ECCV 2026 | MALMÖ | SEPT 8–12

Women in Computer Vision

Frozen CLIP Priors for Robust Self-Supervised Poisson Inverse Problems

Laura C. Diaz-Delgado · **Emmanuel Martinez** · **Henry Arguello**

Department of Computer Science, Universidad Industrial de Santander, Colombia

henarfu@uis.edu.co

High Dimensional Signal Processing Research Group

Universidad Industrial de Santander

code & paper

1 The problem

PHOTON-LIMITED

Photon-limited sensors are corrupted by **signal-dependent Poisson noise**, which acts on the output of the acquisition operator:

$$\mathbf{y} \;=\; \gamma \cdot \mathrm{Poisson}(\mathbf{A}\mathbf{x}/\gamma)$$

$\mathbf{A}$ known linear operator · γ shot-noise level

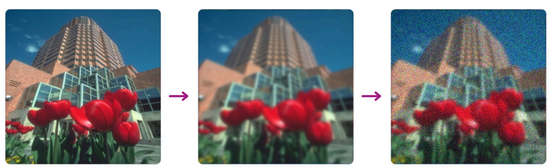
CFA demosaicing Bayer sampling

$$\mathbf{A}\mathbf{x} = \mathbf{M}_R\mathbf{x}_R + \mathbf{M}_G\mathbf{x}_G + \mathbf{M}_B\mathbf{x}_B$$

clean $\mathbf{x}$ → mosaiced $\mathbf{A}\mathbf{x}$ → measured $\mathbf{y}$

Deblurring 9×9 Gaussian kernel

$$\mathbf{A}\mathbf{x} = \big[\mathbf{H}\mathbf{x}_R;\ \mathbf{H}\mathbf{x}_G;\ \mathbf{H}\mathbf{x}_B\big], \qquad \mathbf{H}\mathbf{x}_c = \mathbf{h} * \mathbf{x}_c$$



clean $\mathbf{x}$ → blurred $\mathbf{A}\mathbf{x}$ → measured $\mathbf{y}$

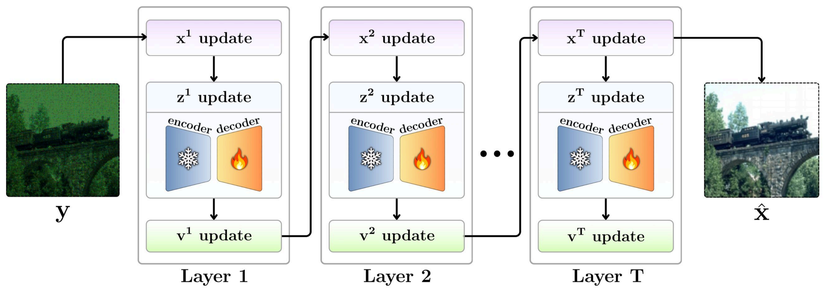
Both rows at $\gamma = 0.05$, the severe photon-limited regime.

WHY THIS IS HARD

- **The noise moves with the signal.** Objectives designed under Gaussian assumptions destabilise.
- **Clean ground truth does not exist** in microscopy, medical and low-light imaging, so supervision has to come from the measurements themselves.
- **Pixel-space priors are fragile.** They fit one training distribution and break under operator and dataset shift.

2 The method

ADMM-UNROLLED PLUG-AND-PLAY



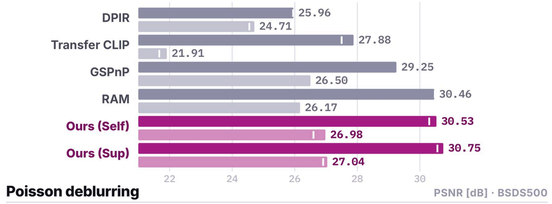
- **x** **Data consistency** — closed form for both operators: one division per pixel under CFA, the same in the Fourier domain under blur.
- **z** **CLIP prior** — frozen encoder, trainable decoder. See **4**.
- **v** **Dual ascent** — hands the split forward.

$T = 2$ layers for demosaicing, $T = 3$ for deblurring. One decoder, shared across all of them.

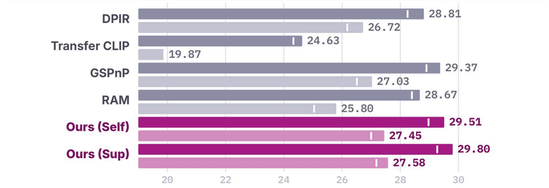
3 Results

QUALITY & COST

Poisson demosaicing PSNR [dB] · BSDS500



Poisson deblurring PSNR [dB] · BSDS500



baselines $\gamma = 0.01$ · baselines $\gamma = 0.05$ · ours $\gamma = 0.01$ · ours $\gamma = 0.05$

DIV2K — trained on BSDS500, tested under dataset shift

Self-supervision costs little. Ours (Self) is never more than **0.31 dB** behind its supervised counterpart across all eight settings, and the margins over the baselines widen where the photon count is lowest.

EFFICIENCY — 3×256×256 ON ONE RTX 4070

0.09	24	10.99
TFLOPs	ms	M trainable
100× vs Transfer CLIP	per image	decoder only

METHOD	PARAMS [M]	TFLOPS	TIME [S]
DPIR	32.64	11.49	0.5193
Transfer CLIP	10.99	9.00	2.4683
GSPnP	17.01	9.12	0.9247
RAM	34.13	0.32	0.8026
Ours	**10.99**	**0.09**	**0.0242**

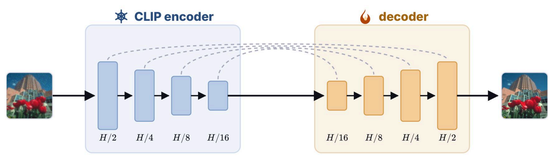
Trainable parameters only; the CLIP encoder is frozen and shared across every unrolled layer. TFLOPs count the denoiser over the maximum iteration count.

4 The CLIP prior

THE ONLY LEARNED PART

The prior step is a proximal update — so instead of writing a regulariser $\mathcal{R}$ down, plug in a denoiser:

$$\mathbf{z}^{t+1} = \mathcal{D}_\theta(\tilde{\mathbf{z}}) = \mathcal{G}_\theta\Big(\mathcal{E}_{\mathrm{CLIP}}(\tilde{\mathbf{z}})\Big)$$



WHY FREEZING WORKS

- **Distortion-invariant** — CLIP RN50 features barely move under corruption, so the prior never fits measurement artefacts.
- **Content-related** — they still organise by scene content, so the decoder has real structure to rebuild.

Fine-tuning the encoder throws both away. Frozen, it leaves a **10.99 M** decoder as the whole task adapter, shared across every unrolled layer.

5 Self-supervision

NO CLEAN IMAGES

$$\mathcal{L}_{\mathrm{self}}(\mathbf{y};\theta) = \mathcal{L}_{\mathrm{GR2R}}(\mathbf{y};\theta) + \tau\,\mathcal{L}_{\mathrm{EI}}(\mathbf{y};\theta)$$

GR2R RE-CORRUPTION

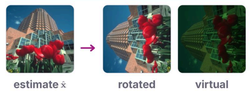
$$\mathcal{L}_{\mathrm{GR2R}} = \big\|\mathbf{A}f_\theta(\mathbf{y}_1) - \mathbf{y}_2\big\|_2^2$$

noisy $\mathbf{y}$ → input $\mathbf{y}_1$ target $\mathbf{y}_2$

The pair $\mathbf{y}_1, \mathbf{y}_2$ is a binomial split of the measurement at $\alpha = 0.2$: unbiased for Poisson data, where Gaussian recipes are not.

EQUIVARIANT IMAGING

$$\mathcal{L}_{\mathrm{EI}} = \big\|\mathbf{x}_v - f_\theta(\mathbf{y}_v)\big\|_2^2$$



estimate $\hat{\mathbf{x}}$ → rotated virtual

Here $\mathbf{x}_v = \mathcal{T}(\hat{\mathbf{x}})$ and $\mathbf{y}_v = \mathbf{A}\mathbf{x}_v$. Rotations for CFA, scalings for blur; consistency alone admits degenerate solutions.

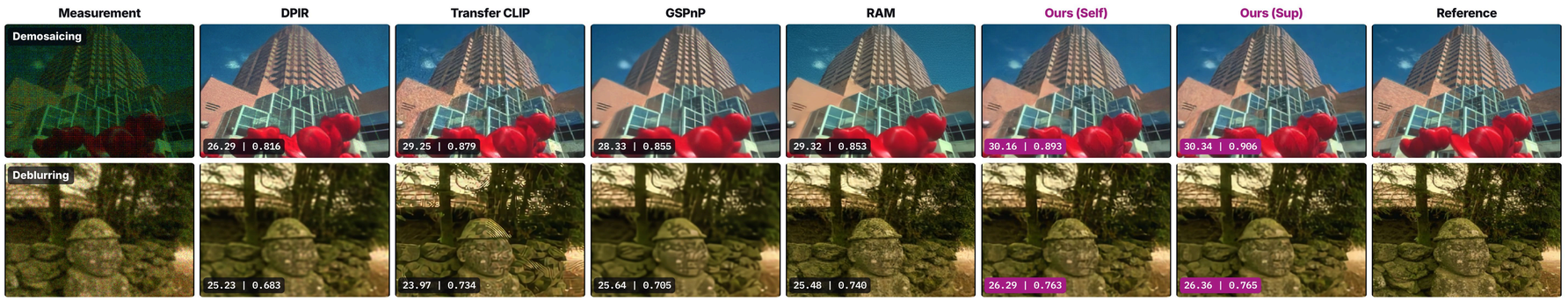
WHAT EACH LOSS TERM BUYS

$\mathcal{L}_{\mathrm{MC}}$	$\mathcal{L}_{\mathrm{GR2R}}$	$\mathcal{L}_{\mathrm{self}}$
21.42	**27.45**	**30.53**
consistency only	+ Poisson re-corruption	+ equivariance

PSNR [dB], BSDS500 demosaicing at $\gamma = 0.01$, same frozen-CLIP prior throughout. Learning that encoder from scratch instead: **21.91**.

6 Qualitative comparison

BSDS500 · $\gamma = 0.01$



7 Real sensor data

SID

Zero-shot adaptation on one low-light raw capture, $\gamma_{\mathrm{rgb}} = (0.018,\ 0.017,\ 0.026)$ per channel — no clean target anywhere.

TAKE AWAY

- A **frozen** encoder beats one trained for the task.
- Poisson re-corruption and equivariance **replace clean supervision**.
- Physics in closed form keeps it **fast enough to be practical**.

The authors acknowledge the VIE of Universidad Industrial de Santander for supporting with "Apoyo a Semilleros de Investigación — Diseño de Codificación para el Muestreo Compresivo de Señales Multidimensionales Utilizando Técnicas Basadas en Aprendizaje Profundo", Project code 4765.