Robustness as Architecture: Designing IQA Models to Withstand Adversarial Perturbations

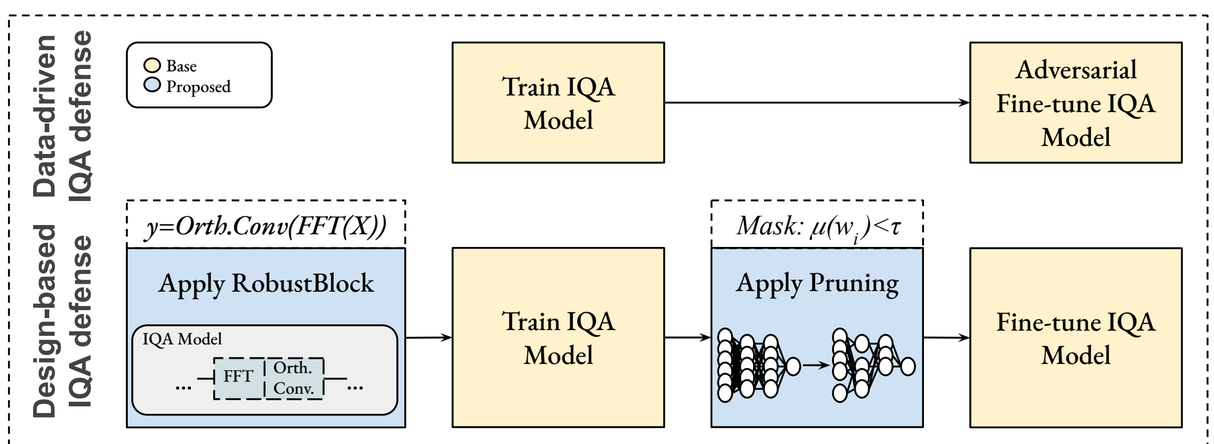
Igor Meleshin[1], Anna Chistyakova[1,2], Anastasia Antsiferova[2,3,4], Dmitriy Vatolin[1,2,3]
[1]Lomonosov Moscow State University, Moscow, Russia
[2]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
[3]MSU Institute for Artificial Intelligence, Moscow, Russia
[4]Laboratory of Innovative Technologies for Processing Video Content, Innopolis University, Innopolis, Russia

Motivation

Image Quality Assessment (IQA) models are critical for modern vision systems, from compression and enhancement to generation and streaming. However, they suffer from fundamental instability and can be easily compromised by minimal, visually imperceptible adversarial perturbations. This vulnerability leads to unreliable quality scores in critical applications and raises concerns about the trustworthiness of automated perceptual evaluation.

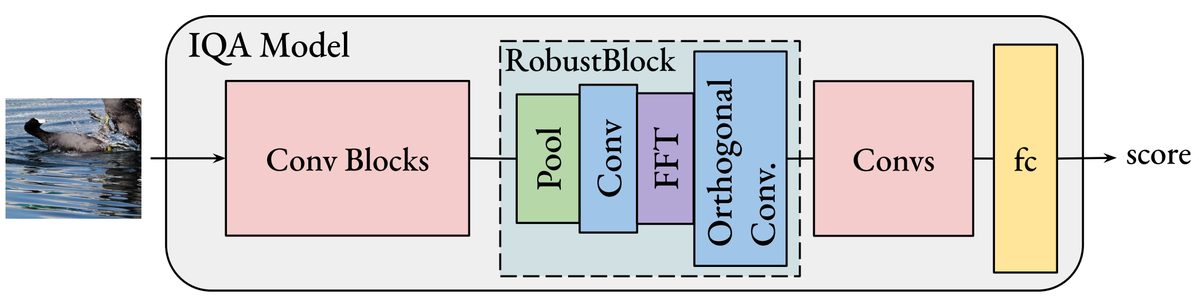
Idea

Traditional defenses rely on expensive and attack-specific data-driven retraining or purification. We propose an architectural modification to increase the robustness of IQA models against adversarial attacks.



Architecture

We introduce a lightweight, modular pipeline built around the RobustBlock: FFT-based feature transformation to frequency domain, orthogonal convolution for norm-preserving mappings, and pruning. RobustBlock receives a squared feature matrix as input, so we use pooling and convolutions to transform the input tensor.



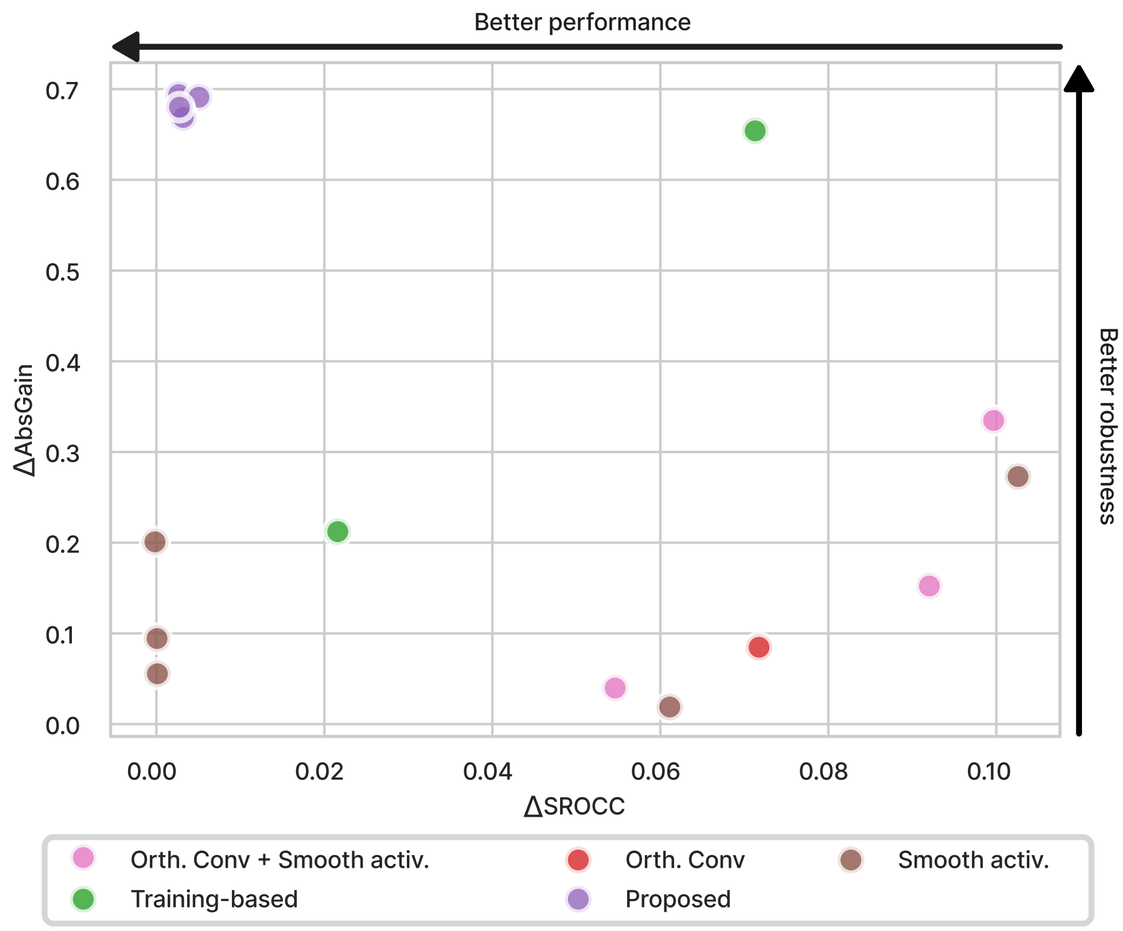
Best position

We compared a few positions of the RobustBlock in the architecture. The optimal placement of the RobustBlocks is near fully connected layers where perturbation amplification is minimal.

Modification	InputDim	SROCC	PLCC	AbsGain	RScore	Time	
						train epoch	test
---	---	**0.926**	**0.936**	0.409	0.012	**100.727**	**31.735**
AT	---	0.855	0.877	**0.209**	0.025	183.682	**32.216**
NT	---	0.904	0.921	**0.265**	**0.026**	229.612	33.028
RobustBlock 1	3×498×664	0.766	0.794	0.415	0.018	131.172	36.787
RobustBlock 2	64×125×166	0.793	0.823	0.508	0.006	**129.313**	63.864
RobustBlock 4	128×63×83	0.854	0.876	0.402	0.016	199.581	103.655
RobustBlock 6	1024×32×42	**0.923**	**0.936**	0.289	**0.031**	181.826	89.703

Comparison

Trade-off between robustness and quality correlation on the NIPS2017 dataset. Our architectural modification (purple) consistently reduces vulnerability to adversarial attacks (lower AbsGainAUC) while maintaining high perceptual alignment (minimal drop in SROCC).



Results

Performance and robustness metrics of IQA models with different architectural variations. AbsGain and R-Score are shown for UAP and stAdv, while area-under-curve (AUC) values are provided for PGD attacks with 1 and 8 iterations. The best result for each metric within each model group is highlighted in bold.

	SROCC	PLCC	AbsGainAUC	RScoreAUC	AbsGain	RScore	AbsGain	RScore
			PGD-1		UAP		stAdv	
Linearity	**0.926**	**0.936**	0.409	0.013	0.465	0.463	0.025	1.732
Linearity+NT	0.904	0.921	**0.305**	**0.026**	**0.190**	**0.825**	**0.015**	**1.937**
Linearity+AT	0.855	0.877	0.365	**0.026**	**0.322**	0.491	**0.013**	**1.985**
Linearity+our	**0.921**	**0.935**	**0.261**	**0.034**	0.360	**0.668**	0.022	1.862
KonCept	**0.915**	**0.929**	0.359	0.015	1.276	0.293	**0.006**	2.176
KonCept+NT	0.794	0.808	**0.122**	0.036	**0.419**	**0.608**	**0.001**	**3.873**
KonCept+AT	0.815	0.850	**0.062**	**0.046**	0.633	**0.488**	0.009	**2.386**
KonCept+our	**0.884**	**0.908**	0.244	**0.040**	**0.396**	0.413	**0.006**	2.324
TReS	**0.918**	**0.929**	**0.300**	**0.027**	0.472	0.181	**0.103**	0.862
TReS+NT	0.887	0.901	**0.307**	**0.029**	**0.273**	**0.395**	0.122	**1.427**
TReS+AT	0.886	**0.904**	0.357	0.026	**0.426**	**0.188**	0.117	0.846
TReS+our	**0.919**	**0.929**	0.314	**0.029**	0.503	0.178	**0.051**	**0.903**

Future work

Although our FFT-based robustness method provides strong gains, it introduces trade-offs like moderate computational overhead and limitations to convolutional networks with square inputs into RobustBlock. We view these constraints as opportunities for future work, focusing on faster spectral approximations and extending the design to new architectures.