When Enhancement Hurts Recognition: Objective Misalignment in Low-Light Scene Text Recognition

Xuanshuo Fu, Lei Kang, Ernest Valveny,
Dimosthenis Karatzas, and Javier Vazquez-Corral

Computer Vision Center, Barcelona, Spain
Universitat Autònoma de Barcelona, Barcelona, Spain
{xuanshuo, lkang, ernest, dimos, javier.vazquez}@cvc.uab.es

Abstract. Low-light scene text recognition is commonly addressed using an enhance-then-recognize paradigm, where a low-light image enhancement (LLIE) module is applied prior to optical character recognition (OCR). A natural extension is to jointly optimize the enhancement network with both perceptual enhancement objectives and OCR-based supervision, with the expectation that recognition performance will improve. In this work, we revisit this assumption and identify a fundamental limitation: joint optimization of LLIE and OCR objectives does not reliably improve recognition accuracy and can lead to degraded performance and unstable training dynamics. We show that this behavior stems from an inherent objective misalignment. LLIE losses typically promote global brightness and perceptual fidelity, while OCR supervision favors discriminative, high-frequency structural cues such as character boundaries and stroke patterns. These objectives induce conflicting optimization signals, leading to representations that are suboptimal for both enhancement quality and text recognition. Through controlled experiments and analysis, we demonstrate that increasing emphasis on perceptual enhancement can suppress fine-grained features critical for OCR, while OCR-driven optimization alone does not yield visually plausible enhancements. Our findings suggest that low-light enhancement for text recognition should not be treated as a purely perceptual restoration problem, but rather as a task-aligned representation learning problem. This work highlights a key failure mode in enhance-then-recognize pipelines and provides insights for designing more effective training strategies for robust low-light OCR.

Keywords: Text Recognition in Low-Light Conditions · Low-light Image Enhancement · Optical Character Recognition · Objective Mismatch.

1 Introduction

Low-light scene text recognition (LLSTR) is an important but still insufficiently understood problem in scene text understanding. Images captured at night or

under poor illumination often suffer from low contrast, sensor noise, blur, color shift, and weakened character boundaries. These degradations substantially reduce the reliability of optical character recognition (OCR) models. Although modern scene text recognizers perform well on normally illuminated benchmarks, their performance can drop sharply under low-light conditions because the input distribution differs significantly from the well-exposed images used during pretraining or fine-tuning.

A common solution is to adopt an enhance-then-recognize pipeline. In this paradigm, a low-light image enhancement (LLIE) module first improves the visual appearance of the input image, and the enhanced image is then passed to an OCR recognizer. This strategy is intuitive and modular, especially because LLIE has been extensively studied. Classical methods such as histogram equalization, gamma correction, and Retinex-based decomposition improve brightness or contrast through hand-crafted image priors. More recent deep LLIE methods, including RetinexNet [12], KinD [16], EnlightenGAN [5], Zero-DCE [3], Zero-DCE++ [6], RUAS [10], SCI [11], GDP [1], and SRDACE [13], have achieved strong visual enhancement under diverse low-light conditions. However, these methods are mainly optimized for human perceptual quality or generic image restoration, rather than for preserving the discriminative visual cues required by OCR.

This distinction is critical for text recognition. OCR does not simply require a brighter image; it requires image evidence that preserves character-level structure. Fine strokes, sharp contours, local contrast, and high-frequency edge patterns are often more important than global illumination. Enhancement methods that improve brightness, smooth illumination, or color fidelity may simultaneously introduce artifacts, suppress local texture, blur stroke boundaries, or distort character shapes. Therefore, a visually enhanced image is not necessarily a recognition-friendly image.

Recent work has begun to address low-light text understanding more directly. In low-light text detection, methods such as spatial–frequency enhancement with classical detectors [15], LIST [9], and Seeing Text in the Dark [14] show that text-centric objectives or illumination-robust representations are necessary under degraded lighting. Other studies focus on enhancing text readability, such as extremely low-light text image enhancement with attention, edge cues, and detector-guided supervision [4], as well as edge-aware enhancement and low-light synthesis for text images [8]. These works demonstrate that generic enhancement is insufficient for text-centric vision tasks. Nevertheless, they often assume that adding task supervision to enhancement will naturally improve downstream performance.

We build on Fu *et al.* [2], a state-of-the-art enhance-then-recognize study that introduces low-light scene text recognition benchmarks and evaluates several LLSTR strategies, including OCR-only adaptation, LoRA-based tuning, frozen LLIE+OCR pipelines, and jointly trained LLIE+OCR models. The study further proposes a re-rendering-based low-light image enhancement module and shows that task-aware enhancement can improve recognition performance. It

also reveals a key failure case: frozen enhancement and frozen recognition modules do not necessarily compose well, and increasing brightness alone does not guarantee better OCR performance. In particular, the results suggest that a mismatch exists between enhancement-oriented processing and recognition-oriented prediction.

Our work takes this failure case as the central object of study. Rather than proposing a new LLIE architecture, dataset, or end-to-end recognition system, we revisit enhance-then-recognize pipelines from the perspective of optimization. We ask why incorporating OCR supervision into an enhancement front-end may still fail, while the OCR recognizer is frozen. This setting is important because it is a natural and practical way to reuse powerful pretrained OCR models: the recognizer remains fixed, while a lightweight enhancement module is trained to adapt low-light inputs to it. At first glance, this appears to align enhancement with recognition. However, we show that this formulation can produce unstable training and degraded recognition accuracy.

We argue that this behavior arises from objective misalignment between LLIE and OCR. LLIE objectives typically encourage global brightness, smooth illumination, chromatic consistency, and perceptual fidelity. OCR supervision, in contrast, favors discriminative local structures that support character classification and sequence prediction. These two objectives can induce conflicting optimization signals. When the recognizer is frozen, its feature extractor and decision boundary cannot adapt to the changing enhanced-image distribution. Consequently, the enhancement module is forced to satisfy both perceptual restoration constraints and the fixed recognizer's learned feature preferences. If these requirements are inconsistent, optimization can shift the enhanced images away from the distribution preferred by the recognizer, even if the images appear visually improved.

This perspective changes how low-light OCR should be understood. The central issue is not merely whether enhancement improves image visibility, but whether enhancement preserves and amplifies the information that OCR models actually use. A brighter image may still be worse for recognition if it weakens stroke-level evidence or alters the recognizer's expected feature statistics. Conversely, an image optimized only for OCR may not correspond to a visually plausible enhancement. Low-light scene text recognition should therefore be treated as a task-aligned representation learning problem, rather than as generic visual restoration followed by recognition.

In this paper, we present a failure-case study of low-light scene text recognition from an optimization perspective. Rather than focusing on which enhancement architecture produces visually better images, we ask a more specific question: *why can enhancement fail to improve recognition when it is optimized in front of a frozen OCR model?* To position this problem, we formulate three representative settings: *OCR only*, *LLIE + frozen OCR*, and *LLIE + trainable OCR*. These settings provide a unified conceptual framework for analyzing the relationship between enhancement and recognition. However, our analysis fo-

cuses on the *LLIE + frozen OCR* setting, where the recognizer remains fixed and only the enhancement module is updated.

In the *LLIE + frozen OCR* setting, the enhancement module is trained using low-light enhancement supervision together with OCR supervision from a fixed recognizer. Although this design appears reasonable, it assumes that the recognition loss produced by a frozen OCR model can serve as a stable supervisory signal for enhancement. We show that this assumption may not hold. Since the recognizer cannot adapt to the changing distribution of enhanced images, the enhancement module may gradually produce outputs that become less compatible with the frozen OCR model.

We further explain this failure from a gradient-level perspective. The LLIE objective and the OCR objective may induce gradients with different directions and magnitudes. When these gradients are inconsistent, the enhancement module can be updated in a direction that is suboptimal, or even harmful, for recognition. As a result, introducing OCR supervision under a frozen recognizer does not necessarily lead to recognition-friendly enhancement.

Our experiments support this analysis. Directly inserting a pretrained enhancement module before a frozen recognizer brings little improvement over directly recognizing low-light inputs. More importantly, optimizing the enhancement module while keeping the recognizer frozen leads to substantial recognition degradation. Experiments with different OCR-loss weights further show that increasing the strength of OCR supervision does not resolve this mismatch, and may instead amplify instability.

Our contributions are summarized as follows:

- We present a failure-case study of enhance-then-recognize low-light scene text recognition, focusing on why enhancement may not improve recognition under a frozen OCR model.
- We provide an optimization-level analysis of the *LLIE + frozen OCR* setting, showing that LLIE and OCR losses may induce conflicting gradients and cause recognition-unfriendly enhancement.
- We experimentally verify that, under a frozen recognizer, both inserting and optimizing an enhancement module can fail to improve recognition, and that stronger OCR supervision does not necessarily alleviate this failure.

2 Methodology

2.1 Problem Setting

To analyze the role of low-light image enhancement in scene text recognition, we do not start from a full system design. Instead, we formulate the problem under three representative optimization settings. Let the low-light input image be denoted by I_{low}, the low-light image enhancement module by $f_\theta(\cdot)$, and the OCR recognizer by $g_\phi(\cdot)$. The enhanced image is given by

$$I_{\text{enh}} = f_\theta(I_{\text{low}}), \tag{1}$$

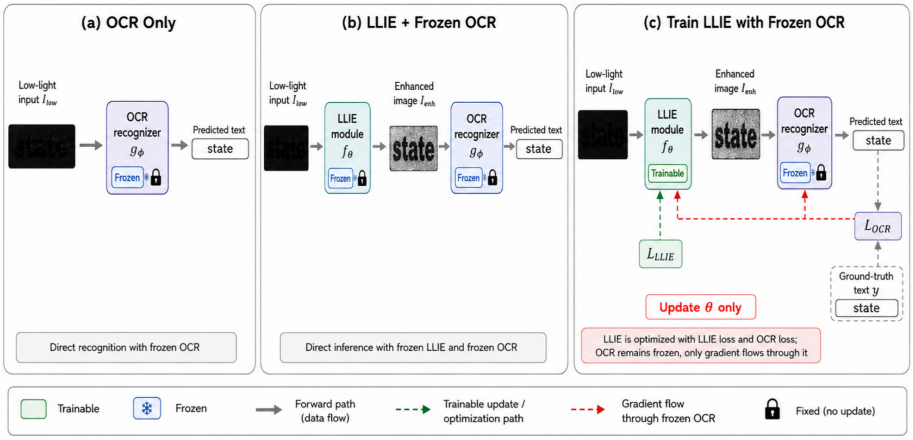

Fig. 1: Illustration of the three optimization settings analyzed in this work. (a) **OCR only**: the low-light input is directly recognized by a frozen OCR model without enhancement or loss supervision. (b) **LLIE + frozen OCR**: both the LLIE module and OCR recognizer are frozen, and recognition is performed by direct enhance-then-recognize inference without training losses. (c) **Train LLIE with frozen OCR**: the LLIE module is optimized using both the LLIE loss and OCR loss, while the OCR recognizer remains frozen. The OCR loss gradient is backpropagated through the frozen OCR model to update only the LLIE parameters.

and the final text prediction is produced by the recognizer. The central question of this work is not which specific enhancement architecture performs best, but rather: *under which optimization setting can enhancement become truly useful for recognition?* This question can be summarized as follows: *Which optimization setting makes enhancement truly useful for recognition?*

The first setting is **OCR only**. In this setting, no front-end enhancement module is introduced, and the low-light image is directly fed into the OCR model for recognition, i.e., the prediction is obtained from $g_\phi(I_{\text{low}})$. This setting corresponds to the most basic formulation of low-light scene text recognition, and serves as a reference for analyzing the other optimization settings. Since low-light images are typically affected by degradations such as low contrast, noise, blur, and color distortion, the **OCR only** setting directly reflects how these factors impair the discriminative capability of the recognizer.

The second setting is **LLIE + frozen OCR**. In this setting, a trainable low-light image enhancement module f_θ is introduced to enhance the input image before recognition, while the parameters ϕ of the OCR model remain fixed, and only the enhancement module is updated. From an optimization perspective, this setting attempts to use the recognition loss from a frozen OCR model as an external supervisory signal to indirectly guide the enhancement module toward producing inputs that are more favorable for recognition. This corresponds to an apparently natural intuition: since a pretrained OCR model already possesses

some text recognition capability, one may freeze it and use it as an evaluator to constrain the enhancement process, so that the enhanced output is gradually adapted to the downstream recognition objective.

Based on the above definitions, our focus is not on comparing the absolute performance of different enhancement modules or recognizers, but on analyzing the essential differences among these three optimization settings, especially under what conditions the enhancement module can genuinely benefit the recognition task. Intuitively, **OCR only** corresponds to directly recognizing low-light inputs, **LLIE + frozen OCR** corresponds to using a fixed recognizer to supervise enhancement. These two settings form the unified framework for the subsequent analysis in this paper.

More importantly, we pay particular attention to the second setting, which raises a critical question: when the OCR module is frozen, can the recognition loss naturally serve as an effective supervisory signal for enhancement? Although this idea appears reasonable in form, it implicitly assumes that the gradient provided by a fixed recognizer can stably characterize what kinds of image transformations are beneficial for recognition. However, for low-light scene text images, this assumption is far from obvious. Low-light enhancement is commonly designed to improve brightness, contrast, or visual visibility, whereas text recognition relies more heavily on the preservation of character boundaries, stroke structures, local textures, and discriminative patterns. As a result, there exists a potential mismatch between the objective of enhancement and the objective of recognition, making **LLIE + frozen OCR** the most critical optimization setting to be examined in this work.

2.2 Theoretical Analysis of Optimization Mismatch

In this subsection, we provide a mathematical analysis to explain why, when the TrOCR model is frozen and only the low-light enhancement module is updated, the combination of OCR supervision and low-light enhancement supervision may lead to optimization mismatch. In particular, the gradients induced by the OCR loss and the low-light enhancement loss may differ in both direction and scale, resulting in gradient conflict during training and, consequently, progressive degradation in text recognition performance.

Basic Setup. Let the low-light enhancement module be denoted by $f_\theta(\cdot)$ with parameters θ, and let the low-light input image be denoted by I_{low}. The enhanced image is defined as

$$I_{\text{enh}} = f_\theta(I_{\text{low}}). \tag{2}$$

The OCR recognizer is represented by a function $g(\cdot)$. In the setting considered here, $g(\cdot)$ is frozen, and only the enhancement module is trainable. The overall objective is written as

$$\mathcal{L}_{\text{total}}(\theta) = \omega\, \mathcal{L}_{\text{OCR}}\big(g(f_\theta(I_{\text{low}}))\big) + \mathcal{L}_{\text{LLIE}}\big(f_\theta(I_{\text{low}}), I_{\text{low}}\big), \tag{3}$$

where $\mathcal{L}_{\mathrm{OCR}}$ denotes the recognition loss computed from the output of the OCR model, and $\mathcal{L}_{\mathrm{LLIE}}$ denotes the enhancement loss used to constrain the low-light enhancement process.

Gradient Decomposition. Since only θ is updated, the gradient of the total loss with respect to θ is

$$\nabla_\theta \mathcal{L}_{\mathrm{total}} = \omega \, \nabla_\theta \mathcal{L}_{\mathrm{OCR}}\big(g(f_\theta(I_{\mathrm{low}}))\big) + \nabla_\theta \mathcal{L}_{\mathrm{LLIE}}\big(f_\theta(I_{\mathrm{low}}), I_{\mathrm{low}}\big). \tag{4}$$

By the chain rule, the OCR-related term can be expanded as

$$\nabla_\theta \mathcal{L}_{\mathrm{OCR}}\big(g(f_\theta(I_{\mathrm{low}}))\big) = \underbrace{\nabla_{I_{\mathrm{enh}}} \mathcal{L}_{\mathrm{OCR}}\big(g(I_{\mathrm{enh}})\big)}_{=:\mathbf{v}} \cdot \nabla_\theta f_\theta(I_{\mathrm{low}}), \tag{5}$$

and the enhancement-related term becomes

$$\nabla_\theta \mathcal{L}_{\mathrm{LLIE}}\big(f_\theta(I_{\mathrm{low}}), I_{\mathrm{low}}\big) = \underbrace{\nabla_{I_{\mathrm{enh}}} \mathcal{L}_{\mathrm{LLIE}}\big(I_{\mathrm{enh}}, I_{\mathrm{low}}\big)}_{=:\mathbf{u}} \cdot \nabla_\theta f_\theta(I_{\mathrm{low}}). \tag{6}$$

Therefore, the total gradient can be written as

$$\nabla_\theta \mathcal{L}_{\mathrm{total}} = \big(\omega \mathbf{v} + \mathbf{u}\big) \cdot \nabla_\theta f_\theta(I_{\mathrm{low}}), \tag{7}$$

where $\mathbf{v}$ and $\mathbf{u}$ represent the gradients on the enhanced image induced by the OCR loss and the LLIE loss, respectively.

Gradient Direction Conflict. Ideally, the two gradients should point in similar directions so that the enhanced image can simultaneously satisfy the requirements of low-light enhancement and text recognition. In this case, the inner product $\langle \mathbf{v}, \mathbf{u} \rangle$ is positive, and both losses contribute consistently to optimization. However, if the angle between $\mathbf{v}$ and $\mathbf{u}$ is large, or if they point in opposite directions, the two objectives become conflicting.

Let the angle between $\mathbf{v}$ and $\mathbf{u}$ be denoted by α. Then

$$\langle \mathbf{v}, \mathbf{u} \rangle = \|\mathbf{v}\| \, \|\mathbf{u}\| \cos \alpha. \tag{8}$$

When $\cos \alpha < 0$, the OCR loss pushes the enhancement module toward one direction, while the LLIE loss pushes it toward another. In that case, the combined update

$$\omega \mathbf{v} + \mathbf{u} = \omega \|\mathbf{v}\| \hat{\mathbf{v}} + \|\mathbf{u}\| \hat{\mathbf{u}} \tag{9}$$

may deviate substantially from the direction favored by OCR optimization. If $\hat{\mathbf{u}} \approx -\hat{\mathbf{v}}$, then

$$\omega \mathbf{v} + \mathbf{u} \approx \big(\omega \|\mathbf{v}\| - \|\mathbf{u}\|\big) \hat{\mathbf{v}}. \tag{10}$$

Thus, when $\|\mathbf{u}\| > \omega \|\mathbf{v}\|$, the overall update may even move in the direction opposite to that encouraged by the OCR loss. Even when the two magnitudes are comparable, partial cancellation may still occur, reducing the effective update strength or skewing the update direction away from the one beneficial to recognition.

This observation suggests that, although the OCR term is explicitly included in the objective, the LLIE loss may still dominate or distort the update if its gradient is stronger or inconsistent with the OCR gradient. As a result, the optimization trajectory of the enhancement module may not align with the objective of improving recognition performance.

Domain Shift Induced by a Frozen Recognizer. The above issue is further amplified by the fact that the OCR model is frozen. Let the OCR model $g_\phi(\cdot)$ be originally optimized on an image distribution $P(I_{\text{gt}})$, e.g., clear or well-conditioned text images, such that

$$\phi^* = \arg\min_{\phi} \mathbb{E}_{I\sim P(I_{\text{gt}})} \left[\mathcal{L}_{\text{OCR}}\big(g_\phi(I)\big)\right]. \tag{11}$$

After passing through the enhancement module, the resulting image distribution becomes $P(I_{\text{enh}})$. Ideally, the enhancement process should make $P(I_{\text{enh}})$ closer to $P(I_{\text{gt}})$. However, if the updates induced by the LLIE loss are not aligned with those required by the OCR objective, then the generated image distribution may gradually drift away from the one expected by the frozen recognizer. This domain discrepancy can be expressed as an increase in

$$\text{dist}\big(P(I_{\text{enh}}), P(I_{\text{gt}})\big), \tag{12}$$

where the distance can be measured by an appropriate distribution metric, such as the Wasserstein distance. Since the OCR model cannot adapt its parameters to compensate for this drift, the recognition performance may deteriorate progressively during training.

Parameter Update Interpretation. The update of the enhancement module parameters is therefore given by

$$\Delta\theta \propto -\nabla_\theta \mathcal{L}_{\text{total}} = -\big(\omega\mathbf{v} + \mathbf{u}\big) \cdot \nabla_\theta f_\theta(I_{\text{low}}). \tag{13}$$

If $\mathbf{v}$ and $\mathbf{u}$ are aligned, the update can simultaneously reduce both losses and improve OCR performance. If they are inconsistent or oppositely directed, the update may deviate from the desired OCR-oriented direction, leading to unstable optimization and degraded recognition results.

Summary. From the derivation above, the gradient of the joint objective can be summarized as

$$\nabla_\theta \mathcal{L}_{\text{total}} = \Big(\omega \nabla_{I_{\text{enh}}} \mathcal{L}_{\text{OCR}} + \nabla_{I_{\text{enh}}} \mathcal{L}_{\text{LLIE}}\Big) \cdot \nabla_\theta f_\theta(I_{\text{low}}). \tag{14}$$

If the gradients induced by the OCR loss and the LLIE loss are not well aligned, the enhancement module will be updated in a direction that is suboptimal for recognition. Since the OCR recognizer is frozen, it cannot adapt to the changing distribution of enhanced images, and the resulting mismatch can progressively worsen the recognition performance. This provides a theoretical explanation for why combining OCR supervision with low-light enhancement supervision does not necessarily improve recognition when only the enhancement module is updated.

LLIE Loss. For low-light enhancement supervision, we adopt an LLIE objective consisting of two parts: a general low-light enhancement loss and an edge-aware term. Specifically, the LLIE loss is defined as

$$\mathcal{L}_{\mathrm{LLIE}} = \mathcal{L}_{\mathrm{LLE}} + \varphi \mathcal{L}_{\mathrm{EC}}, \tag{15}$$

where $\mathcal{L}_{\mathrm{LLE}}$ is the standard low-light enhancement loss used to constrain enhancement quality, and $\mathcal{L}_{\mathrm{EC}}$ is an edge content loss introduced to preserve text-relevant structural information.

The edge content loss is formulated as the mean squared error between the Sobel edge maps of the low-light input I_{low} and the enhanced output I_{out}:

$$\mathcal{L}_{\mathrm{EC}} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left(S(I_{\mathrm{out}})_i - S(I_{\mathrm{low}})_i \right)^2, \tag{16}$$

where N_p denotes the total number of pixels and $S(\cdot)$ is the Sobel operator. The weighting factor φ controls the contribution of the edge term.

OCR Loss. For recognition supervision, we use the standard token-level cross-entropy loss on the decoder logits $z \in \mathbb{R}^{N_t \times V}$:

$$\mathcal{L}_{\mathrm{OCR}} = -\frac{1}{N_t} \sum_{i=1}^{N_t} \log \left(\frac{\exp(z_{i,y_i})}{\sum_{j=1}^{V} \exp(z_{i,j})} \right), \tag{17}$$

where y_i denotes the ground-truth token index for token i, V is the vocabulary size, and N_t is the total number of tokens.

Joint Objective. The final training objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{LLIE}} + \omega \mathcal{L}_{\mathrm{OCR}}, \tag{18}$$

where ω is the weighting factor of the OCR term. Under the frozen-OCR setting, this joint objective reveals that the enhancement module is simultaneously constrained by two potentially inconsistent optimization signals, which forms the basis of the theoretical analysis above.

3 Experiments

In this section, we provide experimental evidence to support the theoretical analysis in Section 3. In particular, we focus on the **LLIE + frozen OCR** setting and examine whether introducing OCR supervision can effectively guide the enhancement module when the recognizer remains fixed. Our analysis is based on both quantitative comparisons and training dynamics under different OCR-loss weights.

3.1 Dataset and Models

We use the recent Low-light Scene Text Recognition (LSTR) dataset [2][1], which contains 11,273 training images, 6,588 validation images, and 17,861 test images of scene text captured under low-light conditions. Since this study aims to investigate optimization misalignment in low-light scene text recognition, we employ the low-light image enhancement (LLIE) model proposed in [2] and the OCR model from TrOCR [7].

3.2 Performance Degradation under Frozen OCR Supervision

Table 1 reports the comparison among three frozen-OCR settings on the LSTR dataset. The **Recognize** setting directly applies the frozen OCR model to the low-light input without enhancement. The **Enhance-Then-Recognize** setting performs direct inference with a frozen LLIE module followed by the frozen OCR model. The jointly supervised **Enhance-Then-Recognize** setting trains the LLIE module with both enhancement and OCR losses while keeping the OCR model frozen.

The results reveal a clear and important phenomenon. Compared with the **Recognize** setting, directly introducing a pretrained LLIE module in the **Enhance-Then-Recognize** setting does not provide meaningful improvement, with CER and WER changing only marginally. More importantly, when the LLIE module is further optimized with both enhancement and OCR supervision while the OCR model remains frozen, recognition performance deteriorates dramatically. The CER increases from 53.16% to 78.64%, and the WER rises from 62.23% to 99.86%. This result indicates that, under a frozen OCR model, the joint objective does not guide the LLIE module toward representations that are more beneficial for recognition. Instead, the optimization process progressively shifts the enhanced images away from the input characteristics preferred by the OCR model, leading to severe recognition degradation.

Table 1: Effect of LLIE and OCR supervision under the frozen-OCR setting. Degraded performance is highlighted in red.

Setting	**LLIE**	$\mathcal{L}_{\text{LLIE}}$	$\mathcal{L}_{\text{OCR}}$	**CER↓**	**WER↓**
Recognize	–	–	–	53.16	62.23
Enhance-Then-Recognize	✓	–	–	53.57	62.41
Enhance-Then-Recognize	✓	✓	✓	78.64	99.86

[1] https://huggingface.co/datasets/lumimusta/Low-light_Scene_Text_Dataset

3.3 Training Dynamics under Different OCR-Loss Weights

To further validate the theoretical analysis, we conduct additional experiments with five different OCR-loss weights, i.e., $\alpha \in \{0.1, 1, 10, 100, 1000\}$, and train each configuration for 20 epochs. Figure 2 shows the evolution of the Character Error Rate (CER) on the test set over training epochs.

As shown in Figure 2, all five curves exhibit an overall upward trend, indicating that the text recognition performance consistently degrades as training proceeds. This observation is highly consistent with the theoretical analysis in Section 3: when the OCR model is frozen, the optimization of the enhancement module under the joint objective does not improve recognition, but instead gradually harms it.

For relatively small values of α, namely $\alpha = 0.1$ and $\alpha = 1$, the increase in CER is comparatively moderate, and the curves become relatively stable in later epochs. Nevertheless, the trend remains clearly upward, which means that even weak OCR supervision is insufficient to make the enhancement module converge to a recognition-friendly solution under the frozen-OCR setting. In addition, the CER curve corresponding to $\alpha = 1$ remains consistently above that of $\alpha = 0.1$, suggesting that increasing the weight of OCR supervision does not alleviate the problem.

When α is further increased to 10, 100, and 1000, the degradation becomes much more severe. Both the initial CER and the final CER after 20 epochs increase with larger α, and the corresponding curves show stronger oscillation and larger fluctuations. In particular, for $\alpha = 100$ and $\alpha = 1000$, the CER rises rapidly to very high levels and does not exhibit any sign of convergence within 20 epochs. These observations suggest that stronger OCR supervision, instead of stabilizing the training process, amplifies the optimization conflict between the OCR objective and the low-light enhancement objective.

This phenomenon can be interpreted from two perspectives. First, the gradients induced by the OCR loss and the LLIE loss are not necessarily aligned, and their conflict becomes more pronounced when the OCR term is assigned a larger weight. Second, since the OCR recognizer is frozen, it cannot adapt to the continuously changing distribution of enhanced images. As a result, stronger OCR supervision may force the enhancement module toward unstable directions that are incompatible with the visual constraints imposed by the enhancement objective, thereby increasing the distribution mismatch between the enhanced outputs and the feature space expected by the recognizer.

3.4 Discussion

Table 1 and Figure 2 jointly provide consistent evidence for the theoretical hypothesis introduced in Section 3. Table 1 shows that optimizing the LLIE module while keeping the OCR model frozen causes a substantial drop in recognition performance. Figure 2 further shows that this degradation is not merely a final-stage artifact, but a systematic behavior that develops throughout training and becomes more pronounced as the weight of the OCR loss increases.

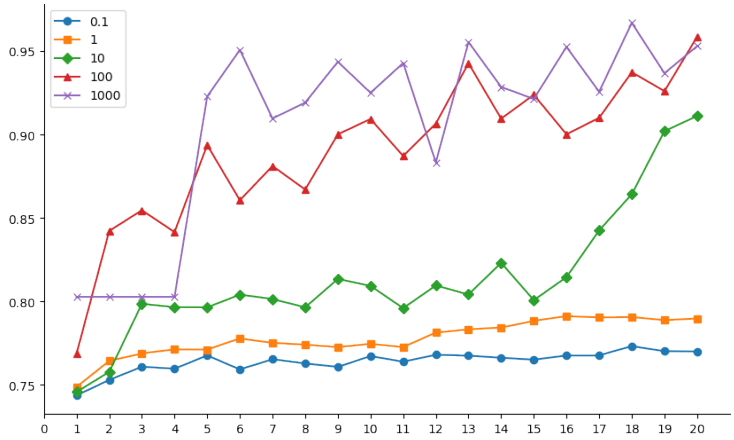

Fig. 2: CER curves on the test set under different OCR-loss weights α in the **LLIE + frozen OCR** setting. The horizontal axis denotes the training epoch, and the vertical axis denotes CER. All curves show an overall upward trend, and the degradation becomes more severe as α increases.

These observations confirm a clear optimization mismatch between the LLIE objective and the OCR objective under the frozen-OCR setting. Because the two objectives favor different image properties, their optimization signals can conflict with each other. Meanwhile, the frozen OCR model cannot adapt to the evolving distribution of enhanced images, resulting in a progressive distribution shift during training. This explains the severe performance degradation observed in the jointly supervised **Enhance-Then-Recognize** setting and supports our central claim: simply adding OCR loss as auxiliary supervision does not necessarily make enhancement more beneficial for recognition.

4 Conclusion

In this work, we revisited low-light scene text recognition from an optimization perspective rather than focusing on architectural design. We show that low-light OCR cannot be reliably improved by simply brightening inputs or by naively adding an OCR loss as an auxiliary objective. By analyzing both direct recognition and enhance-then-recognize settings, we identify a key failure mode: when the OCR recognizer is frozen and only the LLIE module is optimized, the joint objective can introduce gradient conflict and distribution shift, leading to degraded recognition performance. Both theoretical analysis and experimental results support this observation. These findings suggest that effective enhancement for low-light text recognition should be measured not only by visual quality, but also by its alignment with the recognition objective. We hope this work encourages future research to move beyond perceptual enhancement and toward

optimization-aware, recognition-aligned strategies for robust low-light scene text recognition.

5 Acknowledgements

This work has been partially supported by the predoctoral program AGAUR-FI ajuts (2026 FI-3 00470) Joan Oró, which are backed by the Secretariat of Universities and Research of the Department of Research and Universities of the Generalitat of Catalonia, as well as the European Social Plus Fund; the Beatriu de Pinós del Departament de Recerca i Universitats de la Generalitat de Catalunya (2022 BP 00256); the Grant PID2021-128178OB-I00, PID2023-146426NB-100, PID2024-162555OB-I00 funded by MICIU/AEI/10.13039/501100011033, ERDF, EU "A way of making Europe" and by the Generalitat de Catalunya CERCA Program; the European Social Plus Fund, European Lighthouse on Safe and Secure AI (ELSA) from the European Union's Horizon Europe programme under grant agreement No 101070617. JVC also acknowledges the 2025 Leonardo Grant for Scientific Research and Cultural Creation from the BBVA Foundation. The BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors.

References

1. Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang, B., Dai, B.: Generative diffusion prior for unified image restoration and enhancement. In: CVPR. pp. 9935–9946 (2023)
2. Fu, X., Kang, L., Valveny, E., Karatzas, D., Vazquez-Corral, J.: Reading in the dark: Low-light scene text recognition. arXiv preprint arXiv:2604.23685 (2026)
3. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: CVPR. pp. 1780–1789 (2020)
4. Hsu, P.H., Lin, C.T., Ng, C.C., Kew, J.L., Tan, M.Y., Lai, S.H., Chan, C.S., Zach, C.: Extremely low-light image enhancement with scene text restoration. In: ICPR. pp. 317–323. IEEE (2022)
5. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE transactions on image processing **30**, 2340–2349 (2021)
6. Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. IEEE transactions on pattern analysis and machine intelligence **44**(8), 4225–4238 (2021)
7. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 13094–13102 (2023)
8. Lin, C.T., Ng, C.C., Tan, Z.Q., Nah, W.J., Wang, X., Kew, J.L., Hsu, P., Lai, S.H., Chan, C.S., Zach, C.: Text in the dark: Extremely low-light text image enhancement. Signal Processing: Image Communication **130**, 117222 (2025)

9. Liu, H., Yuan, M., Wang, T., Ren, P., Yan, D.M.: List: low illumination scene text detector with automatic feature enhancement. The Visual Computer **38**(9), 3231–3242 (2022)
10. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: CVPR. pp. 10561–10570 (2021)
11. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: CVPR. pp. 5637–5646 (2022)
12. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
13. Wen, J., Wu, C., Zhang, T., Yu, Y., Swierczynski, P.: Self-reference deep adaptive curve estimation for low-light image enhancement. arXiv preprint arXiv:2308.08197 (2023)
14. Xu, C., Fu, H., Ma, L., Jia, W., Zhang, C., Xia, F., Ai, X., Li, B., Zhang, W.: Seeing text in the dark: Algorithm and benchmark. In: Proceedings of the 32nd ACM MM. pp. 2870–2878 (2024)
15. Xue, M., Shivakumara, P., Zhang, C., Xiao, Y., Lu, T., Pal, U., Lopresti, D., Yang, Z.: Arbitrarily-oriented text detection in low light natural scene images. IEEE Transactions on Multimedia **23**, 2706–2720 (2020)
16. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: ACM MM. pp. 1632–1640 (2019)