FAST RANDOMIZED SMOOTHING VIA REPRESENTATION SPACE DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Denoised Smoothing (DDS) is currently a prominent approach for obtaining certified robustness, but its prohibitively high inference cost remains a major barrier to practical deployment. Specifically, each Monte Carlo draw in the smoothing process requires a large diffusion denoiser to reconstruct an image in pixel space, and a separate pretrained model to classify the denoised image. In this paper, we propose Representation-space Denoised Smoothing (**RepDS**), a framework that replaces pixel reconstruction with denoising in the feature space of a pretrained vision encoder. Given a Gaussian-corrupted image, RepDS adapts the encoder to recover representations aligned with those of the corresponding clean image, directly supervising the features used for prediction instead of relying on pixel-space fidelity as an indirect proxy. At inference, RepDS predicts from the recovered features in a single non-generative forward pass per smoothing draw, eliminating the need for both pixel-space reconstruction and an additional encoder pass over the reconstructed image. Extensive experiments show that RepDS achieves state-of-the-art certified accuracy across all reported ImageNet radii while substantially reducing inference cost. At radius 1.5, RepDS-H+ reaches 64.2% certified accuracy compared with 54.6% for our backbone-matched DDS rerun, while running approximately $4\times$ faster under the same timing protocol. Compared with DensePure, RepDS-H+ reduces per-draw inference time by over three orders of magnitude.

1 INTRODUCTION

Certified robustness seeks a formal guarantee that a model's prediction remains unchanged within a prescribed perturbation set, rather than evidence tied to a particular attack (Raghunathan et al., 2018a; Wong & Kolter, 2018). Among methods that scale to modern image classifiers, randomized smoothing constructs a smoothed classifier by voting over Gaussian-corrupted inputs and certifies an ℓ_2 neighborhood from the resulting class probabilities (Lecuyer et al., 2019; Li et al., 2019; Cohen et al., 2019). Its classifier-agnostic formulation makes randomized smoothing broadly applicable, but also exposes a central practical tension: strong certificates require the base classifier to remain accurate under substantial Gaussian noise and, at certification time, to be evaluated over many Monte Carlo draws.

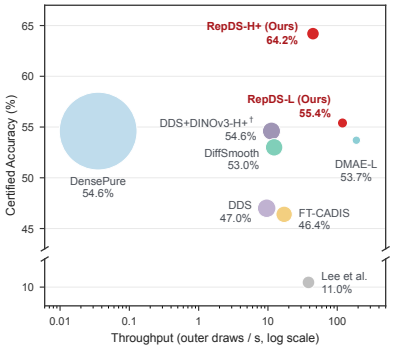


Figure 1: Certified accuracy versus inference throughput on ImageNet at $r = 1.5$ ($N = 10{,}000$). Marker area is proportional to GFLOPs.

Improving the base classifier under Gaussian noise has therefore become a central problem in randomized smoothing. Training-based methods adapt the classifier through Gaussian augmentation (Cohen et al., 2019), robustness-oriented training (Salman et al., 2019; Zhai et al., 2020; Jeong et al., 2021), and consistency regularization (Jeong & Shin, 2020). A complementary line of work, denoised smoothing (Salman et al., 2020), places a learned denoiser before a pretrained classifier. DDS (Carlini et al., 2023) substantially improved certified robustness within this framework by using an off-the-shelf diffusion model as the denoiser. Subsequent methods have used repeated reverse trajecto-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

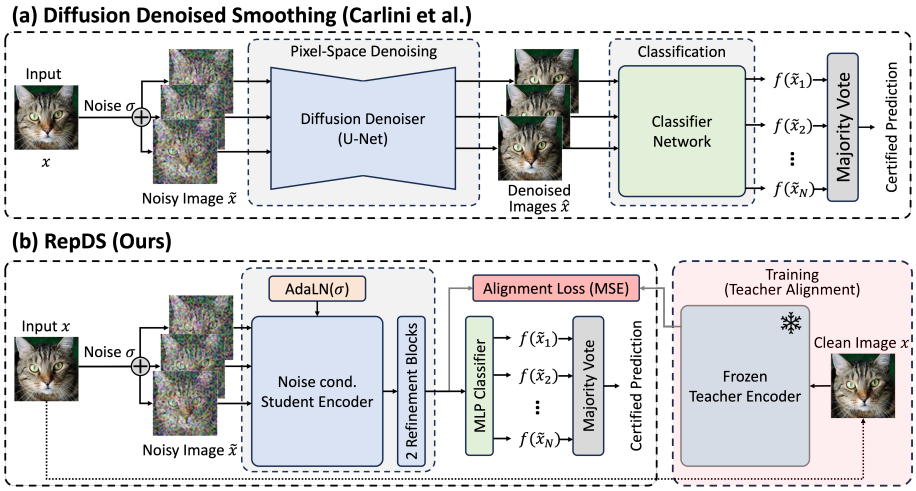


Figure 2: Two inference paths to a certified prediction from the same noisy input under randomized smoothing. (a) Diffusion Denoised Smoothing (DDS) denoises each Monte Carlo sample in pixel space, reconstructing an image that a separate classifier must re-encode. (b) RepDS maps the noisy image directly to clean-aligned patch and class tokens via a fine-tuned, noise-level-conditioned student and lightweight refinement blocks, and classifies their fused representation directly. This eliminates pixel reconstruction and re-encoding from the repeated inference path.

ries and voting (Xiao et al., 2023), local or multi-scale smoothing (Zhang et al., 2023; Jeong & Shin, 2023), and one-step consistency purification (Li et al., 2024). Although these methods differ in their exact pipelines, they retain an image-space intermediate before classification in every smoothing draw. However, these gains come at substantial inference cost. Diffusion denoisers are substantially heavier than the base classifiers of earlier randomized-smoothing pipelines, and the most accurate methods add repeated denoising trajectories with voting (Xiao et al., 2023) or auxiliary local smoothing passes (Zhang et al., 2023). Because randomized smoothing repeats this pipeline thousands of times per input, certifying a single image can require hours for computationally intensive multi-pass pipelines (Xiao et al., 2023; Chen et al., 2024), making them impractical to deploy at scale.

This pixel-space denoise-then-classify design carries two structural weaknesses. *First*, it is computationally redundant: the denoiser reconstructs an image that must then be encoded again by a separate, typically large, classifier (Lei et al., 2025), and randomized smoothing repeats this round trip for every Monte Carlo noise sample. *Second*, pixel-space fidelity is only an indirect proxy for recovering class-relevant features; reconstructions can be visually plausible yet belong to the wrong class (Carlini et al., 2023; Jang et al., 2024). These limitations motivate changing the denoising target from reconstructed pixels to the features used for prediction.

Contributions. In this paper, we propose **RepDS**, a representation-space denoised-smoothing framework that directly recovers the clean representation of a pretrained vision encoder from Gaussian-corrupted inputs. A frozen teacher provides the sample-specific clean target, while a student conditioned on the smoothing scale σ and lightweight refinement blocks learn to recover the corresponding clean representations from corrupted inputs. The recovered representation is classified directly, removing pixel reconstruction and re-encoding from the per-draw inference path. For each fixed smoothing scale, the resulting model serves as a deterministic base classifier under the standard randomized-smoothing certificate of Cohen et al. (2019).

In our experiments, we evaluate RepDS on ImageNet and CIFAR-10 under the standard randomized-smoothing protocol and examine certified zero-shot and cross-task transfer. On ImageNet, RepDS-H+ achieves state-of-the-art certified accuracy across all reported radii. At radius 1.5, it reaches 64.2% certified accuracy compared with 54.6% for our backbone-matched DDS rerun using the same DINOv3-H+/16 backbone, while reducing measured inference time from 89.0 to 22.4 seconds per 1,000 draws. It is also over three orders of magnitude faster than DensePure. On CIFAR-10, RepDS-S achieves 83.2% and 70.0% certified accuracy at radii 0.25 and 0.5, respectively, and

outperforms backbone-matched DDS across all reported radii. Beyond these benchmarks, RepDS supports certified zero-shot transfer with SigLIP2. We further show that the downstream versatility of vision foundation models (Oquab et al., 2024; Siméoni et al., 2025) extends to certified prediction: frozen ImageNet-trained RepDS encoders support certified semantic segmentation on Cityscapes and video classification on UCF101 with only task-specific head training.

2 Related Work

Certified robustness. Certified defenses provide formal guarantees that predictions remain unchanged within a prescribed perturbation set, unlike empirical defenses whose apparent robustness can fail under adaptive attacks (Athalye et al., 2018). Deterministic certification has been developed through complete verification (Katz et al., 2017; Tjeng et al., 2019), convex relaxation and bound propagation (Wong & Kolter, 2018; Zhang et al., 2018; Singh et al., 2019; Gowal et al., 2019), semidefinite programming (Raghunathan et al., 2018b), and Lipschitz-constrained architectures (Cisse et al., 2017; Tsuzuku et al., 2018; Leino et al., 2021). Complete methods provide strong guarantees but scale poorly, while tractable alternatives typically trade certificate tightness or architectural flexibility for scalability. A complementary line of certified defenses uses randomized smoothing to provide probabilistic robustness guarantees for arbitrary base classifiers (Lecuyer et al., 2019; Li et al., 2019; Cohen et al., 2019). RepDS builds on this framework.

Randomized smoothing and certification efficiency. Within randomized smoothing, prior work improves certified robustness through noise-robust training (Cohen et al., 2019; Salman et al., 2019; Zhai et al., 2020; Jeong et al., 2021) and consistency regularization (Jeong & Shin, 2020), as well as tighter estimation and adaptive smoothing schemes (Delattre et al., 2024; Sun et al., 2026). The cost of certification depends on both the number of Monte Carlo evaluations and the cost of each base-classifier evaluation. Input-specific sampling, confidence-sequence procedures, and incremental certification reduce or reuse evaluations (Chen et al., 2022; Voráček, 2024; Ugare et al., 2024), while one-step purification and latent-space classification reduce per-evaluation cost (Li et al., 2024; Lei et al., 2025). RepDS targets the latter bottleneck by eliminating pixel-space reconstruction and re-encoding from every smoothing draw.

Representation learning and denoising. Prior work learns noise-robust representations that support classification without image reconstruction at inference. DMAE (Wu et al., 2023) pretrains an encoder through pixel reconstruction from Gaussian-corrupted and masked images, then removes the decoder for certified classification. rRCM (Lei et al., 2025) learns representation consistency through pairwise instance discrimination between inputs corrupted with shared Gaussian noise at adjacent noise levels, followed by supervised fine-tuning for classification under noise.

Related generative work also incorporates representation objectives into denoising models: REPA (Yu et al., 2025) aligns denoiser features with clean representations from a pretrained encoder, while EPG (Lei et al., 2026) pretrains trajectory-consistent encoder representations before attaching a decoder for pixel-space generation. More closely related, Ditch the Denoiser (Lu et al., 2025) combines a denoised-to-noisy training curriculum with frozen-teacher supervision on self-distillation output scores. RepDS instead directly recovers the clean teacher representations used for prediction, treating the prediction representation itself as the denoising target.

3 Method

We present RepDS, a representation-space denoising framework for randomized smoothing. A noise-level-conditioned student, augmented with a lightweight refinement module, learns to recover clean representations directly from Gaussian-corrupted inputs. A frozen teacher supplies the clean targets during training. At inference, predictions are made directly from the recovered features.

Representation training proceeds in two stages: Stage 1 aligns the recovered features with clean teacher targets, and Stage 2 adds classification and prediction-consistency objectives while retaining the recovery objective (Sections 3.3 and 3.4). We then freeze the student and refinement module and train a classification head on the recovered features. For zero-shot transfer, we instead use only Stage 1 while keeping the pretrained vision–language model's zero-shot readout fixed (Section 3.3).

3.1 Background: Randomized Smoothing

Randomized smoothing (Cohen et al., 2019) constructs, from any *base classifier* $f : \mathbb{R}^d \to \mathcal{Y}$, a classifier with a provable ℓ_2 robustness guarantee. It defines the *smoothed classifier*

$$F(x) \;=\; \arg\max_{c \in \mathcal{Y}} \; \mathbb{P}_{\delta \sim \mathcal{N}(0,\sigma^2 I)}\big[f(x+\delta) = c\big], \tag{1}$$

and certifies $F(x) = c_A$ whenever $p_A = \max_c \mathbb{P}_\delta[f(x+\delta) = c] \geq \frac{1}{2}$. For an adversarially perturbed input x' around x, Lecuyer et al. (2019); Cohen et al. (2019) have shown that the smoothed prediction at x', *i.e.*, $F(x')$, is guaranteed to be the same with $F(x)$ as long as $\|x' - x\|_2$ within the following *certified radius* $R(x)$ for any $\underline{p_A} \leq p_A$, a statistical lower bound of p_A:

$$R(x) \;\geq\; \sigma\, \Phi^{-1}\big(\underline{p_A}\big), \tag{2}$$

where Φ^{-1} is the inverse standard Gaussian CDF. In practice, (1) is estimated by Monte Carlo, with a Clopper–Pearson (1934) lower bound supplying $\underline{p_A}$ (see Appendix A for details).

3.2 Noise-Level-Conditioned Feature Recovery

Pretrained vision encoders are optimized for clean images, and their representations degrade under the strong Gaussian corruption used for randomized smoothing. We therefore initialize the student E_θ from the pretrained encoder and adapt it to recover clean representations from corrupted inputs. Since the recovery process depends on the corruption level, we explicitly condition the student on the noise scale σ. Given a clean image $x \in \mathbb{R}^{H \times W \times C}$, the student receives $\tilde{x} = x + \delta$, where $\delta \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma \geq 0$.

We inject the noise level into every Transformer block through adaptive normalization of the attention and MLP branch inputs, inspired by diffusion transformers (Peebles & Xie, 2023). Zero-initialized modulation projections preserve the pretrained encoder function at initialization; details of the shared noise embedder and block-level modulation are provided in Appendix D.1.

For encoders with a dedicated class token, we append it to the patch-token sequence from the encoder's final normalized output and jointly refine the resulting $(P+1) \times D$ sequence:

$$\begin{aligned} (u, u_{\text{cls}}) &= E_\theta(\tilde{x}, \sigma), \\ [\,\hat{z};\, \hat{z}_{\text{cls}}\,] &= [\,u;\, u_{\text{cls}}\,] + \Delta_\phi\big([\,u;\, u_{\text{cls}}\,], \sigma\big). \end{aligned} \tag{3}$$

Here, $u, \hat{z} \in \mathbb{R}^{P \times D}$ and $u_{\text{cls}}, \hat{z}_{\text{cls}} \in \mathbb{R}^D$, where P is the number of patch tokens and D is the feature dimension. For encoders without a dedicated class token, we instead let $u = E_\theta(\tilde{x}, \sigma)$, refine the patch tokens as $\hat{z} = u + \Delta_\phi(u, \sigma)$, and define $\hat{z}_{\text{cls}} = \text{Norm}(\text{Pool}(\hat{z}))$, where $\text{Pool}(\cdot)$ is the encoder's native pooling/readout operation and $\text{Norm}(v) = v/\|v\|_2$. Δ_ϕ is a lightweight residual refinement module that provides additional capacity to correct noise-induced distortions in the encoder representations. Within Δ_ϕ, we project the student token sequence and add a learnable positional embedding together with a refinement-specific embedding of σ, broadcast across all tokens. A stack of L pre-norm Transformer blocks, followed by LayerNorm and an output projection, produces the additive correction in Eq. (3).

For supervised classification, predictions are produced by a classification head h_ψ, an MLP over the concatenation of the mean-pooled patch tokens and the global representation $\hat{z}_{\text{cls}}$; it is fitted after the representation is trained (Section 3.4). The student, the refinement module, and this head together form the corresponding base classifier that randomized smoothing evaluates at every draw.

3.3 Stage 1: Clean-Feature Recovery

Stage 1 trains the student and refinement module to recover the clean teacher representation from a corrupted view of the same image. For each training image x, we sample $\sigma \sim p_{\text{train}}$, where p_{train} is the training distribution of noise scales, and form the Gaussian-corrupted view $\tilde{x}$.

The frozen teacher provides the corresponding clean patch and global targets $(z^\star, z^\star_{\text{cls}})$ from its final normalized output, where $z^\star \in \mathbb{R}^{P \times D}$ and $z^\star_{\text{cls}} \in \mathbb{R}^D$. For encoders with a dedicated class token, $z^\star_{\text{cls}}$ is the teacher class token; otherwise, we define $z^\star_{\text{cls}} = \text{Norm}(\text{Pool}(z^\star))$. Because the

teacher and student receive the same underlying image and spatial view, their patch tokens remain positionally corresponding. We therefore align each recovered patch token with the teacher token at the same spatial position and minimize

$$\mathcal{L}_{\text{recov}} = \frac{1}{PD} \|\hat{z} - z^\star\|_F^2 + \lambda_{\text{cls}} \frac{1}{D} \|\hat{z}_{\text{cls}} - z^\star_{\text{cls}}\|_2^2 . \tag{4}$$

We minimize mean squared error on the patch and global representations, with λ_{cls} controlling the relative weight of the global term. For ℓ_2-normalized pooled representations, the second term is equivalent to cosine distance up to a constant factor. The teacher remains frozen, while gradients update the student and refinement parameters θ and ϕ.

3.4 Stage 2: Smoothing-Aware Representation Fine-Tuning

Clean-feature recovery alone does not directly encourage consistent class predictions across independent Gaussian draws. Motivated by consistency regularization (Jeong & Shin, 2020), we augment recovery training with classification and prediction-consistency objectives. To ensure that the classification supervision adapts the recovered representation rather than the probes themselves, we fit auxiliary probes on clean teacher features and keep them fixed throughout Stage 2. We use separate probes for the two representation paths: a_{patch} on the mean-pooled patch tokens and a_{cls} on the class token, each implemented as a single-hidden-layer MLP. For a Gaussian-corrupted input, the fixed probes produce logits from the corresponding recovered representations:

$$\begin{gathered} \ell_{\text{patch}} = a_{\text{patch}} \left(\frac{1}{P} \sum_{j=1}^{P} \hat{z}_j \right), \qquad \ell_{\text{cls}} = a_{\text{cls}} \left(\hat{z}_{\text{cls}} \right), \\ p_b = \text{softmax} \left(\ell_b \right), \qquad b \in \{\text{patch}, \text{cls}\}. \end{gathered} \tag{5}$$

For each example in Stage 2, we sample a shared noise scale $\sigma \sim p_{\text{train}}$ and $m \geq 2$ independent perturbations $\delta^{(i)} \sim \mathcal{N}(0, \sigma^2 I)$. Applying the fixed probes to each recovered view yields $\ell_b^{(i)}$ and $p_b^{(i)}$ for $b \in \{\text{patch}, \text{cls}\}$, from which we define the consistency-training loss

$$\mathcal{L}_{\text{cons}} = \sum_{b \in \{\text{patch},\text{cls}\}} \left[\lambda_{\text{CE}} \, \text{CE} \left(\frac{1}{m} \sum_{i=1}^{m} \ell_b^{(i)}, y \right) + \lambda_{\text{cons}} \frac{1}{\binom{m}{2}} \sum_{1 \leq i < j \leq m} D_{\text{SKL}} \left(p_b^{(i)}, p_b^{(j)} \right) \right], \tag{6}$$

where $D_{\text{SKL}}(p, q) = \frac{1}{2} \left[D_{\text{KL}}(p \| q) + D_{\text{KL}}(q \| p) \right]$ is the symmetric KL divergence. Classification uses the view-averaged logits, while consistency averages D_{SKL} over all pairs of noisy views. Stage 2 combines this consistency-training loss with the recovery loss on all noisy views:

$$\mathcal{L}_{\text{S2}} = \frac{2}{m} \sum_{i=1}^{m} \mathcal{L}_{\text{recov}}^{(i)} + \mathcal{L}_{\text{cons}}, \tag{7}$$

where $\mathcal{L}_{\text{recov}}^{(i)}$ is Eq. (4) evaluated on view i. Recovery anchors each noisy view to the clean representation, while classification and consistency gradients propagate through the fixed probes to all views without stop-gradient; only θ and ϕ are updated.

Classification head fitting. Once representation training is complete, we freeze the student and refinement module, discard the auxiliary probes, and fit the deployed head h_ψ directly on the recovered representation. For each example, we sample $\sigma \sim p_{\text{train}}$, draw m independent noise views, and form the feature-axis concatenation $r^{(i)} = [\, \bar{\hat{z}}^{(i)}; \hat{z}_{\text{cls}}^{(i)} \,] \in \mathbb{R}^{2D}$, where $\bar{\hat{z}}^{(i)} = P^{-1} \sum_{j=1}^{P} \hat{z}_j^{(i)}$ is the mean-pooled patch representation. The head-fitting objective is

$$\mathcal{L}_{\text{head}} = \frac{1}{m} \sum_{i=1}^{m} \text{CE}\big(h_\psi(r^{(i)}), y\big) + \lambda_{\text{out}} \frac{1}{\binom{m}{2}} \sum_{1 \leq i < j \leq m} D_{\text{SKL}}\big(p^{(i)}, p^{(j)}\big), \tag{8}$$

where $p^{(i)} = \text{softmax}(h_\psi(r^{(i)}))$. The objective is averaged over the minibatch, with label smoothing and λ_{out} specified in Appendix D.2. A single head is trained over the full noise range and used at every certification scale without receiving σ as input.

3.5 Training and Inference

Overall training. We initialize the student from the same pretrained encoder as the teacher, then train the student and refinement module with the Stage 1 objective in Eq. (4). Before Stage 2, we fit the patch and class-token probes on clean teacher features and freeze them. We then continue representation training with Eq. (7). Finally, we freeze the student and refinement module, discard the probes, and fit the deployed head h_ψ from scratch on recovered features. The teacher remains frozen throughout representation training.

Inference and certification. At certification time, we set σ to the target smoothing scale. For supervised classification, the deterministic base classifier for a noisy input $v = x + \delta$ is

$$f_\sigma(v) = \arg\max_{c \in \mathcal{Y}} \left[h_\psi \left([\, \bar{\hat{z}}(v, \sigma); \hat{z}_{\mathrm{cls}}(v, \sigma) \,] \right) \right]_c , \tag{9}$$

where $\hat{z}(v, \sigma)$ and $\hat{z}_{\mathrm{cls}}(v, \sigma)$ are the recovered patch and global representations from Eq. (3), and h_ψ outputs class logits. For zero-shot classification, we instead apply the frozen readout to $\hat{z}_{\mathrm{cls}}$, selecting the class with highest similarity to the cached text prototypes. In either case, we use the resulting deterministic classifier as $f = f_\sigma$ in the randomized-smoothing formulation of Section 3.1. Each base-classifier evaluation consists of one forward pass through the student and refinement module followed by the corresponding prediction readout. We certify the resulting smoothed classifier using the Monte Carlo procedure of Cohen et al. (2019), with details in Appendix A.

4 Experiments

We follow the standard randomized-smoothing protocol of Cohen et al. (2019) on CIFAR-10 (Krizhevsky, 2009) and ImageNet (Deng et al., 2009), evaluating at noise levels $\sigma \in \{0.25, 0.5, 1.0\}$ with $N_0 = 100$, $\alpha = 0.001$, and $N = 100{,}000$ on CIFAR-10 and $N = 10{,}000$ on ImageNet. For each reported radius, we report the highest certified accuracy across these smoothing scales. Following Cohen et al. (2019), we certify a uniformly sub-sampled subset of 500 test images per dataset and report approximate certified accuracy at ℓ_2 radii ϵ, i.e., the fraction of examples correctly classified and certified with $R \geq \epsilon$ in Eq. (2). We report GFLOPs per draw and remeasure inference time for all methods under a common protocol, using draw batch sizes of 100 on CIFAR-10 and 16 on ImageNet and SigLIP2 transfer. Timings cover the complete prediction route from a noised input to logits. Baseline certified accuracies are quoted from their original evaluations, except for our backbone-matched DDS rerun. Detailed training settings are provided in Appendix D.2.

Architecture. The teacher E_{T} and student E_θ are initialized from the same pretrained encoder: DINOv2 ViT-S/14 (Oquab et al., 2024; Dosovitskiy et al., 2021) for CIFAR-10 (RepDS-S) and DINOv3 ViT-H+/16 (Siméoni et al., 2025) as the main ImageNet model (RepDS-H+); we also evaluate DINOv3 ViT-L/16 (RepDS-L) on ImageNet and transfer tasks.

We fine-tune all Transformer blocks in the student while keeping the patch embedding frozen, use AdaLN noise conditioning as described in Section 3.2, and set the refinement depth to $L = 2$ for the main models. The Stage 2 probes use one 512-dimensional hidden layer with GELU, while the deployed head uses two hidden layers. Gaussian noise is added to clean inputs scaled to $[0, 1]$, without clipping the perturbed inputs; both σ and the certified radius are expressed in these pixel coordinates. CIFAR-10 inputs are noised at their native 32×32 resolution, then bicubically resized to 224×224 and channel-normalized inside the base classifier. For ImageNet, we use 256×256 random crops with horizontal flipping during training and center crops during evaluation, followed by noise addition and channel normalization without further resizing. We use ImageNet channel-normalization statistics for both datasets, and both configurations yield $P{=}256$ patch tokens.

4.1 Main Results

ImageNet. RepDS-H+ achieves state-of-the-art certified accuracy at every reported ImageNet radius (Table 1), exceeding the strongest non-RepDS baseline by 1.2–9.6 percentage points. The largest margin occurs at radius 1.5, where RepDS-H+ reaches 64.2%, compared with 54.6% for both DensePure and backbone-matched DDS. To control for encoder strength, we rerun DDS with

Table 1: Certified accuracy (%) on ImageNet at ℓ_2 radii ϵ using $N = 10{,}000$ smoothing samples. Superscripts denote radius-zero accuracy at the selected smoothing scale. Time is mean $\pm$ sample standard deviation over three fixed synthetic inputs, with 1,000 draws per input, on a single RTX 6000 Ada; GFLOPs are per draw. † Our backbone-matched DDS rerun uses DINOv3-H+/16 with a head trained on DDS outputs as the classifier. Best and second-best results are bold and underlined.

Method	Time (s) ↓	GFLOPs ↓	Certified Accuracy at ϵ				
			0.5	1.0	1.5	2.0	3.0
Lee (2021)	26.0 ± 0.7	325.3	41.0	24.0	11.0	–	–
DMAE-L	5.3 ± 0.0	61.6	$^{81.7}73.6$	$^{77.6}64.6$	$^{77.6}53.7$	$^{65.7}41.5$	$^{65.7}\underline{27.5}$
DDS	104.3 ± 2.1	1,481.8	$^{82.0}74.0$	$^{77.2}59.8$	$^{77.2}47.0$	$^{64.6}31.0$	$^{64.6}19.0$
DDS + DINOv3-H+†	89.0 ± 0.4	1,344.6	$^{86.4}\underline{79.8}$	$^{84.4}\underline{68.4}$	$^{84.4}54.6$	$^{72.0}\underline{42.8}$	$^{72.0}23.8$
FT-CADIS	58.4 ± 0.4	911.9	$^{81.1}71.9$	$^{77.0}60.1$	$^{66.2}46.4$	$^{66.2}39.4$	$^{66.2}22.5$
DiffSmooth	81.4 ± 0.1	1,164.1	$^{83.8}77.2$	$^{79.2}63.2$	$^{73.4}53.0$	$^{61.4}37.6$	$^{50.8}24.8$
DensePure	$28{,}178.2 \pm 92.6$	462,412.5	$^{84.0}77.8$	$^{80.2}67.0$	$^{80.2}54.6$	$^{67.8}42.2$	$^{67.8}25.8$
RepDS-L	8.4 ± 0.1	83.0	$^{85.2}77.4$	$^{80.0}67.4$	$^{80.0}\underline{55.4}$	$^{61.8}35.0$	$^{61.8}20.2$
RepDS-H+	22.4 ± 0.4	225.8	$^{86.8}\mathbf{81.0}$	$^{83.4}\mathbf{73.2}$	$^{83.4}\mathbf{64.2}$	$^{71.0}\mathbf{48.8}$	$^{71.0}\mathbf{29.6}$

Table 2: Certified accuracy (%) on CIFAR-10 at ℓ_2 radii ϵ using $N = 100{,}000$ smoothing samples. Superscript notation, efficiency reporting, and highlighting follow Table 1. † Our backbone-matched DDS rerun uses DINOv2-S/14 as the classifier.

Method	Time (s) ↓	GFLOPs ↓	Certified Accuracy at ϵ			
			0.25	0.5	0.75	1.0
DDS	2.588 ± 0.020	25.95	$^{88.0}73.8$	$^{88.0}56.2$	$^{88.0}41.6$	$^{74.2}31.0$
DDS + DINOv2-S†	1.548 ± 0.014	14.51	$^{87.0}72.6$	$^{87.0}58.6$	$^{87.0}42.8$	$^{75.4}34.4$
DMAE-B + CPT	1.758 ± 0.017	17.56	$\underline{80.8}$	67.1	49.7	37.7
rRCM-B	0.529 ± 0.017	5.69	$^{83.6}73.4$	$^{83.6}61.4$	$^{79.8}48.0$	$^{61.8}39.2$
Diffusion Calibration	2.755 ± 0.011	25.95	$^{90.2}76.4$	$^{90.2}57.2$	$^{90.2}42.6$	$^{90.2}32.4$
FT-CADIS	2.588 ± 0.020	25.95	$^{88.7}80.3$	$^{88.7}\underline{68.4}$	$^{88.7}\mathbf{54.5}$	$^{74.9}\underline{39.9}$
DiffSmooth	1.221 ± 0.001	9.65	$^{82.8}72.0$	$^{82.8}62.8$	$^{73.4}51.2$	$^{68.6}\mathbf{41.2}$
Consistency Purification + FT	4.078 ± 0.025	38.96	$^{90.2}79.4$	$^{90.2}62.4$	$^{90.2}43.8$	$^{90.2}35.4$
DensePure	442.6 ± 3.1	4,057.2	$^{87.6}76.6$	$^{87.6}64.6$	$^{87.6}50.4$	$^{73.6}37.4$
RepDS-S	0.791 ± 0.015	6.21	$^{91.6}\mathbf{83.2}$	$^{91.6}\mathbf{70.0}$	$^{91.6}\underline{52.2}$	$^{78.0}38.4$

the same DINOv3-H+/16 backbone used by RepDS-H+. Under this matched setting, RepDS-H+ remains more accurate at every reported radius while reducing inference time from 89.0 to 22.4 seconds per 1,000 draws and GFLOPs from 1,344.6 to 225.8. The smaller RepDS-L variant provides a lower-cost operating point at 83.0 GFLOPs per draw while retaining 67.4% and 55.4% certified accuracy at radii 1.0 and 1.5, respectively.

CIFAR-10. RepDS-S achieves the highest certified accuracy at radii 0.25 and 0.5, reaching 83.2% and 70.0%, respectively (Table 2), improvements of 2.4 and 1.6 percentage points over the strongest competing results. The backbone-matched comparison shows the same pattern as on ImageNet: with DINOv2-S/14 in both pipelines, RepDS-S exceeds DDS by 4.0–11.4 percentage points across all reported radii while reducing inference time from 1.548 to 0.791 seconds per 1,000 draws and GFLOPs from 14.51 to 6.21. RepDS-S is outperformed at radius 0.75 by FT-CADIS and at radius 1.0 by DiffSmooth, FT-CADIS, and rRCM-B. Taken together, the backbone-matched results suggest that directly recovering the clean representations used for prediction can serve as an effective denoising objective without requiring pixel-space reconstruction as an intermediate step.

4.2 Transfer and Generalization

Zero-shot transfer with SigLIP2. We compare RepDS and DDS for certified zero-shot transfer with SigLIP2 (Table 3) using official WebLI-pretrained checkpoints (Tschannen et al., 2025). RepDS adapts the larger SigLIP2-SO400M/16 encoder on ImageNet-1K for 200,000 updates

Table 3: Certified zero-shot accuracy (%) with SigLIP2 across 13 domains at $\sigma = 0.5$ ($N_0 = 100$, $N = 1{,}000$, $\alpha = 0.001$; 500 class-balanced images per domain). $\epsilon = 0$ denotes radius-zero certified accuracy. Time is mean $\pm$ sample standard deviation over three fixed images per domain (39 total; 1,000 draws each) on one RTX 6000 Ada; GFLOPs are per draw. Bold indicates the better pipeline.

Model	ϵ	Flowers	DTD	CIFAR100	EuroSAT	VOC07-C	SVHN	SUN397	GTSRB	Pets	Food101	Cars	Aircraft	Caltech101	Time (s) ↓ / GFLOPs ↓
DDS-B/16	0.00	71.2	52.6	63.8	**21.2**	83.2	**61.0**	54.2	**45.4**	83.6	63.6	74.8	26.6	90.8	36.20 ± 0.16 / 1,143.2
	0.25	64.4	48.6	56.4	15.8	79.0	52.4	47.6	**41.0**	79.0	55.0	68.2	20.0	89.0	
	0.50	57.8	43.6	49.0	9.0	73.6	46.0	41.0	34.8	73.8	45.8	60.2	12.4	84.8	
	0.75	50.4	38.8	44.0	2.8	67.0	39.4	33.0	30.0	68.0	36.0	49.8	7.4	81.8	
	1.00	41.8	33.8	33.8	1.0	59.4	30.6	25.8	22.4	58.4	26.6	39.8	5.6	77.8	
RepDS-SO400M/16	0.00	**81.8**	**58.6**	**74.4**	20.8	**83.6**	58.8	**62.0**	45.2	**87.0**	**72.4**	**80.0**	**42.6**	**92.2**	$\mathbf{3.67 \pm 0.05}$ / **117.5**
	0.25	**77.4**	**53.4**	**68.8**	**17.6**	**79.6**	**53.0**	**55.0**	40.6	**83.4**	**65.8**	**75.0**	**35.6**	**90.4**	
	0.50	**72.6**	**47.2**	**61.8**	**14.2**	**75.4**	**47.4**	**48.4**	**36.4**	**78.4**	**57.4**	**67.8**	**29.8**	**88.2**	
	0.75	**66.8**	**40.8**	**53.6**	**10.0**	**71.0**	**41.6**	**42.0**	**32.0**	**73.0**	**50.2**	**59.4**	**24.4**	**85.2**	
	1.00	**58.6**	**35.4**	**42.4**	**4.8**	**63.2**	**34.2**	**34.4**	**26.6**	**64.8**	**39.2**	**51.4**	**18.6**	**83.0**	

Table 4: Certified cross-task transfer at $\sigma = 0.5$ with frozen ImageNet-trained RepDS encoders and task-specific heads. (a) Cityscapes: 100 images at $R = 0.337$; abstention is treated as a zero-IoU twentieth class for mIoU. (b) UCF101: RepDS-L on 101 videos; prior results use different evaluation sets and are shown for context. † Reported by prior work. ‡GFLOPs are graph-count estimates from released architectures, counting one fused multiply–add as one operation.

(a) **Cityscapes semantic segmentation**

Method	GFLOPs ↓	Cert. Pixel Acc.	mIoU	Abstain
SegCertify + HRNet (Fischer et al., 2021)	747.5	34.0	6.0	40.0
DenoiseCertify + HRNet (Laousy et al., 2023)	1,867.4	55.0	21.0	41.0
DenoiseCertify + ViT (Laousy et al., 2023)‡	30,127.1	65.0	28.0	36.0
RepDS-H+	7,226.8	**72.5**	**33.2**	**25.9**

(b) **UCF101 video classification**

Method	GFLOPs ↓	ACR	CA@0	CA@0.5
Standard RS-8 (Joshi et al., 2022)†,‡	91.5	0.289	50.88	34.63
Smooth-Mean-(128,16) (Joshi et al., 2022)†,‡	242.9	0.352	51.59	44.17
DINOv3-L	10,543.2	0.633	57.43	56.44
RepDS-L	10,625.4	**0.659**	**59.41**	**58.42**

(25.6M image presentations), whereas DDS applies one-step denoising as in Carlini et al. (2023) with the unconditional ImageNet-256 diffusion model of Dhariwal & Nichol (2021) before a frozen SigLIP2-B/16 encoder. We use the larger encoder for RepDS to reduce the parameter-count gap between the two pipelines; despite using the larger encoder, RepDS remains substantially cheaper in per-draw computation. We use fixed CLIP Benchmark class names and dataset-specific prompt ensembles without fitting a target-domain classifier. We evaluate 500 class-balanced examples per evaluation split; VOC07-C denotes the cropped single-label object task.

Across all 13 target domains in Table 3, RepDS-SO400M/16 improves mean certified accuracy at radius 0.5 from 48.6% to 55.8% over DDS-B/16 while reducing inference time and per-draw GFLOPs by factors of 9.88 and 9.73, respectively. Because the pipelines use different SigLIP2 encoders, these results compare the end-to-end pipelines rather than isolating the effect of representation-space recovery. These results suggest that RepDS learns a reusable noise-robust representation beyond the ImageNet label space: a single model adapted on ImageNet-1K supports certified zero-shot prediction across diverse domains and label spaces without target-domain adaptation.

Transfer beyond image classification. We test whether recovered representations transfer beyond classification by freezing the ImageNet-trained RepDS encoders and fitting task-specific heads to their recovered features: RepDS-H+ for certified Cityscapes segmentation and RepDS-L for certified UCF101 video classification. Table 4 summarizes the results; protocols are provided in Ap-

Table 5: CIFAR-10 ablations at σ=0.5 under a reduced protocol (N=10,000, 100 test images), reporting radius-zero certified accuracy (CA@0, %) and average certified radius (ACR). Panel (a) ablates training components; (b) and (c) vary alignment targets and noise conditioning. Refinement-depth ablations and results at σ=0.25 are reported in Appendix Tables 6 and 7, respectively.

(a) **Training components**

Stage 1	Stage 2		Head Re-fit	CA@0	ACR
	$\mathcal{L}_{\text{recov}}$	$\mathcal{L}_{\text{cons}}$			
✓	✓	✓	✓	78.0	0.749
×	✓	✓	✓	74.0	0.677
✓	×	✓	✓	75.0	0.734
✓	✓	×	✓	80.0	0.719
✓	✓	✓	×	71.0	0.723
×	×	✓	✓	68.0	0.669

(b) **Alignment targets**

Target	CA@0	ACR
CLS	75.0	0.718
Patch	77.0	0.745
Patch + CLS	78.0	0.749

(c) **Noise conditioning**

σ input	CA@0	ACR
No explicit σ	76.0	0.735
Specialist per σ	78.0	0.676
Shared σ-cond.	78.0	0.749

pendices C.1 and C.2. On Cityscapes, RepDS-H+ achieves 72.5% certified pixel accuracy and 33.2 mIoU; Appendix Figure 3 provides a matched qualitative comparison with DDS. On UCF101, RepDS-L achieves an ACR of 0.659 and 58.42% certified accuracy at radius 0.5, compared with 0.633 and 56.44%, respectively, for DINOv3-L. These results indicate that the noise-robust representations learned on ImageNet retain utility beyond the source classification task, supporting certified prediction across distinct downstream tasks with frozen encoders.

4.3 Ablation Study

All training-component ablations in Table 5(a) use 200k optimizer updates for representation training: 150k in Stage 1 and 50k in Stage 2, or all 200k in Stage 2 when Stage 1 is removed. Ablating only a Stage 2 objective preserves the 150k/50k schedule; the final row removes recovery from both stages. Disabling consistency removes both CE and symmetric KL, leaving unit-weight recovery on a single noisy view. Without head refitting, we fit the same head architecture on clean teacher features with the same fitting budget and apply it unchanged to recovered representations.

At $\sigma = 0.5$, the full configuration achieves the highest ACR among the training-component variants: removing Stage 1 causes the largest individual drop in ACR (0.749 to 0.677), while removing Stage 2 recovery, consistency training, or head refitting reduces ACR to 0.734, 0.719, and 0.723, respectively. Removing recovery from both stages further lowers ACR to 0.669. At $\sigma = 0.5$, Patch+CLS alignment and shared σ-conditioning achieve the highest ACR among their respective variants. Appendix B further shows that four refinement blocks yield only a modest ACR gain over two (0.749 to 0.756) while increasing inference time, motivating the two-block configuration.

5 Conclusion

RepDS performs denoised smoothing in the representation space of a pretrained encoder by directly recovering clean-aligned representations from Gaussian-corrupted inputs. This removes pixel reconstruction and post-denoising re-encoding while preserving the standard randomized-smoothing certificate. On ImageNet, RepDS-H+ improves both certified accuracy and per-draw inference efficiency over backbone-matched DDS while achieving state-of-the-art certified accuracy across all reported radii. We further demonstrate certified zero-shot classification with SigLIP2 and cross-task transfer to semantic segmentation and video classification. RepDS depends on the pretrained representation, and we do not systematically study how pretraining or teacher choice affects certified robustness. Certification still requires N Monte Carlo evaluations despite the reduced per-draw cost. Extending the learned representations to broader noise regimes and compatible generative decoders (Lei et al., 2026) is an interesting direction for future work.

AI USE STATEMENT

We used generative AI assistance primarily in debugging our codebase, manuscript editing, literature search and reviewing. We have reviewed all AI-assisted work. AI-generated code was manually inspected and tested, and literature-related suggestions were checked by the authors before inclusion. We take responsibility for the final content of this work, including text, claims, code, experiments, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The experimental setup, architectures, and certification protocol are described in Section 4. Detailed training settings are provided in Appendix D.2. The training objectives are specified in Section 3, and Appendix A provides the full randomized-smoothing certification procedure.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the International Conference on Machine Learning*, pp. 274–283. PMLR, 2018.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (Certified!!) adversarial robustness for free! In *International Conference on Learning Representations*, 2023.

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wGP1tBCP1E.

Ruoxin Chen, Jie Li, Junchi Yan, Ping Li, and Bin Sheng. Input-specific robustness certification for randomized smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6295–6303, 2022.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the International Conference on Machine Learning*, pp. 854–863. PMLR, 2017.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, 2019.

Blaise Delattre, Alexandre Araujo, Quentin Barthélemy, and Alexandre Allauzen. The Lipschitz-variance-margin tradeoff for enhanced randomized smoothing. In *International Conference on Learning Representations*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In *Proceedings of the International Conference on Machine Learning*, pp. 3340–3351. PMLR, 2021.

Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4842–4851, 2019.

Suhyeok Jang, Seojin Kim, Jinwoo Shin, and Jongheon Jeong. Confidence-aware denoised fine-tuning of off-the-shelf models for certified robustness. *Transactions on Machine Learning Research*, 2024.

Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10558–10570, 2020.

Jongheon Jeong and Jinwoo Shin. Multi-scale diffusion denoised smoothing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67374–67397, 2023.

Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, and Jinwoo Shin. SmoothMix: Training confidence-calibrated smoothed classifiers for certified robustness. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30153–30168, 2021.

Ameya Joshi, Minh Pham, Minsu Cho, Leonid Boytsov, Filipe Condessa, J. Zico Kolter, and Chinmay Hegde. Smooth-Reduce: Leveraging patches for improved certified robustness. *arXiv preprint arXiv:2205.06154*, 2022.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Othmane Laousy, Alexandre Araujo, Guillaume Chassagnon, Marie-Pierre Revel, Siddharth Garg, Farshad Khorrami, and Maria Vakalopoulou. Towards better certified segmentation via diffusion models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 1185–1195. PMLR, 2023.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, pp. 656–672. IEEE Computer Society, 2019.

Kyungmin Lee. Provable defense by denoised smoothing with learned score function. In *ICLR Workshop on Security and Safety in Machine Learning Systems*, 2021. URL https://aisecure-workshop.github.io/aml-iclr2021/papers/28.pdf.

Jiachen Lei, Julius Berner, Jiongxiao Wang, Zhongzhu Chen, Chaowei Xiao, Zhongjie Ba, Kui Ren, Jun Zhu, and Anima Anandkumar. Robust representation consistency model via contrastive denoising. In *International Conference on Learning Representations*, 2025.

Jiachen Lei, Keli Liu, Julius Berner, Haiming Yu, Hongkai Zheng, Jiahong Wu, and Xiangxiang Chu. There is no VAE: End-to-end pixel-space generative modeling via self-supervised pre-training. In *International Conference on Learning Representations*, 2026.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yiquan Li, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Jiachen Lei, Bo Li, and Chaowei Xiao. Consistency purification: Effective and efficient diffusion purification towards certified robustness. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105017–105038, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Wenquan Lu, Jiaqi Zhang, Hugues Van Assel, and Randall Balestriero. Ditch the denoiser: Emergence of noise robustness in self-supervised learning from data curriculum. In *Advances in Neural Information Processing Systems*, volume 38, pp. 136057–136089, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018a.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, volume 31, 2018b.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21945–21957, 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019.

Chenhao Sun, Yuhao Mao, and Martin Vechev. Dual randomized smoothing: Beyond global noise variance. In *International Conference on Learning Representations*, 2026.

Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shubham Ugare, Tarun Suresh, Debangshu Banerjee, Gagandeep Singh, and Sasa Misailovic. Incremental randomized smoothing certification. In *International Conference on Learning Representations*, 2024.

Václav Voráček. Treatment of statistical estimation problems in randomized smoothing for adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2018.

QuanLin Wu, Hang Ye, Yuntian Gu, Huishuai Zhang, Liwei Wang, and Di He. Denoising masked autoencoders help robust classification. In *International Conference on Learning Representations*, 2023.

Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. DensePure: Understanding diffusion models for adversarial robustness. In *International Conference on Learning Representations*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. MACER: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jiawei Zhang, Zhongzhu Chen, Huan Zhang, Chaowei Xiao, and Bo Li. DiffSmooth: Certifiably robust learning via diffusion models and local smoothing. In *USENIX Security Symposium*, pp. 4787–4804, 2023.

A Randomized Smoothing Certification Procedure

For completeness, we describe the randomized smoothing certification procedure of Cohen et al. (2019) as applied to our base classifier. Given a base classifier $f : \mathbb{R}^d \to \mathcal{Y}$ (here, the RepDS base classifier described in Section 3.5), the smoothed classifier F returns the majority class under Gaussian perturbations:

$$F(x) = \arg\max_{c \in \mathcal{Y}} \ \mathbb{P}_{\delta \sim \mathcal{N}(0, \sigma^2 I)} \big[f(x + \delta) = c \big]. \tag{10}$$

Robustness guarantee. Let p_A and p_B denote the probabilities of the most likely and second most likely classes, respectively. If $p_A \geq \underline{p_A} \geq \overline{p_B} \geq p_B$ for some bounds $\underline{p_A}, \overline{p_B}$, then $F(x') = F(x)$ for all $\|x' - x\|_2 < R$, where

$$R = \frac{\sigma}{2} \left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}) \right), \tag{11}$$

and Φ^{-1} is the inverse standard Gaussian CDF.

Monte Carlo estimation. Since F cannot be evaluated exactly, Cohen et al. (2019) provide a practical algorithm that estimates p_A with high-probability confidence bounds using N Monte Carlo samples:

1. Draw N_0 samples of $\delta \sim \mathcal{N}(0, \sigma^2 I)$ and use $f(x + \delta)$ votes to identify the top candidate class $\hat{c}_A$.
2. Draw N additional independent samples and count the number k of times $f(x + \delta) = \hat{c}_A$.
3. Obtain a one-sided lower confidence bound $\underline{p_A}$ on p_A from a $\mathrm{Binomial}(N, p_A)$ tail bound at confidence level $1 - \alpha$.
4. If $\underline{p_A} > 1/2$, return $\hat{c}_A$ with certified radius $R = \sigma \Phi^{-1}(\underline{p_A})$; otherwise abstain.

This simplified radius follows from the conservative choice $\overline{p_B} = 1 - \underline{p_A}$. We use $N_0 = 100$, $N = 100{,}000$, and $\alpha = 0.001$ unless otherwise stated.

Applying the certificate. Each evaluation of $f(x+\delta)$ consists of one RepDS forward pass through the noise-level-conditioned student, the two refinement blocks, and the corresponding prediction readout.

B Additional Ablation Results

Table 6 reports the refinement-depth ablation at $\sigma = 0.5$, while Table 7 reports the corresponding ablations at $\sigma = 0.25$.

At $\sigma = 0.5$, increasing the refinement depth from two to four blocks yields only a modest ACR gain (0.749 to 0.756) while increasing inference time from 0.787 to 0.870 seconds per 1,000 draws, supporting two blocks as an accuracy–efficiency trade-off.

Table 6: Refinement-depth ablation on CIFAR-10 at $\sigma = 0.5$ under the same reduced protocol as Table 5. Time is reported in seconds per 1,000 draws.

Blocks	Time (s/1k) ↓	CA@0	ACR
0	0.701 ± 0.011	77.0	0.742
1	0.758 ± 0.011	77.0	0.740
2	0.787 ± 0.013	78.0	0.749
4	0.870 ± 0.012	79.0	0.756

At $\sigma = 0.25$, several rankings change: removing Stage 2 recovery, using the clean-feature head, patch-only alignment, and four refinement blocks slightly exceed the corresponding main configurations in ACR. However, the full configuration attains the highest CA@0.5 within each ablation group, indicating that these small ACR gains do not translate into uniformly better certification across radii. Shared σ-conditioning remains strongest among the conditioning variants, while removing recovery from both stages gives the lowest ACR among the training-component variants.

Table 7: CIFAR-10 ablations at smoothing scale $\sigma = 0.25$. All rows use the same reduced protocol as Table 5: 100 test images, $N_0 = 100$, $N = 10{,}000$, and $\alpha = 0.001$. CA@0 is radius-zero certified accuracy under $\sigma = 0.25$, not base accuracy evaluated at zero input noise. The six training configurations follow the definitions and row order in Table 5(a).

Group	Variant	CA@0	CA@0.25	CA@0.5	ACR
Training components	Full configuration	88.0	77.0	69.0	0.5498
	No Stage 1 recovery	86.0	77.0	65.0	0.5313
	No Stage 2 recovery	82.0	76.0	68.0	0.5526
	No Stage 2 consistency	88.0	78.0	63.0	0.5363
	Clean-feature head	88.0	79.0	66.0	0.5544
	No recovery loss	80.0	71.0	65.0	0.5254
Alignment target	CLS	82.0	75.0	67.0	0.5491
	Patch	86.0	78.0	68.0	0.5527
	Patch + CLS	88.0	77.0	69.0	0.5498
Noise conditioning	No explicit σ	87.0	77.0	64.0	0.5413
	Shared σ-cond.	88.0	77.0	69.0	0.5498
	Specialist per σ	90.0	75.0	62.0	0.5333
Refinement depth	0 blocks	85.0	77.0	68.0	0.5484
	1 block	87.0	76.0	68.0	0.5433
	2 blocks	88.0	77.0	69.0	0.5498
	4 blocks	88.0	78.0	67.0	0.5506

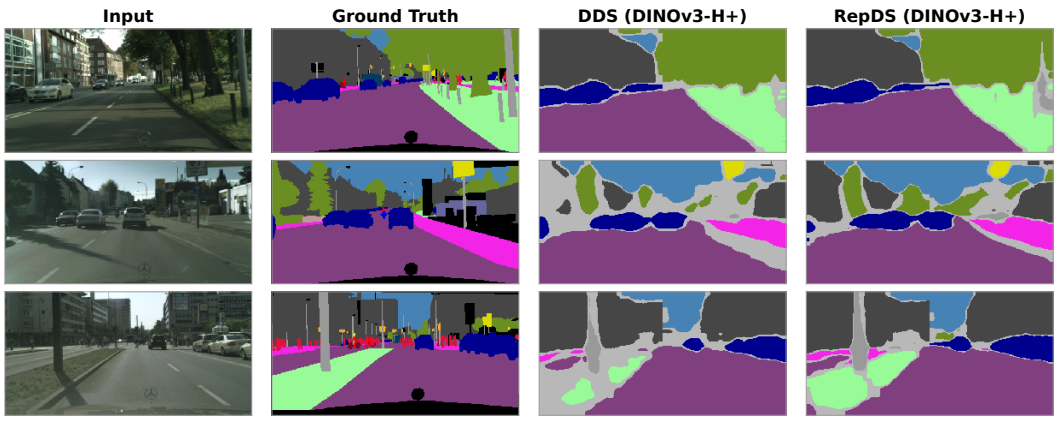


Figure 3: Certified semantic segmentation examples on Cityscapes under a matched DINOv3-H+/16 comparison. From left to right: input, ground truth, DDS, and RepDS-H+; gray denotes abstention. Both pipelines use DINOv3-H+/16 encoders with method-specific heads; DDS applies one-step pixel-space purification. Both methods are evaluated on the same images with matched noise seeds at 256×512, $\sigma = 0.50$, $N_0 = 100$, $N = 1{,}000$, $\tau = 0.75$, and $\alpha = 0.001$, with Holm correction over all 131,072 pixels. The displayed examples are three cases selected for visual readability and scene diversity. This DDS configuration is a matched pixel-space baseline rather than an exact reproduction of DenoiseCertify.

C Additional Transfer Results

C.1 Certified Segmentation Transfer

For the Cityscapes result in Table 4(a), we freeze the ImageNet-trained RepDS-H+ encoder based on DINOv3 ViT-H+/16 and fit only a linear 19-class head on the Cityscapes training set. We evaluate the 100 validation images at indices $0, 5, \ldots, 495$ of the fixed `val.lst`. For this quantitative evaluation, we use $\sigma = 0.5$, $N_0 = 10$, $N = 100$, $\tau = 0.75$, and $\alpha = 0.001$, with certified radius $R = 0.337$ and native 1024×2048 output. Holm correction is applied separately to each image over all 2,097,152 output pixels. Figure 3 provides a matched qualitative comparison under a separate higher-sample confirmation protocol.

C.2 Certified Video Classification

For UCF101, we use the ImageNet-trained RepDS-L checkpoint based on DINOv3 ViT-L/16 after 75k Stage 1 updates and 25k Stage 2 updates. We select 128 consecutive central frames from each video; for shorter videos, we cyclically repeat the frame sequence from the beginning until 128 frames are obtained. The selected frames remain fixed across smoothing draws. Each frame is preprocessed to 112×112 with pixel values in $[0, 1]$. For each draw, we add independent Gaussian noise at $\sigma = 0.5$ to the full clip tensor without clipping, then bilinearly resize each frame to 256×256 before encoding.

For each frame, we concatenate the mean-pooled recovered patch tokens with the recovered class token. We average these features over the 128 frames and apply a single linear 101-class head. The student and refinement module remain frozen, and only this head is trained on all 9,537 videos from the official UCF101 split-1 training set. Head fitting uses cross-entropy for eight epochs, with fresh noise at the same fixed scale $\sigma = 0.5$ sampled for each video in every epoch.

The DINOv3-L baseline uses the frozen pretrained encoder without RepDS adaptation or refinement, together with a separately trained noise-adapted linear head. Both models use the same clip preprocessing, head-training protocol, and certification settings.

Each smoothing draw yields one video-level class prediction, which supplies one vote to the Monte Carlo certification procedure in Appendix A. We evaluate one test video per class from UCF101 split 1 (101 videos), using $N_0 = 10$, $N = 1{,}000$, and $\alpha = 10^{-5}$. The ℓ_2 certificate applies to the

entire $128 \times 3 \times 112 \times 112$ preprocessed clip tensor; resizing to 256×256 is performed inside the base classifier after noise addition.

For the prior video methods, GFLOPs are averaged over their 283-video evaluation subset and include all chunk/subvideo model evaluations; RepDS-L uses fixed 128-frame inputs.

D Implementation Details

D.1 Noise Conditioning

A shared noise-level embedder first encodes $\log(1+\sigma)$,

$$c(\sigma) = \mathrm{MLP}_{\mathrm{noise}}\left(\mathrm{PE}\big(\log(1+\sigma)\big)\right), \tag{12}$$

where PE is a fixed 256-dimensional sinusoidal encoding. Each block ℓ maps this embedding to separate scale and shift parameters for its attention and MLP branches:

$$\left(\gamma_\ell^{\mathrm{attn}}, \beta_\ell^{\mathrm{attn}}, \gamma_\ell^{\mathrm{mlp}}, \beta_\ell^{\mathrm{mlp}}\right) = W_\ell c(\sigma) + b_\ell. \tag{13}$$

These parameters modulate the normalized branch inputs as $(1+\gamma)\odot\mathrm{LN}(\cdot)+\beta$, where LN includes the pretrained affine transformation, whose parameters are fine-tuned with the backbone. We zero-initialize W_ℓ and b_ℓ, so the conditioned encoder initially preserves the pretrained encoder function and gradually learns noise-dependent deviations during fine-tuning.

Refinement module. The refinement blocks match the hidden width and number of attention heads of the corresponding backbone: 384/6 for DINOv2-S/14, 1024/16 for DINOv3-L/16, 1280/20 for DINOv3-H+/16, and 1152/16 for SigLIP2-SO400M/16. All use an MLP expansion ratio of 2 without dropout. For SigLIP2, the output projection is zero-initialized.

D.2 Training Details

Stage 1. For the DINO-based models, Stage 1 minimizes the clean-feature recovery loss in Eq. (4) with $\sigma \sim \mathrm{Uniform}(0,\, 1.25)$ sampled per example and $\lambda_{\mathrm{cls}} = 0.03$. On CIFAR-10, we train Stage 1 for 150k updates with batch size 256 and learning rates of 1×10^{-5}, 1×10^{-4}, and 1×10^{-6} for the refinement module, student noise-conditioning parameters, and trainable backbone parameters, respectively. For RepDS-L, Stage 1 runs for 75k updates with batch size 256; for RepDS-H+, it runs for 275k updates with batch size 64. Both use learning rates of 3×10^{-4}, 3×10^{-5}, and 3×10^{-7} for the same three parameter groups. The student's shared noise embedder uses the same learning rate as the AdaLN modulation heads.

Clean-feature probe fitting. Before Stage 2, we fit both auxiliary probes on clean features from the frozen teacher using cross-entropy without label smoothing or auxiliary losses. All probes use AdamW with a fixed learning rate of 10^{-3} and weight decay 10^{-4}. For CIFAR-10, we select 10,000 fitting images and 5,000 validation images from the official training set and run probe fitting for 10 epochs with batch size 128. For ImageNet, we use the full training set for fitting and a fixed 10,000-image subset for validation, and fit the probes for three epochs with batch size 128.

Stage 2. Stage 2 optimizes Eq. (7) with $m = 2$ independent noise draws at the same sampled σ, with $\lambda_{\mathrm{CE}} = 1$, $\lambda_{\mathrm{cons}} = 5$, and unit weight on each recovery term $\mathcal{L}_{\mathrm{recov}}^{(i)}$.

We train Stage 2 for 50k updates on CIFAR-10 and 25k updates on ImageNet with effective batch size 128; RepDS-H+ uses a physical batch size of 32 with four-step gradient accumulation. All models use Stage 2 learning rates of 3×10^{-6}, 3×10^{-5}, and 3×10^{-7} for the refinement blocks, AdaLN modulation heads, and backbone blocks, respectively. Both stages use AdamW (Loshchilov & Hutter, 2019) with weight decay 10^{-4}, fixed learning rates, and no warmup or learning-rate scheduling.

Classification head fitting. After representation training, we fit the deployed head h_ψ with the student and refinement module frozen.

On CIFAR-10, the head is trained for 600 updates with pair batch size 2048 using AdamW, learning rate 10^{-3}, weight decay 10^{-4}, standard cross-entropy, and symmetric-KL consistency with $\lambda_{\text{out}} = 2$, without label smoothing, warmup, or learning-rate scheduling. On ImageNet, the head is trained for 10 epochs with pair batch size 256 for RepDS-L and 64 for RepDS-H+, using AdamW with learning rate 10^{-3}, weight decay 10^{-4}, 500 warmup steps followed by cosine decay, label smoothing 0.1, and symmetric-KL consistency with $\lambda_{\text{out}} = 1$.

SigLIP2 training. For the zero-shot model, we use only Stage 1 and optimize Eq. (4) with $p_{\text{train}} = \text{Uniform}(0.05,\ 1.05)$. The global representation is the ℓ_2-normalized output of the frozen SigLIP2 attention pooler, and we set $\lambda_{\text{cls}} = D/2 = 576$, which makes the global MSE term equivalent to unit-weight cosine distance. We fine-tune all 27 Transformer blocks and the final LayerNorm of SigLIP2-SO400M/16 together with the noise-conditioning parameters and two refinement blocks, while keeping the patch and position embeddings, attention pooler, and text encoder frozen. We train for 200,000 updates on ImageNet-1K at 256×256 resolution using the same augmentation and noise convention as the DINO models, followed by SigLIP2 normalization with mean and standard deviation 0.5. We use AdamW with weight decay 10^{-4} and batch size 128, with fixed learning rates of 10^{-5}, 10^{-4}, and 10^{-6} for the refinement module, noise-conditioning parameters, and backbone, respectively. Training uses BF16 mixed precision, gradient clipping at norm 1.0, and no warmup or learning-rate scheduling.