软件开发环境国家重点实验室
State Key Laboratory of Software Development Environment

Towards Benchmarking and Assessing Visual Naturalness of Physical World Adversarial Attacks

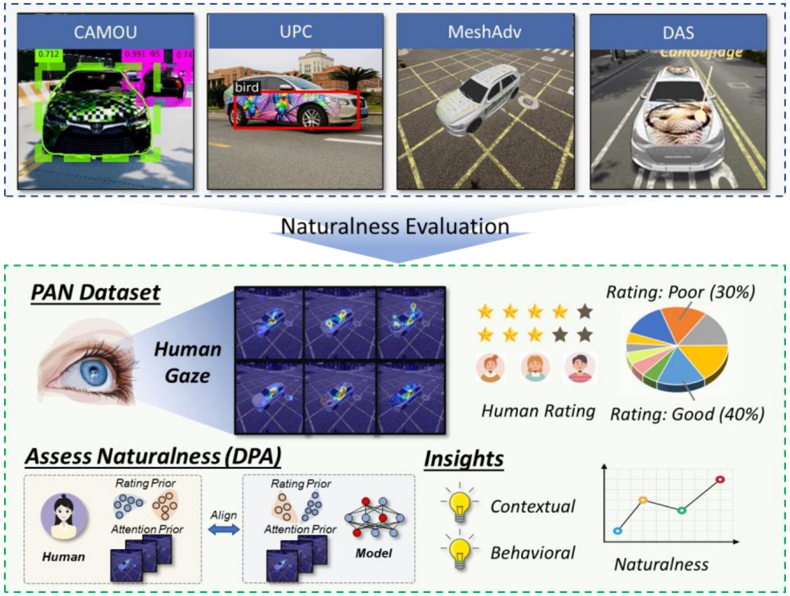
Paper ID 4612, WED-AM-391

Simin Li, Shuning Zhang, Gujun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, Xianglong Liu

Preview

Background

- ■Physical world adversarial attacks are harmful in real world but are conspicuous to human. Many works improve naturalness of attacks.
- ■But **how to evaluate** the naturalness of these attacks?



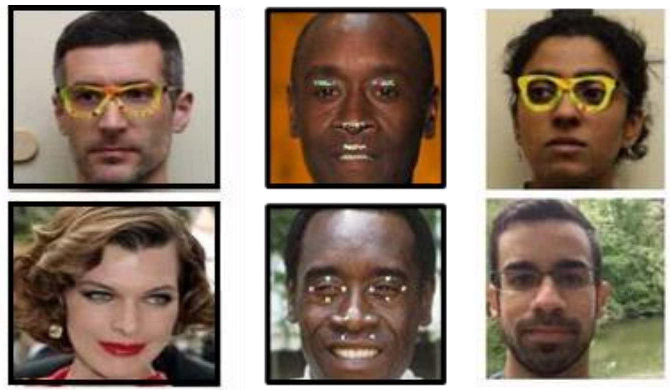
Contribution

- ■We take the first step to evaluate the naturalness of physical world attacks.
- ■We contribute Physical World Naturalness (PAN) dataset, including 2688 images with human *ratings* and human *gaze*.
- ■We unveil how environment and human gaze contribute to naturalness.
- ■We provide algorithms to evaluate naturalness of physical world attacks, by aligning model behavior with human behavior.

Introduction

With prominent success gained by DNNs, physical world attacks can easily fail DNNs by daily artifacts with adversarial capability

Surveillance



Face detection

Autonomous Driving

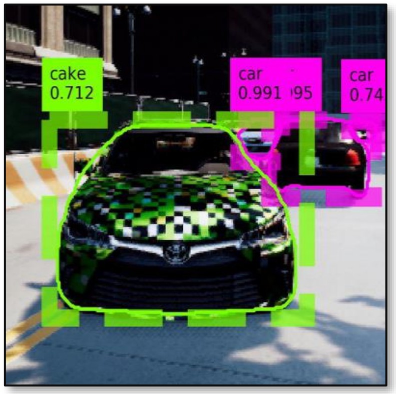
However, physical world attacks are often conspicuous, allowing **human** to easily identify and remove such attacks in real world

- In 48 physical world attack papers we surveyed:
  - 20 papers (42%) emphasize their attack is natural and stealthy.

Natural physical world attack is a critical issue!

Introduction

But how to we assess naturalness?



CAMOU
ICLR 2018

MeshAdv
CVPR 2019

UPC
CVPR 2020

DAS
CVPR 2021

- In 20 papers claimed to be natural:
  - 11 papers perform **no experiment** to validate their claim
  - 11 papers claim their attack closely imitate natural image, **but do this mean naturalness** in human?
  - 5 papers validate naturalness by human experiment, but in a **case-by-case setting**

Introduction

■How to assess the naturalness of physical world adversarial attacks?

- ■Assessing and understanding by human
- ■Automated evaluation by an algorithm

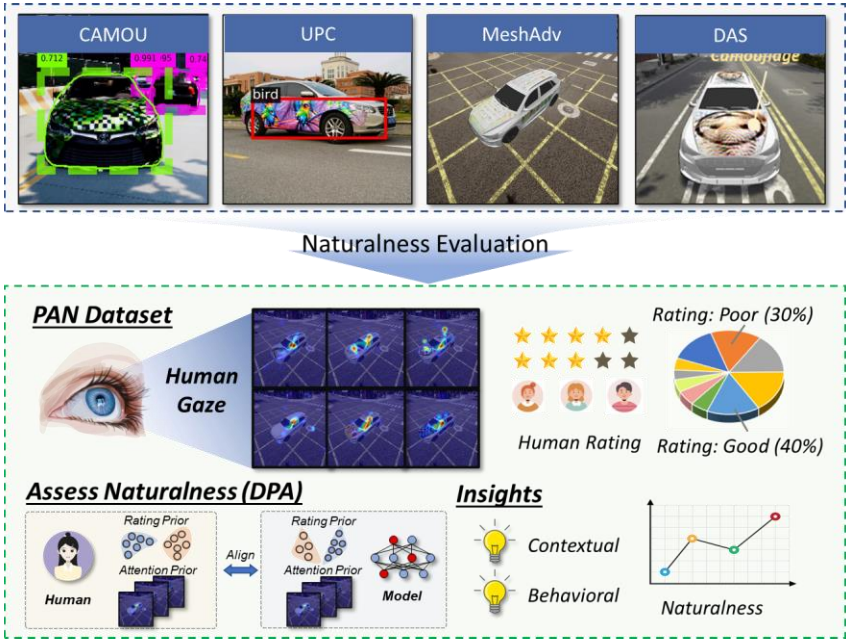
Contribution

■We take the first step to evaluate the naturalness of physical world attacks.

■We contribute Physical World Naturalness (PAN) dataset, including 2688 images with human *ratings* and human *gaze*.

■We unveil how environment and human gaze contribute to naturalness.

■We provide algorithms to evaluate naturalness of physical world attacks, by aligning model behavior with human behavior.



Physical Attack Naturalness (PAN) Dataset

- Image Quality Assessment (IQA) treats human judgement as golden standard.
- However, they focus on different distortion type, image source and evaluated content.

Datasets	Distortion	Image Source	Property
LIVE [49]	Artificial	Kodak Test Set	Quality
TID2008 [41]	Artificial	Kodak Test Set	Quality
CSIQ [29]	Authentic	Kodak Test Set	Quality
LIVE-itW [16]	Authentic	Daily Scenes	Quality
TID2013 [40]	Artificial	Kodak Test Set	Quality
KADID-10k [30]	Artificial	Social Media	Quality
KonIQ-10k [20]	Authentic	MultiMedia	Quality
PAN (Ours)	**Adversarial**	**Autonomous Driving**	**Naturalness**

Previous IQA dataset vs our dataset

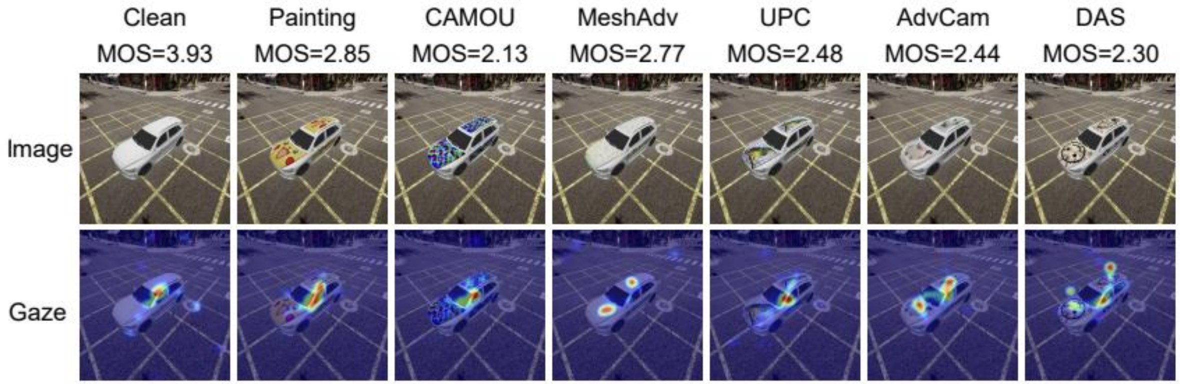
Our Contribution

- Contribute physical attack naturalness (PAN) dataset.
- Contains 2688 images with human ratings and gaze.
- Considers effect of environmental and semantic variations, with enhanced diversity

Human Rating: MOS

Raw Image

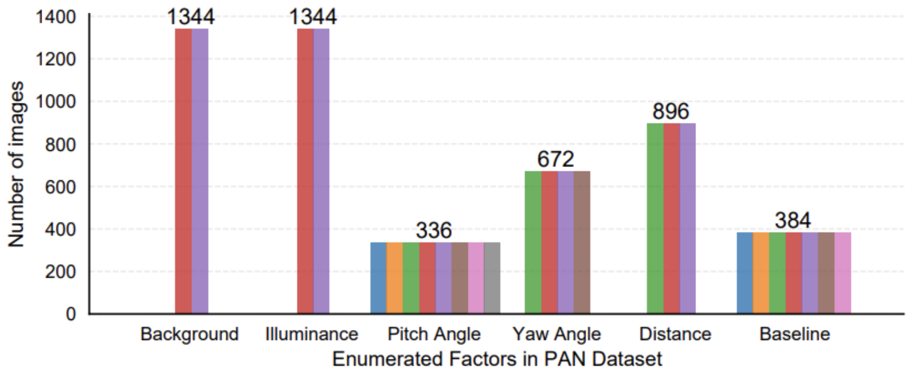
Human Gaze: Heatmap

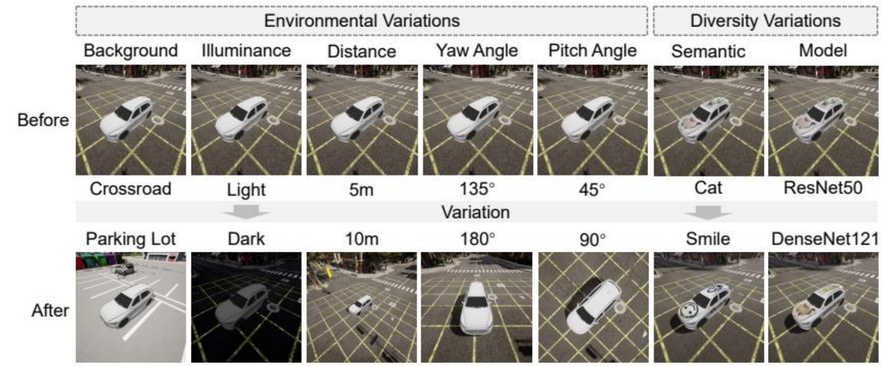


Physical Attack Naturalness (PAN) Dataset

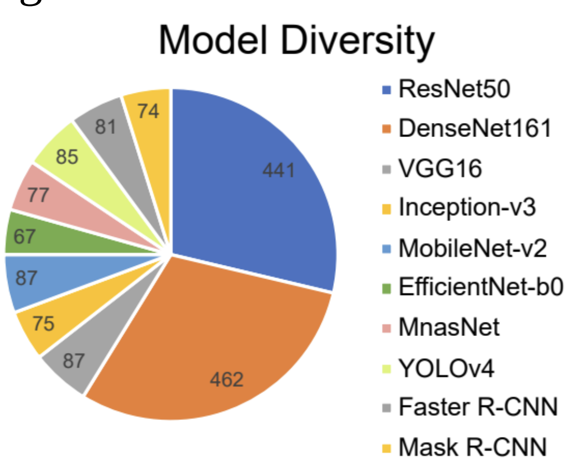
■PAN considers environmental variations, model diversity and semantic diversity

Environment Variations: background, illuminance, pitch/yaw, distance, baselines

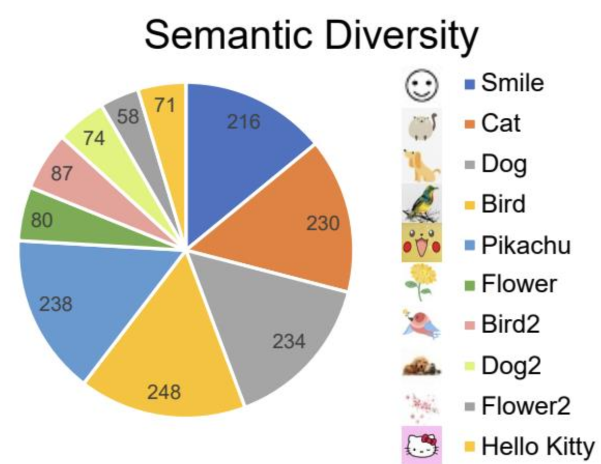




Model diversity:
generate attack on different model



Semantic diversity:
generate attack on different natural image



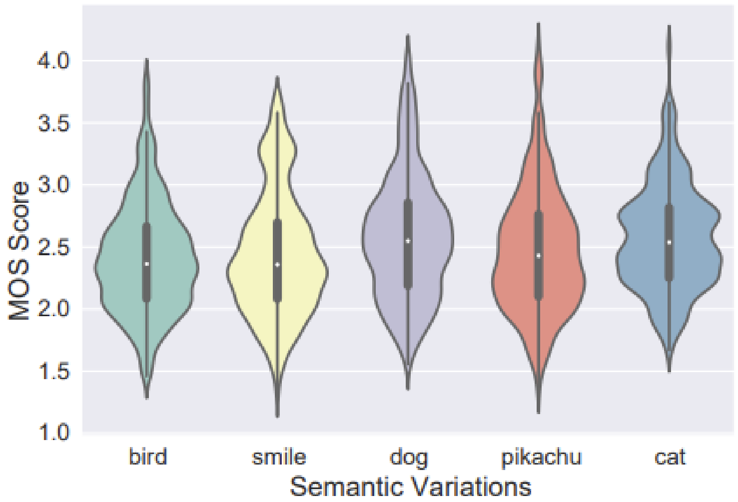
Insights from PAN

- Insight 1: Naturalness is affected by contextual features, including **semantic diversity** and **environmental variations**; Naturalness can be improved by selecting proper contextual features.

Impact of environment factor and baselines.
The effect is significant except background

Factors	Significance
Background	p=0.588, n.s.
Illumination	p<.001
Pitch angle	p<.001
Yaw angle	p<.001
Distance	p<.001
Baselines	p<.001

Impact of semantic diversity.
The effect is significant (p<.001)



Physical world attacks can be *more* stealthy at certain occasions!

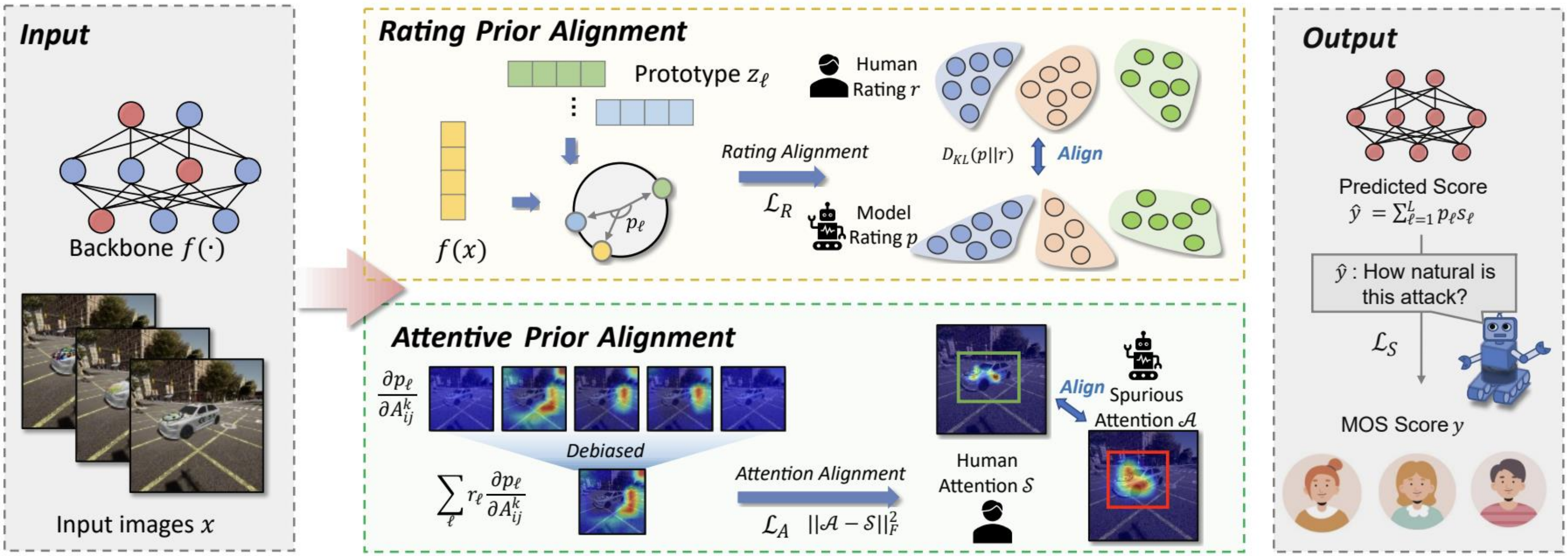
Insights from PAN

- Insight 2: Contextual features have disparate impact on naturalness of different attacks, which can lead to biased evaluation even under identical settings.
  - Different attacks can have different naturalness under certain **conditions**, while still being statistically significant
  - Should report naturalness results on multiple scenarios to avoid randomness

- Insight 3: Naturalness is correlated with behavioral feature (i.e., human gaze). Manipulation of human gaze can be a feasible direction to improve naturalness.
  - Attacks are considered less natural if gaze are more centralized ($p < .05$), or focus more on vehicle ($p < .001$)
  - A way to improve naturalness of attacks is to mislead human gaze.

Assess Naturalness by Dual Prior Alignment

■Human labels are expensive.
■How to automatically assess naturalness, without human participation?



Simple supervised training cannot sufficiently capture human value.
We propose **Dual Prior Alignment (DPA)** algorithm, which:
■Align model *rating distribution* with human *rating distribution*.
■Align model *attention* with human *gaze*.

Experiments

- Do we even need to collect PAN dataset?
  - Can methods trained on existing IQA dataset accurately evaluate naturalness?
  - Train on existing TID 2013 dataset, evaluate on PAN

Category	**Method**	**SROCC** (↑)	**PLCC** (↑)	$\mathbf{S_C}$ (↑)
FR-IQA	PSNR	0.3560	0.3685	-
	SSIM	0.4573	0.3968	-
	LPIPS	0.1056	0.1395	0.0583
	E-LPIPS	0.3990	0.3694	0.0727
Others	GIQA(KNN)	0.1382	0.1133	-
	GIQA(GMM)	0.1537	0.1392	-
NR-IQA	BRISQUE	0.1029	0.0494	-
	ResNet50	0.1149	0.1682	0.1692
	WaDIQaM	-0.0704	-0.1078	0.1821
	RankIQA	0.1809	0.1992	0.0095
	DBCNN	0.1409	0.1167	0.0876
	HyperIQA	0.1639	0.1285	0.2188
	Paq2Piq	0.0320	0.0504	0.2791
	MANIQA	0.2741	0.2717	0.0810
NR-IQA	DPA+PAN (Ours)	**0.7501**	**0.7727**	**0.7178**

- Results:
  - Existing IQA dataset do not solve the problem of naturalness evaluation!

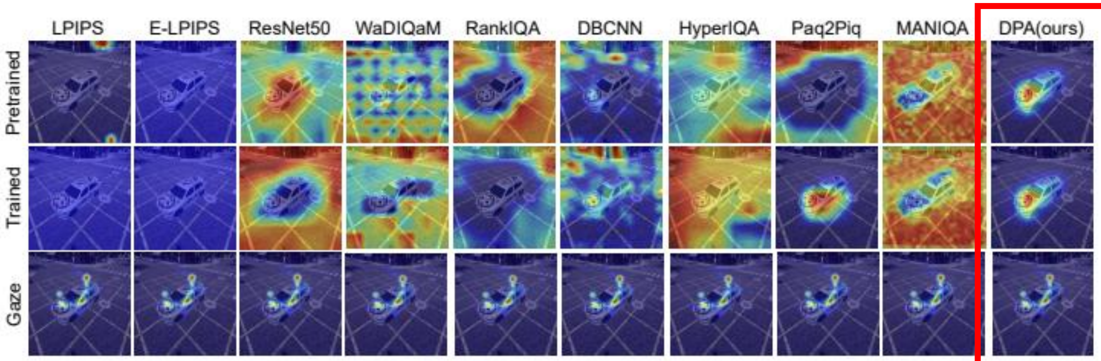
Experiments

- Do we get better result by Dual Prior Alignment?
  - Are human behaviors helpful?

Category	Method	SROCC (↑)	PLCC (↑)	S_C (↑)
FR-IQA	PSNR	0.3560	0.3685	-
	SSIM	0.4573	0.3968	-
	LPIPS	0.0994	0.1114	0.0089
	E-LPIPS	0.4082	0.4064	0.0136
Others	GIQA(KNN)	0.1428	0.1132	-
	GIQA(GMM)	0.0838	-0.0366	-
NR-IQA	BRISQUE	0.4753	0.3777	-
	ResNet50	0.6916	0.7453	0.2066
	WaDIQaM	0.6998	0.6841	0.2130
	RankIQA	0.7227	0.7564	0.1134
	DBCNN	0.6800	0.6621	0.3947
	HyperIQA	0.7253	0.7265	0.1955
	Paq2Piq	0.6044	0.6089	0.2003
	MANIQA	0.7129	0.7331	0.0861
NR-IQA	DPA (Ours)	**0.7501**	**0.7727**	**0.7178**

- Results:

Incorporating human behaviors are indeed helpful!



Model attention are also more aligned with human gaze, while others focus on spurious areas

Experiments

■Do we get better generalization to real world?

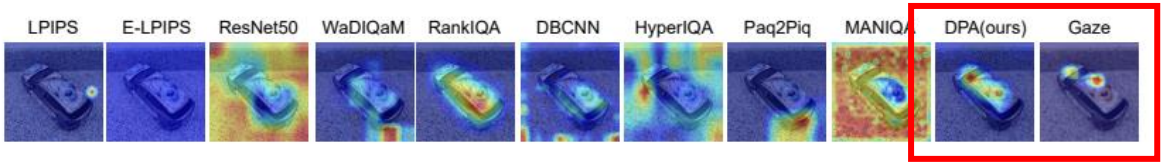
■Can DPA evaluate naturalness of new methods and scenarios?

Category	Method	SROCC (↑)	PLCC (↑)	S_C (↑)
FR-IQA	PSNR	0.3163	0.3009	-
	SSIM	0.3594	0.3558	-
	LPIPS	-0.2659	-0.3540	0.0163
	E-LPIPS	-0.3778	-0.3589	0.1658
Others	GIQA(KNN)	0.0075	0.0275	-
	GIQA(GMM)	0.0747	0.0809	-
NR-IQA	BRISQUE	0.0261	0.0245	-
	ResNet50	0.2874	0.3282	0.1935
	WaDIQaM	-0.1362	-0.1375	0.0329
	RankIQA	-0.1313	-0.1368	0.2942
	DBCNN	0.3907	0.4144	0.3028
	HyperIQA	0.3951	0.4416	0.3645
	Paq2Piq	0.3752	0.3905	0.2244
	MANIQA	0.3673	0.3839	0.2502
NR-IQA	DPA (Ours)	**0.4283**	**0.4652**	**0.4109**

■Results:

Our DPA also gets best generalization!

However, additional domain adaptation approach is required.



Model attention stay aligned with human gaze.

Thanks For Your Interest!