LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CHEM: Estimating and Understanding Hallucinations in Deep Learning for Image Processing

Ines Rosellon Inclan

Doctoral Researcher

Chair of Mathematical Foundations of Artificial Intelligence
Ludwig-Maximilians-Universität München (LMU)

Joint work with

Jianfei Li, Gitta Kutyniok and Jean-Luc Starck

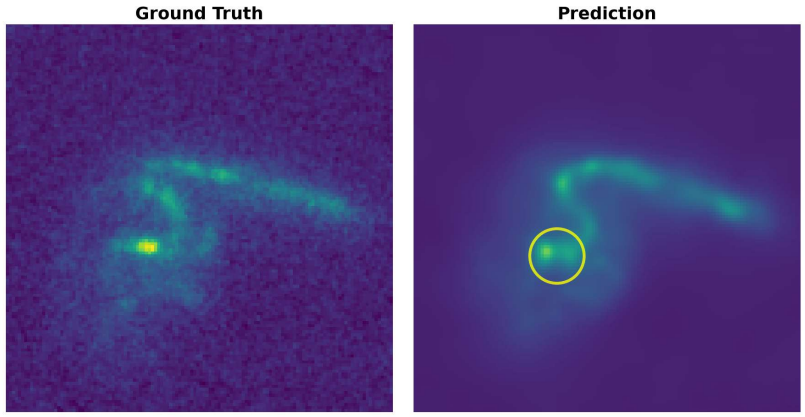
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview

1. Introduction
2. Image Deconvolution in Astronomical Imaging
3. Conformal Hallucination Estimation Metric (CHEM)
4. Experimental Results on CANDELS

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction



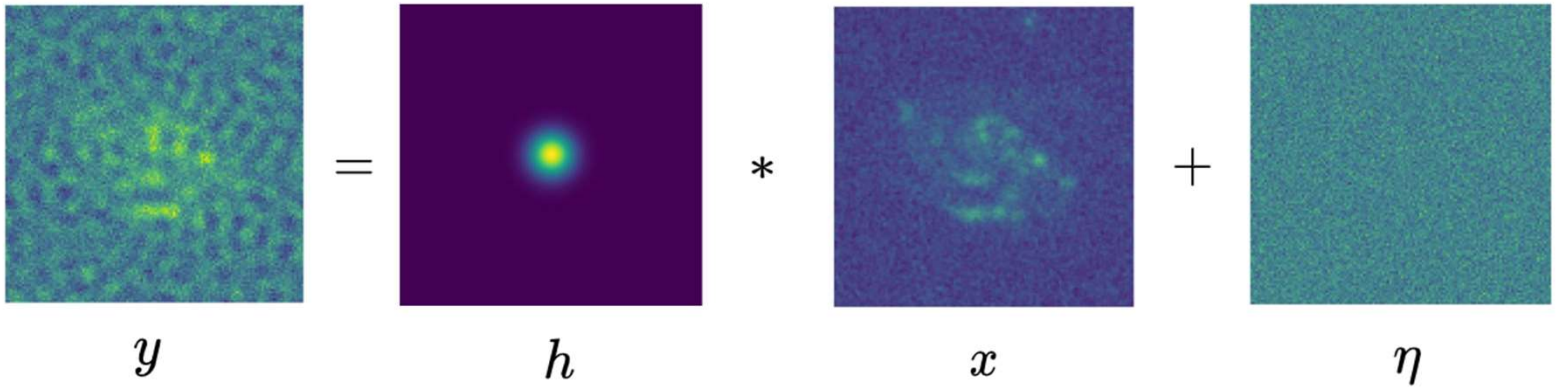
Deep Learning- based Image Deconvolution on image from the CANDELS dataset using U-Net architecture trained with l2-loss, (Akhaury *et al.* 2022).

- Conceptually: „realistic-looking artifacts absent from the ground-truth image“
- Setting: deep-learning–based image deconvolution

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Setting

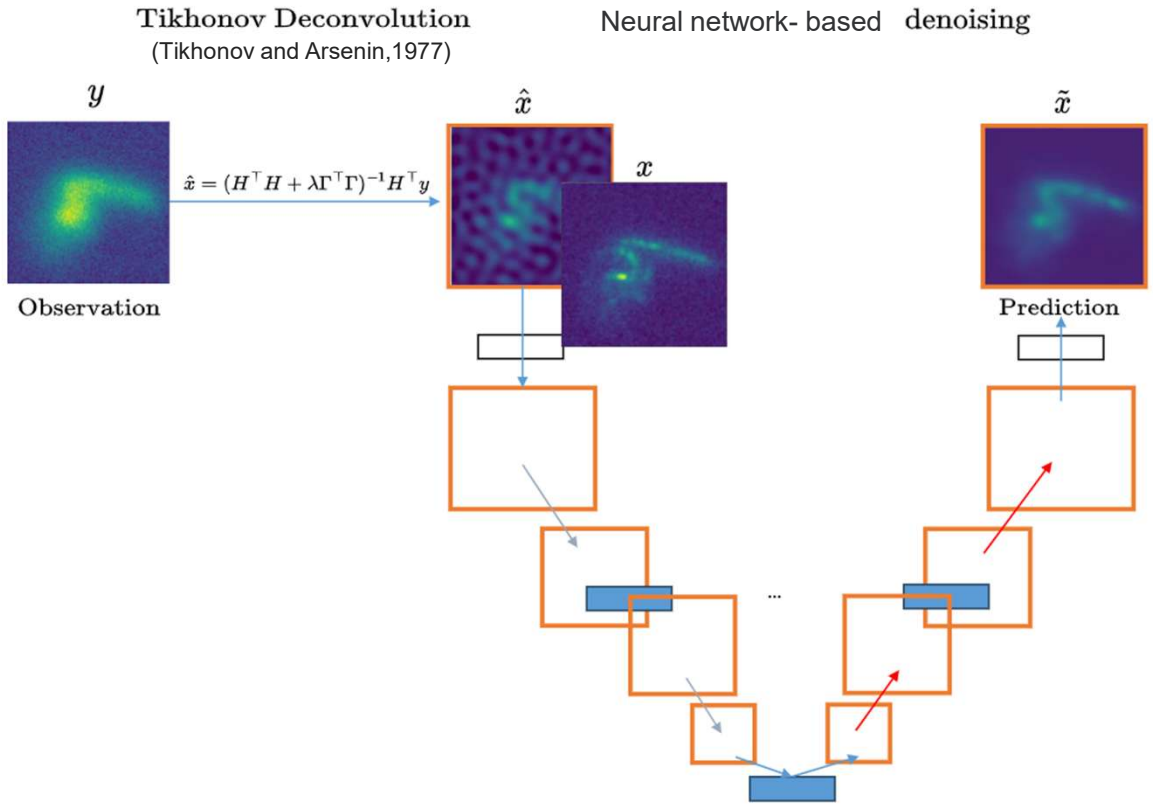
Deconvolution Problem



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Setting

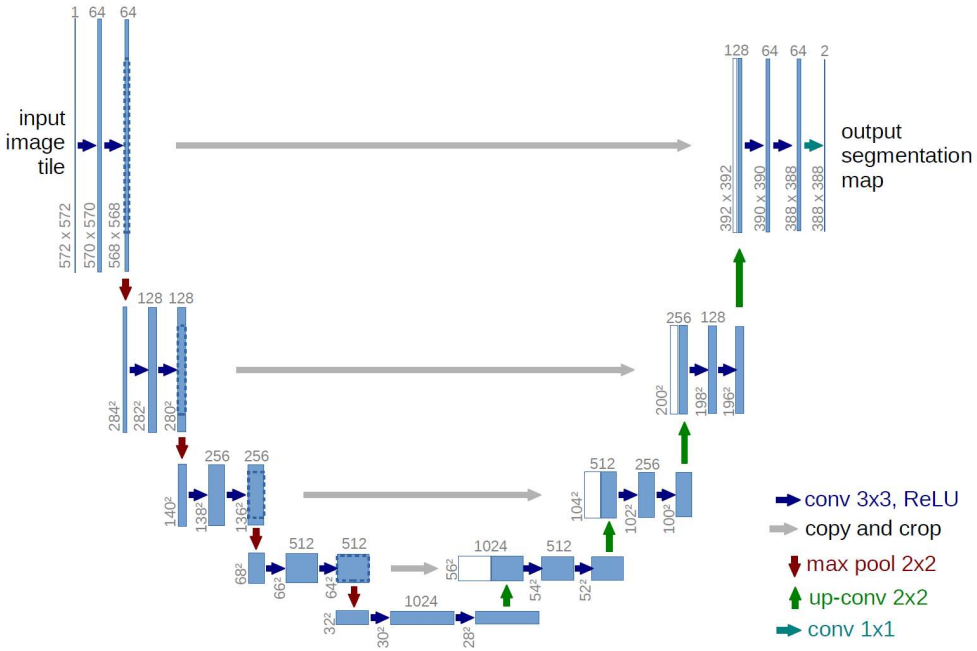
Solution: Two-stage deep deconvolution Method – Tikhonet (Sureau, Lechat and Starck, 2020)



- U-Net, (Sureau, Lechat and Starck, 2020).
- SUNet, (Akhaury *et al.*, 2024).
- Learnlets (Akhaury *et al*. 2022).

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Setting



Method	Loss	No. of parameters	Batch size	Epochs	Training Time [h]
Learnlets	L1	21,673	4	500	65.9
Learnlets	L2	21,673	4	500	61.0
SUNet	L1	99,475,367	4	500	134.0
SUNet	L2	99,475,367	4	500	135.1
U-Net	L1	7,781,761	4	500	75.0
U-Net	L2	7,781,761	4	500	76.4

U-Net, (Ronneberger, Fischer and Brox, 2015).

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Conformal Hallucination Estimation Metric (CHEM)

Hallucination of a Neural Network

Let $\alpha \in (0,1)$ and $\mathcal{P}$ be a probability measure over $\mathcal{X}$. For a learning task $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$, we define the hallucination of a neural network $\Phi(\cdot)$ for a randomly sampled data pair $(X, Y) := (X, \mathcal{M}(X))$ as

$$\mathsf{H}(\Phi; \alpha) := \mathsf{H}(\Phi) := \mathbb{E}\, \mathrm{dist}\left(Y, B_\alpha(\Phi(X))\right), \tag{1}$$

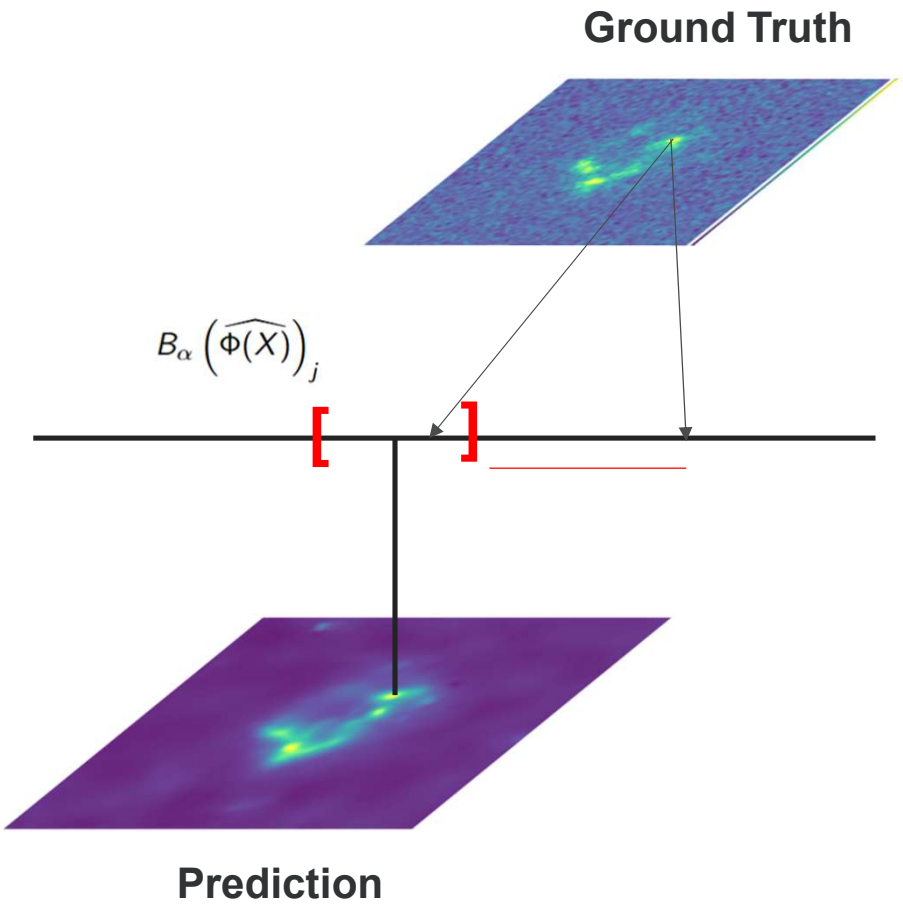
where $B_\alpha(\Phi(X))$ is given by

$$\mathcal{P}\left(Y \in B_\alpha(\Phi(X))\right) \geq 1 - \alpha. \tag{2}$$

- How do we construct $B_\alpha\big(\Phi(x)\big)$?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Conformal Hallucination Estimation Metric (CHEM)



- Extend to wavelet/shearlets to get information at different frequency bands
- Estimate interval using Conformalized Quantile Regression (Leterme, Fadili and Starck, 2025).

Advantages:

- Model-agnostic
- Distribution-free
- Compatible with different representations

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Conformal Hallucination Estimation Metric (CHEM)

input image — ground truth

Let $X \in \mathcal{X} \subset \mathbb{R}^{t_1}$, $Y \in \mathcal{Y} \subset \mathbb{R}^{t_2}$ and let W be the discrete wavelet transform (DWT) or the discrete shearlet transform (DST). Set $\hat{X} := WX \in \mathbb{R}^{\hat{t}_1}$ and $WY \in \mathbb{R}^{\hat{t}_2}$, $j \in \{1, \ldots, \hat{t}_2\}$. Let

input image after transform

$$B_\alpha\left(\widehat{\Phi(X)}\right)_j := \left[\widehat{\Phi(X)}_j - \hat{R}(X)_j, \widehat{\Phi(X)}_j + \hat{R}(X)_j\right]. \tag{3}$$

output image after transform

Goal: find $\hat{R}(\cdot)$ such that (2) is satisfied.

→ Conformalized Quantile Regression (Leterme, Fadili and Starck, 2025).

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Conformal Hallucination Estimation Metric (CHEM)

CHEM

For an upper bound $\theta > 0$, we define

$$H^{\theta}(X, Y)_j := \min\left\{\left(\left|\widehat{\Phi(X)}_j - (\hat{Y})_j\right| - \hat{R}(X)_j\right)_+, \theta\right\}$$

and

$$\mathrm{H}^{\theta}(\Phi) \leftarrow \frac{1}{M}\sum_{m=1}^{M}\frac{1}{\hat{t}_2}\sum_{j=1}^{\hat{t}_2} H^{\theta}(X_m, Y_m)_j$$

for a validation dataset $\mathcal{D}_1 = \{(X_m, Y_m)\}_{m=1}^{M}$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Theoretical Results

Theorem D.2.

"We can approximate the actual CHEM value with averaging over the pixels and the validation set."

$$\left| \mathrm{H}^{\theta}(\Phi) - \frac{1}{M} \sum_{m=1}^{M} \frac{1}{\hat{t}_2} \sum_{j=1}^{\hat{t}_2} H^{\theta}(X_m, Y_m)_j \right| \leq \frac{\sqrt{\theta^2 \log(2/\delta)}}{\sqrt{2M}}.$$

Proposition 3.3

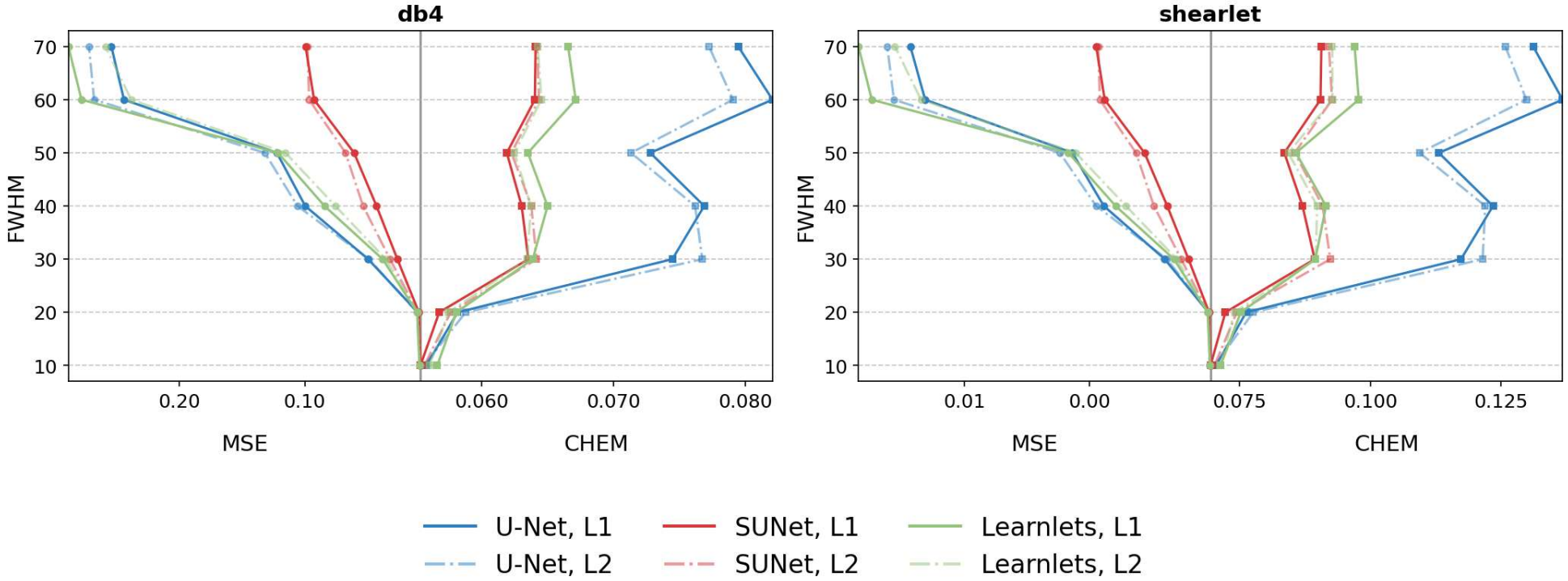
$$\mathrm{H}_{\text{high}} \geq (1-\mu) R_{\text{high}} \mathcal{E}(\hat{Y} - \widehat{\Phi(X)}) \qquad \mathrm{H}_{\text{low}} \leq (1+\nu)(1-R_{\text{high}}) \mathcal{E}(\hat{Y} - \widehat{\Phi(X)}).$$

Theorem 4.3

An upper bound for the prediction error $\mathcal{E}$ of U-shaped architectures on discretized functions (images) on a finite number of points, under some other assumptions ...

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

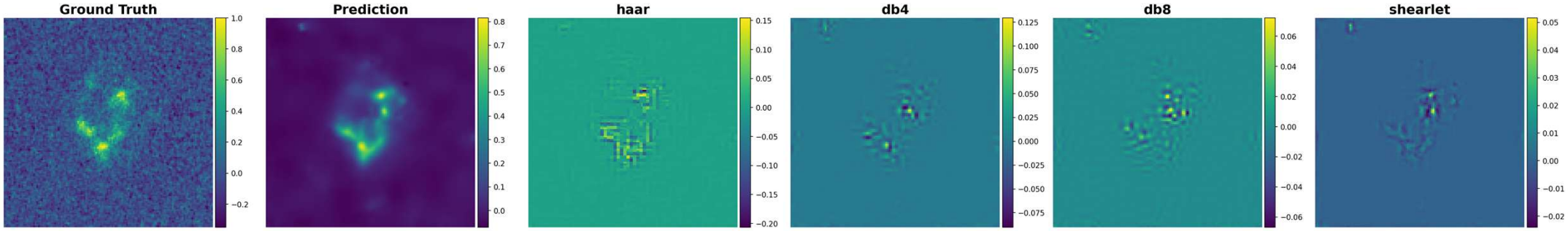
Experimental Results – Model Comparison



- U-Net and Learnlets perform similarly in MSE, yet Learnlets generate fewer hallucinations.
- Model assessment should explicitly account for hallucination artifacts.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

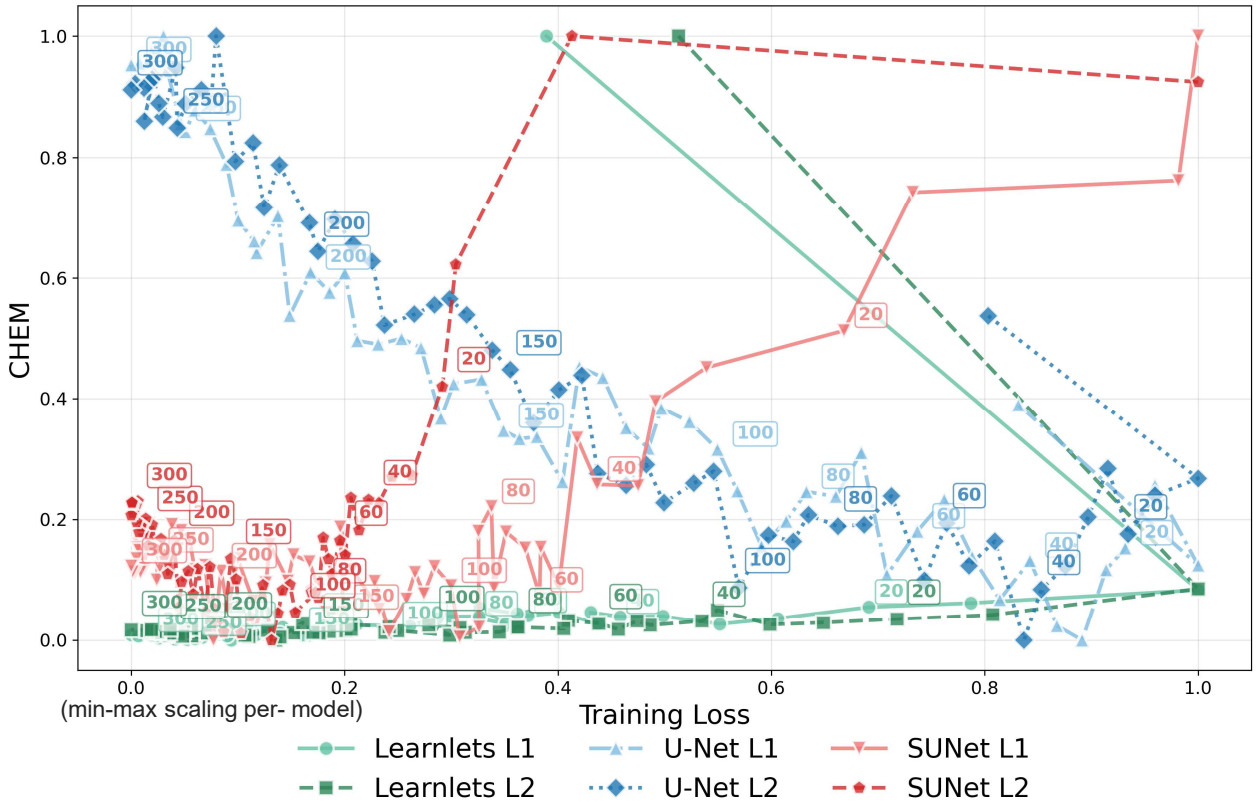
Experimental Results – Varying Dictionaries



- Wavelets and shearlets help capture multiscale information, which is crucial for astronomical imaging (Starck and Murtagh, 2006).
- Visually, shearlets and db8 provide a clearer representation of the hallucinations in the prediction.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Experimental Results – Hallucination- performance Tradeoff



- Beyond a certain number of epochs, further reductions in training loss come at the cost of increased hallucination.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Next Steps

- Use CHEM to penalize hallucination artifacts during training of new models
- Validation of our model on other safety-critical tasks
- Develop a mathematical explanation for the trade-off phenomenon and identify “optimal trade-off”

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

References

[1] F. Sureau, A. Lechat and J.-L. Starck (2020). Deep learning for a space-variant deconvolution in galaxy surveys A&A, 641 A67, DOI: https://doi.org/10.1051/0004-6361/201937039

[2] Akhaury, U., Starck, J.-L., Jablonka, P., Courbin, F., & Michalewicz, K. (2022). Deep learning-based galaxy image deconvolution. *Frontiers in Astronomy and Space Sciences*, *9*. https://doi.org/10.3389/fspas.2022.1001043

[3] Akhaury, U., Jablonka, P., Starck, J.-L., & Courbin, F. (2024). Ground-based image deconvolution with Swin Transformer UNet. *Astronomy & Astrophysics*, *688*, A6. https://doi.org/10.1051/0004-6361/202449495

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation* (No. arXiv:1505.04597). arXiv. https://doi.org/10.48550/arXiv.1505.04597

[5] Ramzi, Z., Michalewicz, K., Starck, J.-L., Moreau, T., & Ciuciu, P. (2023). Wavelets in the Deep Learning Era. *Journal of Mathematical Imaging and Vision*, *65*(1), 240–251. https://doi.org/10.1007/s10851-022-01123-w

[6] Fan, C.-M., Liu, T.-J., & Liu, K.-H. (2022). *SUNet: Swin Transformer UNet for Image Denoising*. 2333–2337. https://doi.org/10.1109/ISCAS48785.2022.9937486

[7] Leterme, H., Fadili, J., & Starck, J.-L. (2025). Distribution-free uncertainty quantification for inverse problems: Application to weak lensing mass mapping. *Astronomy & Astrophysics*, *694*, A267. https://doi.org/10.1051/0004-6361/202451756

[8] Starck, J.-L., & Murtagh, F. (2006). *Astronomical Image and Data Analysis*. Springer. https://doi.org/10.1007/978-3-540-33025-7

[9] Tikhonov, A. N., & Arsenin, V. Y. 1977, Solutions of Ill-posed problems, ed. W. H. Winston

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Ines Rosellon Inclan
089-2180-4678
rosellon@math.lmu.de