Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 3, 2026, 291-299.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

A Siamese Network Based on InceptionV3 with Custom Loss Functions for Document Image Quality Assessment

Mohammad Hossein Khosravi*

Electrical and Computer Engineering Faculty, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 04 August 2025
Revised 15 September 2025
Accepted 06 June 2026

DOI: 10.22044/jadm.2026.16621.2788

Keywords:
Document Image Quality Assessment, Siamese Network, InceptionV3, Deep Learning, Custom Loss Functions.

**Corresponding author: mohokhosravi@birjand.ac.ir (M. H. Khosravi).*

Abstract

Document Image Quality Assessment (DIQA) is critical for ensuring the reliability of downstream applications such as Optical Character Recognition (OCR), digital archiving, and automated document workflows. In this paper, we propose a deep learning-based DIQA framework using a Siamese neural network architecture with an InceptionV3 backbone. Our model leverages a composite loss function that combines linear regression loss with a monotonic ranking constraint to jointly optimize for score-level accuracy and perceptual consistency. Unlike prior works that rely on handcrafted features or narrow degradation types, our approach generalizes across diverse distortions commonly observed in scanned and photographed documents. Experimental results on the SOC and SmartDoc-QA datasets demonstrate that the proposed model exhibits a strong correlation with OCR accuracy, achieving SROCC values of 0.952 and 0.873, respectively, and outperforming several state-of-the-art DIQA methods.

1. Introduction

The digitization of documents has become indispensable in modern information systems, driven by the need for efficient storage, retrieval, and automated processing [1]. Document Image Quality Assessment plays a pivotal role in ensuring the usability of scanned or camera-captured documents, particularly for OCR, archival systems, and automated workflows. Traditional quality assessment methods often rely on human evaluation or handcrafted features, which are time-consuming, subjective, and fail to generalize across diverse distortion types [2]. With the exponential growth of digitized documents, there is a pressing need for automated, robust, and scalable DIQA methods that can mimic human perception while adapting to varied document conditions.

Despite advancements in image quality assessment (IQA) for natural scenes, document images present unique challenges due to their structural properties and distortion profiles. Unlike natural images, document distortions—such as blur, noise, uneven lighting, and compression artifacts—directly impact textual readability and OCR performance. Existing IQA metrics like PSNR or SSIM are inadequate for DIQA, as they ignore semantic content and focus solely on pixel-level discrepancies. Moreover, OCR-based DIQA methods, while practical, are computationally expensive and struggle with non-textual document elements (e.g., tables, figures). These limitations underscore the need for a specialized approach that bridges the gap between human perception and computational efficiency.

Recent deep learning techniques, particularly convolutional neural networks (CNNs), have shown promise in IQA by learning hierarchical features directly from data [3]. However, most CNN-based DIQA methods adopt generic architectures (e.g., VGG, ResNet) designed for natural images, neglecting document-specific features like stroke consistency, character clarity, and background uniformity. Furthermore, few studies address the ordinal nature of quality scores, where the relative ranking of images matters as much as absolute values [4]. A robust DIQA system must not only predict quality scores accurately but also preserve the monotonic relationship between distortions and their perceived severity—a nuance often overlooked in current models.

To bridge the limitations of previous approaches, this study introduces a Siamese neural network architecture built upon the InceptionV3 backbone, complemented by custom-designed loss functions specifically tailored for DIQA. A Siamese architecture is adopted because it effectively supports pairwise learning of perceptual quality differences, enabling the model to learn relative quality rankings between document image patches without requiring a pristine reference image. This design leverages the Siamese network's strength in similarity and ranking learning, which aligns naturally with the subjective and comparative nature of document quality assessment.

While a prior study [4] employed a ResNet-based Siamese model, we hypothesize that the architectural characteristics of InceptionV3—particularly its multi-scale convolutional filters—are better suited for modeling diverse document degradation patterns that occur across multiple spatial scales, thereby motivating its selection as the backbone for our Siamese framework. The proposed Siamese network is trained on individual pairs of document patches to learn patch-level quality relationships. During inference, multiple patches are sampled from each document image, and their predicted scores are aggregated to produce a single document-level quality estimate. This strategy enables the model to capture both local distortions and overall structural quality. The learning process is guided by a composite loss function consisting of a linear regression term for precise score prediction and a monotonic ranking term to preserve perceptual consistency. This dual-objective formulation improves alignment with both Pearson and Spearman correlation measures relative to OCR accuracy. Experimental evaluations on the SmartDoc-QA and SOC datasets confirm that our method outperforms most existing DIQA techniques, achieving stronger correlation with subjective and task-driven quality metrics, while maintaining robustness against real-world document distortions. Overall, this work contributes to the advancement of DIQA by integrating powerful deep feature extraction with perceptual modeling in a scalable and practical framework.

2. Related works

Early approaches to DIQA primarily relied on handcrafted features and heuristic models [2]. Traditional metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), adapted from general Image Quality Assessment (IQA), proved inadequate for documents because they operate purely at the pixel level and fail to capture text-specific degradations such as blur, uneven illumination, and contrast loss [2].

Alaei et al. proposed a full-reference DIQA method using Local Binary Patterns (LBP) to predict mean human opinion scores (MHOS) [5], and later introduced a no-reference approach based on a quality-aware bag of visual words constructed using k-means clustering [6]. Another method based on HAST derivations [7] showed competitive performance on JPEG2000-compressed documents. Although these handcrafted approaches provided interpretable models, they were sensitive to feature design, limited in robustness to diverse distortions, and failed to generalize across varying document layouts and text densities.

To overcome these limitations, OCR-based DIQA methods were proposed, inferring quality indirectly from OCR performance. For example, Li et al. employed MSER-based character patch detection combined with gradient features to predict document quality [8]. , and Obafemi-Ajayi et al. designed an MLP-based ensemble framework trained on human perceptual rankings [9]. These approaches better aligned with human-perceived readability but suffered from two main drawbacks: (1) dependence on OCR engines made them computationally expensive and domain-specific, and (2) their performance degraded on documents containing non-textual elements such as tables, graphics, or handwritten content.

The advent of deep learning introduced data-driven solutions that automated feature extraction and improved generalization across diverse document types [3]. CNN-based architectures, such as ResNet [10], were fine-tuned for DIQA tasks, while Kang et al. [3] used CNNs to predict OCR accuracy from informative patches, achieving strong performance on benchmark datasets. Li et al. [11] proposed TextNet, an encoder–decoder model that jointly detects text lines and predicts their quality, while Lu et al. [12] employed two-stage transfer learning for OCR-oriented DIQA. Peng et al. introduced a Siamese deep network with customized loss functions to improve linearity and monotonicity [4], and Gao et al. presented DeQA-Doc, a multi-modal large language model integrating visual–linguistic understanding for generalizable quality assessment [13].

Despite progress, critical gaps remain in DIQA research. First, few studies explicitly model the monotonicity of quality scores—a key aspect of human perception. Second, most deep learning methods use generic architectures that lack document-optimized design, such as multi-scale feature extraction. Our work addresses these gaps by (1) leveraging InceptionV3's hierarchical filters for document-specific feature learning, (2)

introducing custom loss functions to enforce score monotonicity. By unifying these innovations, we advance DIQA toward practical, human-aligned automation.

3. Proposed method

Our proposed document image quality assessment framework consists of three key components: (1) a Siamese network architecture with a shared InceptionV3 backbone, (2) multi-scale feature extraction and fusion, and (3) dual custom loss functions for quality prediction optimization. Figure 1 shows the overall architecture of our system.

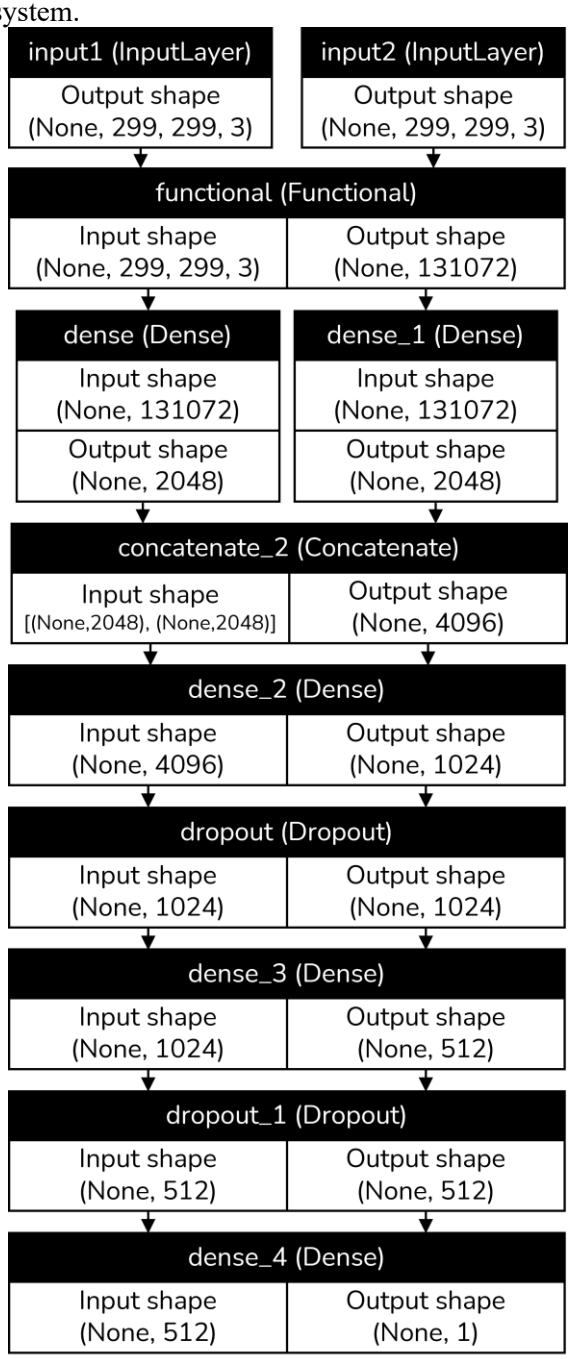


Figure 1. Layer-wise structure of proposed Siamese model.

3. 1. Siamese Network Architecture

The proposed Siamese network is designed for no-reference document image quality assessment, where the goal is to predict perceptual quality without access to a pristine reference image. The network consists of two identical branches with shared weights. For each document image, multiple non-overlapping patches of size $299 \times 299 \times 3$ pixels normalized to the $[0,1]$ range are randomly extracted from different regions to capture diverse local characteristics, including textual, graphical, and background areas. These patches are fed simultaneously into the parallel branches of the Siamese network, all sharing identical weights. Unlike traditional Siamese networks that compare pairs of distinct images, our architecture focuses on multi-patch processing within a single document. This intra-document setup allows the model to integrate information from spatially diverse regions, enabling robust quality prediction even in the presence of localized distortions.

Each branch employs a modified InceptionV3 backbone for feature extraction, in which the classification head is removed, a global average pooling layer is added, and batch normalization and dropout (rate = 0.5) are applied after each convolutional block to enhance generalization.

During training, the network receives pairs of patches (x_i, x_j) with known relative quality rankings, derived from OCR scores. Both branches independently extract features f_i and f_j, which are concatenated in a fusion layer to form a joint representation $[f_i, f_j]$. This fused feature vector passes through a regression head composed of two fully connected layers (1024 and 512 neurons with ReLU activations) and a final sigmoid-activated neuron producing the predicted quality score.

The learning objective combines a regression loss (e.g., MSE between predicted and target scores) with a monotonicity loss that penalizes violations of the expected ranking order between paired samples. This Siamese-based comparative learning enforces score consistency and encourages the model to capture subtle perceptual differences among document patches, even without a reference image. The shared-weight design ensures uniform feature extraction while improving computational efficiency.

During training, the network learns to predict the document-level quality score by jointly optimizing over multiple patches. At inference, the final document quality is obtained by aggregating the predicted scores from the sampled patches.

The selection of InceptionV3 as the foundational architecture for our Siamese network is motivated by the following principal considerations, each substantiated through empirical evidence and architectural analysis:

1) **Multi-Scale Feature Representation**: The inception modules' parallel convolutional pathways (1×1 , 3×3 , and 5×5 kernels) enable simultaneous extraction of both local and global document features. This hierarchical processing is particularly advantageous for DIQA, where text legibility depends on stroke-level details (captured by smaller kernels) while overall quality assessment requires layout-level analysis (addressed by larger receptive fields).
2) **Computational Efficiency**: Through dimensionality-reducing bottleneck layers (1×1 convolutions) and aggressive feature map compression, InceptionV3 achieves superior parameter efficiency (23.8M parameters) compared to conventional architectures (VGG16: 138M or ResNet152: 60.2M). This optimization is critical for our Siamese implementation, where weight sharing across multiple branches necessitates lightweight operations.
3) **Image Analysis Provenance**: Prior research has established InceptionV3's efficacy in image and specifically document processing tasks, including document classification [14], handwritten keyword spotting (KWS) [15], human fingerprint verification [16], handwritten numeral recognition [17] and ancient character recognition [18]. These results validate its capacity to handle document-specific distortions through learned hierarchical representations.
4) **Siamese Architecture Compatibility**: InceptionV3 possesses several architectural characteristics that make it particularly well-suited for use within Siamese networks. First, it consistently produces 2048-dimensional feature vectors, which allows for direct and meaningful comparison between the outputs of the two network branches. Second, its modular structure facilitates efficient weight sharing across branches, a crucial requirement in Siamese architectures. Lastly, the built-in batch normalization layers contribute to the stability of the training process by normalizing intermediate feature distributions, thereby improving convergence in multi-branch settings.

The combined advantages of multi-scale processing, computational efficiency, and proven document analysis capabilities establish InceptionV3 as the optimal backbone for our quality assessment framework. Quantitative validation of this selection is presented in Section 4.3 through comparative experiments with alternative architectures. Figure 2 shows the overall InceptionV3 architecture.

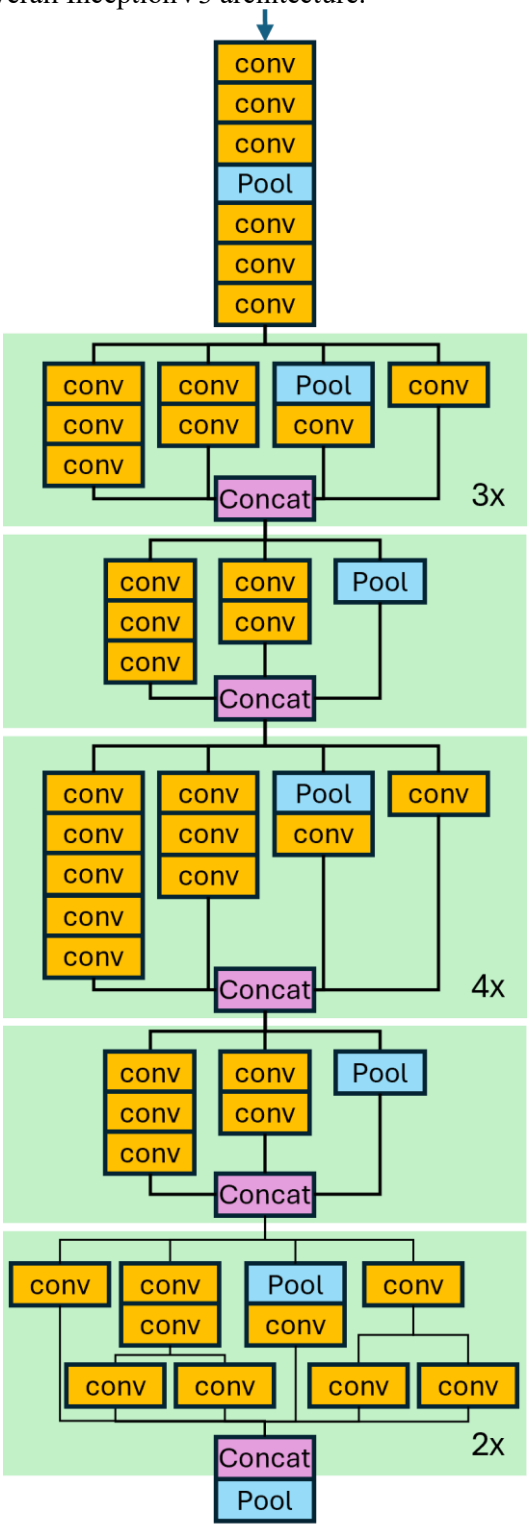


Figure 2. InceptionV3 Architecture.

3. 2. Custom Loss Functions

The proposed framework employs a dual-loss objective function to simultaneously optimize for

precise quality score prediction and consistent ordinal ranking. This design addresses two critical aspects of DIQA: (1) accurate regression to ground-truth quality scores, and (2) preservation of monotonic relationships between quality levels.

3. 2. 1. Linear Regression Loss

The primary loss component is formulated as:

$$L_{linear} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i\right)^2 + \lambda\|\theta\|^2 \qquad (1)$$

where y_i and y_i denote the ground-truth and predicted quality score for the i^{th} image patch, respectively, and λ controls the L2 regularization strength. This mean squared error (MSE) term ensures direct minimization of the deviation between predicted and OCR-based quality scores, while the regularization term prevents overfitting by penalizing large parameter values θ.

3. 2. 2. Monotonic Ranking Loss

To maintain ordinal consistency in quality predictions, we introduce a pairwise ranking loss:

$$L_{mono} = \frac{1}{N(N-1)} \times \sum_{i=1}^{N}\sum_{j=1}^{N}\max\left(0, -\left(y_i - y_j\right)\left(y_i - y_j\right)\right) \qquad (2)$$

This term enforces that for any pair of images (i, j), if $y_i > y_j$ in the ground truth, the predicted scores must satisfy $y_i > y_j$. The hinge function ensures zero penalty when the predicted ranking matches the ground truth, while applying a linearly increasing penalty for ranking violations.

3. 2. 3. Combined Optimization Objective

The total loss function combines these components through a balancing parameter α:

$$L_{total} = \alpha L_{linear} + (1-\alpha) L_{mono} \qquad (3)$$

where $\alpha \in [0,1]$ determines the relative weighting between absolute score accuracy and ranking preservation. The adjustable α parameter allows customization for different applications, whether prioritizing precise scores (e.g., quality grading) or relative rankings (e.g., best-quality selection).

This approach mirrors human quality assessment behavior, where evaluators consider both absolute quality judgments and relative comparisons between documents. Through empirical validation (Section 4.3), we set $\alpha = 0.8$ to prioritize score prediction while maintaining robust ranking performance. The combined loss demonstrates superior performance compared to single-objective alternatives, as quantified in Section 4.5.

4. Experiments

This section presents the experimental design, datasets, training setup, evaluation methodology, and quantitative results for the proposed Siamese-InceptionV3 model developed for document image quality assessment.

4. 1. Datasets

To evaluate the performance of the proposed model, we used two frequently used public datasets, including SmartDoc-QA [19] and SOC (Sharpness-OCR-Correlation)[20]. The SmartDoc-QA dataset comprises 4,260 document images, with a resolution of 3096×4128 pixels, taken from 30 pristine documents, by two distinct mobile phone camera models that introduce varying forms of distortion. This dataset has two OCR scores for each image, provided by FineReader and Tesseract OCR engines. In our experiment, we used the mean of the two OCR scores as the image quality score.

The construction of the SOC dataset involved photographing 25 English documents using smartphone cameras, with each document captured 6–8 times at different focal lengths to introduce controlled distortion variations. This process yielded 175 images, each measuring 1840×3264 pixels. In our experiment, we relied on consensus averaging across three OCR engines, namely FineReader, Tesseract, and Omnipage, which collectively assess the OCR performance of individual images.

4. 2. Evaluation Metrics

We used two popular evaluation criteria to assess the performance of the proposed model: Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC). The range of both criteria is between 0 and 1, and higher values indicate better performance. PLCC describes the accuracy of prediction and is defined as

$$PLCC = \frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}} \qquad (4)$$

where N is the number of test images, y_i and y_i are the ground truth and prediction result of the i^{th}

test image respectively, and $\bar{y}$ and $\bar{y}$ are the mean values of all ground truth and predictions.
To better understand the PLCC, consider a simple example with *N*=5 test samples. The ground-truth quality scores and the predicted scores are given as:

$$y = \{0.8, 0.6, 0.9, 0.4, 0.7\}$$

$$y = \{0.75, 0.65, 0.85, 0.45, 0.72\}$$

The mean values are: $\bar{y} = 0.68$ and $\bar{y} = 0.684$. After substituting these values into the PLCC formul, we obtain PLCC=0.995. This value, being very close to 1, indicates a strong linear correlation between the predicted and ground-truth quality scores. For comparison, if the predicted values were poorly correlated with the ground truth, e.g. $y = \{0.4, 0.8, 0.3, 0.9, 0.5\}$, the PLCC would decrease to approximately 0.2, reflecting a weak linear relationship.
The SROCC assess the monotony of the predicted scores and is defined as:

$$SROCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)} \quad (5)$$

where d_i is the rank difference between the prediction result of the i^{th} test image and the ground truth.

To illustrate the computation of the SROCC, we use the same data as in the previous example. First, ranks are assigned to both sets (the smallest value receives rank 1):

Rank of $y = \{4, 2, 5, 1, 3\}$

Rank of $y = \{4, 2, 5, 1, 3\}$

Then, we compute the difference between the ranks of each pair:

$$d_i = rank\left(y_i\right) - rank\left(y_i\right) = \{0, 0, 0, 0, 0\}$$

Substituting into Eq. (5):

$$SROCC = 1 - \frac{6\sum_{i=1}^{5} d_i^2}{5\left(5^2 - 1\right)} = 1 - \frac{0}{120} = 1$$

This result indicates a perfect monotonic relationship between the ground-truth and predicted scores.

4. 3. Implimentation Details

To train the proposed Siamese network based on the InceptionV3 backbone, we employed a two-stage training procedure and carefully tuned the network and optimization settings to ensure effective learning and generalization. All experiments were conducted using the Python Keras library with TensorFlow framework on a workstation equipped with an NVIDIA GeForce RTX 4050 GPU (24G).
Each branch of the Siamese network was initialized with InceptionV3 weights pretrained on ImageNet. We removed the final classification layer and added a Global Average Pooling (GAP) layer followed by Batch Normalization and a dropout layer with a probability of 0.5. The outputs from each branch were concatenated and passed through a regression head consisting of two fully connected layers (1024 and 512 units) with ReLU activation, followed by a final single-unit output with sigmoid activation to predict the normalized document quality score in the range [0, 1].
In our experiments, we randomly selected 80% of the data as training set, and leave the remaining 20% as validationa and test sets (each 10%). In the training phase, document images were divided into 8 non-overlapping patches of size 299×299, matching the InceptionV3 input requirements. During training, the random rotation within ±15 degrees as data augmentation technique were applied to each patch to improve robustness and mitigate overfitting. All pixel values were normalized to the [0,1] range. Training was performed in two stages, namely, Warm-up and Fine-tuning. In the first stage, the InceptionV3 backbone was frozen, and only the regression head was trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-3}. In second stage, all layers of the network were unfrozen and jointly fine-tuned for up to 100 epochs with a reduced learning rate of 1×10^{-4}. The learning rate was further reduced by a factor of 0.1 if the validation loss plateaued for more than 5 consecutive epochs. Although training is performed at the patch level to learn fine-grained quality differences, document-level quality prediction is achieved by averaging the scores of multiple randomly sampled patches, ensuring robustness to spatial quality variations.
As we mentioned before, the total loss was defined as a weighted combination of two components. In our experiment, the weight were set to $\alpha = 0.8$ based on empirical tuning using the validation set.

4. 4. Experiment Results

To evaluate the effectiveness of our proposed Siamese-InceptionV3 model, we conducted comprehensive experiments on SOC and SmartDoc-QA datasets. Our results are compared against several state-of-the-art DIQA approaches from various methodological families.

Table 1 summarizes the performance of all competing models. On the SOC dataset, our model achieved a PLCC of 0.981 and SROCC of 0.952, outperforming strong baselines such as CNN-based method [3], Attention Based RNN [21], and even the well-established Siamese-ResNet model [4].

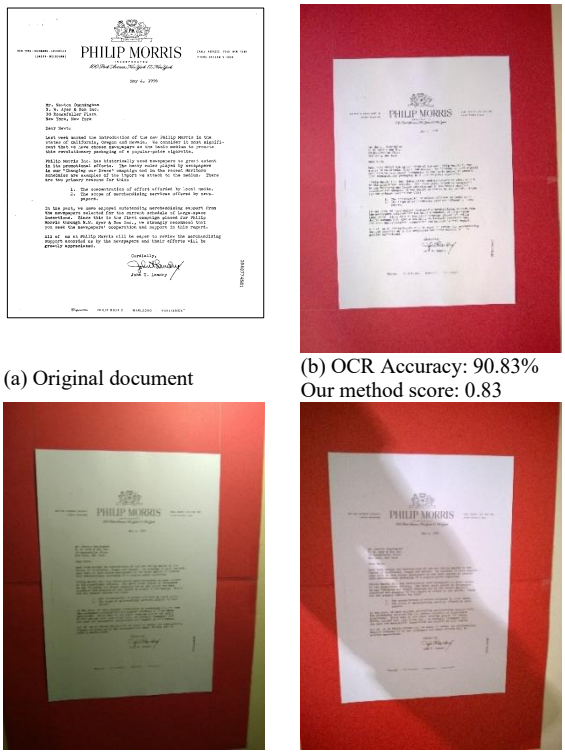
Table 1. PLCC and SROCC results on both two datasets.

Loss Function Components	SOC		SmartDoc-QA	
	PLCC	SROCC	PLCC	SROCC
CNN-based [3]	0.950	0.898	NA	NA
Attention Based RNN [21]	0.956	0.916	0.814	0.865
Sparce Representation [23]	0.935	0.928	NA	NA
Text-line-based [24]	0.914	0.857	0.719	0.746
Feature fusion-based [22]	**0.991**	**0.968**	**0.956**	0.854
CG-DIQA [8]	0.906	0.856	0.625	0.631
Transfer learning-based [12]	0.914	0.872	0.743	0.757
Siamese-Resnet model [4]	0.975	0.936	0.927	0.788
Proposed Siamese-InceptionV3 model	0.981	0.952	0.946	**0.873**

While the feature fusion-based method [22] slightly surpasses our model with a PLCC of 0.991 and SROCC of 0.968, it relies on high-dimensional feature fusion of local and deep semantic representations, which may increase model complexity and limit scalability.

On the SmartDoc-QA dataset, our model also demonstrated strong performance with a PLCC of 0.946 and SROCC of 0.873, surpassing methods like the Siamese-ResNet model (PLCC: 0.927, SROCC: 0.788), CG-DIQA [8], and transfer learning-based approaches [12]. Compared to these models, our method shows greater stability across datasets and maintains high SROCC values, which are particularly critical in DIQA where ordinal quality consistency is paramount. Although the feature fusion-based method again achieved a slightly higher PLCC (0.956), its SROCC lagged behind (0.854 vs. 0.873), suggesting that our model better captures perceptual ranking, a core goal in practical quality assessment.

Figure 3 shows three samples from Smartdoc-QA with different degradation and illumination types. in addition, The ground truth OCR accuracies, beside the proposed method scores for each image are shown. It can be seen that the provided scores track the OCR accuracies well.



(a) Original document

(b) OCR Accuracy: 90.83%
Our method score: 0.83

(c) OCR Accuracy: 45.31%
Our method score: 0.52

(d) OCR Accuracy: 27.65%
Our method score: 0.3

Figure 3. The original document #12 from SmartDoc-QA dataset and three samples with different degradation. The OCR Accuracy and proposed method score are shown, for each image.

The improvements observed can be attributed to the multi-scale feature capturing ability of the InceptionV3 backbone and the Siamese architecture's capacity to model relative comparisons between image pairs. This architecture enables the network to learn discriminative features that generalize across different types of degradations such as blur, noise, and compression. The custom composite loss function, which enforces both absolute score prediction and monotonic ranking constraints, further enhances alignment with human opinion scores. Notably, comparative analysis on SmartDoc-QA revealed an 11% improvement in SROCC over single-scale backbones like ResNet (0.873 vs. 0.788), validating the advantage of multi-scale design in handling diverse document layouts and distortions.

Taken together, these results affirm that the proposed Siamese-InceptionV3 model not only rivals but often outperforms existing DIQA methods in terms of correlation with subjective quality, while maintaining robustness across different datasets and degradation types. Its relatively compact design, end-to-end learning capability, and strong perceptual alignment make it a practical and effective solution for real-world document quality assessment applications.

4. 5. Ablation Study

To assess the effectiveness of our composite loss function, which combines linear regression and monotonic ranking losses, we conducted ablation experiments on both the SmartDoc-QA and SOC datasets. Results are shown in Table 2.

Table 2. Contribution of each loss function components on PLCC and SROCC results on both two datasets.

Loss Function Components	SOC		SmartDoc-QA	
	PLCC	SROCC	PLCC	SROCC
Only $L_{linear}(\alpha=1)$	0.962	0.776	0.913	0.763
Only $L_{mono}(\alpha=0)$	0.872	0.769	0.843	0.751
$\alpha L_{linear}+(1-\alpha)L_{mono}$ with ($\alpha=0.8$)	0.981	0.952	0.946	0.873

The full model, trained with both losses, achieved the highest performance: PLCC/SROCC of 0.946/0.873 on SmartDoc-QA and 0.981/0.952 on SOC.

This demonstrates that integrating absolute and relative supervision yields stronger alignment with human judgments.

When trained with only the regression loss, the model's ability to predict numerical quality scores remained strong (PLCC: 0.962 and 0.913), but its ability to capture perceptual ranking dropped (SROCC: 0.776 and 0.763). Conversely, using only the ranking loss preserved some ordinal consistency (SROCC: 0.769 and 0.751) but led to significantly reduced numerical accuracy (PLCC: 0.872 and 0.843).

These results confirm that both loss components contribute to the model's learning: regression enhances score precision, while ranking enforces perceptual consistency. Their combination produces a more accurate and perceptually aligned document quality assessment model.

5. Discussion And Conclusion

This paper introduced a general-purpose deep learning framework for DIQA, designed to address a broad range of real-world degradations such as blur, noise, compression, scanning artifacts, and lighting inconsistencies. Unlike traditional DIQA methods that often rely on heuristic rules or are tailored to specific distortion types, our model adopts a pairwise comparison strategy through a Siamese neural network architecture, enabling it to learn quality rankings and score estimations that better reflect human visual perception.

At the core of our method is a composite loss function that fuses two complementary objectives: a linear regression loss that aligns predicted quality scores with human Mean Opinion Scores (MOS), and a monotonic ranking loss that enforces consistent quality ordering across image pairs. This dual-objective training approach allows the model to not only predict accurate absolute scores but also preserve perceptual relationships between samples—an essential trait for robust DIQA performance across diverse document types and conditions.

Experimental evaluations on the SmartDoc-QA and SOC datasets demonstrated that our model consistently outperforms several state-of-the-art approaches, including CNN-based classifiers, attention-driven models, and earlier Siamese architectures using simpler backbones like ResNet. Ablation studies further revealed the complementary impact of the regression and ranking losses: removing either led to measurable drops in correlation with human judgment, validating the synergy of the composite loss design.

Although the proposed method shows strong generalization and performance, future work can extend its scope by incorporating multilingual and handwritten document datasets, modeling more complex distortions (e.g., skew, bleed-through), or integrating OCR feedback directly into the training process. Moreover, lightweight variants of the architecture could be explored to enable real-time quality assessment on mobile capture devices and scanning hardware.

In summary, the proposed Siamese-InceptionV3 architecture provides a robust, scalable, and perceptually aligned solution to the general DIQA problem, with promising applications in document digitization pipelines, archival systems, and intelligent scanning environments.

References

[1] M. Askari, M. Asadi, A. Asilian Bidgoli, and H. Ebrahimpour, "Isolated Persian/Arabic handwriting characters: Derivative projection profile features, implemented on GPUs," *Journal of AI and Data Mining,* vol. 4, no. 1, pp. 9–17, 2016, doi: 10.5829/idosi.JAIDM.2016.04.01.02.

[2] A. Alaei, V. Bui, D. Doermann, and U. Pal, "Document Image Quality Assessment: A Survey," *ACM computing surveys,* vol. 56, no. 2, pp. 1–36, 2023.

[3] L. Kang, P. Ye, Y. Li, and D. Doermann, "A deep learning approach to document image quality assessment," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014: IEEE, pp. 2570–2574.

[4] X. Peng and C. Wang, "Camera captured DIQA with linearity and monotonicity constraints," in *Document*

Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14, 2020: Springer, pp. 168–181.

[5] A. Alaei, D. Conte, M. Blumenstein, and R. Raveaux, "Document image quality assessment based on texture similarity index," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016: IEEE, pp. 132–137.

[6] A. Alaei, D. Conte, C. Martineau, and R. Raveaux, "Blind document image quality prediction based on modification of quality aware clustering method integrating a patch selection strategy," *Expert Systems with Applications,* vol. 108, pp. 183–192, 2018.

[7] A. Alaei, "A new document image quality assessment method based on hast derivations," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019: IEEE, pp. 1244–1249.

[8] H. Li, F. Zhu, and J. Qiu, "CG-DIQA: No-reference document image quality assessment based on character gradient," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018: IEEE, pp. 3622–3626.

[9] T. Obafemi-Ajayi, G. J. I. T. o. S. Agam, Man,, C.-P. A. Systems, and Humans, "Character-based automated human perception quality assessment in document images," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* vol. 42, no. 3, pp. 584–595, 2011.

[10] N. S. Rani, K. Akshatha, and K. Koushik, "Quality assessment model for handwritten photo document images," *Procedia Computer Science,* vol. 218, pp. 133–142, 2023.

[11] H. Li, J. Qiu, and F. Zhu, "TextNet for text-related image quality assessment," in *International Conference on Artificial Neural Networks*, Cham, 2018: Springer International Publishing, in Artificial Neural Networks and Machine Learning – ICANN 2018, pp. 275–285.

[12] T. Lu and A. Dooms, "A deep transfer learning approach to document image quality assessment," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019: IEEE, pp. 1372–1377.

[13] J. Gao *et al.*, "DeQA-Doc: Adapting DeQA-Score to Document Image Quality Assessment," *arXiv preprint arXiv:2507.12796,* 2025.

[14] J. Aissam, M. Hain, and A. Chergui, "Handwritten Documents Validation Using Pattern Recognition and Transfer Learning," *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT),* vol. 17, no. 5, pp. 1–13, 2022, doi: 10.4018/IJWLTT.20220901.oa1.

[15] S. Kundu, S. Malakar, Z. W. Geem, Y. Y. Moon, P. K. Singh, and R. Sarkar, "Hough Transform-Based Angular Features for Learning-Free Handwritten Keyword Spotting," *Sensors,* vol. 21, no. 14, p. 4648, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4648.

[16] M. Talebian, K. Kiani, and R. Rastgoo, "A Deep Learning-based Model for Fingerprint Verification," *Journal of AI and Data Mining,* vol. 12, no. 2, pp. 241–248, 2024, doi: 10.22044/jadm.2024.14298.2531.

[17] A. Fateh, R. T. Birgani, M. Fateh, and V. Abolghasemi, "Advancing multilingual handwritten numeral recognition with attention-driven transfer learning," *IEEE Access,* vol. 12, pp. 41381–41395, 2024.

[18] X. Liu, X. Tang, and S. Chen, "Learning a similarity metric discriminatively with application to ancient character recognition," in *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part I 14*, 2021: Springer, pp. 614–626.

[19] N. Nayef, M. M. Luqman, S. Prum, S. Eskenazi, J. Chazalon, and J.-M. Ogier, "SmartDoc-QA: A dataset for quality assessment of smartphone captured document images-single and multiple distortions," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015: IEEE, pp. 1231–1235.

[20] J. Kumar, P. Ye, and D. Doermann, "A dataset for quality assessment of camera captured document images," in *Camera-Based Document Analysis and Recognition: 5th International Workshop, CBDAR 2013, Washington, DC, USA, August 23, 2013, Revised Selected Papers 5*, 2014: Springer, pp. 113–125.

[21] P. Li, L. Peng, J. Cai, X. Ding, and S. Ge, "Attention based RNN model for document image quality assessment," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 1: IEEE, pp. 819–825.

[22] W. Wang, Z. Yan, and H. Lin, "A Document Image Quality Assessment Method Based on Feature Fusion," in *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, 2022: Springer, pp. 889–899.

[23] X. Peng, H. Cao, and P. Natarajan, "Document image quality assessment using discriminative sparse representation," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016: IEEE, pp. 227–232.

[24] H. Li, F. Zhu, and J. Qiu, "Towards document image quality assessment: a text line based framework and a synthetic text line image dataset," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019: IEEE, pp. 551–558.

یک شبکهٔ سیامی مبتنی بر InceptionV3 با توابع زیان سفارشی برای ارزیابی کیفیت تصاویر اسناد

محمدحسین خسروی*

دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران.

ارسال ۲۰۲۵/۰۸/۰۴؛ بازنگری ۲۰۲۵/۰۹/۱۵؛ پذیرش۲۰۲۶/۰۶/۰۶

چکیده:

ارزیابی کیفیت تصویر اسناد (DIQA) نقش مهمی در تضمین کارایی و قابلیت اعتماد کاربردهای پایین‌دستی، از جمله تشخیص نوری نویسه‌ها(OCR) ، بایگانی دیجیتال و سامانه‌های خودکار پردازش اسناد ایفا می‌کند. در این پژوهش، چارچوبی مبتنی بر یادگیری عمیق برای ارزیابی کیفیت تصاویر اسناد ارائه شده که بر پایه معماری شبکه عصبی سیامی با استفاده از شبکه InceptionV3 به‌عنوان استخراج‌کننده ویژگی طراحی شده است. مدل پیشنهادی از یک تابع هزینه ترکیبی بهره می‌برد که با تلفیق خطای رگرسیون خطی و قید رتبه‌بندی یکنواخت، به‌طور هم‌زمان دقت پیش‌بینی امتیاز کیفیت و سازگاری ادراکی میان تصاویر را بهینه می‌کند. برخلاف بسیاری از روش‌های پیشین که بر ویژگی‌های دست‌ساخته یا انواع محدودی از اعوجاج‌ها متکی هستند، روش پیشنهادی قابلیت تعمیم به طیف گسترده‌ای از تخریب‌های رایج در اسناد اسکن‌شده و تصاویر ثبت‌شده با دوربین را دارد. نتایج آزمایش‌ها بر روی مجموعه‌داده‌های SmartDoc-QA و SOC نشان می‌دهد که مدل پیشنهادی همبستگی بالایی با دقت سامانه‌های OCR دارد و به‌ترتیب مقادیر ۰/۹۵۲ و ۰/۸۷۳ را برای ضریب همبستگی رتبه‌ای اسپیرمن (SROCC) به دست آورده است. همچنین، این روش در مقایسه با چندین روش پیشرفته موجود در حوزه ارزیابی کیفیت تصاویر اسناد، عملکرد برتری از خود نشان می‌دهد.

کلمات کلیدی: ارزیابی کیفیت تصاویر سند، شبکه سیامز، InceptionV3، یادگیری عمیق، تابع هزینه.