Supplementary Materials of "Towards Generalized Representations for Low-Light Understanding: When Signal Constancy Meets Semantic Enrichment"

Yifan Li[1] Haofeng Huang[1] Wenhan Yang[2] Jiaying Liu[1*]
[1] Wangxuan Institute of Computer Technology, Peking University [2] Peng Cheng Laboratory
liyifan02@stu.pku.edu.cn, yangwh@pcl.ac.cn, {hhf, liujiaying}@pku.edu.cn

Contents

1. Qualitative Comparison and Visualization of Our Test-Time Adaptive Enhancer

Fig. 1 shows more visual comparisons for the low-light segmentation task. We select representative state-of-the-art (SoTA) methods for comparison, including DAI-Net [2], the current SoTA in zero-shot low-light adaptation. To further highlight the effectiveness of our Test-Time Adaptive (TTA) enhancer, we also visualize the results of a leading machine-oriented enhancement method, GEFU [8].

The visual results clearly demonstrate the limitations of these prior approaches. In terms of segmentation, DAI-Net [2] struggles with complex scenes with intensive overexposed light sources, erroneously classifying regions near overexposed lights as purple ground rather than correct cyan sky. Meanwhile, GEFU [8], which proposes a CycleGAN-based model for machine-oriented enhancement, is also limited by unrealistic lighting intensity and amplified noise artifacts, which in turn degrades the segmentation performance, such as the mis-recognization of the wall pointed by the red arrow. In contrast, our TTA enhancer produces a well-exposed enhancement that suppresses overexposed regions to preserve and restore semantics. Our method achieves a significantly more accurate segmentation map, demonstrating our superior generalization ability.

*Corresponding author.

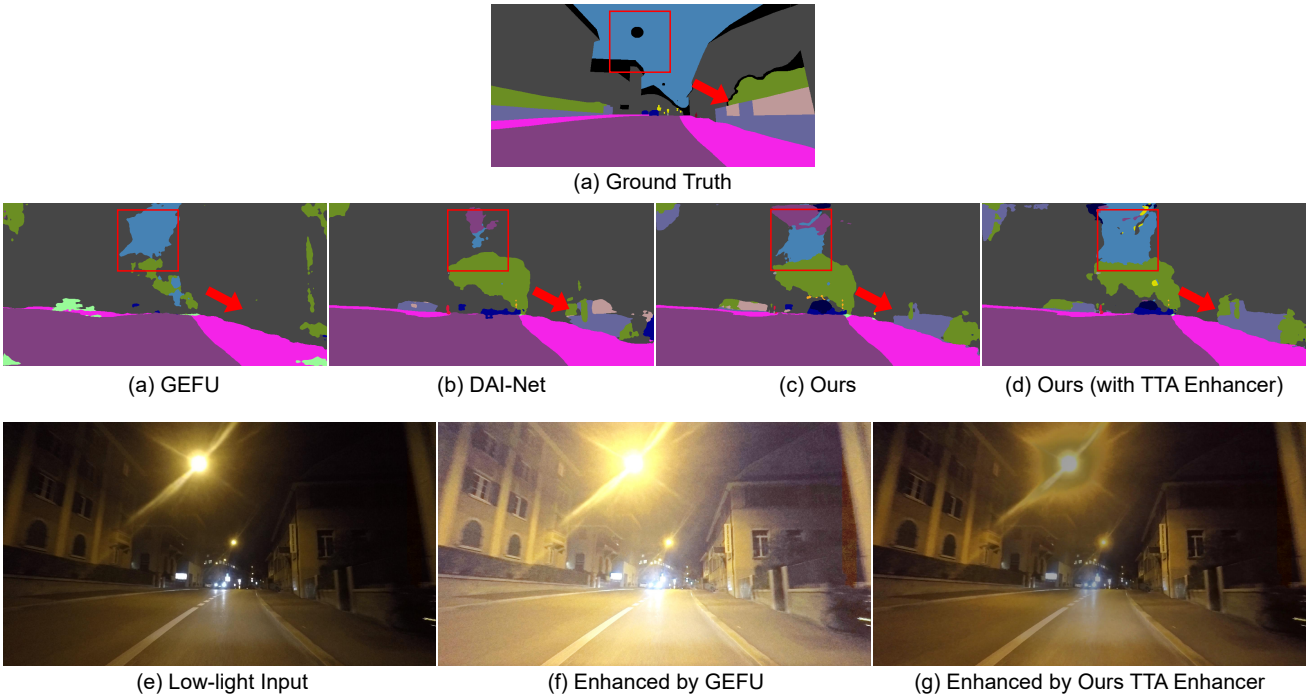

Figure 1. Visualization of segmentation results and enhanced images.

2. More Visualization on Reliability of Signal Priors

We observe that QuadPrior [40] may be unstable (*e.g.* noises and missing structures), which hampers generalized machine understanding. In contrast, our design improves robustness of the prior by using a **machine**-oriented objective as an **information bottleneck** to suppress noise and preserve semantic structures, as shown in Fig. 2.

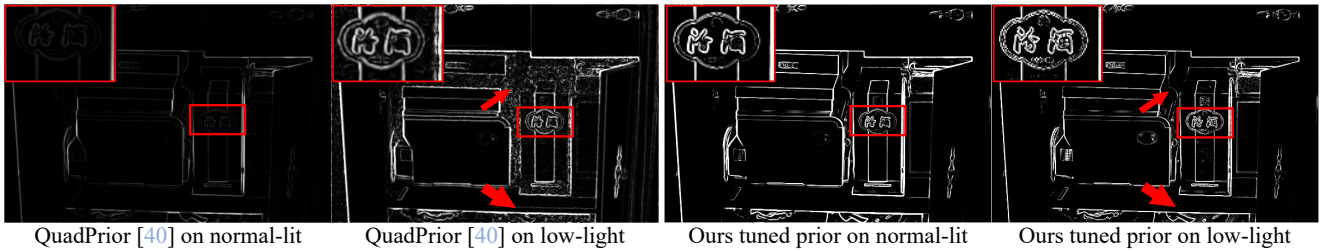

Figure 2. Our priors are more **consistent and precise** compared to QuadPrior [9].

We present more quantitative demonstrations in Tab. 1, we first calculate SSIM of the visualized prior images with paired LOL [10] normal/low-light data. Our prior achieves **0.8020** SSIM on paired LOL (*vs.* 0.6614 for QuadPrior). **Crucially**, beyond the improved stability over [9], our innovative designs effectively maintain the cross-illumination robustness of the **deep network**. Feature MSE on LOL reduces from 0.0156/0.3857 (baseline w/o prior) to **0.0150/0.2824** (ours) in deep layers, confirming that our network maintains the whole model stability.

Table 1. **Prior structure and deep feature consistency on paired LOL [10] data.** Our prior benefits from the information bottleneck constructed based on the machine-oriented target, enjoying a more stable and robust illumination representation.

Metrics	Ours	Baseline
SSIM on visualized prior ↑	**0.8020**	0.6614
MSE on the 3rd ResBlock of ResNet-18 ↓	**0.2824**	0.3857
MSE on the 4th ResBlock of ResNet-18 ↓	**0.0150**	0.0156

3. Additional Ablation Experiments

3.1. Hyper-parameter Sensitivity Analysis

We present a sensitivity analysis in Tab. 2 for the key hyper-parameters of our TTA enhancer and the Outlier-Robust loss represented in Eq. (1,2) and Eq. (11) of main text. The results demonstrates that our method's performance is stable across a wide range of hyper-parameters. Among multiple choices of hyper-parameters, our method always surpass existing SoTA, validating the robustness of our framework to effectively extract robust low-light representations for machine understanding.

Table 2. Verification of our insensitiveness to different hyper-parameters.

(a) Ablation of different hyper-parameters on our TTA Enhancer.

Basic Visual Loss $[\mathcal{L}_{\text{spa}}, \mathcal{L}_{\text{TV}}, \mathcal{L}_{\text{exp}}, \mathcal{L}_{\text{sparse}}]$	Machine-oriented Loss $[\lambda_{\text{clip-inter}}, \lambda_{\text{consis}}, \lambda_{\text{LE}}]$	Classification Top-1
without TTA Enhancer		74.52
1,20,10,1	0.1, 1, 0.01	**75.72**
	0.05, 1, 0.01	74.46
	1, 1, 0.01	75.31
	0.1, 0.1, 0.01	75.29
	0.1, 0.5, 0.01	75.10
	0.1, 1, 0.05	74.71
	0.1, 1, 0.1	74.67

(b) Ablation of different hyper-parameters on our Outlier-Robust loss.

	α	Face Detection $\text{mAP}_{0.5}$
$\mathcal{L}_{\text{iip-consis}}$	0.95	30.2
	0.9 (Ours)	**31.3**
	0.8	31.1
	0.7	30.9
Eq. (11) in the enhancer	0.95	32.5
	0.9 (Ours + TTA Enhancer)	**33.6**
	0.8	33.2
	0.7	32.8

3.2. Comparison of CLIP-based Intermediate Loss and Its Intuitive Variant

An intuitive variant of our proposed CLIP-based intermediate loss is to maximize the distance between the CLIP embedding of enhanced image and the ones of low-light text descriptions (*push away*), while minimizing such distance between enhanced image and normal-light text descriptions (*pull together*). In this design, we define two semantically opposing CLIP embedding subspaces based on illumination attributes (e.g., 'daytime' vs. 'nighttime'). Our objective then compels the enhanced image's embedding to cluster within the target subspace, while simultaneously maximizing its distance from the opposing subspace However, we found that such loss is not as effective as the "intermediate loss" which pushes both normal-lit and low-light distance, as shown in Tab. 3. As we analyze, such a phenomenon is probably caused by the selection of text descriptions. Although we have tried to enrich these prompts through powerful LLMs [3, 11], they cannot represent all lightning conditions in the real world. Furthermore, the text embedding space is much more sparse than the image embedding space. Therefore, the aforementioned "push-pull" loss may cause *overfitting* to certain text embeddings, leading to "semantic distortion" and hampering further generalizabilty.

3.3. Comparison of CoLIE [1] and Our Proposed TTA Enhancer

We conduct additional ablation experiments in Tab. 4 to demonstrate the superiority of our TTA Enhancer compared to CoLIE, which is adpoted as our enhancement backbone. Our TTA enhancer presents a superior cost-performance trade-off. Its machine-oriented objectives converge in just 5 iterations. The original CoLIE, conversely, demands at least 20 iterations and a massive computational load (662 GFLOPs at 224×224), yet still lags significantly behind our method in performance.

3.4. Architecture Variants of Signal Prior Insertion Module

In Sec.3.2 in the main text, we introduce our simple yet effective prior-insertion module with only a lightweight convolution layer to merge auxiliary signal constancy prior. To further demonstrate the effectiveness of our framework, we replace such single convolution layer with a frontload convolution, which compress the channel dimension from 9 to 3, while maintaining the original architectures. As evidenced in Tab. 5, this modified architecture further reduces parameters and computational cost, yet its classification accuracy remains significantly higher than the previous SOTA. This result demonstrates that our framework is robust to model capacity and achieves marginal performance gains with only minor additional parameters.

4. Implementation Details

4.1. Hyper-parameter Settings for Reproduction

We provide details of our hyper-paramter settings in Tab. 6.

Table 3. Ablation on CLIP-intermediate Loss. By examine an intuitive variant, *i.e.*, CLIP-push-pull Loss which pushes the enhanced images away from low-light texts and pulls the ones with normal-lit texts together, we further demonstrate the effectiveness of our design.

Variants	Classification Top-1 Acc	Face Detection $mAP_{0.5}$
Ours (w/o Enhancer)	74.52	31.3
CLIP-push-pull Loss	73.98	30.2
CLIP-intermediate Loss (Ours Enhancer)	**75.72**	**33.6**

Table 4. Comparison on cost-performance between CoLIE and Ours TTA Enhancer.

Methods	Classification (Top-1)	Face Detection ($mAP_{0.5}$)	Segmentation-ND (mIoU)	Segmentation-ACDC (mIoU)
CoLIE-100 steps	57.36	32.7	33.2	26.8
CoLIE-50 steps	57.19	31.4	33.1	26.2
CoLIE-20 steps	56.92	32.1	32.6	26.3
Ours TTA Enhancer	75.72	33.6	48.6	28.5

Table 5. Ablation on prior-insertion architecture. Our method improves performance significantly with high parameter efficiency.

arch	Params	FLOPs	Classification Top-1
ResNet-18 [4] (Baseline)	11.70 M	1.80 G	53.32
DAI-Net [2] (Previous SOTA)	11.70 M	1.80 G	68.44
Ours: `Conv(3,9)+Conv(9,64)`	11.72 M	2.04 G	74.52
`Conv(9,3)+Conv(3,64)`	11.70 M + 243	1.82 G	74.12

Table 6. Implementation details on nighttime classification, face detection and semantic segmentation.

		Classification	Semantic Segmentation	Face Detection
Baselines		ResNet-18 [4]	RefineNet [6]	DSFD [5]
High-level Feature Augmentation				
Basics	Train Set	CoDaN	Cityscapes	WIDER Face
	Batch Size	64	16	8
	Learning Rate	0.0001	0.0005	0.001
	Learning Rate of Signal Prior	0.1 lr	0.1 lr	0.01 lr
	Learning Rate Schedule	Cosine	No	Cosine
	λ_{task}	1		
Signal Prior	$\lambda_{\text{iip-consis}}$	2	5	2
	α in $\lambda_{\text{iip-consis}}$	0.9		
	$\lambda_{\text{iip-decode}}$	2	0.01	2
Enriched Semantics	Anchor Select Strategy in $\mathcal{L}_{\text{contra}}$	the point with maximum response in similarity matrix	middle point of bbox and its neighbour points	centroid of several hard classes
	λ_{contra}	0.01	1	0.01
	λ_{ca}	0.1	0.01	0.01
Low-level Signal Enhancement				
Basic Visual Loss	$[\mathcal{L}_{\text{spa}}, \mathcal{L}_{\text{TV}}, \mathcal{L}_{\text{exp}}, \mathcal{L}_{\text{sparse}}]$	1,20,10,1		
Machine-oriented	$\lambda_{\text{clip-inter}}$	0.1		
	λ_{consis}	1		
	α in λ_{consis}	0.9		
	λ_{LE}	0.01		

4.2. Generation of the Prompt Pool in Low-light Enhancement

We utilize LLMs [3, 11] to generate the prompt pool used in our machine-oriented low-light enhancement module. We request LLMs to generate 500 text descriptions about illumination environments without any description about specific objects under

normal-light and low-light scenarios respectively. We then use the same criterion to ask LLMs check their generated prompts again. The final size of prompt pool holds 500 prompts for normal-light and low-light conditions respectively. We also upload the complete list of these prompts in the `zip` file.

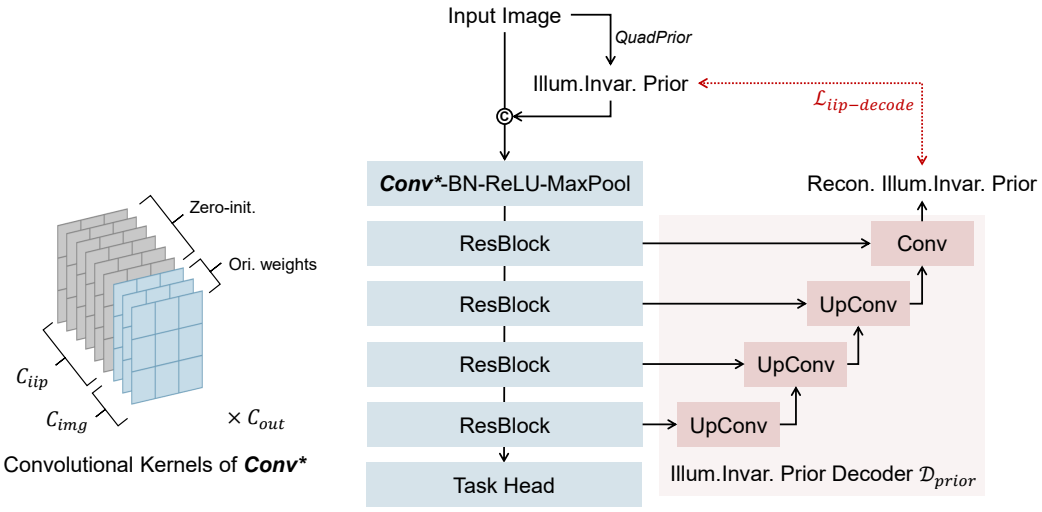

Figure 3. Architecture illustration of our proposed signal constancy mechanism. We devise a simple but effective "conv-merge-in" module, with zero-initialization strategy to preserve the original backbone at the very beginning of the training. In pursuit of more compact and robust representation, we propose a decoding-based regularization, aiming to reconstruct the illumination-invariant prior.

4.3. Implementation of Decoding-based Regularization

In Sec.3.2 in the main text, we propose a decoder-based regularization mechanism, which enforces the intermediate features of our backbone to maintain suffcient information to reconstruct the illumination-invariant prior, thereby achieving compact and cross-illumination robust representation. We provide more details in Fig. 3. Specifically, each feature extracted from the middle layers are processed with a upsample module and a convolution layer, then concatenated with the former one.

4.4. Implementation of Signal Constancy Prior

In the main text Sec. 3.2, we proposed a signal prior-conditioned architecture to stablize the feature distribution against adverse low-light environment. In detail, we adopt QuadPrior [9] as such physical-informed prior which is pretrained on COCO [7] with low-light enhancement task. The prior includes 6-channel, four components, providing a stable characterization of the saturation (one channel), hue (one channel), chroma (one channel), and color orders (three channels). Please refer to the original paper of QuadPrior for detailed derivations, explanations, and visualizations. For more general adaptability, we attach an `Instance Norm` layer after this prior.

4.5. Implementation of Test-Time Adaptive Enhancement

In Sec.3.4 of the main text, we introduce our machine-oriented designs for sample-adaptive enhancement, which dynamically characterizes complex and unseen low-light distributions. We will introduce more implementation details here.

4.5.1. Sample-adaptive Learning

CoLIE [1] presents a zero-shot low-light enhancement method, modeling each input low-light image with an individual implicit neural network (INR), thereby improves generalizability to unseen low-light distribution. Formally, given a low-light image $I \in \mathcal{R}^{h\times w\times 3}$, CoLIE first converts it to HSV color space and learns an INR model f_θ to modulate the V channel I_V. Such explicit decoupling ensures color consistency and structural stability. After this, the INR model will take the coordinates of each pixels as input, embed them with `sin` and learnable MLP transformation, and output a corresponding luminance residual to compensate the original under-exposed pixels:

$$\hat{I}_V = \frac{I_V}{I_V + f_\theta(x, \mathbf{P})}, \tag{1}$$

where $\mathbf{P}$ denotes the set of pixel coordinates. This modulation is performed iteratively until the following losses to be converged. Such INR-based TTA framework naturally remains the original chorminance and produce minor artifacts, which is crucial for low-light machine understanding.

4.5.2. Human Visual Loss

CoLIE [1] proposes four human visual losses to maintain visual plausibility, including:

Spatial loss $\mathcal{L}_{\text{spa}}$ is designed for structural consistency, which computed the mean square error (MSE) between the original luminance and the one after modulation:

$$\mathcal{L}_{\text{spa}} = ||\hat{I}_V - I_V||_2^2. \tag{2}$$

Note that this term can be interpreted as a *regularization term* to avoid overfitting to specific pattern or artifacts.

Total variance loss $\mathcal{L}_{\text{TV}}$ is utilized to smooth the generated illumination, which is computed as the L_2 norm of x-/y-deviation of the luminance image through vertical and horizontal gradient operations:

$$\mathcal{L}_{\text{TV}} = \left(||\nabla_x(\hat{I}_V)||_2 + ||\nabla_y(\hat{I}_V)||_2\right)^2. \tag{3}$$

Sparisty loss $\mathcal{L}_{\text{sparse}}$ is leveraged to mitigate the risk of over-exposure and preserve fidelity:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{M}\sum_{i=1}^{M}\hat{I}_V(i), \tag{4}$$

where i denotes pixel indexs, and M denotes total pixel numbers.

Refined Exposure Loss. The original exposure loss exploited by CoLIE [1] aims to minimize the discrepancy between a fixed value and the mean value across patches. However, we observe that this setting enhances the dark regions effectively, but can also lead to over-exposed and severe information loss when encounter light source or car flares which are common in nighttime road scenarios.

For further suppression on light effect, we proposed a refined exposure loss. We refine the target intensity with a curve, which enhances the value in dark, while lower the requirements in light effect regions, which is indicated by the light effect region map $\mathcal{M}_{LE}$ as introduced in Sec.3.4 of the main text. Specifically, we design a curve to balance the brightness target:

$$I_{gt-V} = \frac{b}{a}\hat{I}_V^2 + (1 - b - \frac{b}{a})\hat{I}_V + b, \tag{5}$$

where a, b denotes hyper-parameters, I_{gt-V} denotes the refined target. After this, we compute the exposure loss as follows:

$$\mathcal{L}_{\text{exp}} = |\mathcal{P}(\hat{I}_V) - \mathcal{P}(I_{gt-V})|_2^2, \tag{6}$$

where $\mathcal{P}$ means patchify operation.

References

[1] Tomáš Chobola, Yu Liu, Hanyi Zhang, Julia A. Schnabel, and Tingying Peng. Fast context-based low-light image enhancement via neural implicit representations. In *Eur. Conf. Comput. Vis.*, 2024. 1, 3, 5, 6

[2] Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 4

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 3, 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2016. 4

[5] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. DSFD: dual shot face detector. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2019. 4

[6] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017. 4

[7] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 5

[8] Sen Wang, Shao Zeng, Tianjun Gu, Zhizhong Zhang, Ruixin Zhang, Shouhong Ding, Jingyun Zhang, Jun Wang, Xin Tan, Yuan Xie, and Lizhuang Ma. From enhancement to understanding: Build a generalized bridge for low-light vision via semantically consistent unsupervised fine-tuning. In *Int. Conf. Comput. Vis.*, 2025. 1

[9] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2024. 2, 5

[10] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *Brit. Mach. Vis. Conf.*, 2018. 2

[11] Jin Xu, Zifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv:2509.17765*. 3, 4