Bi-Level Optimization for Self-Supervised AI-Generated Face Detection

Mian Zou[1,2], Nan Zhong[2], Baosheng Yu[3], Yibing Zhan[4], Kede Ma[2]

[1]Jiangxi University of Finance and Economics [2]City University of Hong Kong
[3]Nanyang Technological University [4]Yunnan United Vision Technology

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Background

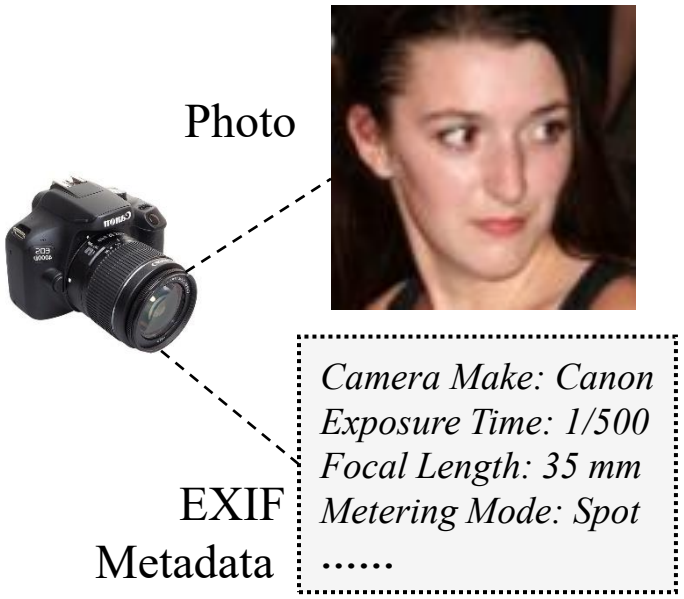
- AI-Generated Faces
  - These faces closely resemble photographic ones
  - Their misuse poses risks, such as spreading misinformation and tampering with media integrity

Red: Photographic face images

Motivation

- Drawbacks of Current Works
  - Supervised detectors often overfit to specific generators
    - visual artifacts
    - architectural principles
    - generation process
    - model-dependent: limited generalization to unseen or emerging generative techniques
  - Self-supervised detectors use pretext tasks
    - not aligned with the goal of AI-generated face detection
    - suboptimal performance

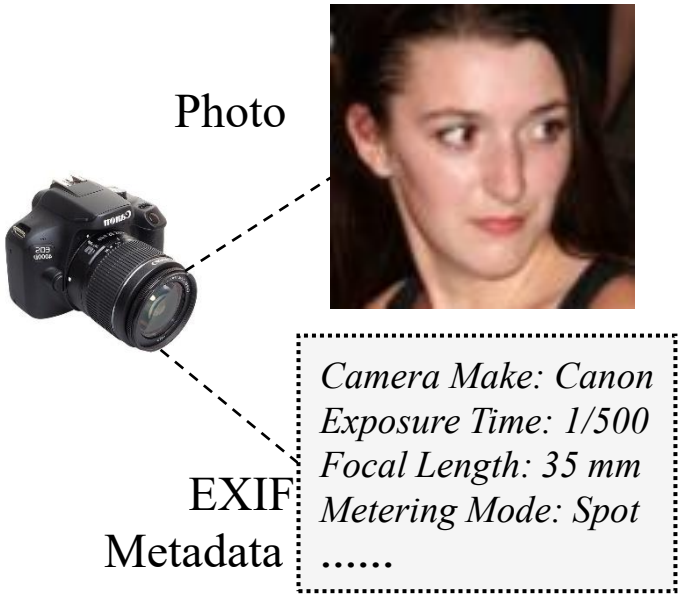
Motivation



Tag	Type	Example Value
Flash	Categorical	Flashfired, Flashauto
Make	Categorical	Canon, FUJIFILM
Metering Mode	Categorical	Multi-segment, Spot
Model	Categorical	EOS5DMarkII, EOS7D
Scene Capture Type	Categorical	Standard, Landscape
Exposure Mode	Categorical	Auto, Manual
White Balance Mode	Categorical	Auto, Manual
Aperture	Ordinal	F2.8, F4
Exposure Bias	Ordinal	0 EV, -1 EV
Exposure Time	Ordinal	1/60 sec, 1/200 sec
F-Number	Ordinal	F2.0, F3.2
Focal Length	Ordinal	5.8 mm, 35 mm
ISO Speed	Ordinal	400, 100
Shutter Speed	Ordinal	1/63 sec, 1/79 sec

The Camera Metadata Perspective!

Aligning Pre-training Pretext Tasks with Downstream Main Task



EXIF-based Pretext Tasks → Goal → AI-generated Face Detection

Aligning?

Bi-Level Optimization for AI-Generated Face Detection

$$\min_{\boldsymbol{\lambda}} \sum_{\boldsymbol{x} \in \mathcal{B}_{\text{val}}} \ell_1 \left(\boldsymbol{x}; \boldsymbol{\theta}^\star \right)$$

$$\text{s.t.} \quad \boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \in \mathcal{B}_{\text{tr}}} \sum_{i=1}^{K} \lambda_i \ell_i (\boldsymbol{x}; \boldsymbol{\theta})$$

- Inner-Loop Optimization $\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \in \mathcal{B}_{\text{tr}}} \sum_{i=1}^{K} \lambda_i \ell_i (\boldsymbol{x}; \boldsymbol{\theta})$
  - Feature extractor training via linear weighted pretext tasks
- Outer-Loop Optimization $\min_{\boldsymbol{\lambda}} \sum_{\boldsymbol{x} \in \mathcal{B}_{\text{val}}} \ell_1 \left(\boldsymbol{x}; \boldsymbol{\theta}^\star \right)$
  - Task weighting tuning guided by the downstream objective

Photographic Faces
vs.
AI-Generated Faces

Inferior Results!

Surrogate Task for AI-Generated Face Detection

- Face Manipulation Detection as the Surrogate Task
  - Artificial Face Manipulations
  - Blind to AI-Generated Faces
  - Good Performance

Horizontal
Eye Flipping

Horizontal
Mouth Flipping

Vertical
Mouth Flipping

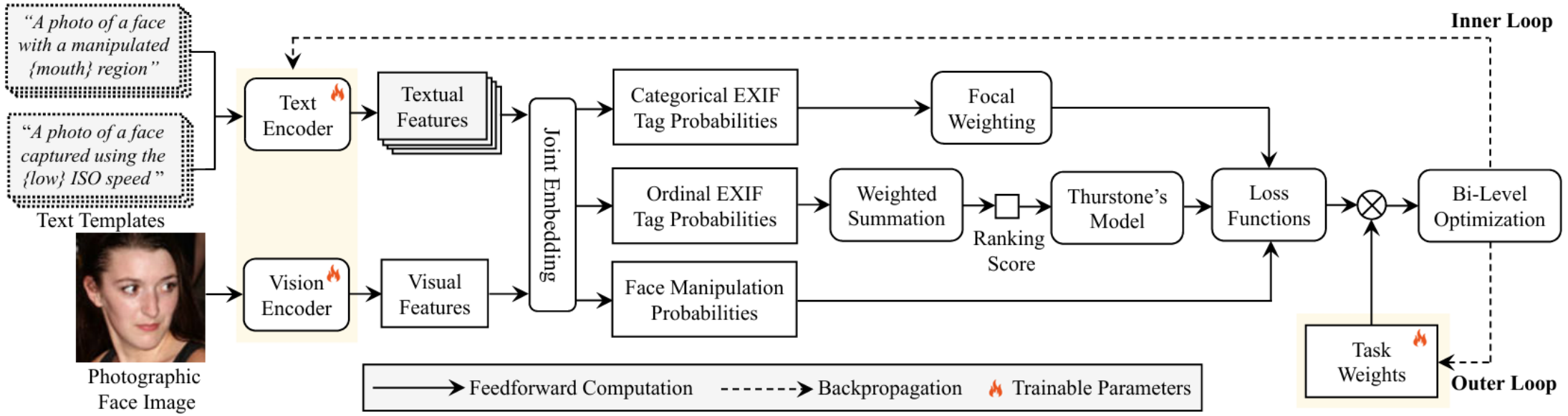
Global Affine
Transformation

Original
Face Image

Surrogate Task for AI-Generated Face Detection

- Two Advantages
  - Avoiding Reliance on Generative Models
  - Aligning Conceptually with the Task of AI-Generated Face Detection
    - Distinguishing between photographic faces and non-photographic ones, manipulated or generated
    - Joint embedding space for providing an explicit binary distinction: *"A photo of a {photographic, manipulated} face"*

BLADES: Bi-Level AI-Generated Face DEtector with Self-Supervision



BLADES

- Coarse-grained Face Manipulation Detection
  - Binary classification on original photographic faces and manipulated ones

Horizontal
Eye Flipping

Horizontal
Mouth Flipping

Vertical
Mouth Flipping

Global Affine
Transformation

Original
Face Image

BLADES

- Categorical EXIF classification
  - E.g., given the tag of make
  - The photo was taken with a Canon camera or a Fujifilm camera ?
  - Long-tailed distribution (e.g., Canon images may outnumber Apple by 10×)
    - Focal version (Lin et al., 2017) of the multi-class classification loss that down-weights well-classified samples

Tag	Type	Example Value
Flash	Categorical	Flashfired, Flashauto
Make	Categorical	Canon, FUJIFILM
Metering Mode	Categorical	Multi-segment, Spot
Model	Categorical	EOS5DMarkII, EOS7D
Scene Capture Type	Categorical	Standard, Landscape
Exposure Mode	Categorical	Auto, Manual
White Balance Mode	Categorical	Auto, Manual
Aperture	Ordinal	F2.8, F4
Exposure Bias	Ordinal	0 EV, -1 EV
Exposure Time	Ordinal	1/60 sec, 1/200 sec
F-Number	Ordinal	F2.0, F3.2
Focal Length	Ordinal	5.8 mm, 35 mm
ISO Speed	Ordinal	400, 100
Shutter Speed	Ordinal	1/63 sec, 1/79 sec

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Doll´ar. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

BLADES

- Ordinal EXIF ranking
  - Numerical relationships are important
  - Given two images $(\boldsymbol{x}, \boldsymbol{x}')$, $\widehat{\text{tag}}_i(\boldsymbol{x}) \geq \widehat{\text{tag}}_i(\boldsymbol{x}')$?
  - Pairwise ranking effectively captures the latent ordinal structure in EXIF-based features, improving the model's sensitivity to subtle visual differences
  - To address tokenizers fragmenting numeric relationships, raw EXIF values are discretized into low/medium/high, with continuous estimates recovered via weighted averaging

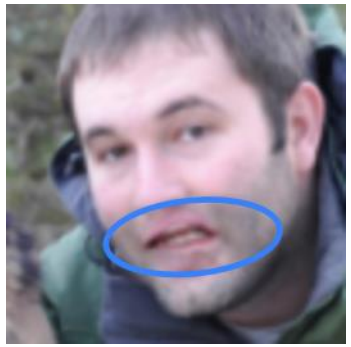
Tag	Type	Example Value
Flash	Categorical	Flashfired, Flashauto
Make	Categorical	Canon, FUJIFILM
Metering Mode	Categorical	Multi-segment, Spot
Model	Categorical	EOS5DMarkII, EOS7D
Scene Capture Type	Categorical	Standard, Landscape
Exposure Mode	Categorical	Auto, Manual
White Balance Mode	Categorical	Auto, Manual
Aperture	Ordinal	F2.8, F4
Exposure Bias	Ordinal	0 EV, -1 EV
Exposure Time	Ordinal	1/60 sec, 1/200 sec
F-Number	Ordinal	F2.0, F3.2
Focal Length	Ordinal	5.8 mm, 35 mm
ISO Speed	Ordinal	400, 100
Shutter Speed	Ordinal	1/63 sec, 1/79 sec

BLADES

- Fine-grained Face Manipulation Detection
  - Identifying which face regions have been manipulated

Horizontal
Eye Flipping

Horizontal
Mouth Flipping



Vertical
Mouth Flipping

Global Affine
Transformation

Original
Face Image

Experiments & Results

- Cross-Generator Detection

Method	StyleGAN2	VQGAN	LDM	DDIM	SDv2.1	FreeDoM	HPS	Midjourney	SDXL	Average
CNND [52]	50.61	99.89	53.07	56.55	50.51	58.62	50.31	51.66	54.49	58.41
GramNet [34]	51.16	99.92	53.25	50.09	50.23	51.59	50.26	52.91	53.63	57.00
RECCE [3]	66.64	**100.0**	70.91	73.10	71.62	77.52	64.14	62.19	65.29	72.38
LNP [31]	80.06	99.43	67.35	50.10	54.13	50.06	50.55	50.40	50.59	61.41
LGrad [47]	52.94	99.99	**99.80**	64.91	57.59	66.58	60.14	76.59	74.03	72.43
DIRE [53]	72.48	69.81	98.92	77.80	58.84	89.05	62.50	90.75	87.79	78.66
Ojha23 [38]	65.45	83.40	70.06	72.25	72.76	78.55	56.21	54.96	58.01	67.96
AEROBLADE [43]	48.69	51.99	69.50	42.53	46.19	90.05	77.40	82.16	81.95	65.61
FatFormer [32]	**98.91**	98.30	97.82	95.63	68.88	81.04	**90.28**	88.20	88.08	89.68
Zou25 [66]	76.88	74.59	93.83	93.63	78.62	95.31	83.79	91.29	91.71	86.63
BLADES-OC (Ours)	76.75	76.78	93.63	**96.05**	**80.70**	**96.09**	84.82	92.79	**94.48**	88.01
BLADES-BC (Ours)	94.22	97.24	96.95	94.33	74.83	95.73	84.40	**95.19**	93.84	**91.86**

- OC: One-class anomaly detection via a Gaussian mixture model fitted to photographic face embeddings, where samples below a likelihood threshold are flagged as AI-generated
- BC: Binary classification using a lightweight MLP trained on learned photographic and AI-generated embeddings

Experiments & Results

- Sensitivity and Specificity Analysis

Method	Photographic (%)		AI-generated (%)		F-score↑
	TNR↑	FPR↓	TPR↑	FNR↓	
LGrad [47]	**99.98**	**0.02**	44.97	55.03	0.60
DIRE [53]	99.96	0.04	54.37	45.63	0.65
Ojha23 [38]	75.49	24.51	60.39	39.61	0.64
FatFormer [32]	97.61	2.38	81.74	18.26	0.87
BLADES-BC	94.64	5.36	**88.97**	**11.03**	**0.91**

- For photographic faces, we report true negative rate(TNR) and false positive rate (FPR)
- For AI-generated faces, true positive rate (TPR) and false negative rate (FNR)
- The F-score summarizes over all detection performance

Experiments & Results

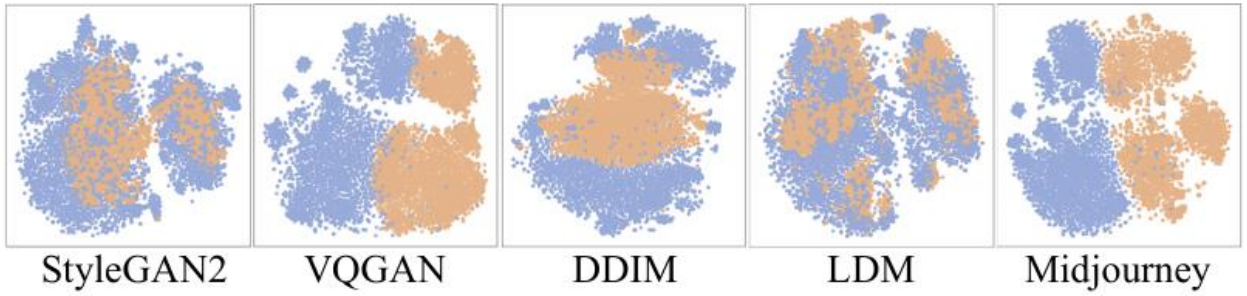
- Feature Separability Comparison

Method	StyleGAN2	VQGAN	LDM	DDIM	SDv2.1	FreeDoM	HPS	Midjourney	SDXL	Average
CLIP [41]	33.99	60.20	55.49	82.85	90.15	85.39	91.22	92.21	93.66	76.13
FaRL [62]	34.35	55.91	47.26	85.10	**95.24**	79.91	**93.97**	89.72	94.61	75.12
EAL [61]	69.71	72.41	85.81	84.98	74.21	97.92	87.33	91.65	93.04	84.12
Hu21 [22]	50.00	49.99	49.99	49.99	49.99	49.99	49.99	50.00	49.99	49.99
LNP [31]	37.55	63.28	71.12	69.54	65.64	66.40	66.64	66.70	67.25	63.79
Zou25 [66]	85.69	84.31	**98.66**	97.96	88.29	99.79	91.92	97.07	97.23	93.43
BLADES-OC	**87.89**	**88.31**	98.24	**99.32**	91.72	**99.92**	93.93	**97.80**	**98.36**	**95.05**

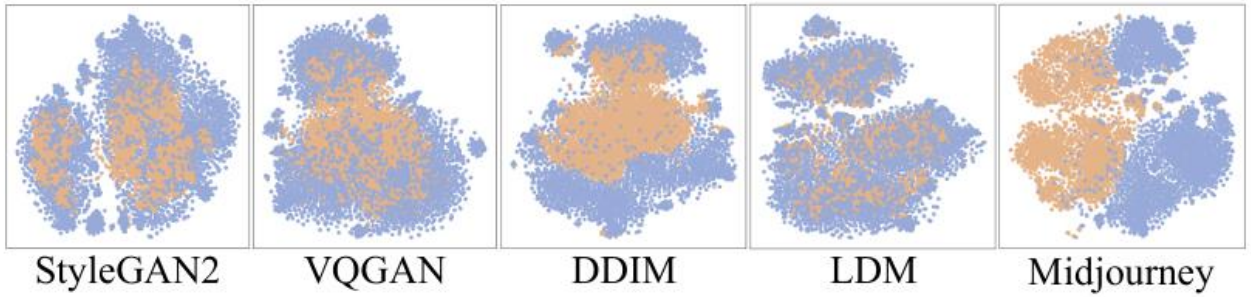
- One-class classification (AUC (%))

Experiments & Results

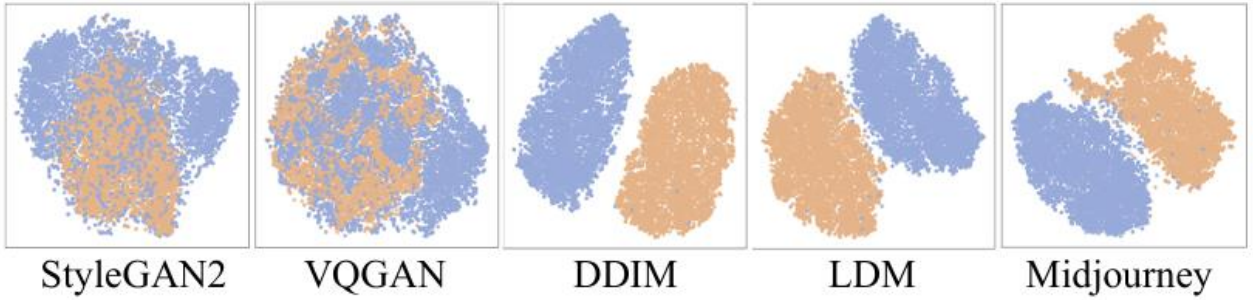
- Feature Separability Comparison
  - t-SNE visualizations



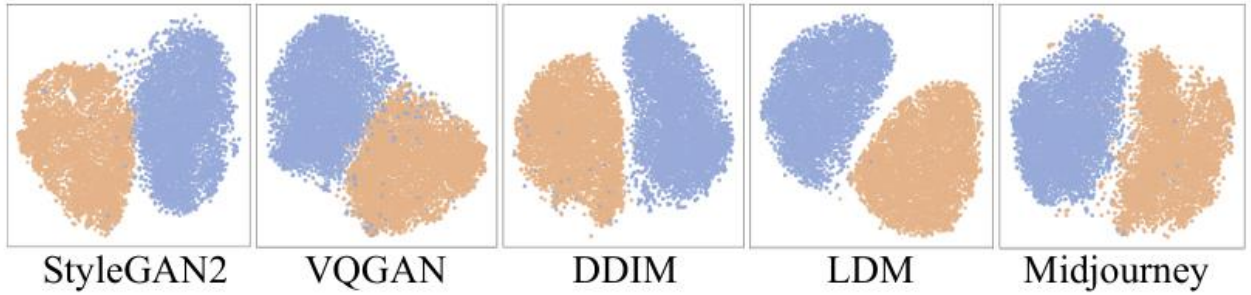
(a) CLIP [41]



(b) FaRL [62]



(c) EAL [61]

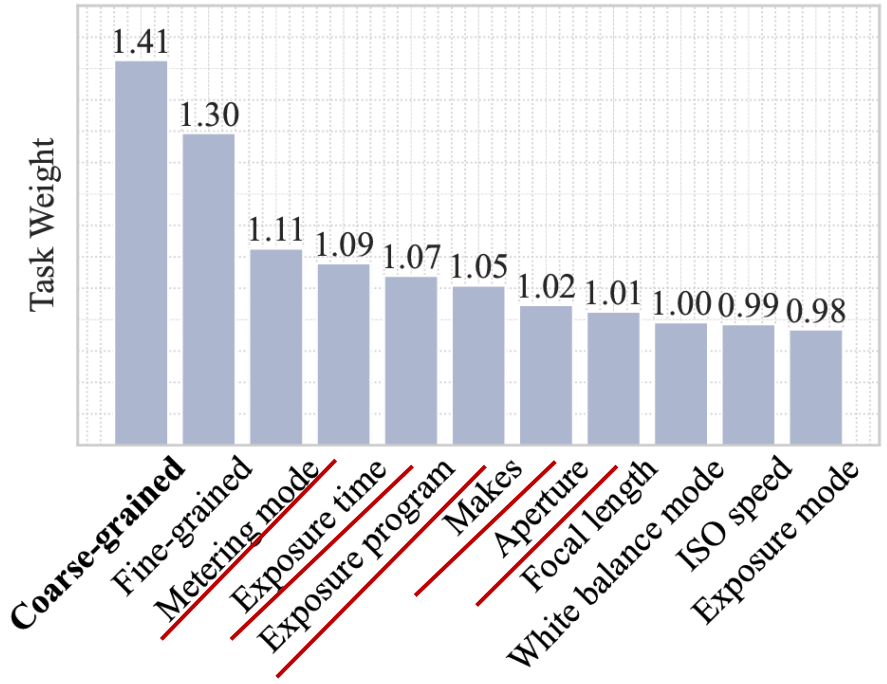


(d) BLADES-OC

➢ photographic (yellow) and AI-generated (blue)

Experiments & Results

- Learned Task Weights



- Face-specific Tasks are Prioritized
- Exposure-related EXIF Tags are More Informative

Summary

- We present BLADES, a bi-level optimization scheme that explicitly steers self-supervised pretraining toward AI-generated face detection
- We implement BLADES using joint embedding that incorporates EXIF-based and manipulation-based pretext/surrogate tasks to detect AI-generated faces
- We demonstrate state-of-the-art performance in both one-class and binary classification evaluations, with strong cross-generator generalization

Thanks