Supplementary Material for Predictive Quality Assessment for Mobile Secure Graphics

Cas Steigstra
Scantrust / University of Amsterdam
cas.steigstra@gmail.com

Sergey Milyaev
Scantrust
sergey.milyaev@gmail.com

Shaodi You
University of Amsterdam
s.you@uva.nl

S1. Overview

This supplementary document provides additional details, figures, and tables to complement our main paper, "Predictive Quality Assessment for Mobile Secure Graphics." The content herein is intended to offer a deeper insight into our methodologies and provide the complete, unabridged results of our empirical evaluation.

S2. Detailed Model Descriptions

Here, we provide further implementation details for each of the model paradigms evaluated in our work.

Baselines. These simple heuristics establish a performance floor. The `Random` model draws a score from a uniform distribution $Q_{random} = U(0, 1)$. The `Sharpness` metric is the weighted average of gradient magnitudes above an Otsu-derived threshold: $Q_{\text{sharpness}}(I_S) = (\sum_{g>\tau_{otsu}} g \cdot H(g))/(\sum_{g>\tau_{otsu}} H(g))$, where $H(g)$ is the gradient histogram. The `Blur` metric is a specialized measure of effective edge width relative to the QR code's cell size: $Q_{\text{blur}}(I_S) = u/(t \cdot csp)$, where u is the number of gray pixels, t is the number of black-to-white transitions, and csp is the cell size in pixels.

General-Purpose IQA Models. The `BRISQUE`$^{\text{NSS}}$ model uses the standard implementation from [1]. It extracts a 36-dimensional feature vector based on the distribution of Mean Subtracted Contrast Normalized (MSCN) coefficients and feeds them into a pre-trained SVR. The `CLIP-IQA` models use the ViT-L/14 version of CLIP [3]. The semantic version uses prompts "Good photo" and "Bad photo". The attribute-based version uses engineered prompts: "A high-resolution scan of a qr-code with crisp, clear, distinct details" (positive) and "A low-resolution scan of a qr-code with blurry, washed-out, indistinct details" (negative). The score is the softmax probability of similarity to the positive prompt.

Unsupervised (Task-Adapted) Models. The `NIQE (SG)` model adapts the NIQE framework [2] by building a reference Multivariate Gaussian (MVG) model of BRISQUE features from a corpus of high-quality Secure Graphic (SG) scans (those with $M > 0.95$ in our training set). The quality score is the Mahalanobis distance to this reference model. `NIQE-LBP (SG)` follows the same principle but replaces the NSS-based features with histograms of Local Binary Patterns (LBP), hypothesizing that texture features are more suitable for non-NSS patterns.

Supervised (Handcrafted) Models. The `BRISQUE (SG|M)` model extracts the 36D BRISQUE feature vector for each image in our training set and trains a Support Vector Regressor (SVR) with an RBF kernel to map these features to the ground-truth score M. The `LBP (SG|M)` model uses a more sophisticated feature set based on locally weighted statistics of uniform LBP codes, adapted from [4]. An extensive hyperparameter search found that a configuration of $P = 8$ neighbors at a radius $R = 8$ yielded a 10-dimensional feature vector with the best separability, which was then used to train an SVR.

Supervised (End-to-End) Models. The shallow `CNN-3x32 (SG|M)` model consists of three convolutional blocks (3x3 Conv, ReLU, 2x2 MaxPool) followed by a regression head of two fully-connected layers. The `MobileNet (SG|M)` model uses the `MobileNetV2` architecture with randomly initialized weights, trained from scratch. Our main proposal, `MobileNet`$^{\text{IN}}$`(SG|M)`, uses the `MobileNetV2` backbone pre-trained on ImageNet, with the top classification layer replaced by our regression head, and the entire network is fine-tuned on our SG dataset.

S3. Full Evaluation Tables

The main paper presents a summarized version of the key results. Table S1 provides the complete, unabridged per-

Table S1. Full unabridged performance results for all models on both test sets. Performance is measured by ΔpAUC over the 0-70% discard rate range. Lower values are better. Best performance in each column is in **bold**, with second-best in *italics*.

Method	Digital (In-Domain)		Offset (Cross-Domain)	
	FNMR ΔpAUC	ISRR ΔpAUC	FNMR ΔpAUC	ISRR ΔpAUC
Baselines				
`Random`	0.2826	0.2967	0.2774	0.3000
`Sharpness`	0.0412	0.0398	0.0871	0.0894
`Blur`	0.0441	0.0414	0.0936	0.1003
General-Purpose IQA				
`BRISQUE`^{NSS}	0.3504	0.3862	0.2243	0.2468
`CLIP-IQA (Semantic)`	0.3810	0.4003	0.3970	0.4318
`CLIP-IQA (Attribute)`	0.2601	0.2701	0.2407	0.2530
Unsupervised (Task-Adapted)				
`NIQE (SG)`	0.0841	0.0828	0.1010	0.1027
`NIQE-LBP (SG)`	0.0273	0.0270	**0.0367**	**0.0356**
Supervised (Handcrafted)				
`BRISQUE (SG|M)`	0.0301	0.0295	0.0574	0.0607
`LBP (SG|M)`	0.0185	0.0173	*0.0390*	*0.0411*
Supervised (End-to-End)				
`CNN-3x32 (SG|M)`	0.0086	0.0086	0.3441	0.3662
`MobileNet (SG|M)`	*0.0063*	*0.0064*	0.1765	0.1788
`MobileNet`^{IN}`(SG|M)`	**0.0042**	**0.0042**	0.0800	0.0788

formance results for all evaluated models on both the in-domain (Digital) and cross-domain (Offset) test sets.

S4. Qualitative Results

This section provides qualitative examples to visually complement the quantitative results, see Figure S1.

S5. Ablation Study: Network Probing

This section provides the complete results for the network probing analysis described in Experiment 3 of the main paper.

S5.1. Probe Architectures

To test different hypotheses about the feature space, we designed two lightweight probe architectures.

Linear Probe (`lin`). This simplest probe tests the linear separability of features at a given layer. Its architecture is a Global Average Pooling (GAP) layer followed by a single fully-connected (Dense) layer that maps the feature vector to the final quality score.

`conv + lin` Probe. This probe has slightly more capacity. It consists of a 1×1 Convolutional layer, which acts as a channel-wise feature recombiner, followed by the same GAP and Dense layer structure as the linear probe. This allows the model to learn an optimal linear combination of feature channels before the final regression.

S5.2. Choice of Probe Layer

Table S2 provides the complete numerical results for both probe architectures and the fine-tuned model, attached to all investigated intermediate blocks of the frozen `MobileNetV2` backbone. This data provides a granular view of where predictive information resides within the network and highlights the trade-off between specialization and generalization, comparing in-domain and cross-domain performance against the fully fine-tuned model.

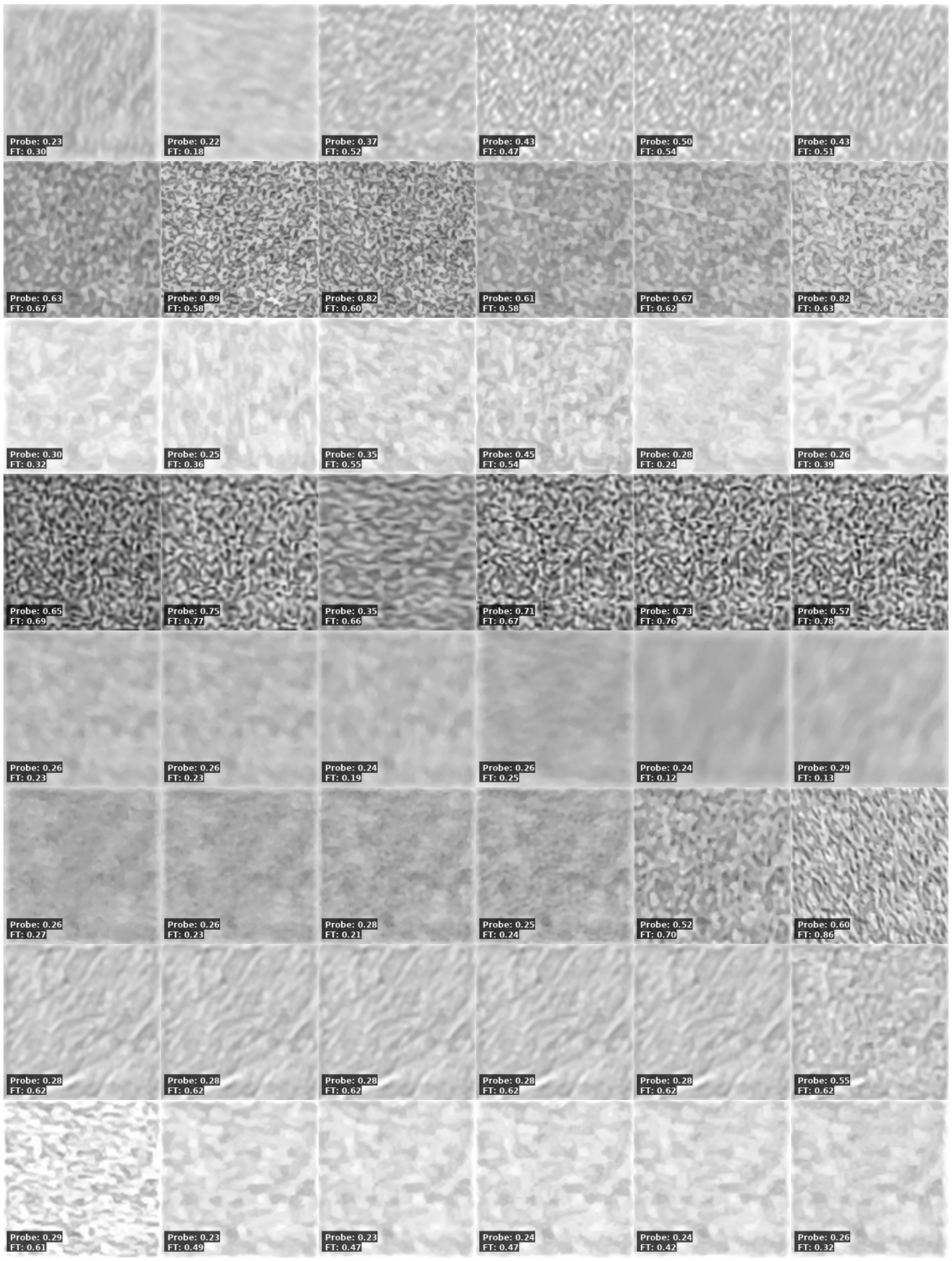

Figure S1. **Qualitative Comparison of Probe vs. Fine-Tuned (FT) Models on the Cross-Domain (Offset) Test Set.** This grid reveals a key failure mode of the FT model. When Offset prints are captured with slight blur or low detail (e.g., rows 3, 4, 7), their texture can incidentally resemble the source (Digital) domain's characteristics. The FT model, having overfitted to these source-domain artifacts, often assigns these defective frames an erroneously high quality score. Conversely, the Probe model proves more robust; it correctly assigns low scores to these degraded frames and, unlike the FT model, reserves its highest scores for the sharpest, highest-fidelity scans (e.g., rows 1 and 3). This highlights the probe's superior generalization in tracking true fidelity.

Table S2. Complete Unabridged Probe Performance vs. Fine-Tuned Model. This table compares probe performance against the fully fine-tuned model. Lower ΔpAUC values are better. The best result in each column is shown in **bold**.

Probe Arch.	Probed Layer	Digital (In-Domain)		Offset (Cross-Domain)	
		FNMR ΔpAUC	ISRR ΔpAUC	FNMR ΔpAUC	ISRR ΔpAUC
conv + lin Probes					
conv + lin	IB 1	0.0073	0.0077	0.0832	0.0944
conv + lin	IB 3	0.0042	0.0046	0.0371	0.0403
conv + lin	IB 6	**0.0041**	0.0044	0.0289	0.0323
conv + lin	IB 7	0.0042	0.0046	0.0262	0.0297
conv + lin	IB 10	0.0058	0.0060	0.0238	0.0269
conv + lin	IB 13	0.0054	0.0057	0.0290	0.0328
conv + lin	IB 14	0.0058	0.0065	0.0444	0.0529
conv + lin	IB 17	0.0063	0.0066	0.0380	0.0431
conv + lin	Conv 18	0.0071	0.0074	0.0471	0.0559
lin Probes					
lin	IB 1	0.0221	0.0218	0.0805	0.0851
lin	IB 3	0.0097	0.0100	0.0378	0.0407
lin	IB 6	0.0051	0.0055	0.0323	0.0364
lin	IB 7	0.0048	0.0052	0.0302	0.0325
lin	IB 10	0.0051	0.0054	0.0259	0.0284
lin	IB 13	0.0068	0.0071	**0.0231**	**0.0255**
lin	IB 14	0.0067	0.0070	0.0334	0.0372
lin	IB 17	0.0067	0.0070	0.0375	0.0435
lin	Conv 18	0.0084	0.0086	0.0383	0.0427
Fully Fine-Tuned Model (Reference)					
MobileNet^{IN}(SG|M)		0.0042	**0.0042**	0.0800	0.0788

References

[1] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 1

[2] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2013. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021. 1

[4] Qingbo Wu, Zhou Wang, and Hongliang Li. A highly efficient method for blind image quality assessment. In *IEEE Int. Conf. Image Process.*, pages 339–343, 2015. 1