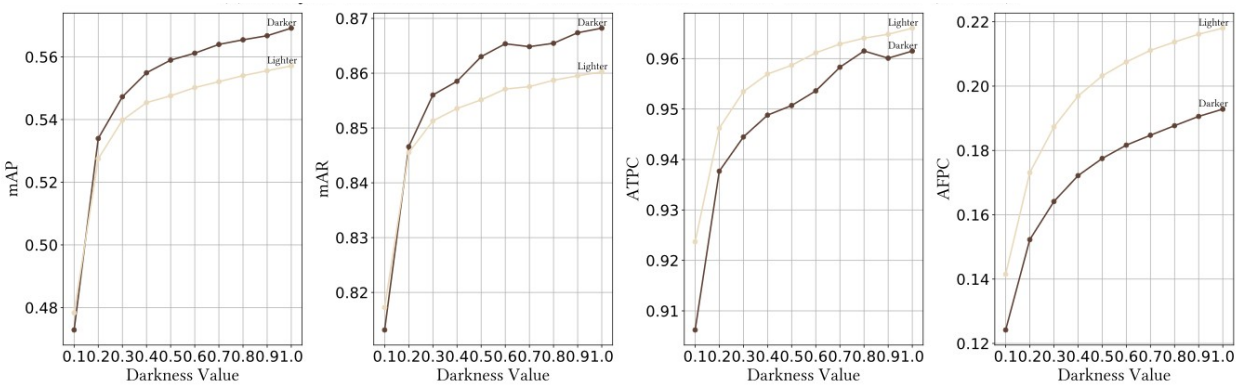
Experiments

- **Trap of Confounding Factors**
- Hypothesis : Filtering out mixed groups will reveal the true effect of skin tone.
- Result(Fig.10) : Filtered only images with people of a single skin tone.
  - Darker skin tones outperformed lighter ones in AP and AR.
  - ATPC/AFPC still remained higher for lighter skin tones.

  => Simulation is necessary to isolate true bias.



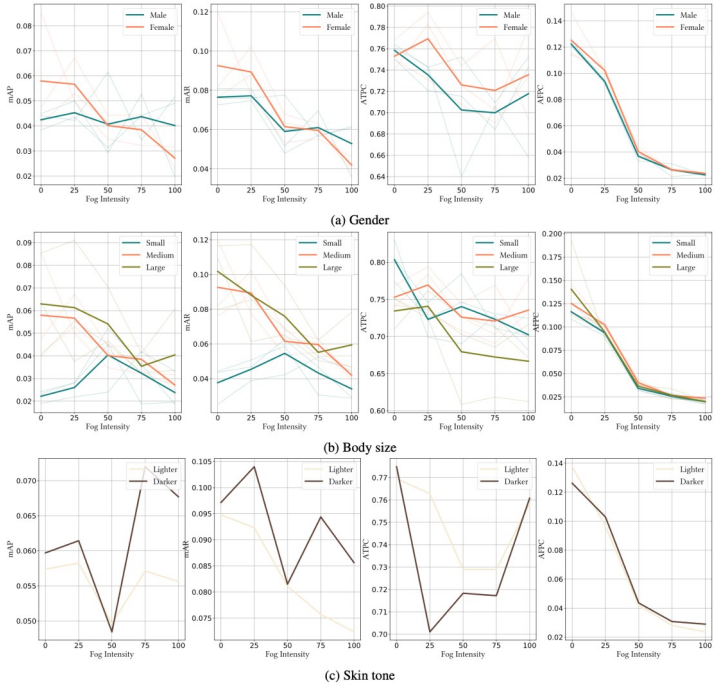
(a) Analysis on filtered skin tone annotations with ambient darkness levels (0=dark).



(b) Analysis on filtered and grouped skin tone annotations with ambient darkness levels (0=dark).

Experiments

- **Protected Attributes vs Weather**
- Compare Gender, Body size, Skin tone
  - Body size had the most significant effect on detection.
  - Children consistently showed much lower AR than adults in all weather conditions.
  - Gender and skin tone showed almost no meaningful difference in the DETR model.
    - DETR focuses on full body structure, not skin patches.



(a) Gender

(b) Body size

(c) Skin tone

Experiments

● Protected Attributes vs Weather

- Paradoxical fairness effect (Table 1):
  - 0% to 100% : (Δworst) between gender and body size decreases.
  - Extremely poor conditions, all groups perform worse, which gives the illusion of fairness.

=> Risk of misinterpreting fairness metrics without context.

	Gender			Body size		
Fog levels	$\Delta_{\text{worst}}mAR$	$\Delta_{\text{best}}mAR$	W_{mAR}	$\Delta_{\text{worst}}mAR$	$\Delta_{\text{best}}mAR$	W_{mAR}
0%	4.71	0.02	0.05	9.45	0.22	0.42
25%	2.73	0.16	0.02	7.85	0.28	0.21
50%	2.42	0.19	0.01	5.14	0.08	0.05
75%	1.43	0.11	0.00	3.24	0.03	0.03
100%	2.27	0.29	0.02	4.95	0.07	0.07

Conclusion

- Bias exists and complex
  - Transformer models are not free from bias
- Context matters
- ATPC, AFPC making them useful for early bias detection.
- Causal evaluation design based on simulation
- consider the **confidence scores of predicted bounding boxes** to accurately evaluate fairness in object detection.