Adaptive Diffusion Denoised Smoothing : Certified Robustness via Randomized Smoothing with Differentially Private Guided Denoising Diffusion

Frederick Shpilevskiy [1] **Saiyue Lyu** [*] [1] [2] **Krishnamurthy Dj Dvijotham** [2] **Mathias Lécuyer** [1] **Pierre-André Noël** [2]

Abstract

We propose Adaptive Diffusion Denoised Smoothing, a method for certifying the predictions of a vision model against adversarial examples, while adapting to the input. Our key insight is to reinterpret a guided denoising diffusion model as a long sequence of adaptive Gaussian Differentially Private (GDP) mechanisms refining a pure noise sample into an image. We show that these adaptive mechanisms can be composed through a GDP privacy filter to analyze the end-to-end robustness of the guided denoising process, yielding a provable certification that extends the adaptive randomized smoothing analysis. We demonstrate that our design, under a specific guiding strategy, can improve both certified accuracy and standard accuracy on ImageNet for an ℓ_2 threat model.

1. Introduction

Rapid advances in deep learning have enabled models to filter toxic content, assist in patient triage, and steer autonomous vehicles. Despite their remarkable accuracy, these models still remain alarmingly brittle: a few carefully chosen, almost invisible pixel changes can force an image classifier to mislabel the input. Such adversarial attacks have been demonstrated against deep learning systems in medical diagnosis (Finlayson et al., 2019), autonomous driving (Eykholt et al., 2018), and AI model jailbreaks (Carlini et al., 2023a), underscoring the critical need for stringent safety and security protection.

Randomized Smoothing (RS) (Lécuyer et al., 2019; Cohen et al., 2019) provides robustness guarantees against adversarial attacks for large models by averaging predictions over noisy versions of the input at test time. To alleviate the negative impact of noise, recent work on diffusion denoised smoothing (DDS) (Carlini et al., 2023b; Xiao et al., 2023; Zhang et al., 2023; Jeong & Shin, 2023) adds one (Carlini et al., 2023b) or several (Xiao et al., 2023) step(s) of diffusion denoising into RS, empirically improving classification performance with the same robustness guarantees.

However, the Gaussian noise required for RS still induces a steep trade-off between robustness and accuracy: to this day, RS can only certify against small attacks, and doing so lowers the accuracy on legitimate instances. To empirically improve utility, previous work on adversarial purification (Wang et al., 2022; Wu et al., 2022; Bai et al., 2024) has explored injecting guidance during the diffusion denoising process, adaptively steering the denoising trajectories towards a better quality final image. However, these methods do not provide a theoretical analysis of the certification guarantees. While such adaptive techniques can be hard to rigorously analyze (Croce et al., 2022; Alfarra et al., 2022), recent work on Adaptive Randomized Smoothing (ARS) (Lyu et al., 2024) enables test-time adaptivity with rigorous guarantees using GDP composition. Still, this setup does not cover diffusion denoising models, and the authors focus on a specially designed two-step model.

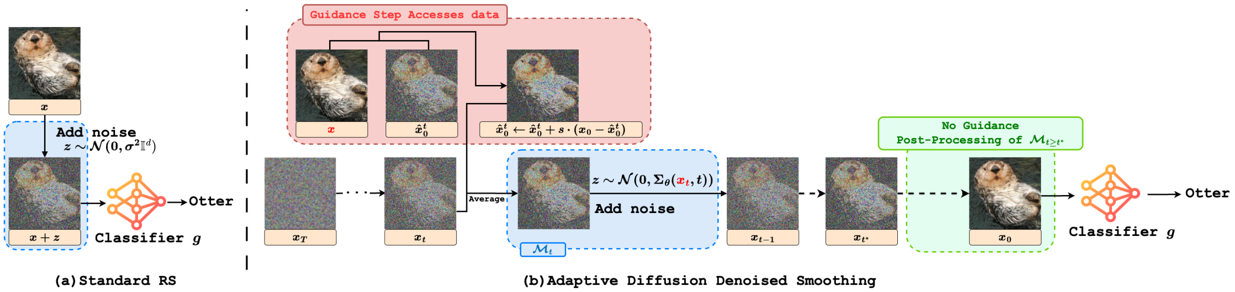
We propose **Adaptive Diffusion Denoised Smoothing (ADDS)**, a more general design for adaptive RS models with a large number of steps based on denoising diffusion models, as illustrated in Figure 1. Our key insight is to see guided denoising diffusion models as a long sequence of GDP mechanisms with data-adaptive variance: a pure noise sample is iteratively refined by injecting input-dependent guidance at each step. We show how to extend ARS with privacy filters for variance adaptive composition, providing an end-to-end certification analysis of guided diffusion denoising for RS. We evaluate our method on ImageNet.

2. Background and Related Work

Adversarial Examples (Szegedy et al., 2014) of radius r in the L_p threat model are perturbed inputs $\boldsymbol{x}+\boldsymbol{e}$, with perturbation $\boldsymbol{e} \in B_p(r)$ in the L_p ball of radius r, that make a classifier g misclassify the input $\boldsymbol{x}$, i.e., $g(\boldsymbol{x}+\boldsymbol{e}) \neq g(\boldsymbol{x})$. These attack inputs are made against classifiers at test time.

Randomized Smoothing (RS) (Lécuyer et al., 2019; Cohen et al., 2019) is a scalable approach to certify model predictions against any adversarial attacks under L_2 norm, which

[*] Work done at ServiceNow Research as an intern. [1] University of British Columbia, Vancouver, Canada [2] ServiceNow Research. Correspondence to: Frederick Shpilevskiy <fshipil@cs.ubc.ca>, Mathias Lécuyer <mathias.lecuyer@ubc.ca>.


Guidance Step Accesses data
$\boldsymbol{x}$
$\hat{\boldsymbol{x}}_0^t$
$\hat{\boldsymbol{x}}_0^t \leftarrow \hat{\boldsymbol{x}}_0^t + s \cdot (\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0^t)$
Add noise
$\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}^d)$
$\boldsymbol{x}+\boldsymbol{z}$
Classifier g
Otter
$\boldsymbol{x}_T$
$\boldsymbol{x}_t$
Average
$\boldsymbol{z} \sim \mathcal{N}(0, \Sigma_\theta(\boldsymbol{x}_t, t))$
Add noise
$\mathcal{M}_t$
$\boldsymbol{x}_{t-1}$
$\boldsymbol{x}_{t^*}$
No Guidance
Post-Processing of $\mathcal{M}_{t \geq t^*}$
$\boldsymbol{x}_0$
Classifier g
Otter

(a) Standard RS

(b) Adaptive Diffusion Denoised Smoothing

Figure 1. (a) Randomized Smoothing adds noise to the input to create a smooth model from a base model g. (b) Our Adaptive Diffusion Denoised Smoothing starts with a pure noise image $\boldsymbol{x}_T$, and guides a reverse-diffusion process at steps $T \leq t \leq t^*$ (the red box, showing one guiding step) towards reconstructing the target image $\boldsymbol{x}_0 \approx \boldsymbol{x}$, for a final prediction by g. Starting from $\boldsymbol{x}_t$, the pretrained diffusion model predicts a less noisy version of the input $\boldsymbol{x}_0^t$, which is then updated by guiding towards $\boldsymbol{x}$. The guiding step is then combined with $\boldsymbol{x}_t$, leading to a less noisy intermediary image (in the blue box). Finally, the intermediary image is re-noised to output $\boldsymbol{x}_{t-1}$. We leverage this re-noising step for our DP guarantees at step t. After timestep t^*, there is no guidance thus does not require access to data. Using GDP composition and privacy filters enable end-to-end analysis over the T steps.

randomizes a base model g by adding spherical Gaussian noise to its input, and produces a smoothed classifier $\mathcal{M}_s$ that returns the class with highest expectation over the noise: $\mathcal{M}_S(\boldsymbol{x}) \triangleq \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathbb{P}_{\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}^d)}\big(g(\boldsymbol{x}+\boldsymbol{z}) = \boldsymbol{y}\big)$. Calling $\underline{p_+}, \overline{p_-} \in [0,1]$ the lower bound on $\mathcal{M}_s$'s top class prediction and the upper bound on each other prediction, i.e. $\mathbb{P}(\mathcal{M}(\boldsymbol{x}) = \mathcal{M}_S(\boldsymbol{x})) \geq \underline{p_+} \geq \overline{p_-} \geq \max_{\boldsymbol{y} \neq \mathcal{M}_S(\boldsymbol{x})} \mathbb{P}(\mathcal{M}(\boldsymbol{x}) = \boldsymbol{y})$, and with Φ^{-1} the inverse Gaussian CDF, the certificate size is $r_{\boldsymbol{x}} = \frac{\sigma}{2}\big(\Phi^{-1}(\underline{p_+}) - \Phi^{-1}(\overline{p_-})\big)$. That is, $\forall \boldsymbol{e} \in B_2(r_{\boldsymbol{x}}), \mathcal{M}_S(\boldsymbol{x}+\boldsymbol{e}) = \mathcal{M}_S(\boldsymbol{x})$.

Adaptive Randomized Smoothing (ARS) (Lyu et al., 2024) extends RS to sequences of data-dependent steps. The method leverages Gaussian Differential Privacy (GDP) (Dong et al., 2022), an extension of (ε, δ)-Differential Privacy, to analyze the end-to-end composition of several steps in terms of GDP, with regards to L_2-norm adversarial attacks. Formally, consider k randomized Gaussian mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$ that apply to an input $\boldsymbol{x}$ and the results of all previous mechanisms, i.e., $m_i \sim \mathcal{M}_i(\boldsymbol{x} \mid m_{<i})$ for $i = 1, 2, \ldots, k$. A final classifier applies to the outputs of all steps $g(m_1, m_2, \ldots, m_k) = \boldsymbol{y} \in \mathcal{Y}$. Together, these create a composed randomized mechanism $\mathcal{M} : \boldsymbol{x} \rightarrow g(m_1, m_2, \ldots, m_k)$. Define the smoothed classifier $\mathcal{M}_S$ as $\mathcal{M}_S(\boldsymbol{x}) \triangleq \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathbb{P}(\mathcal{M}(\boldsymbol{x}) = \boldsymbol{y})$. According to Lyu et al. (2024, Theorem 2.3), if for all $r \geq 0$ each $\mathcal{M}_i$ is $\frac{r}{\sigma_i}$-GDP in a $B_2(r)$-neighbourhood, then the smoothed classifier is robust to all perturbations $\boldsymbol{z} \in B_2(r_{\boldsymbol{x}})$, such that $\mathcal{M}_S(\boldsymbol{x}) = \mathcal{M}_S(\boldsymbol{x}+\boldsymbol{z})$, with

$$r_{\boldsymbol{x}} = \frac{1}{2\sqrt{\sum_{i=1}^{k} \frac{1}{\sigma_i^2}}} \left(\Phi^{-1}(\underline{p_+}) - \Phi^{-1}(\overline{p_-})\right). \tag{1}$$

Notice that in ARS (Equation (1)) each step $\mathcal{M}_i$ is r/σ_i-GDP for an attack of size r, with Gaussian noise variance σ_i^2 fixed in advance. In practice, one might want to adapt the variance based on the results of previous steps.

Privacy Filters (Rogers et al., 2016) support the composition of privacy mechanisms with adaptive (dependent on previous steps) privacy guarantees, as long as their composition always remains below an upper-bound fixed in advance. Specifically, GDP privacy filters (Smith & Thakurta, 2022; Koskela et al., 2022) state that, for a parameter μ fixed in advance, any composition of μ_i-GDP mechanisms (where μ_i can be chosen based on the results of previous mechanisms) such that $\sum_i \mu_i^2 \leq \mu^2$ is μ-GDP.

Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020; Song et al., 2021) is a type of generative model that learns to denoise a noisy sample. The forward process of DDPM constructs a Markov chain $\{\boldsymbol{x}_0, \cdots, \boldsymbol{x}_T\}$ from a clean image to pure noise. At each time step, it adds noise to the previous state $\boldsymbol{x}_{t-1}$ via $\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \mathcal{N}(0, \beta_t \mathbf{I})$, where $0 < \beta_1 < \beta_2 < \cdots < \beta_T < 1$ control the variances of the diffusion process. Denoting $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s = \prod_{s=1}^{t}(1-\beta_t)$, we can obtain $q(\boldsymbol{x}_t|\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}, (1-\bar{\alpha}_t)\mathbf{I}\big)$. At each time step of the reverse process, Ho et al. (2020, Equation 11) outputs a reverse Markov chain $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\big(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\big)$, where $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\big(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\big)$ and $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is a learned predictor. Following Nichol & Dhariwal (2021), models predict a diagonal covariance matrix $\Sigma_\theta(\boldsymbol{x}_t, t)$ that depends on $\boldsymbol{x}_t$, and where the noise added to the image can differ by pixel. One denoising step follows:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) + \mathcal{N}(0, \Sigma_\theta(\boldsymbol{x}_t, t)). \tag{2}$$

Diffusion Denoised Smoothing (DDS) (Carlini et al., 2023b) utilizes off-the-shelf high-fidelity diffusion models (Dhariwal & Nichol, 2021) as powerful denoisers with no extra training, to remove added smoothing Gaussian noise via a one-shot denoising step. The algorithm first matches the RS perturbed data point $\boldsymbol{x}_{\text{rs}} = \boldsymbol{x} + \mathcal{N}(0, \sigma^2 \mathbf{I})$ with the noised data point from DDPM forward process $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x} + \mathcal{N}(0, (1-\bar{\alpha}_t)\mathbf{I})$ to output a unique time step t^* that satisfies the match. Next it embeds a RS sample on the DDPM trajectory by $\boldsymbol{x}_{t^*} = \sqrt{\bar{\alpha}_{t^*}} \cdot \boldsymbol{x}_{\text{rs}}$. Then the one-shot denoising step outputs $\hat{\boldsymbol{x}}_0^{t^*} := \frac{1}{\sqrt{\bar{\alpha}_{t^*}}}\big(\boldsymbol{x}_{t^*} - (\sqrt{1-\bar{\alpha}_{t^*}})\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_{t^*}, t^*)\big)$ to feed it to the classifier. DensePure (Xiao et al., 2023) achieves better robustness results by repeating a multi-hop denoising process multiple times and taking a majority vote over purified outputs as the final prediction.

3. Adaptive Diffusion Denoised Smoothing (ADDS)

Current DDS certificates do not support adaptive choices during denoising, which leaves potential robustness untapped. To close this gap we propose ADDS, an adaptive version of DDS which demonstrates that privacy-inspired adaptive analysis can tighten certificates for diffusion based defences. In what follows, we describe a GDP filter (Algorithm 1) based guided DDPM sampling algorithm in (Algorithm 2). We first analyze the individual pixel sensitivity for one denoising step (Proposition 3.1), and use it to compose the analysis over many steps (Proposition 3.2). This yields an end-to-end robustness analysis (Theorem 3.3).

As noted in Equation (12) of Song et al. (2021), one can generate a sample $\boldsymbol{x}_{t-1}$ from a sample $\boldsymbol{x}_t$ by utilizing the predicted original image $\hat{\boldsymbol{x}}_0^t$. Specifically, we substitute $\hat{\boldsymbol{x}}_0^t := \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\boldsymbol{x}_t - (\sqrt{1-\bar{\alpha}_t})\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\big)$ and rearrange Equation (2) to obtain an expression for the denoising step in terms of the state $\boldsymbol{x}_t$ and prediction $\hat{\boldsymbol{x}}_0^t$ only,

$$\begin{aligned}
\boldsymbol{x}_{t-1} &= \tfrac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \tfrac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) + \mathcal{N}(0, \Sigma_\theta(\boldsymbol{x}_t, t)) \\
&= \tfrac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \tfrac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\big(\tfrac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t - \tfrac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{x}}_0^t\big)\right) \\
&\quad + \mathcal{N}(0, \Sigma_\theta(\boldsymbol{x}_t, t)) \\
&= \tfrac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} \cdot \hat{\boldsymbol{x}}_0^t + \tfrac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} \cdot \boldsymbol{x}_t \\
&\quad + \mathcal{N}(0, \Sigma_\theta(\boldsymbol{x}_t, t)) \qquad (3)
\end{aligned}$$

Inspired by the Backward Universal Guidance of Bansal et al. (2023), we introduce guidance in Equation (3) via shifting $\hat{\boldsymbol{x}}_0^t$ towards $\boldsymbol{x}$ with scale s, yielding $\hat{\boldsymbol{x}}_0^t \leftarrow \hat{\boldsymbol{x}}_0^t + s \cdot (\boldsymbol{x} - \hat{\boldsymbol{x}}_0^t)$. We can thus view the guided $\hat{\boldsymbol{x}}_0^t$ as a convex combination $(1-s)\hat{\boldsymbol{x}}_0^t + s\boldsymbol{x}$, which becomes standard DDPM sampling when $s = 0$.

Using definitions above, we can formulate the guided denoising process as a sequence of GDP mechanisms $\mathcal{M}_t : \boldsymbol{x} \to \mathcal{A}_t(\boldsymbol{x}) + \boldsymbol{z}, \boldsymbol{z} \sim \mathcal{N}(0, \Sigma_\theta(\boldsymbol{x}_t, t))$ where $\mathcal{A}_t(\boldsymbol{x}) = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\big((1-s) \cdot \hat{\boldsymbol{x}}_0^t + s \cdot \boldsymbol{x}\big) + \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} \cdot \boldsymbol{x}_t$. Note that Σ_θ is a diagonal covariance matrix, such that each pixel of $\boldsymbol{x}$ gets an independent draw with potentially different variance (Nichol & Dhariwal, 2021). We now perform our certified guarantee analysis first on one pixel and then derive the overall guarantee for the whole image based on the pixel-wise results.

Algorithm 1 PrivacyFilter

```
Input: Per pixel budget $\Lambda, s, t$.
  $\Lambda' \leftarrow \Lambda - s^2 \cdot \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2 \cdot \sigma_t^2}$
  Return: $\Lambda$, no if $\Lambda' \leq 0$
  Return: $\Lambda'$, ok otherwise
```

Algorithm 2 Clean Image Guided Denoising

```
Input: original image $\boldsymbol{x}$, guidance scale $s$, total RS variance $\sigma^2$
  Initialize $\boldsymbol{x}_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
  Initialize $\mu = 1/\sigma, \Lambda \leftarrow \mu\mathbb{1}^d$        ▷ Vector of all $\mu$
  for $t$ from $T$ to $1$ do
    $\hat{\boldsymbol{x}}_0^t \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\boldsymbol{x}_t - (\sqrt{1-\bar{\alpha}_t})\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\big)$
    $\Lambda$, filter $\leftarrow$ PrivacyFilter($\Lambda, s, t$)        ▷ Alg.1 filter
    if filter == ok then
      $\hat{\boldsymbol{x}}_0^t \leftarrow \hat{\boldsymbol{x}}_0^t + s \cdot (\boldsymbol{x} - \hat{\boldsymbol{x}}_0^t)$
    end if
    $\boldsymbol{x}_{t-1} \sim \mathcal{N}(\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} \cdot \hat{\boldsymbol{x}}_0^t + \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} \cdot \boldsymbol{x}_t, \Sigma_\theta(\boldsymbol{x}_t, t))$
  end for
Output: $\boldsymbol{x}_0$
```

Consider a fixed pixel i ($i \in \{1, d\}$) of $\boldsymbol{x}$, with an adversarial change of size r_i. Denote $\Sigma_\theta(\boldsymbol{x}_t, t) = \text{diag}(\sigma_{t,1}^2, \cdots, \sigma_{t,i}^2, \cdots, \sigma_{t,d}^2)$. At step t, the denoising process at pixel i is a GDP mechanism $\mathcal{M}_{t,i} : \boldsymbol{x} \to \mathcal{A}_{t,i}(\boldsymbol{x}) + \boldsymbol{z}, \boldsymbol{z} \sim \mathcal{N}(0, \sigma_{t,i}^2)$, with sensitivity:

$$\begin{aligned}
\Delta\mathcal{A}_{t,i}(\boldsymbol{x}) &= \max_{e_i \in B_2(r_i)} \|\mathcal{A}_{t,i}(\boldsymbol{x} + e_i) - \mathcal{A}_{t,i}(\boldsymbol{x})\| \\
&\leq r_i \cdot s \cdot \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t},
\end{aligned}$$

where e_j is a vector of zeros except for pixel i. This directly yields (Dong et al., 2022):

Proposition 3.1 (One step denoising budget)**.**

$$\mu_{t,i}^2 = \frac{\Delta\mathcal{A}_{t,i}^2}{\sigma_{t,i}^2} = \frac{r_i^2}{\sigma_{t,i}^2} \cdot s^2 \cdot \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2}.$$

Proposition 3.2 (End-to-end Pixel GDP)**.** *Consider as neighbours any two inputs differing in pixel i, $\boldsymbol{x}$ and $\boldsymbol{x} + e_i$, by a size at most r_i ($e_i \in B_2(r_i)$). Under this neighbouring definition, Algorithm 2 is $\frac{r_i}{\sigma}$-GDP.*

Proof. Algorithm 1 ensures that over any run of Algorithm 2 we have $\sum_t \frac{1}{\sigma_i^2} \cdot s^2 \cdot \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \leq \mu^2$. This in turns implies that $\sum_t \mu_{t,i}^2 \leq r_i^2\mu^2 = \left(\frac{r_i}{\sigma}\right)^2$. This is a valid GDP filter (Smith & Thakurta, 2022; Koskela et al., 2022), ensuring that Algorithm 2 is $\frac{r_i}{\sigma}$-GDP. □

Table 1. Certified accuracy (at $r = 0$) for ImageNet

σ	1.0	1.5	2.0
Carlini et al.	**62.0**	38.4	26.8
DensePure	57.6	40.0	25.6
DensePure w/ 5 votes	61.6	45.6	31.2
ADDS (Ours)	58.8	40.8	27.6
ADDS w/ 5 votes	60.4	**46.8**	**32.0**
ADDS (w/o unguided denoising)	61.2	44.8	31.2

Table 2. Clean accuracy for ImageNet

σ	1.0	1.5	2.0
Carlini et al.	69.6	55.2	46.8
DensePure	68.4	58.0	46.4
DensePure w/ 5 votes	68.4	55.2	45.2
ADDS (Ours)	68.8	58.0	47.6
ADDS w/ 5 votes	68.8	57.2	46.8
ADDS (w/o unguided denoising)	**70.0**	**60.0**	**48.0**

Based on the pixel wise results we can obtain the certified guarantee for the whole image:

Theorem 3.3 (Adaptive Diffusion Denoised Smoothing)**.** *Consider the guided DDPM denoising process from Algorithm 2, with total Randomized Smoothing variance σ^2, coupled with a predictive model $g(\cdot)$. Consider the associated smoothed model $M_S : \boldsymbol{x} \to \arg\max_{y \in \mathcal{Y}} \mathbb{P}_{\boldsymbol{x}_0 \sim Alg.2}(g(\boldsymbol{x}_0) = y)$.*

Let $y_+ \triangleq M_S(\boldsymbol{x})$ be the prediction on input $\boldsymbol{x}$, and let $\underline{p_+}, \overline{p_-} \in [0, 1]$ be such that $\mathbb{P}(g(\boldsymbol{x}_0) = y_+) \geq \underline{p_+} \geq \overline{p_-} \geq \max_{y_- \neq y_+} \mathbb{P}(g(\boldsymbol{x}_0) = y_-)$.

Then $\forall e \in B_2(r_x)$, $M_S(\boldsymbol{x} + e) = M_S(\boldsymbol{x})$, with:

$$r_{\boldsymbol{x}} = \frac{\sigma}{2} \left(\Phi^{-1}(\underline{p_+}) - \Phi^{-1}(\overline{p_-}) \right).$$

Proof. Consider any adversarial change $e \in B_2(r_x)$. Denoting r_i the change at each pixel i, we have that $\sum_i r_i^2 = r_x^2$. By Proposition 3.2, we also know that for any i, considering only the pixel i mechanism is $\frac{r_i}{\sigma}$-GDP. By concurrent composition (Haney et al., 2023) of the GDP mechanisms over each pixel, Algorithm 2 is $\frac{\sqrt{\sum_i r_i^2}}{\sigma} = \frac{r_x}{\sigma}$-GDP. Applying Corollary 2.2 of Lyu et al. (2024) concludes the proof. □

4. Experiments

We evaluate certified ℓ_2 robustness on ImageNet (Deng et al., 2009). We follow Carlini et al. (2023b), using the unconditional 256 × 256 guided diffusion model of Dhariwal & Nichol (2021) and a pre-trained BEiT large model (Bao et al., 2022) as classifier (88.6% top-1 validation accuracy). We use three noise levels $\sigma \in \{1.0, 1.5, 2.0\}$ and randomly select 250 samples (each of a different class) from the ImageNet validation set for certification. We denoise sequentially over 20 evenly-spaced timesteps from the original 1000 denoising steps (i.e. 999, 949, 899, . . .). We compare the certified accuracy at $r = 0$ and the clean accuracy of several methods: ADDS with 1 and 5 votes, Carlini et al. (2023b), and DensePure (Xiao et al., 2023) with 1 and 5 votes. We also evaluate ADDS without doing unguided denoising past t^*, where we perform one-shot sampling of the final image (like in Carlini et al. (2023b)) once out of budget ($\Lambda = 0$ in Algorithm 2).

Table 1 shows the certified accuracy at $r = 0$, and Table 2, the clean accuracy. ADDS without unguided denoising performs best in clean accuracy. In certified accuracy, Carlini et al. (2023b) performs best at $\sigma = 1.0$, and ADDS with 5 votes is best at larger noise. ADDS without unguided denoising is competitive across all noise levels. We make three important observations that explain these results.

First, unguided denoising (in DensePure and ADDS after the budget is exhausted) increases variance. Indeed, conditioned on the robustness noise draws (i.e., DensePure's diffusion noising process or the noise during ADDS guiding), additional denoising steps introduce more noise, and hence variance. We can see on Table 2 that this variance systematically degrades clean accuracy. As a result, Carlini et al. (2023b) outperforms in low noise ($\sigma = 1$) and ADDS without unguided denoising is always best.

Second, and perhaps surprisingly, going from 1 vote to 5 votes, which alleviates the variance increase, degrades the clean accuracy of both DensePure and ADDS, while increasing certified accuracy. This is because voting concentrates predictions on the top class, conditioned on robustness noise draws. On "easy" images that are often well classified, this increases the top (correct) probability, thereby improving certified accuracy. However, on harder images where several labels typically have high probability, majority voting can bias predictions toward the incorrect label, reducing pure accuracy (see details in Appendix B and Figure 5).

Third, compared to Carlini et al. (2023b), ADDS makes a trade-off between original image fidelity and noise: each guided denoising step mixes the predicted image $\hat{\boldsymbol{x}}_0^t$ with $\boldsymbol{x}$ such that the signal component of $\boldsymbol{x}_t$ is not composed entirely of the original image, but also of predictions from the denoising model. As a result, when ADDS runs out of budget, it reaches a state $\boldsymbol{x}_{t^*}$ with less noise that the starting point of Carlini et al. (2023b) (which uses all of the budget to sample $\boldsymbol{x}_t$ from $q(\boldsymbol{x}_t|\boldsymbol{x})$). This translates to an increased level of detail in generated images, in exchange for the potential of deviating from the original (e.g. a blurry dog might become a detailed cat). The tradeoff is more pronounced at higher noise levels ($\sigma = 1.5, 2.0$), at which ADDS without unguided denoising also outperforms Carlini et al. (2023b) in robust accuracy. At lower noise ($\sigma = 1.0$) the noise scale sampled by Carlini et al. (2023b) is already small enough for good one-shot denoising in most cases (see details in Appendix B and Figures 6 to 8).

Acknowledgements

FS was partially supported by a UBC Four Year Fellowship and SL was partially supported by ServiceNow and Mitacs. We are grateful for the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [reference number RGPIN-2022-04469]. This research was enabled by computational support provided by the Digital Research Alliance of Canada (alliancecan.ca), by the University of British Columbia's Advanced Research Computing (UBC ARC), and by ServiceNow.

References

Alfarra, M., Bibi, A., Torr, P. H., and Ghanem, B. Data dependent randomized smoothing. In *Uncertainty in Artificial Intelligence*, pp. 64–74. PMLR, 2022.

Bai, M., Huang, W., Li, T., Wang, A., Gao, J., Caiafa, C. F., and Zhao, Q. Diffusion models demand contrastive guidance for adversarial purification to advance. In *International Conference on Machine Learning*. PMLR, 2024.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Koh, P. W. W., Ippolito, D., Tramer, F., and Schmidt, L. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36:61478–61500, 2023a.

Carlini, N., Tramèr, F., Dvijotham, K., Rice, L., Sun, M., and Kolter, Z. (certified!!) adversarial robustness for free! In *International Conference on Learning Representations*, 2023b.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Croce, F., Gowal, S., Brunner, T., Shelhamer, E., Hein, M., and Cemgil, T. Evaluating the adversarial robustness of adaptive test-time defenses. In *International Conference on Machine Learning*, pp. 4421–4435. PMLR, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2022.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

Haney, S., Shoemate, M., Tian, G., Vadhan, S., Vyrros, A., Xu, V., and Zhang, W. Concurrent composition for interactive differential privacy with adaptive privacy-loss parameters. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1949–1963, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jeong, J. and Shin, J. Multi-scale diffusion denoised smoothing. *Advances in Neural Information Processing Systems*, 36:67374–67397, 2023.

Koskela, A., Tobaben, M., and Honkela, A. Individual privacy accounting with gaussian differential privacy. *arXiv preprint arXiv:2209.15596*, 2022.

Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pp. 656–672. IEEE, 2019.

Lyu, S., Shaikh, S., Shpilevskiy, F., Shelhamer, E., and Lécuyer, M. Adaptive randomized smoothing: Certified adversarial robustness for multi-step defences. *Advances in Neural Information Processing Systems*, 37:134043–134074, 2024.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Rogers, R. M., Roth, A., Ullman, J., and Vadhan, S. Privacy odometers and filters: Pay-as-you-go composition. *Advances in Neural Information Processing Systems*, 29, 2016.

Smith, A. and Thakurta, A. Fully adaptive composition for gaussian differential privacy. *arXiv preprint arXiv:2210.17520*, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Wang, J., Lyu, Z., Lin, D., Dai, B., and Fu, H. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.

Wu, Q., Ye, H., and Gu, Y. Guided diffusion model for adversarial purification from random noise. *arXiv preprint arXiv:2206.10875*, 2022.

Xiao, C., Chen, Z., Jin, K., Wang, J., Nie, W., Liu, M., Anandkumar, A., Li, B., and Song, D. Densepure: Understanding diffusion models for adversarial robustness. In *International Conference on Learning Representations*, 2023.

Zhang, J., Chen, Z., Zhang, H., Xiao, C., and Li, B. {DiffSmooth}: Certifiably robust learning via diffusion models and local smoothing. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 4787–4804, 2023.

A. Extended Related Work and Background

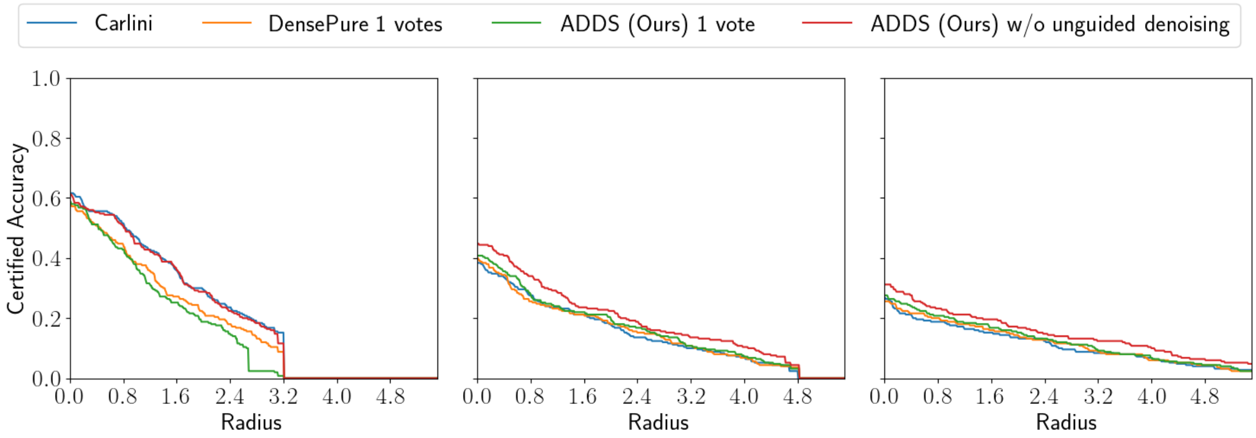
Gaussian Differential Privacy (GDP or f-DP) (Dong et al., 2022) is an extension of (ε, δ)-Differential Privacy that defines privacy according to the power of any hypothesis test to differentiate a Gaussian-distributed output from any of its neighbours. Consider a Gaussian mechanism $\mathcal{M} : \boldsymbol{x} \rightarrow \mathcal{A}(\boldsymbol{x}) + \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(0, \frac{r^2}{\mu^2}\mathbf{I})$ and $\mathcal{A}$ is some model. According to Dong et al. (2022, Theorem 2.7), for any neighbouring inputs $\boldsymbol{x}, \boldsymbol{x'}$ such that $\|\mathcal{A}(\boldsymbol{x}) - \mathcal{A}(\boldsymbol{x'})\|_2^2 \leq r$, $\mathcal{M}$ is μ-GDP.

DDS Variations Concurrently to Xiao et al. (2023), DiffSmooth (Zhang et al., 2023) proposes a one-shot denoising procedure that locally smooths the output by taking a majority vote of the predictions for each purified sample over multiple Gaussian noise samples. This process relies on much stronger assumptions than the RS approach we, Xiao et al. (2023), and Lyu et al. (2024) build on, which we believe is risky in the adversarial setting we consider.

Adding Guidance during Denoising. Recent work (Dhariwal & Nichol, 2021; Bansal et al., 2023) has explored different techniques of injecting guidance into diffusion models to steer the denoising trajectories towards task-specific semantics, such that controllable high-fidelity outputs are sampled without retraining the underlying generator. To further improve robustness, various methods (Wang et al., 2022; Wu et al., 2022; Bai et al., 2024) has utilized guidance with diffusion in adversarial purification. Although obtaining empirical accuracy improvement, these methods do not provide provable certification. Inspired by (Bansal et al., 2023), our adaptive pipeline takes advantage of the guidance during denoising, while the certification analysis stays sound with the benefit from ARS.

B. Extended Experiments

Implementation Details. We did all our experiments on a 250 image subset of the 50,000 image ImageNet (Deng et al., 2009) validation set, sampling every 200th image. We computed the top class lower bound over 1000 samples and the certified radius of 10,000 samples. For ADDS (every version), we chose guiding scales of 0.8 at $\sigma = 1.0, 1.5$ and 0.9 at $\sigma = 2.0$ and fixed 20 denoising steps (i.e. 999, 949, 899, $\ldots$). We tuned these parameters in the $\sigma = 1.0, 1.5$ settings on ADDS (1 vote). If we are unable to do a guiding step at the guiding scale we chose, we compute the largest possible guiding scale such that all of the budget is expended. For Carlini et al. (2023b), we found that noising the image by σ performed better than by the noise scale corresponding to the timestep that is at least as noisy as σ (i.e. for timestep t, the noise scale would be $\sqrt{1 - \bar{\alpha}_t} \geq \sigma$) – DensePure also noises the image by σ rather than $\sqrt{1 - \bar{\alpha}_t}$.



a) $\sigma = 1.0$ b) $\sigma = 1.5$ c) $\sigma = 2.0$

Figure 2. **Certified Test Accuracy of 1 Vote Methods on ImageNet.** Certified accuracy of ADDS with 1 vote is competitive with Xiao et al. (2023) in (a) and with Carlini et al. (2023b) in (b) and (c). Although at radius 0, we showed in Table 1 that we improve on Carlini et al. (2023b) by two percentage points, (b) and (c) show that for larger radius, the gap tightens. ADDS without unguided de-noising is competitive with Carlini et al. (2023b) in (a) and performs better than both in (b) and (c).

In the non-voting setting, Figure 2 shows that ADDS without unguided denoising performs the best at $\sigma = 1.5, 2.0$ at both radius 0 and larger radii. At $\sigma = 1.0$, Carlini et al. (2023b) performs the best and is matched by ADDS without unguided denoising. The superior performance of these methods at $\sigma = 1.0$ is because they do not do unguided denoising. At the timestep in the denoising process where the noise has scale $\sigma = 1.0$, the predictions of the denoiser are clear and unambiguous. As a result, the classifier is able to accurately predict on the denoised images. This is reflected by the high pure accuracy of Carlini et al. (2023b) and ADDS without unguided denoising in Table 2. And since these methods do not do unguided denoising – which introduces an additional source of variance – they have less variance in classifications than DensePure and ADDS. As a result, they are able to certify more correctly predicted images, and consequently obtain the highest certified accuracies. The reason why Carlini et al. (2023b) performs better than ADDS without unguided denoising is because ADDS still introduces some additional variance because the guiding scale $s < 1$.

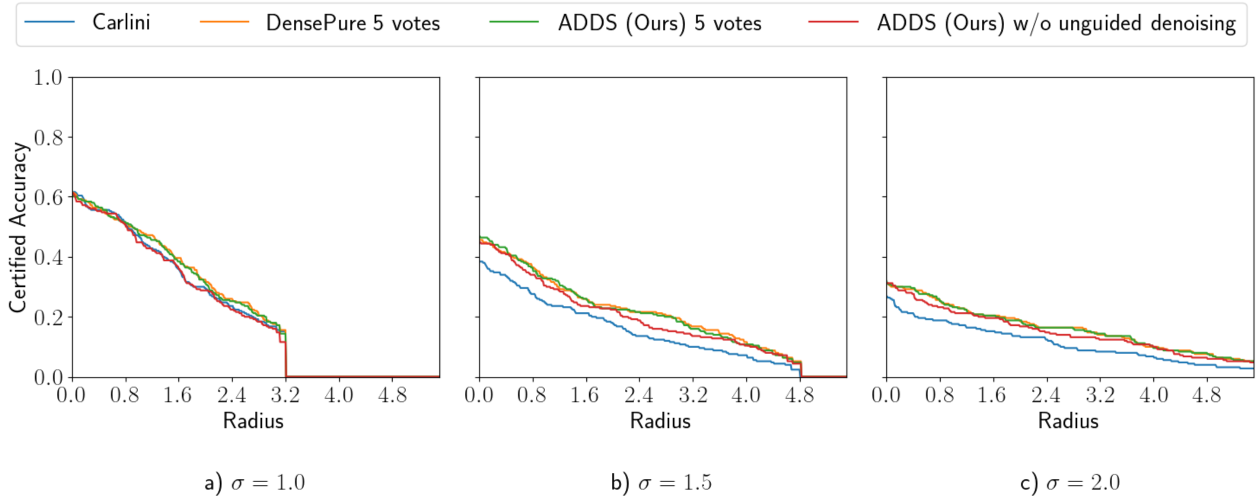


Figure 3. **Certified Test Accuracy of 5 Vote Methods on ImageNet.** Certified accuracy of ADDS with 5 vote is competitive with Xiao et al. (2023) with 5 votes at every σ. ADDS without unguided de-noising is outperformed by Xiao et al. (2023) and ADDS with 5 votes, but the gap tightens as σ increases.

In Figure 3, we see that majority voting over 5 unguided trajectories increases the certified accuracy of DensePure and ADDS. Both DensePure and ADDS with 5 votes match and slightly exceed the certified accuracy of Carlini et al. (2023b) and ADDS without unguided denoising. In Figure 5, we see the effect of majority voting on a selection of images where voting causes a change in classification or certification. For example, at $\sigma = 1.5$, image 12400 changes from abstained to certified and image 23200 changes from misclassified to correctly classified (although still abstained). It should be noted that Figure 5 does not reflect that the vast majority of these changes are abstains that become certified. Voting concentrates counts on the most frequent classes, conditioned on robustness noise draws. The count of each highest frequency class increases – shown as the class becoming more yellow. By further concentrating classification towards the highest frequency classes, voting is able to increase the certified radius – reflected in Table 1 by an improvement in certified accuracy. We can see that all misclassifications as a result of voting occur in abstains (e.g. images 18200, 46000, 47000), where multiple classes have high counts. In these cases, certified accuracy stays the same, but pure accuracy reduces. Figure 4 shows that voting comes with the drawback of being more computationally expensive.

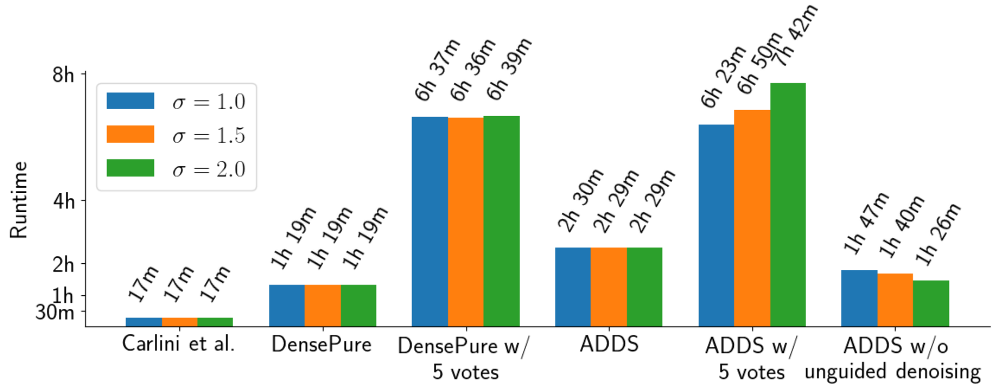


Figure 4. **Runtime Plot.** The time it takes each algorithm from Table 1 and Table 2 to certify one image over 10,000 certification samples. We run each algorithm on one A100-40GB GPU on the Narval Compute Canada cluster.

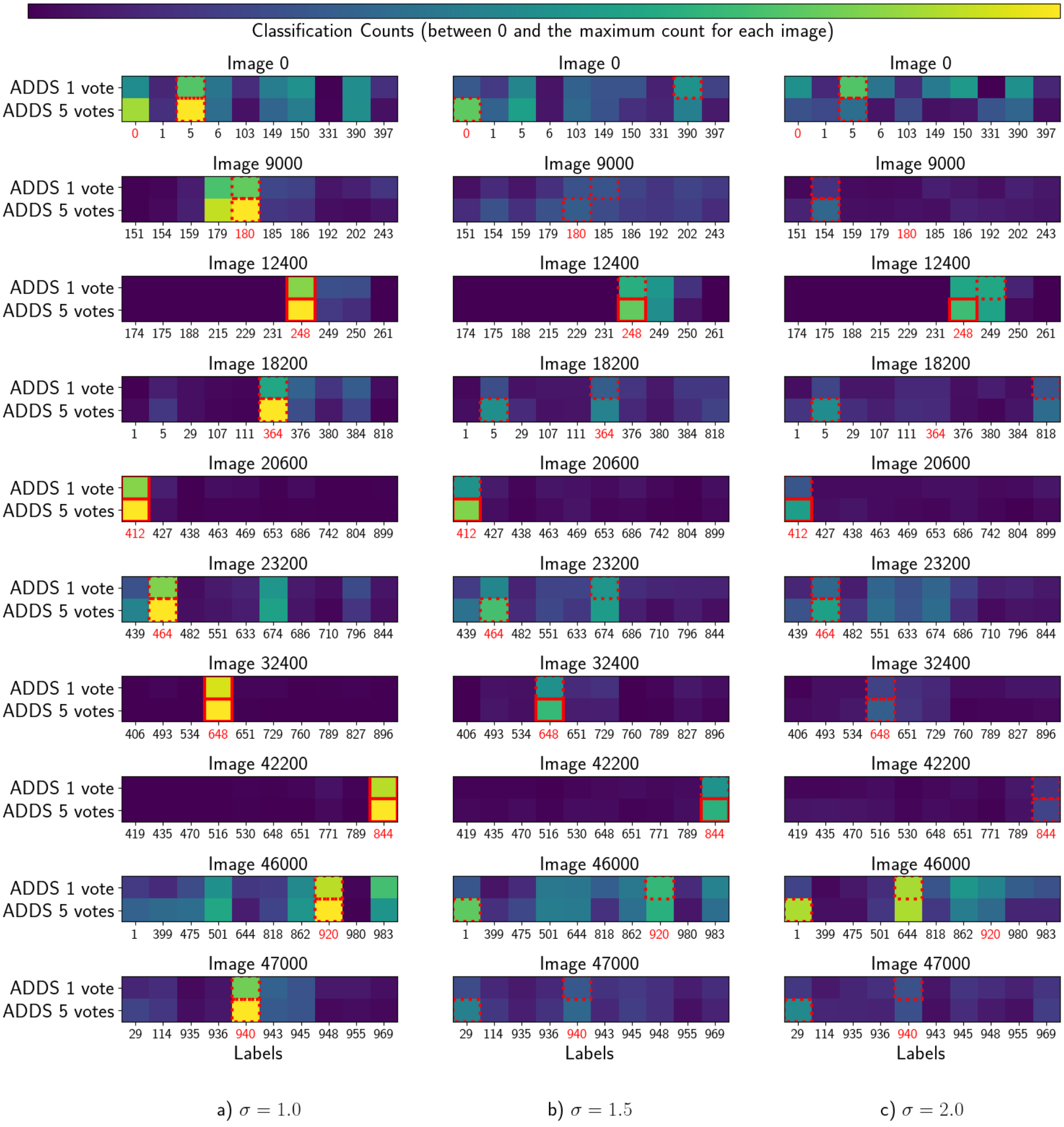


Figure 5. **Distribution of Classification Counts of ADDS over 1 and 5 Votes.** The plot shows the classification counts of ADDS at 1 and 5 votes over 1000 noise seeds on a selection of 10 images. Each column corresponds to a different privacy budget. For each image, we show the 10 most frequent labels, with the label in red being the true label. The solid red box corresponds to a certified prediction and the dashed red box corresponds to a prediction that is not certified. The selected images were chosen because they exhibited a change in classification correctness or certification as a result of voting. The proportion of these changes is not reflected. In particular, the number of occurrences where voting causes an abstained classification to become certified (for example, image 42200 at $\sigma = 1.5$) far exceeds the number of correct classifications that become incorrect (image 18200 at $\sigma = 1.5$).

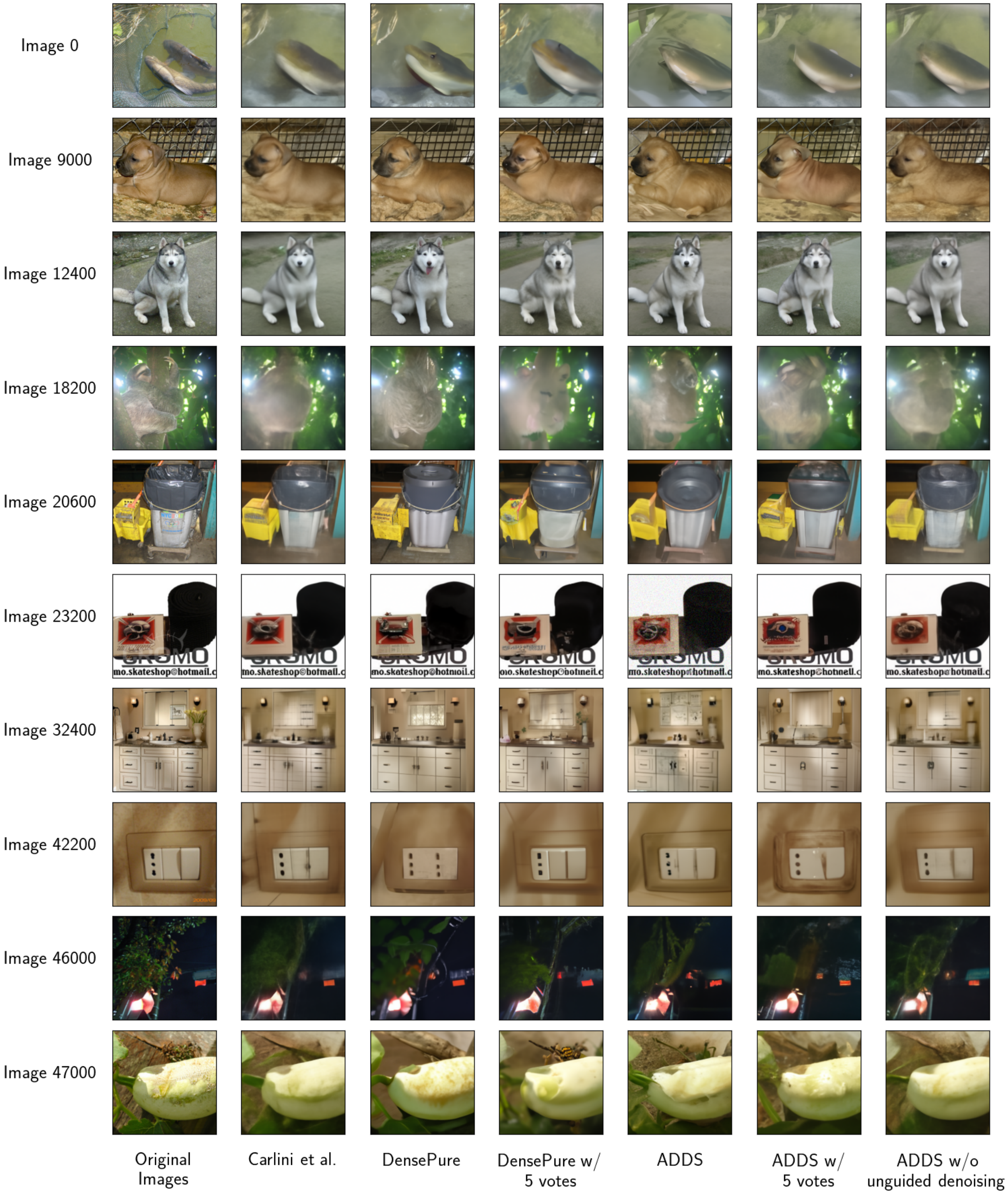


Figure 6. **Selection of Denoised Images at $\sigma = 1.0$.**

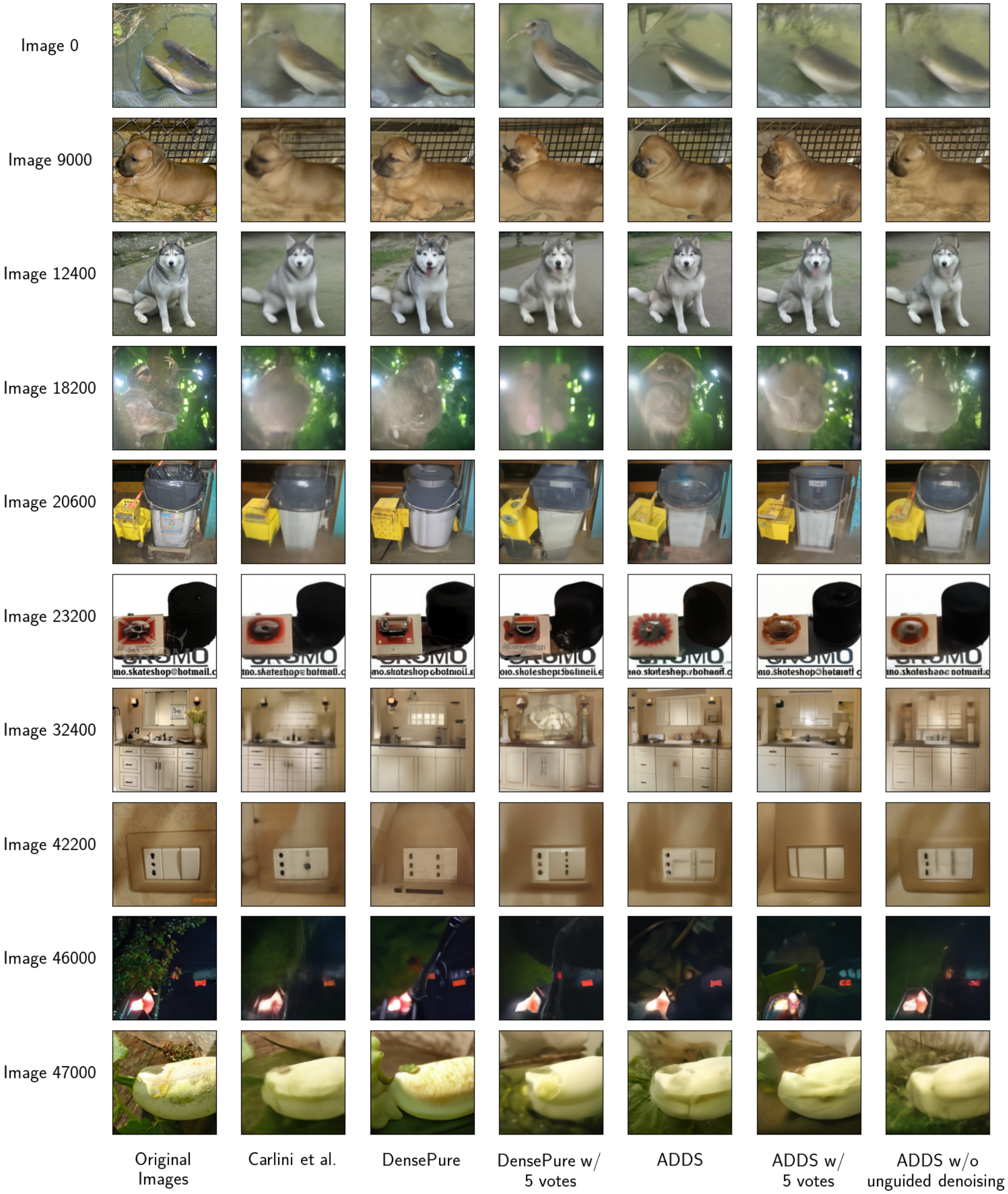


Figure 7. **Selection of Denoised Images at $\sigma = 1.5$.**

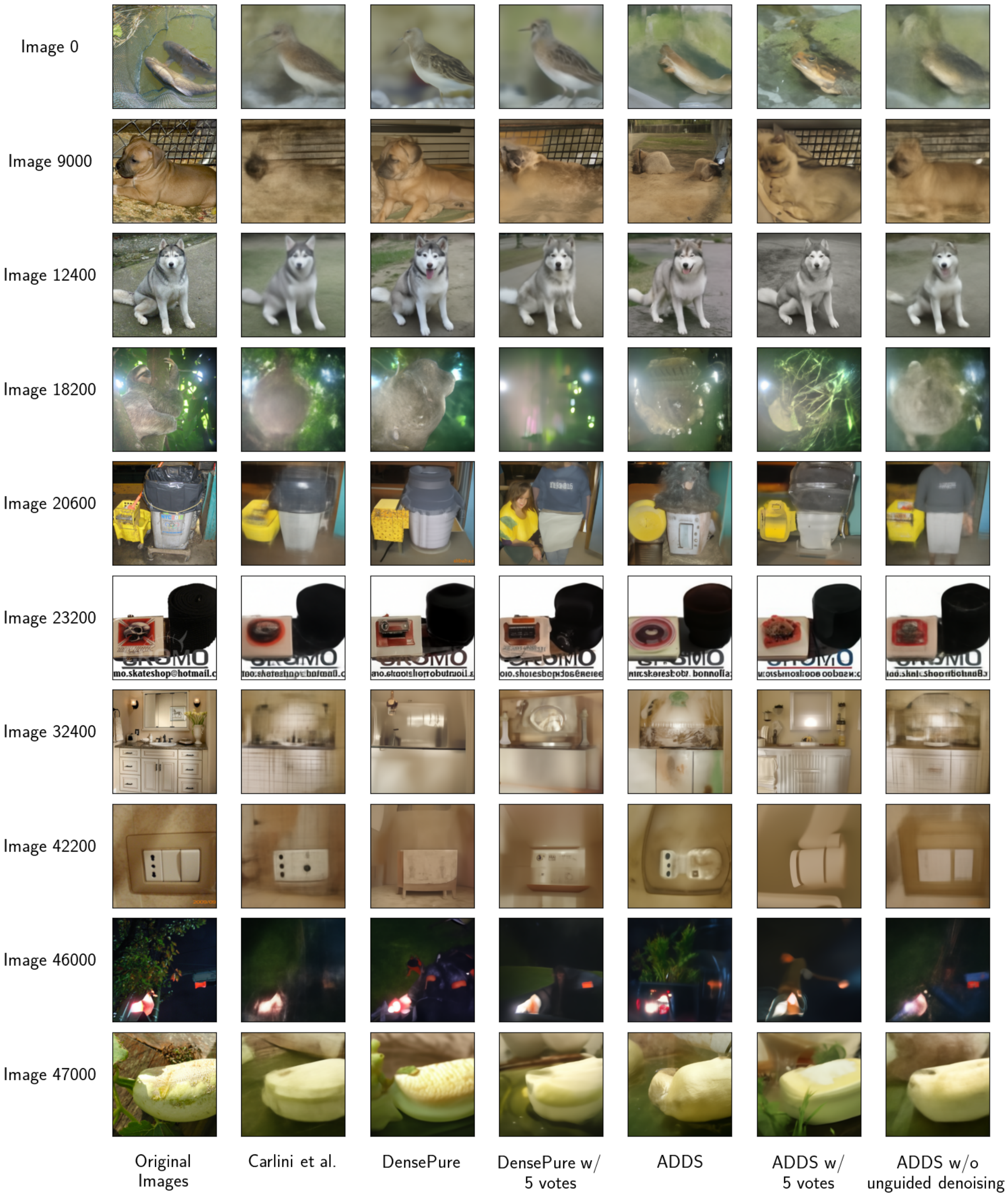


Figure 8. **Selection of Denoised Images at $\sigma = 2.0$.**