

Brief Paper:

LTE-QA: Luminance and Texture Entropy for CCTV Image Quality Assessment

Yujin Han¹, Taewan Kim^{1*}

Abstract: Real world CCTV imagery is often perceived as low quality even without synthetic distortion, since operational conditions such as poor illumination and sparse texture limit visibility. Conventional no-reference image quality assessment (NR-IQA) methods, designed for natural images with artificial distortions, do not capture this surveillance-specific behavior well. We analyze human perceptual quality on a CCTV IQA benchmark and find that two interpretable cues, luminance entropy and texture entropy, are strongly correlated with subjective mean opinion scores (MOS). Motivated by this, we propose LTE-QA, a transformer-based NR-IQA model that takes spatial maps of these entropy measures as direct input, achieving higher correlation with subjective ratings than representative NR-IQA baselines.

Key Words: Image Quality Assessment, CCTV, Luminance Entropy, Texture Entropy.

I. INTRODUCTION

Predicting how humans perceive image quality is a central goal of no-reference image quality assessment (NR-IQA), which estimates perceived quality without an undistorted reference [1]. Most NR-IQA progress has relied on benchmarks with artificially applied distortions such as blur, noise, or compression [2-3], assumptions that rarely hold for footage from deployed closed-circuit television (CCTV) systems, which is typically captured without post-hoc distortion yet still perceived as unclear or hard to interpret [4-5], since low illumination flattens contrast, sparse backgrounds suppress texture, and limited salient content reduces semantic interpretability.

In this paper, we model human perceptual quality for distortion-free CCTV imagery. Using CQA-DB, a CCTV IQA benchmark with subjective mean opinion scores (MOS) [6], we find that the entropy of local luminance and local texture account for much of the variation in human ratings. Motivated by this, we propose luminance and texture entropy based quality assessment (LTE-QA), a transformer-based NR-IQA model that takes spatial maps of these two entropy measures as direct input, achieving higher agreement

with subjective ratings than representative baselines on real CCTV images. Perceived quality is also operationally consequential, since surveillance footage feeds downstream analytics such as abnormal behavior recognition [7] and personalized anomaly alarm systems [8], whose reliability degrades on frames that human operators already find hard to read.

II. RELATED WORK

2.1. Conventional IQA Benchmarks

Early IQA benchmarks such as LIVE [1], TID2013 [2], and KADID-10k [3] apply synthetic distortions to clean references, while in-the-wild datasets such as KonIQ-10k [9] and SPAQ [10] capture authentic degradations in consumer photographs; these remain distinct from operational surveillance conditions. Recent transformer-based methods [11-12] further improve accuracy on such benchmarks by jointly modeling local distortion and semantic context, an idea also pursued through long-range interaction between local perceptual responses [13] and extended to video- and action-level quality prediction [14]. Beyond the images themselves, the subjective protocol used to label a benchmark also shapes what it measures; continuous and multimodal scoring has been shown to yield more stable subjective quality data than single-stimulus rating [15].

2.2. Surveillance IQA

CCTV imagery presents distinct challenges for perceptual modeling. Unlike natural photographs, CCTV frames are captured under operational rather than photographic conditions, and their perceptual variability arises from environmental factors such as uneven illumination and sparse texture rather than explicit corruption [4], so distortion-centric IQA models do not capture them well even without synthetic distortion. Unlike prior transformer-based NR-IQA models [11-12] that learn distortion cues implicitly, LTE-QA explicitly encodes luminance and

Manuscript received July 13, 2026; Revised August 25, 2026; Accepted August 28, 2026. (ID No. JMIS- 26M-07-032)

Corresponding Author (*): Taewan Kim, +82-2-940-4751, kimtwan21@dongduk.ac.kr

¹Department of Data Science, Dongduk Women's University, Seoul, Korea, hanyujinius@gmail.com, kimtwan21@dongduk.ac.kr

texture entropy as spatial maps fed directly into the transformer, targeting CCTV-specific degradations rather than the synthetic corruptions such models typically assume. Surveillance-specific quality benchmarks have only recently begun to appear, including CQA-DB for general CCTV scenes [6] and HIQA-DB for hospital surveillance [5], and both report that entropy-like statistics track subjective ratings more closely than distortion-oriented features do.

III. DEFINING PERCEPTUAL CUES FOR LTE-QA

3.1. Luminance Entropy

Given an image I , the luminance entropy H_{lum} is defined over the normalized histogram of its luminance component as

$$H_{\text{lum}}(I) = -\sum_{l \in L} p_l \log_2 p_l. \quad (1)$$

A higher value indicates a broader distribution of brightness levels, typically corresponding to better-lit, higher-contrast scenes, while underexposed or low-contrast frames common in night-time surveillance yield smaller values. We verify this relationship empirically on CQA-DB, obtaining a PLCC of 0.65 and an SRCC of 0.64 between luminance entropy and subjective MOS, as shown in Fig. 1 (left). The luminance histogram uses 256 bins over the

standard 8-bit range, and coarser binning did not noticeably affect the resulting entropy values.

3.2. Texture Entropy

To characterize structural complexity, the texture entropy H_{tex} is computed from the responses of a steerable pyramid decomposition modeled by a generalized Gaussian distribution (GGD), and is defined as

$$H_{\text{tex}}(I) = \sum_{i=1}^N \left(\frac{1}{\gamma_i} - \log \left(s \cdot \gamma_i \cdot \Gamma \left(\frac{1}{\gamma_i} \right) \right) \right). \quad (2)$$

Larger values reflect richer fine structures such as edges, foliage, or cluttered backgrounds, while flat or empty scenes yield lower values, texture entropy likewise correlates positively with subjective MOS, yielding a PLCC of 0.63 and an SRCC of 0.62, as shown in Fig. 1 (right). The pyramid decomposition uses three scales and four orientations, and the fitted GGD closely matches the empirical response histograms, supporting its use here. Entropy-based descriptors have proven perceptually informative in adjacent quality domains as well, including oculomotor demand in short-form video [16] and kinematic diversity in action quality assessment [14].

IV. LTE-QA MODEL

LTE-QA is a transformer-based NR-IQA model that takes luminance and texture entropy maps as direct input rather than raw pixel intensities, explicitly encoding the two perceptual cues that govern human ratings on CCTV imagery. The overall pipeline is illustrated in Fig. 2.

4.1. Entropy Map Generation

Given an input CCTV image I , we compute two spatially resolved entropy maps that encode brightness diversity and structural richness. The luminance map $\mathbf{M}_{\text{lum}}$ is obtained by evaluating H_{lum} over disjoint local patches of I ,

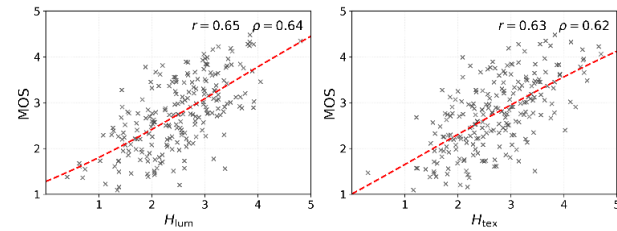


Fig. 1. Scatter plots of luminance entropy H_{lum} (left) and texture entropy H_{tex} (right) against subjective MOS on CQA-DB.

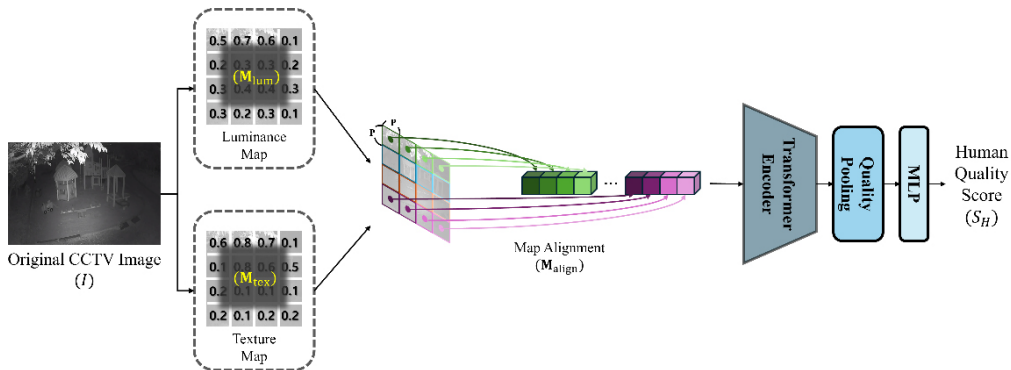


Fig. 2. Overview of LTE-QA. An input CCTV image is converted into luminance and texture entropy maps, tokenized and encoded by a transformer, then pooled and mapped to a quality score by an MLP regressor.

producing a per-patch measure of brightness diversity. The texture map $\mathbf{M}_{\text{tex}}$ is computed in the same manner using H_{tex} , capturing local structural richness. Both maps are normalized and resized to a common spatial resolution, and are then concatenated along the channel dimension to form the aligned representation as

$$\mathbf{M}_{\text{align}} = [\mathbf{M}_{\text{lum}} \parallel \mathbf{M}_{\text{tex}}], \quad (3)$$

which preserves the spatial correspondence between the two cues and serves as a unified input tensor for the subsequent transformer. Both maps are resized to 224×224 and combined via min-max normalization, which gave more stable training than z-score normalization.

4.2. Transformer Architecture

The aligned tensor $\mathbf{M}_{\text{align}}$ is divided into disjoint patches and linearly projected into a sequence of token embeddings. For the p -th patch, the embedding is computed as

$$\mathbf{z}_p = \mathbf{W}_p \text{Flatten}(\mathbf{M}_{\text{align}}^{(p)}) + \mathbf{b}_p, \quad (4)$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are learnable parameters of the linear projection. Following the standard Vision Transformer (ViT) formulation, and consistent with prior evidence that long-range attention benefits perceptual quality regression [13], learnable position embeddings are added to the resulting tokens before they are passed through a transformer encoder. A quality pooling operation aggregates the encoded tokens into a single feature vector, which is then mapped to a scalar quality score S_H by a multilayer perceptron (MLP) regressor. The model is trained end to end using the mean squared error (MSE) loss between the predicted score and the ground truth MOS as

$$\mathcal{L} = (S_H - S_H^*)^2. \quad (5)$$

We further add a ranking-aware pairwise loss alongside MSE, which improves accuracy as shown in Table 1. We also examine design choices, where element-wise addition

Table 1. Training and regressor comparison.

Category	Method	PLCC $\uparrow$	SRCC $\uparrow$
Optimizer	SGD with momentum	0.852	0.840
	AdamW (ours)	0.912	0.901
Loss function	MSE	0.881	0.869
	MSE+Ranking loss (ours)	0.912	0.901
Regressor	Linear regressor	0.864	0.851
	MLP regressor (ours)	0.912	0.901

and a convolutional-stem embedding underperform our design, and 8×8 patches trade accuracy for roughly four times the compute. Removing the position embedding lowers performance, and attention pooling outperforms the CLS token and mean pooling.

V. EXPERIMENTS

5.1. Experimental Setup

We evaluate LTE-QA on the CQA-DB benchmark, which provides authentic CCTV images with subjective MOS labels. The model uses a ViT-Base (ViT-B/16) backbone in PyTorch, with entropy maps resized to 224×224 and partitioned into 16×16 patches. We train with the AdamW optimizer (learning rate 1×10^{-4} , weight decay 1×10^{-5}) for 200 epochs, and report SRCC and PLCC to measure agreement between predicted scores and MOS. LTE-QA also generalizes beyond CCTV, outperforming the strongest baseline, MUSIQ, on KonIQ-10k, as shown in Table 2.

5.2. Quantitative Comparison

We compare LTE-QA with two classical natural scene statistics based methods, BRISQUE [17] and NIQE [18], and three learning-based methods, DBCNN [19], HyperIQA [20], and MUSIQ [21]. LTE-QA attains the best correlation among all methods, showing that entropy-guided representations align more closely with human ratings than handcrafted or generic learned features. A lightweight ViT-Tiny backbone remains competitive with our full ViT-Base model, and AdamW with a ranking-aware loss term improves accuracy and lowers the root mean squared error (RMSE) over SGD and MSE alone, as shown in Table 1.

5.3. Ablation Study

We find that combining luminance and texture entropy consistently outperforms using either alone, as shown in

Table 2. Quantitative comparison on CQA-DB and KonIQ-10k.

Method	Backbone	CQA-DB			KonIQ-10k	
		PLCC	SRCC	RMSE	PLCC	SRCC
BRISQUE [17]	-	0.624	0.601	12.45	0.665	0.651
NIQE [18]	-	0.512	0.498	15.12	0.542	0.531
DBCNN [19]	ResNet-50	0.814	0.802	8.32	0.854	0.841
HyperIQA [20]	ResNet-50	0.845	0.833	7.64	0.872	0.859
MUSIQ [21]	ResNet-50	0.871	0.860	6.88	0.891	0.880
LTE-QA (ours)	ViT-Tiny	0.865	0.854	7.01	0.884	0.872
LTE-QA (ours)	ViT-Base	0.912	0.901	5.12	0.924	0.915

Table 3, confirming that the two capture complementary aspects of perceived quality. Similarly, a linear regressor and single-scale entropy maps both underperform our full configuration, as shown in Table 1 and Table 3.

Texture entropy contributes more to performance than luminance entropy, likely because CCTV scenes vary more in structural clutter than in overall brightness.

5.4. Qualitative Analysis

We inspect representative CCTV scenes to verify how LTE-QA responds to varying entropy patterns. Scenes with high H_{lum} and H_{tex} receive high predicted S_H as shown in Fig. 3, while dark or texture-poor frames receive lower scores. Asymmetric cases, where one entropy is high and the other low, produce intermediate S_H . Performance remains stable across conditions in Table 4. The modest drop under night conditions (IN/ON) likely reflects reduced luminance dynamic range, partially compensated by texture entropy.

VI. CONCLUSION

In this paper, we addressed human perceptual quality assessment for distortion-free CCTV imagery, showing on CQA-DB that luminance entropy and texture entropy are strongly correlated with subjective MOS. Building on this, we proposed LTE-QA, a transformer-based NR-IQA model

Table 3. Ablation on entropy components.

Configuration	PLCC $\uparrow$	SRCC $\uparrow$
LTE-QA (Full, multi-scale)	0.912	0.901
w/o luminance entropy	0.854	0.841
w/o texture entropy	0.832	0.820
LTE-QA (single-scale)	0.879	0.864

Table 4. Robustness across illumination and scene conditions.

Condition	PLCC $\uparrow$	SRCC $\uparrow$
Indoor / day (ID)	0.921	0.910
Indoor / night (IN)	0.894	0.881
Outdoor / day (OD)	0.915	0.902
Outdoor / night (ON)	0.891	0.879

that directly encodes these entropy maps as input, achieving higher agreement with human ratings than representative NR-IQA baselines. We note that the present framework is scoped to luminance and texture entropy as interpretable perceptual cues for authentic, distortion-free CCTV imagery, and does not yet incorporate other perceptual factors such as color fidelity or motion blur. Extending LTE-QA to additional perceptual cues, and to video-level settings where temporal entropy cues have proven informative [16,22], remains a direction for future work.

ACKNOWLEDGEMENT

This study was supported by the Dongduk Women's University grant.

REFERENCES

- [1] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, 2006.
- [2] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, and J. Astola, et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57-77, 2015.

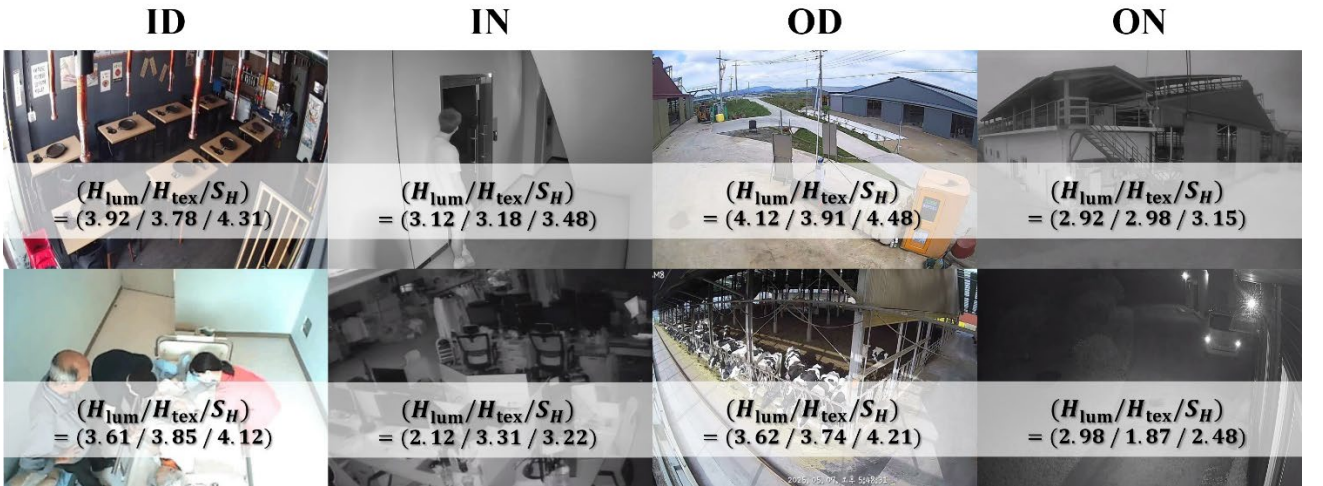


Fig. 3. Qualitative examples of LTE-QA across indoor-day (ID), indoor-night (IN), outdoor-day (OD), and outdoor-night (ON) scenes. Each image is annotated with H_{lum} , H_{tex} , and S_H .

- [3] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1-3.
- [4] D. Chen, T. Wu, K. Ma, and L. Zhang, "Toward generalized image quality assessment: Relaxing the perfect reference quality assumption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 12742-12752.
- [5] Y. Han and T. Kim, "HIQA-DB: A benchmark dataset for image quality assessment in hospital surveillance," in *Proceedings of the APSIPA Annual Summit and Conference*, 2025.
- [6] Y. Han, J. Kang, S. Lee, and T. Kim, "Understanding perceptual quality in CCTV images: A benchmark dataset and entropy-based insights," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Oct. 2025, pp. 3412-3419.
- [7] H. Song, J. Kang, and T. Kim, "Real-time abnormal behavior recognition for patient monitoring in hospitals," in *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2024, pp. 1-8.
- [8] H. Song, J. Kang, and T. Kim, "Continual learning based personalized abnormal behavior recognition alarm system," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, pp. 1-29, 2024.
- [9] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041-4056, 2020.
- [10] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3677-3686.
- [11] J. Shi, P. Gao, and A. Smolic, "Blind image quality assessment via transformer predicted error map and perceptual quality token," *IEEE Transactions on Multimedia*, vol. 26, pp. 4641-4651, 2024.
- [12] Q. Mao, S. Liu, Q. Li, G. Jeon, H. Kim, and D. Camacho, "No-reference image quality assessment: Past, present, and future," *Expert Systems*, vol. 42, p. e13842, 2025.
- [13] H. Oh, J. Kim, T. Kim, and S. Lee, "Convolved quality transformer: Image quality assessment via long-range interaction between local perception," *IEEE Access*, vol. 10, pp. 102968-102980, 2022.
- [14] D. Kim, T. Kim, I. Lee, and S. Lee, "Kinematic diversity and rhythmic alignment in choreographic quality transformers for dance quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5677-5692, 2024.
- [15] T. Kim, J. Kang, S. Lee, and A. C. Bovik, "Multimodal interactive continuous scoring of subjective 3D video quality of experience," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 387-402, 2014.
- [16] T. Kim, "Quality is not comfort: Oculomotor demand entropy for short-form videos," *IEEE Signal Processing Letters*, vol. 33, pp. 3019-3023, 2026.
- [17] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695-4708, 2012.
- [18] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209-212, 2013.
- [19] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36-47, 2020.
- [20] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, and J. Sun, et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667-3676.
- [21] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148-5157.
- [22] Y. Han, Y. Kim, and T. Kim, "SF-DanceQA: A benchmark dataset for dance quality assessment in short-form videos," in *Proceedings of the APSIPA Annual Summit and Conference*, 2026.

AUTHORS



Yujin Han is a fourth-year student in the Division of Interdisciplinary Studies in Cultural Intelligence at Dongduk Women's University, Seoul, South Korea. Currently, she is pursuing her B.S. degree in Data Science. Her research interests include computer vision and machine learning.



Taewan Kim received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2008, 2010, and 2015, respectively. From 2015 to 2021, he was with the Vision AI Laboratory, SK Telecom, Seoul. In 2022, he joined as a faculty with the Division of Future Convergence (Data Science Major), Dongduk Women's Uni-

versity, Seoul, where he is currently an Assistant Professor. His research interests include computer vision and machine learning including continual and online learning.