ICPR 2026 · ORAL PRESENTATION

LLIE-CvT : A Convolutional Vision Transformer for Low-Light Image Enhancement

Debanjan Goswami · Bishal Bashyal · Shayok Chakraborty

Department of Computer Science, Florida State University

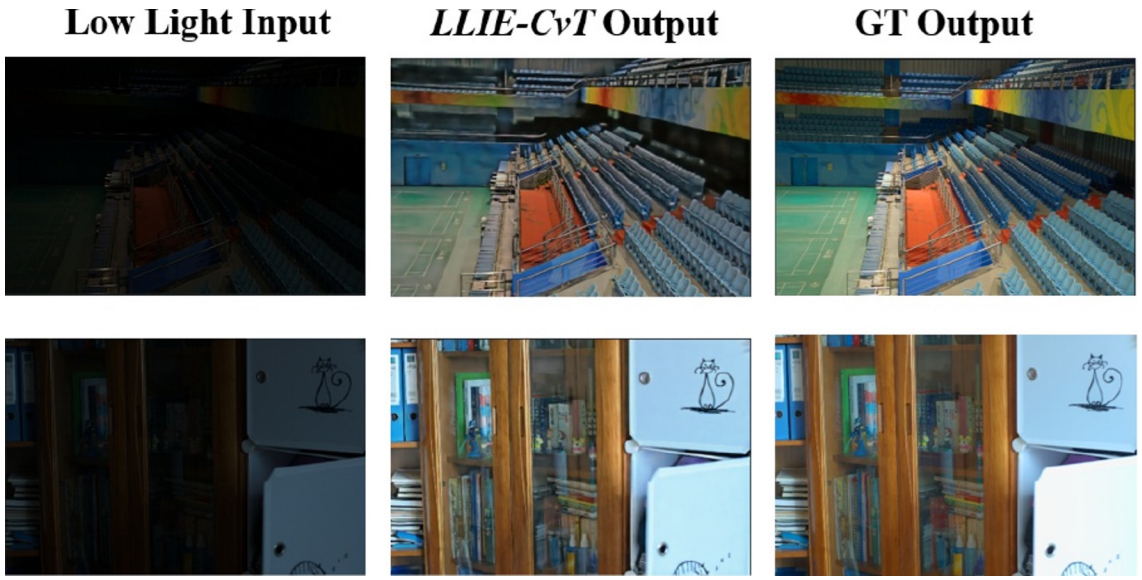
"Making the dark visible, bringing exposure-aware attention for low-light scenes"

LOW-LIGHT INPUT

LLIE-CvT OUTPUT

Why low-light image enhancement?

- **Cameras fail in the dark:** Poor ambient light, small ISO, and short exposures produce dark, noisy, low-contrast images.
- **Downstream vision suffers:** Detection, tracking, and recognition degrade sharply on under-exposed inputs.
- **Broad application space:** Visual surveillance, autonomous driving at night, computational photography, remote sensing, underwater imaging.
- **The core challenge:** Brighten dark regions without amplifying noise or washing out well-exposed areas.



LLIE-CvT on the LOL benchmark: enhanced outputs closely match ground truth.

BACKGROUND

Where existing methods fall short

Traditional

Histogram equalization, gamma correction, Retinex variants.

Limitation: Ignore illumination context; poor generalization and robustness.

CNN-based

Retinex decomposition + deep learning: RetinexNet, KinD, KinD++.

Limitation: Convolutions struggle with long-range dependencies and non-local self-similarity.

Generative

EnlightenGAN; diffusion models Diff-Retinex, PyDiff.

Limitation: Tedious, time-consuming training; unstable convergence; slow sampling.

Transformer-based

SNR-Net, LLFormer, Retinexformer.

Limitation: Fixed tokenization and attention treat dark and bright regions identically.

LLIE-CvT in three ideas

llie-cvt.github.io

1 Convolutional vision transformer backbone

Convolutional token embeddings + multi-head self-attention: fine local detail and global illumination context in one hierarchical encoder–decoder.

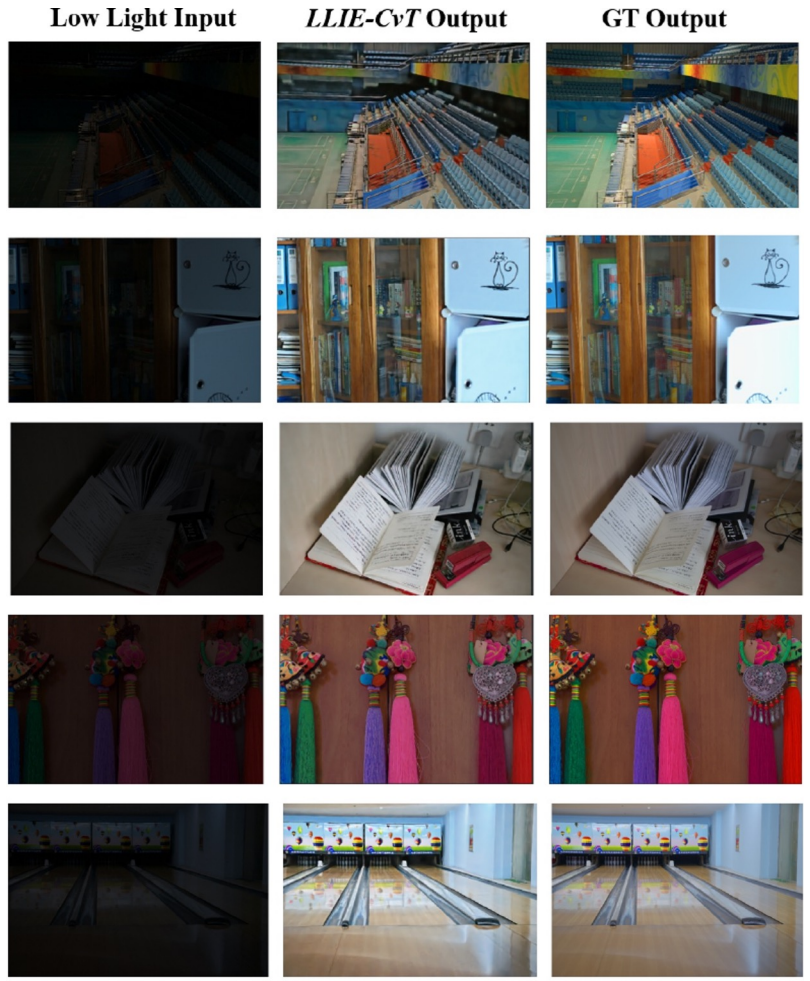
2 Illumination-aware attention

An exposure map computed from the input drives two novel modules — Illumination-Adaptive Tokenization (IAT) and Exposure-Gated Self-Attention (EG-SA) — so the model concentrates on dark, under-exposed regions.

3 Skip attention for efficiency

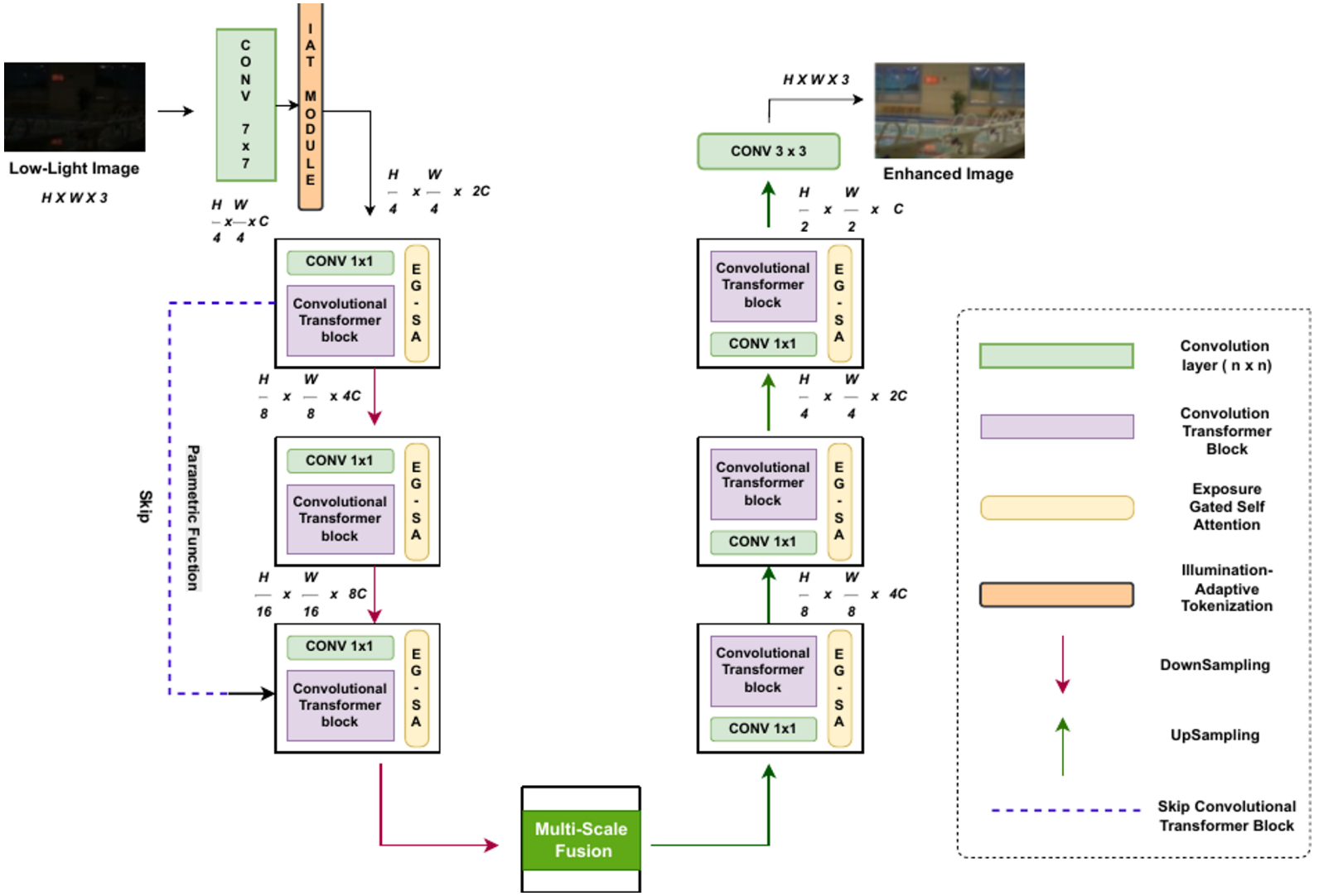
Redundant transformer blocks are detected on the fly and their MHSA is replaced by a lightweight parametric surrogate which cuts FLOPs by ~32%.

First framework to combine convolutional vision transformers, adaptive attention modulation, and skip attention for LLIE.



Architecture overview

llie-cvt.github.io



- **7×7 conv stem (stride 4)**
  then IAT re-tokenizes by exposure
- **3-stage CvT encoder**
  C = 64 → 128 → 256, EG-SA in every block
- **Multi-scale fusion**
  aligns and fuses encoder outputs
- **Decoder + skips**
  upsampling blocks reconstruct the image

Illumination-Adaptive Tokenization (IAT)

llie-cvt.github.io

Idea: let the exposure of each region decide its token density instead of the fixed stride used by every prior ViT/CvT encoder.

Per-pixel exposure from log-luminance:

$$Y = \log(0.2126\ R + 0.7152\ G + 0.0722\ B + 10^{-6})$$

normalized to $E \in [0, 1]$ (low E = dark pixel)

Result: a variable-resolution token grid where dark regions get a dense grid, bright regions are sampled sparsely. Exposure-driven, computed at the input stem itself.

2 — **DARK** $E' < 0.35$ — **stride 2 → dense tokens**

4 — **MID** $0.35 \le E' < 0.70$ — **stride 4 → regular grid**

8 — **BRIGHT** $E' \ge 0.70$ — **stride 8 → sparse tokens**

Coarse discrete strides {2, 4, 8} balance efficiency and fidelity (sensitivity study in supplemental).

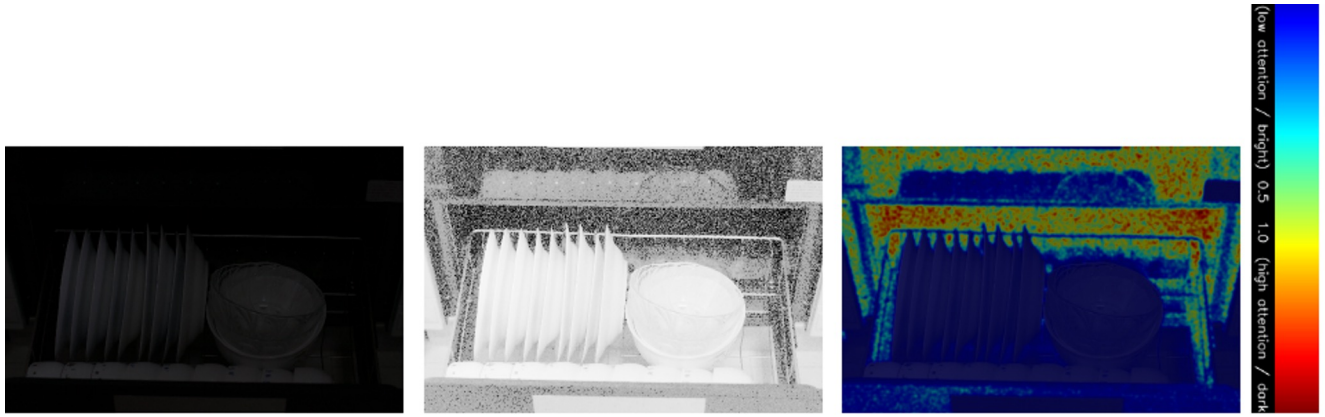
Exposure-Gated Self-Attention (EG-SA)

llie-cvt.github.io

Idea: conventional attention treats all tokens equally. EG-SA rescales queries and keys with a smooth gate so dark tokens attract more attention.

Stable by design: **P**rior guided-attention methods inject pre-computed maps and are unstable. Our gate is differentiable, bounded, and α is annealed 0 → 3 over the first 10% of training.

$$g(E') = \sigma(\alpha \cdot (1 - E'))$$
$$\hat{Q} = Q \odot g(E'), \quad \hat{K} = K \odot g(E')$$
$$A(X) = \text{softmax}(\hat{Q}\hat{K}^T / \sqrt{d_k})\, V$$



Low Light Image | Log Luminance Exposure | Attention heatmap-RGB Overlay

Low-light input → exposure map → attention heatmap: dark regions (red) receive the highest attention.

llie-cvt.github.io

Skip attention : paying less for redundant blocks

1 **Detect redundancy on the fly**

If a block's self-attention map has cosine similarity $\gtrsim 0.9$ with the previous block's, its MHSA is redundant.

2 **Skip the expensive MHSA**

The block's quadratic attention computation is bypassed entirely.

3 **Substitute a parametric surrogate**

Two FC layers + depth-wise conv + ECA, with learnable scale A and bias B to re-amplify the retained attention scores.

2 / 6

transformer blocks skipped in practice

−32%

FLOPs: 34.15 G → 23.16 G

−0.22 dB

marginal PSNR cost (29.04 → 28.82)

Overall complexity stays $\mathcal{O}(n^2d)$, like a standard CvT with a much smaller constant.

llie-cvt.github.io

Loss function & experimental setup

A four-term loss balancing fidelity, perception, structure, and smoothness

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}\_L1 + \lambda_2 \cdot \mathcal{L}\_perc + \lambda_3 \cdot \mathcal{L}\_SSIM + \lambda_4 \cdot \mathcal{L}\_TV \qquad (\lambda = 1.0,\ 0.1,\ 0.5,\ 0.01)$$

L1
pixel-wise fidelity to ground truth

Perceptual
VGG-16 feature-space similarity

SSIM
preserves edges, contours, structure

Total variation
suppresses noise in flat regions

Benchmarks 7 datasets: LOL-v1, LOL-v2 (real + synthetic), SID, SMID, SDSD (indoor + outdoor), MIT-Adobe FiveK — plus ExDark for downstream detection.

Training Adam, lr $10^{-4} \rightarrow 10^{-6}$ cosine decay, 128×128 inputs with random flips, batch 16, 240 epochs. Metrics: PSNR and SSIM.

State of The Art Performance on LLIE Benchmarks

llie-cvt.github.io

Methods	Complexity FLOPS (G)	Complexity Params (M)	LOL-v1 PSNR	LOL-v1 SSIM	LOL-v2-real PSNR	LOL-v2-real SSIM	LOL-v2-syn PSNR	LOL-v2-syn SSIM	SID PSNR	SID SSIM	SMID PSNR	SMID SSIM	SDSD-in PSNR	SDSD-in SSIM	SDSD-out PSNR	SDSD-out SSIM
SID [5]	13.73	7.76	14.35	0.436	13.24	0.442	15.04	0.610	16.97	0.591	24.78	0.718	23.29	0.703	24.90	0.693
3DLUT [50]	0.075	0.59	14.35	0.445	17.59	0.721	18.04	0.800	20.11	0.592	23.86	0.678	21.66	0.655	21.89	0.649
DeepUPE [34]	21.10	1.02	14.38	0.446	13.27	0.452	15.08	0.623	17.01	0.604	23.91	0.690	21.70	0.662	21.94	0.698
RF [16]	46.23	21.54	15.23	0.452	14.05	0.458	15.97	0.632	16.44	0.596	23.11	0.681	20.97	0.655	21.21	0.689
DeepLPF [25]	5.86	1.77	15.28	0.473	14.10	0.480	16.02	0.587	18.07	0.600	24.36	0.688	22.21	0.664	22.76	0.658
IPT [7]	6887	115.31	16.27	0.504	19.80	0.813	18.30	0.811	20.53	0.561	27.03	0.783	26.11	0.831	27.55	0.850
UFormer [37]	12.00	5.29	16.36	0.771	18.82	0.771	19.66	0.871	18.54	0.577	27.20	0.792	23.17	0.859	23.85	0.748
RetinexNet [38]	587.47	0.84	16.77	0.560	15.47	0.567	17.13	0.798	16.48	0.578	22.83	0.684	20.84	0.617	20.96	0.629
Sparse [45]	53.26	2.33	17.20	0.640	20.06	0.816	22.05	0.905	18.68	0.606	25.48	0.766	23.25	0.863	25.28	0.804
EnGAN [13]	61.01	114.35	17.48	0.650	18.23	0.617	16.57	0.734	17.23	0.543	22.62	0.674	20.02	0.604	20.10	0.616
RUAS [22]	0.83	0.003	18.23	0.720	18.37	0.723	16.55	0.652	18.44	0.581	25.88	0.744	23.17	0.696	23.84	0.743
FIDE [41]	28.51	8.62	18.27	0.665	16.85	0.678	15.20	0.612	18.34	0.578	24.42	0.692	22.41	0.659	22.20	0.629
DRBN [44]	48.61	5.27	20.13	0.830	20.29	0.831	23.22	0.927	19.02	0.577	26.60	0.781	24.08	0.868	25.77	0.841
KinD [52]	34.99	8.02	20.86	0.790	14.74	0.641	13.29	0.578	18.02	0.583	22.18	0.634	21.95	0.672	21.97	0.654
Restormer [48]	144.25	26.13	22.43	0.823	19.94	0.827	21.41	0.830	22.27	0.649	26.97	0.758	25.67	0.827	24.79	0.802
MIRNet [49]	785	31.76	24.14	0.830	20.02	0.820	21.94	0.876	20.84	0.605	25.66	0.762	24.38	0.864	27.13	0.837
SNR-Net [42]	26.35	4.01	24.61	0.842	21.48	0.849	24.14	0.928	22.87	0.625	28.49	0.805	29.44	0.894	28.66	0.866
Retinexformer [2]	15.57	1.61	25.16	0.845	22.80	0.840	25.67	**0.930**	24.44	0.680	**29.15**	**0.815**	29.77	0.896	**29.84**	**0.877**
PyDiff [54]	-	-	25.17	0.856	24.01	**0.876**	19.60	0.878	24.97	0.673	28.15	0.708	**29.86**	**0.910**	28.35	0.806
LLIE-CvT	23.16	10.6	**25.93**	**0.860**	**24.03**	0.861	**26.38**	0.808	**25.39**	**0.683**	28.84	0.811	29.20	0.906	28.86	0.874

https://llie-cvt.github.io/

Ablation Studies

llie-cvt.github.io

A and B	Convolutional Projection Layer	PSNR	SSIM
✗	✓	27.63	0.849
✓	✗	22.11	0.636
✓	✓	**28.82**	**0.924**

***Table 1**: Ablation study results showing the effects of the parameters A and B and the convolutional projection layer on the LOL dataset. Best results are shown in **bold**.*

Table 2. All Components	IAT	EG-SA	PSNR	SSIM
✓	✓	✗	28.73	0.905
✓	✗	✓	27.71	0.852
✓	✓	✓	**28.82**	**0.924**

***Table 2**: Ablation study on the effects of the IAT and EG-SA modules on the LOL dataset. Best results are shown in **bold**.*

$\mathcal{L}_{\text{L1}}$	$\mathcal{L}_{\text{perceptual}}$	$\mathcal{L}_{\text{SSIM}}$	$\mathcal{L}_{\text{TV}}$	PSNR	SSIM
✓	✗	✓	✓	28.77	0.902
✓	✓	✗	✓	28.66	0.873
✓	✓	✓	✗	28.53	0.881
✓	✓	✓	✓	**28.82**	**0.924**

***Table 3:** Ablation study results showing the effects of different terms in the loss function on the LOL dataset. Best results are shown in **bold**.*

Performance on ExDark Dataset

llie-cvt.github.io

Methods	Bicycle	Bus	Car	Cat	Chair	Cup	Motor	People	Table	Mean
MIRNet [10]	71.8	81.4	71.1	58.8	58.9	61.3	52.0	68.8	45.5	63.2
RetinexNet [7]	73.8	84.9	80.8	53.4	57.2	68.3	51.3	65.9	43.1	64.3
RUAS [3]	72.0	72.9	78.1	57.3	**62.4**	61.8	61.5	69.4	46.8	64.6
Restormer [9]	76.2	84.0	76.3	59.2	53.0	58.7	62.9	68.6	45.0	64.8
KinD [12]	72.2	83.7	74.5	55.4	61.7	61.3	63.0	**70.5**	47.8	65.5
ZeroDCE [2]	75.8	84.9	77.2	56.3	53.8	59.0	64.0	68.3	46.3	65.1
SNR-Net [8]	75.3	85.3	77.5	59.1	54.1	59.6	65.2	69.1	44.6	65.5
SCI [5]	74.6	**85.4**	76.3	59.4	57.1	60.5	63.9	69.1	45.9	65.8
Retinexformer [1]	76.3	84.7	77.6	61.2	53.5	60.7	63.4	69.5	46.0	65.9
LLIE-CvT	**82.1**	79.8	**84.7**	**68.5**	62.1	**77.4**	**82.5**	68.9	**77.1**	**75.9**

Table 4*. Object detection performance results on the ExDark dataset. The best AP scores in each category are marked in bold and the second best scores are underlined. Our LLIE-CvT achieves the highest AP in 6 object categories and the highest mean AP across all categories.*

User Study on LOL Dataset

llie-cvt.github.io

Method	Score
RetinexNet [7]	2.32 ± 0.11
MIRNet [10]	2.57 ± 0.13
KinD++ [11]	3.76 ± 0.16
Retinexformer [1]	3.64 ± 0.26
LLIE-CvT	$\mathbf{4.22} \pm 0.16$

***Table 5**: User study results on the LOL dataset. Best results are shown in **bold**.*

For each low light image, the enhanced images produced by the 5 different methods(4 baselines and our method) were presented to each subject (without revealing the names of the methods). 10 human subjects were requested to provide a score of the visual quality of the enhanced images independently. The scores range from 1 (worst) to 5 (best).

The subjects were asked to rate the images based on their contrast / sharpness, whether they contain over/under-exposed regions and noisy artifacts, and the overall quality of the image enhancement.

Qualitative comparison on LOL

llie-cvt.github.io

MIRNet & RetinexNet: blurry. KinD++: residual under-exposure. LLIE-CvT: clean, well-lit, retaining accurate color.

CONCLUSION

llie-cvt.github.io

LLIE-CvT: exposure-aware attention, efficient by construction

1. **Illumination-aware by design:** IAT densifies tokens in dark regions and EG-SA boosts their attention To the best of our knowledge, This is the first explicit illumination-dependent attention scoring in transformer LLIE.

2. **Efficient by construction:** Skip attention removes redundant MHSA blocks: −32% FLOPs for −0.22 dB.

3. **Results:** Best or second-best in 12 of 14 metrics on 7 benchmarks; best mAP on ExDark; first in human preference.

Code is open-sourced at https://llie-cvt.github.io/

Thank you. We'd love to answer any questions you may have.