Knowledge Distillation for HDR Image Denoising

Compressing a 19.5M-Parameter Model by "almost" 100× (87x)

Ryan Chen

April 2026

Teacher: 19.5M params **Student:** 223K params **Speedup:** 25× FPS

Outline

1. Motivation & Background
2. Teacher Model Training
3. Knowledge Distillation
4. Evaluation Results
5. Conclusions
6. Hybrid Student Model
7. Optimization Execution
8. Final Hybrid Model Results
9. Why denoising HDR is fundamentally hard

Problem Statement

HDR Imaging Challenge

- Real HDR sensors capture images corrupted by **Poisson-Gaussian noise**
- Noise model: $\sigma^2 = 14 \cdot X_{clean} + \varepsilon$, $\varepsilon \sim \mathcal{U}(135, 160)$
- Input: demosaicked noisy RAW $\rightarrow$ Output: clean HDR

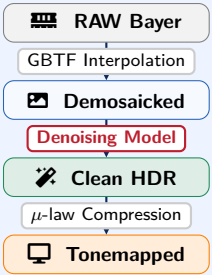
Deployment Problem

The best-performing teacher model has **19.5M parameters** and runs at only $\sim$**1 FPS** — too slow for real-world use. (Nvidia RTX3080 used)

Goal

100$\times$ smaller model
with minimal quality loss

Inference Pipeline



Linear Raw Input (I_{DN})



Tonemapped Noisy Observation (I_{DN}^{γ})



Model Architecture: Restormer-based U-Net

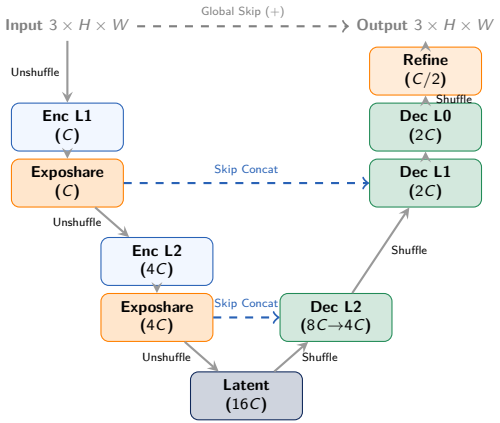
Key Design Choices

- **Backbone:** Restormer blocks (MDTA attention)
- **Structure:** 3-level U-Net with skip connections
- **Attention:** Channel-wise (transposed), not spatial
- **Novel:** `Exposhare` deformable conv modules for cross-exposure alignment

Restormer MDTA Block

$$\text{Attn} = \text{softmax}\left(\frac{Q^\top K}{\tau}\right) \quad (\text{C/h} \times \text{C/h map})$$

Window-free $\Rightarrow$ handles arbitrary resolution



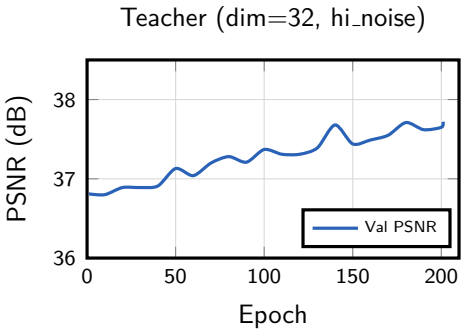
Teacher Model: Training Results

Teacher Configuration

Parameter	Value
Dim (C)	32
Blocks	[4, 4, 4, 4]
Refinement blocks	4
Total params	19.53M
Training epochs	201
Batch size	8
LR	2×10^{-4}

Peak Teacher PSNR (training)

37.72 dB @ epoch 201



Knowledge Distillation Framework

Distillation Loss Formulation

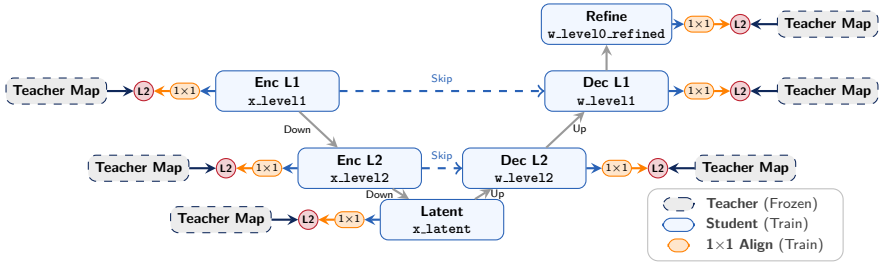
$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\,\mathcal{L}_{\text{out}} + \beta\,\mathcal{L}_{\text{feat}} + \gamma\,\mathcal{L}_{\text{LPIPS}}$

- $\mathcal{L}_{\text{task}}$: Tonemap MSE (Student vs. GT)
- $\mathcal{L}_{\text{out}}$: Tonemap L1 (Student vs. Teacher)
- $\mathcal{L}_{\text{LPIPS}}$: Perceptual Loss (Student vs. GT)
- $\mathcal{L}_{\text{feat}}$: Total MSE across 6 U-Net maps

Spatial Feature Alignment ($\mathcal{L}_{\text{feat}}$)

1×1 convs project Student features ($C_s = 8$) to Teacher capacity ($C_t = 32$), computing loss at specific stages:

- **Enc/Latent**: `x_level1`, `x_level2`, `x_latent`
- **Dec**: `w_level2`, `w_level1`, `w_level0`



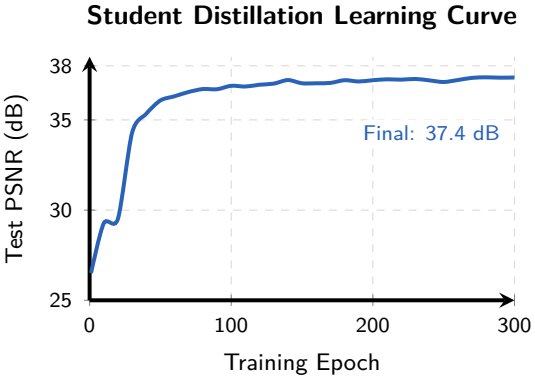
Student Architecture & Size Reduction

Model Comparison

	Teacher	Student
Dim (C)	32	8
Blocks	[4,4,4,4]	[2,2,2,1]
Params	19.53M	0.55M
Size	74.5 MB	2.1 MB
Ratio	**1**×	**35**× **smaller**

Student Training

- 300 epochs, Adam, lr=2×10^{-4}
- Batch size 8, patch 128×128
- PSNR: **37.4 dB** @ 300 epochs



Quantitative Evaluation Results

Test Set Performance Across Noise Levels

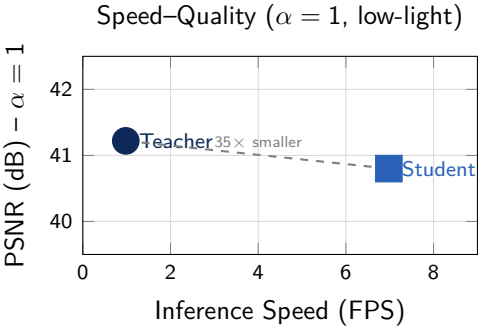
Model	**Params**	$\alpha = 1$ **(low-light)** PSNR / LPIPS	$\alpha = 3$ **(medium)** PSNR / LPIPS	$\alpha =$ *full* **(dynamic)** PSNR / LPIPS
Teacher	19.53M	41.22 / 0.477	35.67 / 0.381	32.43 / 0.315
Student	0.55M	**40.80** / 0.523	**35.45** / 0.422	**32.24** / 0.341
PSNR gap		−0.42 dB	−0.22 dB	−0.18 dB

35×
fewer parameters

7×
faster (FPS)

<0.5 dB
PSNR degradation

Speed vs. Quality Trade-off



Speed Summary

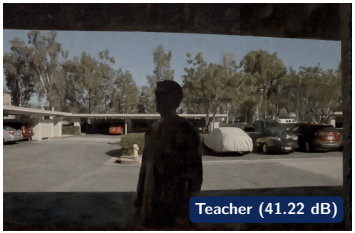
	Teacher	Student
$\alpha = 1$	0.98 FPS	**6.98 FPS**
$\alpha = 3$	1.09 FPS	**7.12 FPS**
$\alpha = 18$	1.34 FPS	**7.29 FPS**
Speedup	1×	**7.1**×

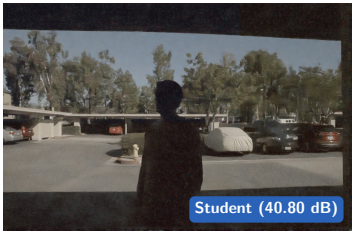
Key Observation

The student achieves **real-time-class** inference (>6 FPS at 512×512 patches) while retaining >99% of the teacher's PSNR.

Visual Comparison: Teacher vs. Student

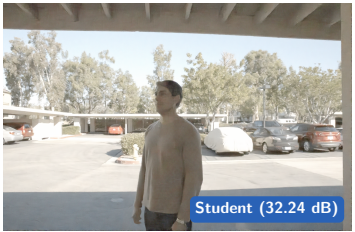
Low-Light ($\alpha = 1$)





Dynamic Range ($\alpha = 18$)





Pushing the Limit: Over-Compressing to dim=4

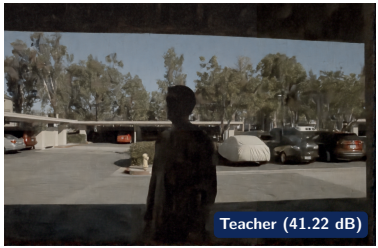
The dim=4 Student

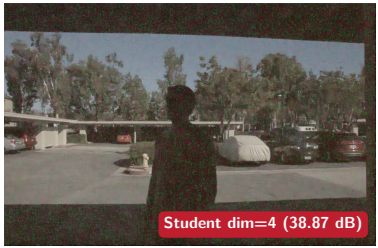
- **Extreme Compression:** 158K params (**122**× smaller)
- **Speed:** ∼12 FPS (vs. 7 FPS for dim=8)
- **Quantitative Drop:** For $\alpha = 1$, PSNR drops from 41.22 dB (Teacher) to **38.87 dB**.
- **LPIPS Degradation:** Perceptual error jumps from 0.477 to 0.615.

Visual Artifacts

While exceptionally fast, the feature capacity ($C = 4$) is too small to cleanly map the noise distribution, resulting in visible noise and color shifts in low-light ($\alpha = 1$) conditions.

Visual Comparison ($\alpha = 1$)





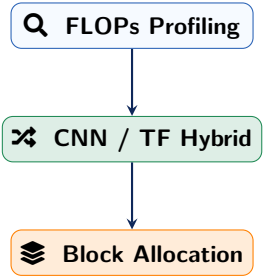
Hybrid CNN-Transformer Student Model

Network Profiling

- Perform detailed **computational profiling** to identify exact FLOPs bottlenecks across the current architecture.

Hybrid Architecture Design

- Investigate swapping heavy attention modules with more computationally efficient structures like **CNNs** where global spatial context is less critical.
- Determine the **most efficient allocation** of Transformer blocks across the different levels of the U-Net to maximize capacity while minimizing compute.



Goal

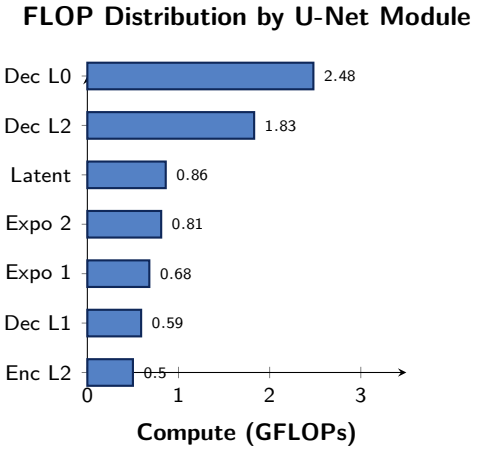
Maximize FPS while maintaining current PSNR levels.

Profiling Insights: Theoretical Compute (FLOPs)

fvcore Analysis

Total compute per 512×512 patch: **8.57 GFLOPs**. The distribution reveals severe spatial inefficiencies:

- **The High-Res Trap:** `decoder_level_0` executes 4 Transformer blocks at full resolution, consuming **2.48 GFLOPs** (29% of total).
- **Alignment Overhead:** The combined `Exposhare` modules account for another **1.48 GFLOPs** (17%).



Architectural Redesign: The Hybrid Strategy

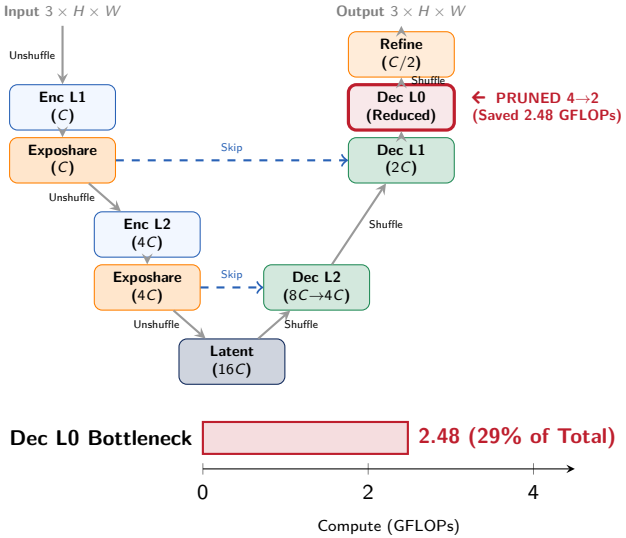
CNN-Transformer Sandwich

We shifted from a pure Restormer U-Net to a Hybrid architecture:

- **High-Res Stages:** Handled by fast, local CNN operations.
- **Low-Res Latent:** Restormer attention reserved for the bottleneck.

Aggressive Pruning

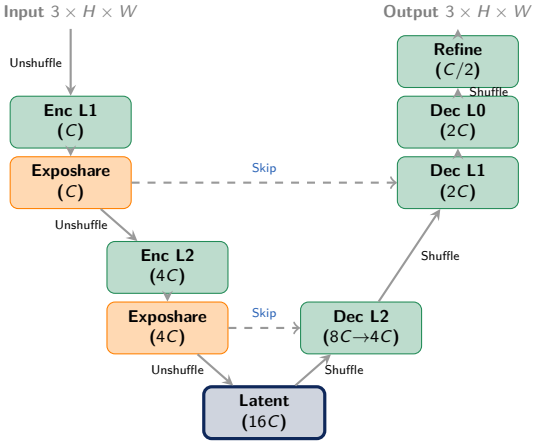
- Reduced final `num_refinement_blocks` in `Dec L0` from 4 down to 2.
- Directly eliminated the largest **2.48 GFLOP** bottleneck found in profiling.



Architectural Redesign: Module Replacements

Lightweight Substitutions

- **CNN Blocks:** Replaced heavy Restormer blocks in Encoder L1/L2 and Decoder L2/L1/L0 with Lightweight CNN blocks (inverted bottlenecks).
- **Fast Exposhare:** Swapped slow Deformable Convolutions (DCNs) for 5×5 depthwise-separable convolutions, maintaining the receptive field at a fraction of the hardware cost.



Lightweight CNN

Restormer (Transformer)

Depthwise Conv (Fast Exposhare)

The Hybrid Breakthrough: 87× Compression

Final Model Comparison

	Teacher	Hybrid Student
Dim (C)	32	8
Params	19.53M	223K
Inference	∼1.2 FPS	∼**32 FPS**
Ratio	**1**×	**87**× **smaller**

Training Milestone

- Successfully converged to a peak validation PSNR of **37.51 dB**.

87×
fewer parameters

25×
faster (Real-Time)

223,475
Total Parameters

Quantitative Evaluation: Hybrid Performance

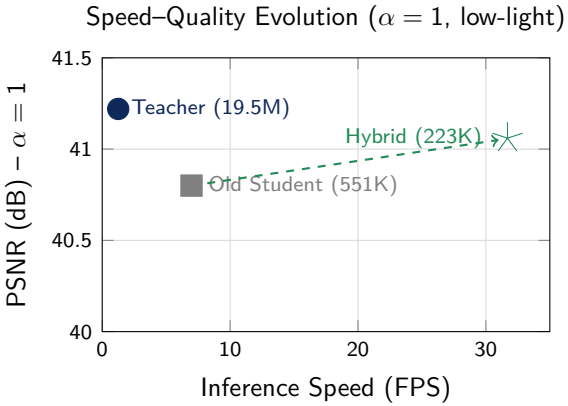
Test Set Performance Across Noise Levels (Hybrid Student)

Model	**Speed**	$\alpha = 1$ **(low-light)** PSNR / LPIPS	$\alpha = 3$ **(medium)** PSNR / LPIPS	$\alpha =$ *full* **(dynamic)** PSNR / LPIPS
Teacher	1.2 FPS	41.22 / 0.477	35.67 / 0.381	32.43 / 0.315
Old Student	7.1 FPS	40.80 / 0.523	35.45 / 0.422	32.24 / 0.341
Hybrid	**32.0 FPS**	**41.06** / **0.461**	**35.44** / **0.373**	**32.23** / **0.311**
Gap to Teacher		−0.16 dB	−0.23 dB	−0.20 dB

The Ultimate Trade-off Solution

The Hybrid CNN-Transformer architecture runs at **31.6 to 33.1 FPS** while strictly improving upon the perceptual quality (LPIPS) of the original 551K student. It successfully captures the teacher's denoising manifold at a fraction of the hardware cost.

Speed vs. Quality: Crossing the Real-Time Threshold



Model Comparison ($\alpha = 1$)

Metric	Teacher	Old Std	**Hybrid**
Params	19.5M	551K	**223K**
PSNR	41.22	40.80	**41.06**
FPS	1.25	6.98	**31.69**
Speedup	1×	5.6×	**25.3**×

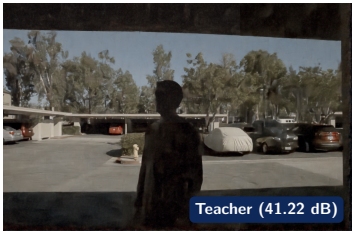
Efficiency Gains

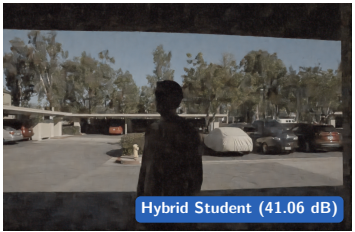
- **4.5**× more FPS than the Old Student.
- **+0.26 dB** PSNR over Old Student.
- **Real-time** performance (∼31ms).

Performance measured on **NVIDIA RTX 3080**

Visual Comparison: Teacher vs. Hybrid Student

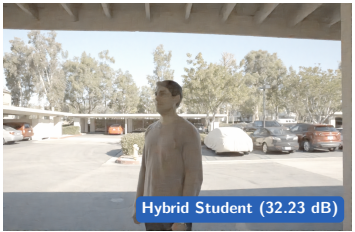
Low-Light Regime ($\alpha = 1$)





Full Dynamic Range ($\alpha =$ full)





Why Denoising HDR is Fundamentally Hard

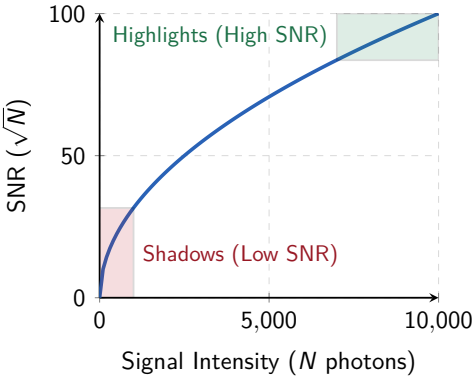
The Physics of Noise: Photon Shot Noise

- Light arrives as discrete photons following a **Poisson distribution**.
- Let N be the number of collected photons.
- Noise standard deviation (σ) is fundamentally bound by $\sqrt{N}$.

Fundamental SNR Equation

$$\text{SNR} = \frac{\mu}{\sigma} = \frac{N}{\sqrt{N}} = \sqrt{N}$$

Implication: Signal reliability is intensely *heteroscedastic*. Brighter regions (highlights) are significantly cleaner than darker regions (shadows).



The "Signal" in SNR for LDR

Signal Definition: The Display Value

8-bit Integer ($0 - 255$)

- **Perceptual Mapping:** Optimized for human vision and standard displays (sRGB).
- **Non-Linearity:** Gamma-encoded ($S^{1/\gamma}$) to match the logarithmic sensitivity of the human eye.
- **Processed Data:** The signal is a product of an Image Signal Processor (ISP).

Noise Model: Complex & Spatially Correlated

Traditional Assumption: AWGN ($n \sim \mathcal{N}(0, \sigma^2)$).
Real In-Camera Reality: The ISP drastically alters the raw sensor noise:

- **Gamma Amplification:** Non-linear tone mapping amplifies variance in the shadows.
- **Spatial Correlation:** Demosaicking spreads isolated pixel noise across local neighborhoods, creating low-frequency color blotches.
- **Intractability:** Distribution is mathematically intractable.

The "Signal" in SNR for HDR

Physical Forward Model (Rigorous)

$$I_{\text{DN}}(\mathbf{x}) = \text{CG}_{\text{ADC}}\Big\{\text{CG}_{\text{col}}\{\text{CG}_{\text{FD}} \times \text{Poisson}\{\text{QE} \cdot \theta(\mathbf{x}) + \theta_{\text{dark}}\} + \text{Gauss}(0, \sigma^2_{\text{pix}})\} + \text{Gauss}(0, \sigma^2_{\text{col}}) + \text{Gauss}(0, \sigma^2_{q})\Big\}$$

Noise Model

The physical cascade collapses into an additive heteroscedastic model:

$$I_{noisy} = X_{dn} + \sqrt{14 \cdot X_{dn} + \epsilon} \cdot \mathcal{N}(0,1), \quad \epsilon \sim U(135, 160)$$

- **Clean Signal (X_{dn}):** Expected digital value $\mathbb{E}[I_{DN}]$. Absorbs irradiance (θ) and Quantum Efficiency (QE).
- **System Gain ($K = 14$):** Measured variance slope. Aggregates all physical conversion gains ($\text{CG}_{\text{FD}} \cdot \text{CG}_{\text{col}} \cdot \text{CG}_{\text{ADC}}$).
- **Read Noise Floor (ϵ):** Aggregates dark current variance (θ_{dark}) and all Gaussian readout errors ($\sum \sigma^2$). Modeled as $U(135, 160)$ to capture frame-to-frame thermal/baseline drift.

LDR vs. HDR

LDR Workflow

Low Dynamic Range (8-bit)

- **Shadows:** Crushed to black (0). Low-SNR data is discarded, hiding the noise floor.
- **Highlights:** Clipped at max value (255). Variance flatlines.
- **Result:** Noise variance appears relatively uniform across the midtones.

HDR Workflow

High Dynamic Range (16/32-bit)

- **Shadows:** Lifted via digital gain to reveal detail. Fragile noise floor is heavily amplified.
- **Highlights:** Protected via multi-exposure, retaining massive photon counts.
- **Result:** Extreme signal-dependent noise requiring spatially variant processing.

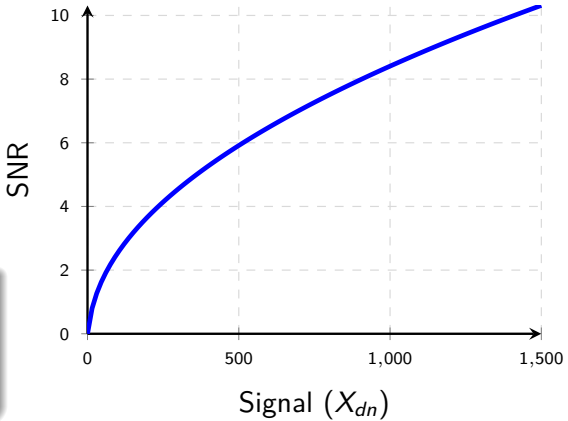
The Physics of Noise: Sensor-Specific Model

- Let X_{dn} be average signal intensity.
- Total variance model:

  $\sigma^2 = 14 \cdot X_{dn} + \epsilon$
- **Shot Noise:** $14 \cdot X_{dn}$ Signal-dependent.
- **Read Noise:** Electronic floor, $\epsilon \sim U(135, 160)$.

Sensor SNR Equation

$$\text{SNR} = \frac{X_{dn}}{\sqrt{14 \cdot X_{dn} + \epsilon}}$$



Implication: SNR crashes in shadows ($14X_{dn} \ll \epsilon$), dominated by the read noise floor.

Next Step: Multiple lightweight Denoisers

The Challenge: A single network cannot optimally handle the extreme SNR variance of HDR data.

Our Spatially-Adaptive Approach:

- **Specialized Experts:** Denoisers are partitioned by their explicit **SNR Operating Range**.
- **Efficiency:** Each 223K student expert only learns the noise statistics bounded within its specific, tightly-controlled SNR threshold.

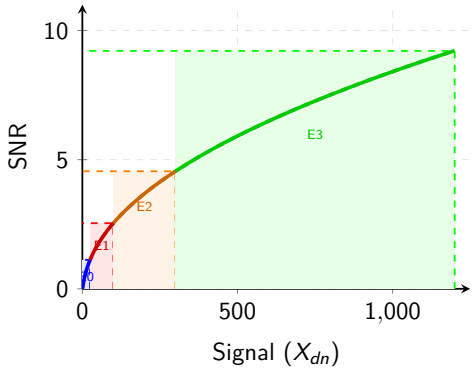


Figure: Specialized student experts (E0-E3) bound by explicit SNR thresholds, dense mapping near zero signal.