Knowledge Distillation and Composite Loss as a Low-Resource Strategy for LDCT Image Denoising

Raíssa X. Contassot[1], Augusto O. Ughini[1], Jerusa D. Finatto[1], Kristen K. Arguello[1], Rafaela C. Ravazio[1], Otávio Parraga[1], Christian Mattjie[1], Lucas S. Kupssinskü[1], Rodrigo C. Barros[1,2]

[1]Machine Learning Theory and Application Lab,
Pontifical Catholic University of Rio Grande do Sul,
90619-900 – Porto Alegre, RS, Brazil

[2]Kunumi Institute, Brazil

{raissa.contassot, augusto.ughini, jerusa.finatto,
kristen.arguello, rafaela.ravazio, otavio.parraga,
christian.oliveira95}@edu.pucrs.br
{lucas.kupssinsku, rodrigo.barros}@pucrs.br

Abstract. *Low-dose CT (LDCT) lowers radiation exposure at the cost of noisier images that can hinder diagnosis. Deep learning restores quality, but state-of-the-art models often need tens of millions of parameters, out of reach for institutions without dedicated GPU infrastructure. We propose knowledge distillation (KD) as a low-resource strategy: a compact 1.11M-parameter RED-CNN-SE student distilled from a 12M-parameter AttentionUNet-Pro teacher, trained with a composite Charbonnier, Sobel, and Stationary Wavelet Transform loss. In a single forward pass the student reaches 33.72 dB PSNR, within 0.02 dB of the teacher and at 99.6% of the PSNR reported by the state of the art, with roughly 10× fewer parameters. Optional test-time augmentation adds a further 0.04 dB at eight times the inference cost and is reported separately, since it conflicts with the low-resource premise of this work.*

1. Introduction

Ionizing radiation is widely used in medical imaging, with X-rays and Computed Tomography (CT) being the most common modalities [Junda et al. 2021]. Unlike X-rays, CT provides volumetric data, enabling detailed visualization of internal structures. However, this comes at the cost of a higher radiation dose [dos Santos Junior et al. 2020], raising concerns about cumulative exposure and long-term risk. Low-Dose CT (LDCT) has emerged as a practical approach to address this concern. By reducing the radiation dose to roughly 25% of that used in Normal-Dose CT (NDCT), LDCT maintains diagnostic utility while limiting patient exposure [Hu et al. 2021, Jo et al. 2023, Demir et al. 2023].

Nevertheless, radiation dose optimization comes at the cost of image quality. The increased noise in LDCT images is visually noticeable, potentially leading to loss of diagnostic information [Demir et al. 2023]. To overcome this challenge, image-to-image translation methods offer an alternative that can eliminate noise from LDCT images without exposing the patient to additional radiation [Chen et al. 2024, Eulig et al. 2024].

Residual Encoder-Decoder Convolutional Neural Network (RED-CNN) [Chen et al. 2017] is among the most widely explored architectures for LDCT

denoising, achieving competitive results in both simulated and clinical settings. However, its purely convolutional design limits its receptive field, thereby restricting its ability to model long-range spatial dependencies. Attention-based approaches with multi-scale feature extraction, such as AttentionUNet-Pro [Huang et al. 2024], address this by incorporating global context, but at the cost of a high parameter count, making them impractical in low-resource settings. Knowledge Distillation (KD) [Hinton et al. 2015] offers a middle ground: by supervising a compact student with the teacher's outputs, it provides an additional training signal that improves generalization beyond what hard labels alone can achieve, thereby retaining performance while reducing computational cost.

This work does not claim novelty in knowledge distillation itself, which is well established for image restoration. Our contribution is empirical and threefold. First, we quantify how far a heavily compressed RED-CNN based student can be pushed on the paired AAPM-Mayo benchmark under an explicit low-resource constraint, reaching 33.72 dB PSNR with 1.11M parameters and staying within 0.02 dB of a 12M-parameter teacher trained under identical conditions. Second, we attribute the gain component by component, showing that the composite Charbonnier, Sobel, and SWT loss accounts for most of the improvement over the baseline, while output-level KD contributes a small but decisive increment that breaks a training plateau longer training alone does not overcome. Third, we characterise the resulting model in deployment terms, reporting parameters and per-slice computational cost, so that the low-resource claim is measurable rather than asserted. Our code and experiments are available online[1].

2. Related Work

The release of the paired LDCT/NDCT dataset from the AAPM/Mayo Clinic Low-Dose CT Grand Challenge [American Association of Physicists in Medicine 2016] established a shared benchmark for supervised deep learning denoising methods. RED-CNN [Chen et al. 2017] defined the residual encoder-decoder CNN baseline on this dataset and remains the standard point of comparison. Later CNN variants added domain-specific priors: wavelet decomposition for frequency-band supervision [Kang et al. 2017], adversarial and perceptual losses for texture preservation [Yang et al. 2018], and edge-based objectives for structural sharpness [Liang et al. 2020]. These models operate in the same 1–2M parameter range as RED-CNN and incrementally improve PSNR.

Incorporating attention and multi-scale feature extraction into encoder-decoder architectures has improved denoising further. MSPMnet [Huang et al. 2024] uses a multi-head decomposable convolution that splits channels across depth-wise kernels of different sizes, combined with a receptive field-ramp that widens the receptive field with network depth. It reports 33.85 dB PSNR on the AAPM-Mayo benchmark with roughly 11M parameters and defines the current state of the art. Transformer-based [Wang et al. 2023a] and diffusion-based [Gao et al. 2023] models reach comparable quality but carry higher inference costs, which limits their use in resource-constrained clinical settings.

Knowledge distillation (KD) was introduced by Hinton et al. [Hinton et al. 2015] for image classification, where a compact student is trained to match the soft output dis-

[1] https://anonymous.4open.science/r/knowledge-distillation-ldct/

tribution of a larger frozen teacher. Adapting KD to dense image-to-image regression is less direct, since the output is a full-resolution image rather than a class distribution. FitNets [Romero et al. 2015] showed that supervising intermediate feature maps produces better-calibrated students in dense prediction, and FAKD [He et al. 2020] found that feature-level distillation for super-resolution requires spatial alignment between teacher and student representations to avoid degraded outputs.

Several methods apply KD directly to LDCT denoising. SGKD [Wang et al. 2023b] targets the unpaired setting, where co-registered LDCT/NDCT slices are unavailable. UNIQ [Zhou et al. 2023] incorporates uncertainty estimation into the distillation objective to improve prediction reliability. More recent methods [Wang and Pang 2026, Zhao 2025] explore feature-level and frequency-domain distillation with different accuracy-to-compactness trade-offs. For natural image restoration, retargeting-based supervision [Zou et al. 2022] and the dual-teacher SLKD [Yang 2025] confirm that KD transfers effectively to this task class. None of these methods target the paired AAPM-Mayo benchmark under an explicit low-resource constraint using a RED-CNN-based student. This work examines how far a heavily compressed student can close the gap to state-of-the-art performance on the standard AAPM-Mayo benchmark, using output-level KD and a composite reconstruction loss.

Throughout this paper, SOTA refers to the 33.85 dB PSNR reported by MSPMnet in its original publication, which we did not reproduce under our protocol since no implementation is publicly available. Our teacher is an in-house architecture inspired by MSPMnet and does not reproduce those numbers, so the only fully controlled comparisons reported here are those between the baseline RED-CNN, the student, and our teacher, all trained and evaluated under identical splits, preprocessing, and metrics.

3. Methodology

This work proposes a knowledge distillation (KD) strategy with a composite loss for LDCT denoising in a low-resource scenario. The pipeline has three stages: training a teacher network, training a compact student network through distillation, and applying test-time augmentation at inference. The remainder of this section describes the data, the two networks, the loss, the distillation procedure, the training setup, and the evaluation protocol.

3.1. Dataset

We used the 2016 Low-Dose CT Grand Challenge from the American Association of Physicists in Medicine (AAPM) and the Mayo Clinic [American Association of Physicists in Medicine 2016]. Of the 30 abdominal scans in the challenge, only 10 training patients were released as paired NDCT and LDCT, with the LDCT at 25% of the normal dose. We split these 10 patients following MSPMnet [Huang et al. 2024], using 9 for training and 1 (L506) for both checkpoint selection and final evaluation, consistent with prior work on this benchmark. All experiments used a slice thickness of 3 mm and a medium reconstruction kernel (B30).

3.2. Data Preprocessing

All images were clipped to the abdominal soft-tissue window of $[-160, 240]$ Hounsfield Units (window level 40, width 400) and linearly mapped to $[0, 1]$.

Data augmentation used three operations applied identically to each (LDCT, NDCT) pair: a horizontal flip with probability 0.5, a vertical flip with probability 0.5, and a rotation of $90 \times k$ with $k \sim \mathcal{U}\{0, 1, 2, 3\}$.

3.3. Teacher Network

The teacher is a custom AttentionUNet-Pro inspired by MSPMnet [Huang et al. 2024]; as no public implementation was available, it was developed in-house and trained from scratch on the AAPM-Mayo training split with no externally pre-trained weights and the same partitions as the student. It has roughly 12 million parameters, a base channel width of 26, GroupNorm normalization, and GELU activations throughout.

The encoder follows a four-level U-Net structure with residual convolutional blocks, and the deeper levels add large-kernel convolutional blocks to enlarge the receptive field. SE blocks [Hu et al. 2018] with a reduction factor of 16 modulate every skip connection between the encoder and decoder. The decoder uses PixelShuffle upsampling.

The bottleneck stacks two Multi-Head Dilated Convolution (MHDC) blocks. Each block has four parallel branches (dilated 3×3 convolutions and global-context pooling), which are concatenated and projected back to the bottleneck width.

The teacher was optimized with a hybrid scheme: Muon for the convolutional weights and AdamW for the remaining parameters (normalization scales and biases, and 1×1 projections). Training used a progressive patch-size schedule and EMA weight averaging. The loss combined three terms: Charbonnier, MS-SSIM ($\lambda = 0.15$), and gradient ($\lambda = 0.10$). Figure 1 shows the teacher architecture.

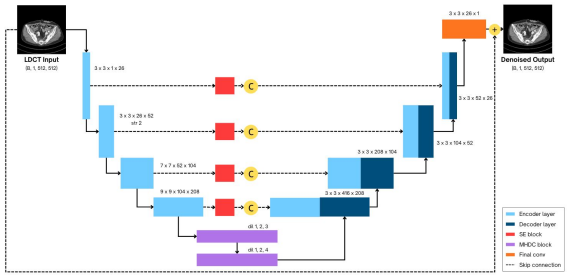


Figure 1. Architecture of the teacher model. The encoder-decoder blocks are defined as kernel (k) x kernel x input channels (i) x output channels (o).

3.4. Student Network

The student is based on RED-CNN [Chen et al. 2017], a compact encoder-decoder CNN that operates at the original resolution without pooling or striding, using 5×5 kernels. The model learns only the noise residual through a global skip connection that adds the input to the final layer. We keep this residual formulation without BatchNorm or GroupNorm.

We extend the original design by adding SE blocks to the skip connections, yielding RED-CNN-SE. The model uses a base channel width of 74, compared to 96 in the

original RED-CNN, reducing the parameter count from 1.85M to 1.11M. Figure 2 shows the student architecture.

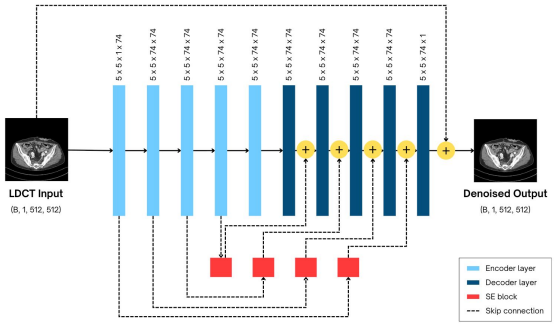


Figure 2. Architecture of the student model. The encoder-decoder blocks are defined as k x k x i x o.

Table 1 shows where the order-of-magnitude gap between the networks is concentrated. The teacher holds 44.8% of its parameters in the MHDC bottleneck (5.37M, or 4.8× the entire student), while the student splits its budget evenly between encoder and decoder.

Table 1. Parameter distribution of student and teacher networks. Components reflect each architecture's natural structure and are not strictly aligned across columns.

	Student (RED-CNN-SE)		Teacher (AttentionUNet-Pro)	
Component	**Params**	**%**	**Params**	**%**
Encoder	549,820	49.5	809,432	6.7
Bottleneck (MHDC)	–	–	5,369,728	44.8
Decoder	549,747	49.5	5,805,020	48.4
SE blocks	10,656	1.0	7,280	0.1
Final conv	–	–	235	0.0
Total	**1,110,223**	**100.0**	**11,991,695**	**100.0**

3.5. Loss

The loss combines four components: Charbonnier, KD, Sobel, and Stationary Wavelet Transform (SWT). Each one supervises a different aspect of the reconstruction, as shown in Equation 1.

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{charbonnier} + \lambda_{grad} \cdot \mathcal{L}_{grad} + \lambda_{swt} \cdot \mathcal{L}_{swt}}_{\mathcal{L}_{hard}} + \lambda_{\text{kd}} \cdot \mathcal{L}_{\text{kd}} \tag{1}$$

Charbonnier loss is common in computer vision and produces clean edges and sharp details, which helps in medical imaging [Gajera et al. 2021]. It is defined in Equation 2, with ε set to 10^{-3}.

$$\mathcal{L}_{\text{Charb}}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(\hat{y}_i - y_i)^2 + \varepsilon^2} \tag{2}$$

Sobel loss, also called gradient loss, uses Sobel filters to compute the image gradient and pushes the network toward sharp, structured edges [Zhang et al. 2022]. It is defined in Equation 3, with λ_{grad} set to 0.15.

$$\mathcal{L}_{\text{grad}} = \|\nabla_x \hat{y} - \nabla_x y\|_1 + \|\nabla_y \hat{y} - \nabla_y y\|_1 \tag{3}$$

SWT loss works on the multi-scale decomposition of image frequencies [Korkmaz and Tekalp 2024]. Haar wavelets decompose the image into four sub-bands (one approximation, LL, and three detail bands, LH, HL, HH) without reducing the resolution. The SWT loss sums the L1 distances between the student and ground-truth wavelet coefficients across all four sub-bands, as defined in Equation 4, with λ_{swt} set to 0.10.

$$\mathcal{L}_{\text{SWT}} = \|\text{LL}_{\hat{y}} - \text{LL}_y\|_1 + \sum_{s \in \{\text{LH, HL, HH}\}} \|s_{\hat{y}} - s_y\|_1 \tag{4}$$

The KD loss measures the distance between the student and teacher outputs, as defined in Equation 5. The teacher output $\hat{y}_T$ is computed with gradients disabled, since the teacher is frozen during distillation. The weight λ_{KD} follows a stepwise schedule, described in Subsection 3.6.

$$\mathcal{L}_{\text{kd}} = \|\hat{y}_S - \hat{y}_T\|_1 \tag{5}$$

Supervising the student with L1 distance on the teacher's outputs, rather than on its training objective, encourages the student to learn the teacher's denoised distribution instead of imitating its optimization trajectory.

3.6. Knowledge Distillation

We adopted soft-label distillation. AttentionUNet-Pro was the teacher, pretrained and fully frozen during distillation. RED-CNN-SE was the student, learning simultaneously from NDCT images as ground truth via $\mathcal{L}_{\text{hard}}$ and from the teacher's output via $\mathcal{L}_{\text{kd}}$. The student was warm-started from a non-distilled RED-CNN-SE checkpoint at 33.67 dB.

The distillation weight λ_{KD} followed a three-stage schedule that reflects the role of distillation at different training phases. From epoch 1 to 20, a low warm-up weight ($\lambda_{KD} = 0.10$) let the student stabilize on the ground-truth signal, moving from 33.56 to 33.57 dB. From epoch 21 to 120, a stronger weight ($\lambda_{KD} = 0.30$) applied maximal teacher guidance once the student could absorb it, carrying it to 33.69 dB. From epoch 121 onward, the weight returned to $\lambda_{KD} = 0.10$ to avoid over-regularizing toward the teacher as the student approached its asymptote, ending at 33.72 dB. Convergence was monotonic across the three stages, with the improvement rate decreasing as the model approached its asymptote. These values were selected by PSNR on L506.

3.7. Training Details

Training ran on an NVIDIA TITAN Xp 12 GB GPU for the full 200 scheduled epochs. The cosine learning-rate schedule decayed from 1×10^{-4} at epoch 1 to 1×10^{-7} at epoch 200. The best validation PSNR (33.72 dB at epoch 194) was selected as the final checkpoint. Training used a progressive patch schedule at roughly constant pixels per batch: 128×128 with batch size 32 for epochs 1 to 39, 192×192 with batch size 16 for epochs 40 to 79, and 256×256 with batch size 8 for epochs 80 to 200.

We used AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay 1×10^{-4}. The first 5 epochs used a linear warm-up, after which the learning rate followed the cosine schedule above. An exponential moving average (EMA) of the student parameters was kept during training, following $\theta_{\text{EMA}}^{(t)} = \alpha \cdot \theta_{\text{EMA}}^{(t-1)} + (1 - \alpha) \cdot \theta^{(t)}$ with $\alpha = 0.999$, updated after every optimizer step. EMA weights were used for all reported metrics, including validation and the final L506 test-set evaluation. Training used automatic mixed precision with dynamic loss scaling, and steps with a non-finite unscaled gradient norm were skipped. Mixed precision reduced activation memory during distillation.

3.8. Test-Time Augmentation (TTA)

At inference, we optionally apply the eight elements of the dihedral group D_4 (four rotations and their horizontal flips). Each transformed input is processed under the overlap-tile protocol of Subsection 3.9, the inverse transformation realigns the output, and the eight predictions are averaged.

3.9. Evaluation

We used the Structural Similarity Index (SSIM) and the Peak Signal-to-Noise Ratio (PSNR), as both are common in medical imaging studies. PSNR was computed as $10 \cdot \log_{10}(R^2/\text{MSE})$ with $R = 1.0$, since images were normalized to $[0, 1]$, and SSIM was computed in local windows to avoid inflated values. To evaluate the full 512×512 slices, we used an overlap-tile strategy. With a stride of 128 pixels and 50% overlap, a 256×256 patch is extracted at each position and passed through the model. The final result is a Gaussian-weighted average over the overlapping regions.

4. Results

Table 2 reports quantitative and efficiency results together. In a single forward pass, the student (RED-CNN-SE + KD) achieved 33.72 dB PSNR and 0.9279 SSIM on the test patient, within 0.02 dB of the teacher and at 99.6% of the PSNR reported by MSPMnet, using $10.8\times$ fewer parameters than the teacher. Optional TTA-8 raises this to 33.76 dB and 0.9281 SSIM, at eight times the inference cost.

Parameter count alone does not determine deployability, since a model operating at full resolution can cost more per image than a larger model with aggressive downsampling. The overlap-tile protocol of Subsection 3.9 processes $2.25\times$ the image area, so the effective evaluation cost is proportionally higher for every model in the table. Even so, the student reaches a PSNR/M ratio of 30.38, nearly $11\times$ higher than the teacher, showing that parameter efficiency and denoising quality are not mutually exclusive.

The training dynamics of the KD student are illustrated in Figure 3. The two upward shifts in $\mathcal{L}_{\text{hard}}$ at epochs 40 and 80 coincide precisely with the progressive crop

Table 2. Quantitative and efficiency results. MSPMnet values are taken from the original publication and were not reproduced under our protocol.

Method	Param. (M)	PSNR (dB)	SSIM	PSNR/M	Δ to MSPMnet
RED-CNN (Baseline)	1.85	32.45	0.9023	17.54	-1.40
RED-CNN-SE + KD (Student)	1.11	33.72	0.9279	30.38	-0.13
+ TTA-8 (optional)	1.11	33.76	0.9281	30.41	-0.09
AttentionUNet-Pro (Teacher)	11.99	33.74	0.9289	2.81	-0.11
+ TTA-8 (optional)	11.99	33.77	0.9290	2.82	-0.08
MSPMnet (reported)	$\sim$11.0	33.85	0.9309	3.08	–

transitions. Larger patches introduce more heterogeneous content, raising $\mathcal{L}_{hard}$ without degrading performance, while $\mathcal{L}_{KD}$ decreases steadily by 30.3% as the student converges toward the teacher on KD loss (from 0.00670 on epoch 1 to 0.00467 on epoch 200). Within each patch-size phase, a linear fit of $\mathcal{L}_{hard}$ has a negative slope (-1.23, -1.82 and -0.50×10^{-5} per epoch for the 128, 192 and 256 phases respectively), ruling out training instability and attributing the upward steps to the change in evaluation pool.

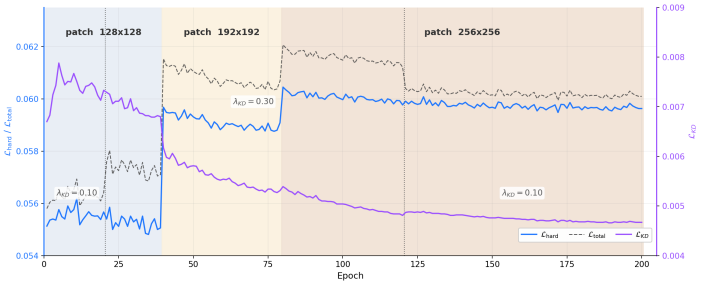


Figure 3. Training dynamics of the KD student model. Left axis: $\mathcal{L}_{\text{hard}}$ and $\mathcal{L}_{\text{total}}$ (blue); right axis: $\mathcal{L}_{KD}$ (purple).

Figure 4 presents a visual comparison between LDCT, denoising output, and ground truth (NDCT) images across different slices from the test patient. In addition, it presents the corresponding residual and noise maps. It is evident that the technique significantly removed noise from LDCT images, indicating a positive path for clinical applications.

4.1. Ablation Study

Table 3 presents the PSNR results for each improvement applied to the model. Upgrade A refers to the Enhanced Training (loss, optimizer and EMA); B is related to the addition of SE blocks, refined loss weights and gradient clipping (fine-tuned from the A checkpoint); C refers to the KD; and D is the optional TTA-8 step. Enhanced training (composite loss, optimizer, and EMA) was the main contributor, accounting for 54% of the total gain. Architectural refinement (SE blocks, loss weight tuning, and gradient clipping) ranked second, representing 39% of the improvement.

Although KD accounts for only 3.8% of the total gain, Figure 5 shows it was essential to break through the baseline plateau at 33.67 dB, a ceiling not achievable via longer training alone. With KD, the student reaches 33.72 dB in a single forward pass,

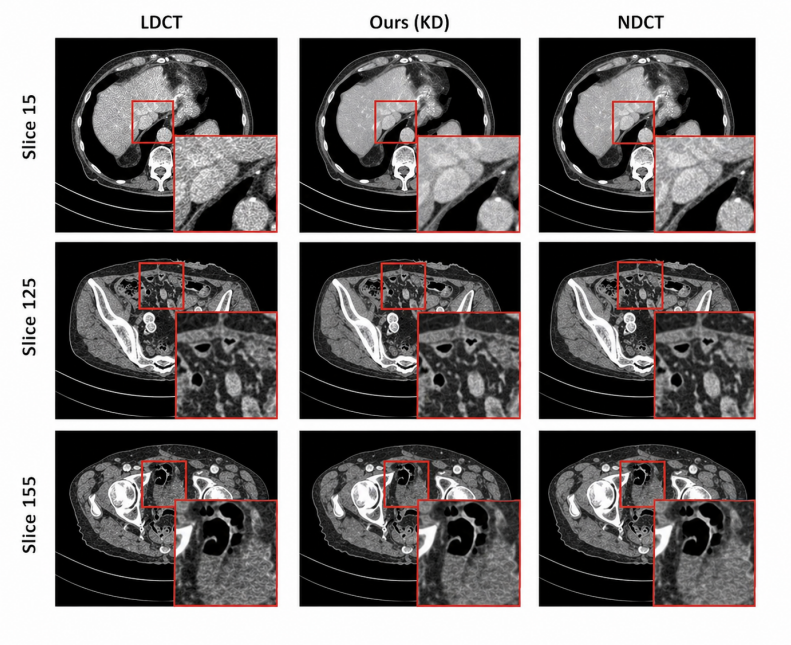


Figure 4. Visual comparison between LDCT, denoising output and ground truth (NDCT) images.

Table 3. Contribution of each component to the study.

Upgrade	PSNR (dB)	Accumulated
Baseline RED-CNN	32.45	-
+ A	33.16	+ 0.71
+ B	33.67	+ 1.22
+ C	33.72	+ 1.27
+ D (optional)	33.76	+ 1.31

within 0.13 dB of the PSNR reported by MSPMnet despite a $10.8\times$ compression ratio relative to the teacher. The SE blocks added approximately 11K parameters at negligible cost, and their inclusion in upgrade B suggests channel attention in skip connections is a high-leverage design choice.

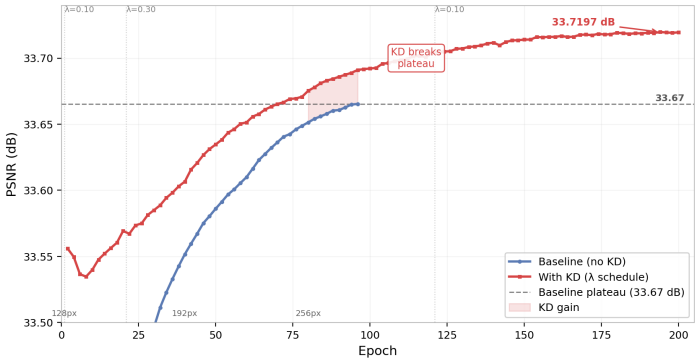


Figure 5. Convergence curves illustrating the power of the KD strategy. Shaded area marks the KD gain.

4.2. Limitation

The main limitation of this work is the same constraint that inspired it: limited computational resources prevented the exploration of larger teacher capacities and finer slice thicknesses. As no public implementation of MSPMnet was available, distillation was performed from an in-house reimplementation, and comparisons against its reported numbers are therefore indicative rather than controlled; access to the original weights would likely raise the student's performance ceiling. Another limitation was the dataset size: the model was trained on 9 patients and tested on only 1, which is the standard for this benchmark but limits generalization across scanners and acquisition protocols. Finally, TTA-8 roughly multiplies inference cost by 8, which runs counter to the low-resource premise of this work. We therefore treat it as strictly optional: it adds only 0.04 dB over the 33.72 dB single-pass student, and every claim in this paper holds without it.

5. Conclusion

This work proposed a KD strategy for LDCT denoising in limited-resource settings. The student (RED-CNN-SE + KD) reached 33.72 dB PSNR on the L506 test patient in a single forward pass, $10.8\times$ more compact than the teacher and within 0.02 dB of it, at 99.6% of the PSNR reported by MSPMnet. Test-time augmentation raises this to 33.76 dB but multiplies inference cost by eight, and is therefore reported as an optional add-on rather than as the headline result. The composite loss and training setup contributed most of the gain over the baseline, while KD broke a plateau that longer training could not overcome. Together, these results show that a compact model can approach the denoising quality of much larger architectures. Future work should explore 1 mm slice-thickness data and alternative teacher architectures, aiming to surpass the state of the art while keeping the parameter count low. By lowering the hardware needed for LDCT denoising, this strategy can extend the reach of dose-reduction technology to institutions without dedicated GPU infrastructure.

Acknowledgments

This study was supported by the Center for Innovation and Artificial Intelligence in Health (CI-IA Saúde), with partial funding from the São Paulo Research Foundation (FAPESP), Grant No. 2020/09866-4; by the Minas Gerais Research Foundation (FAPEMIG), Grant No. PPE-00030-21; by UNIMED Belo Horizonte; by the Coordination for the Improvement of Higher Education Personnel (CAPES), Finance Code 001; by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant No. 443072/2024-8; and by Kunumi Institute. The authors thank the institution for its financial support and commitment to advancing scientific research.

References

American Association of Physicists in Medicine (2016). Low Dose CT Grand Challenge. `https://www.aapm.org/grandchallenge/lowdosect/`. Accessed: Apr. 24, 2026.

Chen, H. et al. (2017). Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535.

Chen, H. et al. (2024). Deep learning-based algorithms for low-dose ct imaging: A review. *European Journal of Radiology*, 172:111355.

Demir, A. et al. (2023). Low-dose ct image enhancement using deep learning. *arXiv preprint arXiv:2310.20265*.

dos Santos Junior, J. A., JansonNey, M. S. C., and de Souza Fonseca, G. V. (2020). Dose efetiva de radiação nos exames de tomografia computadorizada: um estudo retrospectivo e descritivo. *Diagnóstico e Tratamento*, 25(2):46–51.

Eulig, E., Ommer, B., and Kachelrieß, M. (2024). Benchmarking deep learning-based low-dose ct image denoising algorithms. *Medical physics*, 51(12):8776–8788.

Gajera, B. et al. (2021). Ct-scan denoising using a charbonnier loss generative adversarial network. *IEEE Access*, 9:84093–84109.

Gao, Q. et al. (2023). Corediff: Contextual error-modulated generalized diffusion model for low-dose ct denoising and generalization. *IEEE Transactions on Medical Imaging*, 43(2):745–759.

He, Z. et al. (2020). FAKD: Feature-affinity based knowledge distillation for efficient image super-resolution. In *IEEE ICIP*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Hu, Q.-j. et al. (2021). Prospective study of low-and standard-dose chest ct for pulmonary nodule detection: a comparison of image quality, size measurements and radiation exposure. *Current Medical Science*, 41:966–973.

Huang, J., Zhong, A., and Liu, Y. (2024). Multi-scale perceptual modulation network for low-dose computed tomography denoising. *Quantitative Imaging in Medicine and Surgery*, 14(12):9290–9305.

Jo, G. D. et al. (2023). 75% radiation dose reduction using deep learning reconstruction on low-dose chest ct. *BMC Medical Imaging*, 23(1):121.

Junda, M., Muller, H., and Friedrich-Nel, H. (2021). Local diagnostic reference levels for routine chest x-ray examinations at a public sector hospital in central south africa. *Health SA Gesondheid (Online)*, 26:1–8.

Kang, E., Min, J., and Ye, J. C. (2017). A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Medical Physics*, 44(10):e360–e375.

Korkmaz, C. and Tekalp, A. M. (2024). Training transformer models by wavelet losses improves quantitative and visual performance in single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6661–6670.

Liang, T. et al. (2020). EDCNN: Edge enhancement-based densely connected network with compound loss for low-dose CT denoising. *arXiv preprint arXiv:2011.00139*.

Romero, A. et al. (2015). Fitnets: Hints for thin deep nets.

Wang, D., Fan, F., Wu, Z., Liu, R., Wang, F., and Yu, H. (2023a). Ctformer: convolution-free token2token dilated vision transformer for low-dose ct denoising. *Physics in Medicine & Biology*, 68(6):065012.

Wang, J. et al. (2023b). A self-supervised guided knowledge distillation framework for unpaired low-dose CT image denoising. *Computerized Medical Imaging and Graphics*, 107:102237.

Wang, X. and Pang, B. (2026). Low-dose CT image denoising based on a progressive fusion distillation network with pixel attention. *Applied Sciences*, 16(7):3292.

Yang, e. a. (2025). Knowledge distillation for image restoration: Simultaneous learning from degraded and clean images. *arXiv preprint arXiv:2501.09268*.

Yang, Q. et al. (2018). Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Transactions on Medical Imaging*, 37(6):1348–1357.

Zhang, Z. et al. (2022). Mkanet: An efficient network with sobel boundary loss for land-cover classification of satellite remote sensing imagery. *Remote Sensing*, 14(18):4514.

Zhao, e. a. (2025). Distillation guided deep unfolding network with frequency hierarchical regularization for low-dose CT image denoising. *Knowledge-Based Systems*.

Zhou, Z. et al. (2023). Patient-specific uncertainty and bias quantification of non-transparent convolutional neural network model through knowledge distillation and bayesian deep learning. *Proceedings of SPIE Medical Imaging*, 12463.

Zou, B. et al. (2022). Toward efficient image denoising: A lightweight network with retargeting supervision driven knowledge distillation. In *ACCV*.