LDP-Slicing: Local Differential Privacy for Images via Randomized Bit-Plane Slicing

Supplementary Material

This supplementary material provides additional details about the proposed LDP-Slicing method. Specifically, we provide:

- Formal definitions of the reconstruction threat model (Sec. 3.1).
- Derivation of the utility-aware budget optimization.
- Complete proofs for the pixel-level Local Differential Privacy (LDP) guarantee and Total Variation bound.
- Training configurations and hyperparameters for all benchmarks.
- Additional Experiments.
- Ethical discussion.

A. Formal attack definition

We formalize the adversarial attack models discussed in our threat model in Sec. 3.1.

Definition 4 (Reconstruction attack [6])**.** *Let* $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ *be the sensitive dataset, and let* $\tilde{S} = (\tilde{X}, \tilde{Y}) \leftarrow \mathcal{M}(S)$ *be the privatized dataset revealed to an adversary* $\mathcal{A}$*. Given* $\tilde{S}$*,* $\mathcal{A}$ *outputs a candidate record* $x^* \in \mathcal{X}$*. We say that* $\mathcal{A}$ *succeeds in a reconstruction attack if*

$$\exists\, i \in [n] \text{ such that } d(x^*, x_i) \leq r,$$

where $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ *is a context-dependent metric and* $r > 0$*.*

In the context of face images, a successful reconstruction means that x^* preserves identifiable features of the original person (*e.g.*, eyes and nose). An ideal privacy mechanism should make it computationally infeasible for any adversary to produce such an x^* from $\tilde{x}$ with high confidence.

B. Derivation of utility-aware budget optimization

In Sec. 4.3 of the main paper, we formulate the following constrained optimization problem:

$$\begin{aligned} &\underset{\{\varepsilon_{c,b}\}}{\text{minimize}} && \sum_{c,b} \frac{W_{c,b}}{\varepsilon_{c,b}} \\ &\text{subject to} && \sum_{c,b} \varepsilon_{c,b} = \varepsilon_{\text{total}}, \qquad \varepsilon_{c,b} > 0, \end{aligned}$$

where $W_{c,b} > 0$ represents the importance weight for bit-plane b of channel c, and $\varepsilon_{c,b}$ is the allocated privacy budget.

We solve this using the method of Lagrange multipliers:

$$\mathcal{L}(\{\varepsilon_{c,b}\}, \lambda) = \sum_{c,b} \frac{W_{c,b}}{\varepsilon_{c,b}} + \lambda \left(\sum_{c,b} \varepsilon_{c,b} - \varepsilon_{\text{total}} \right),$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier. Taking partial derivatives with respect to each decision variable $\varepsilon_{c,b}$ and setting them to zero yields:

$$\frac{\partial \mathcal{L}}{\partial \varepsilon_{c,b}} = -\frac{W_{c,b}}{\varepsilon_{c,b}^2} + \lambda = 0.$$

Solving for $\varepsilon_{c,b}$, we obtain:

$$\varepsilon_{c,b} = \frac{\sqrt{W_{c,b}}}{\sqrt{\lambda}}. \tag{8}$$

Since $W_{c,b} > 0$ and $\varepsilon_{c,b} > 0$, this implies $\lambda > 0$. Substituting (8) into the budget constraint $\sum_{c,b} \varepsilon_{c,b} = \varepsilon_{\text{total}}$, we obtain:

$$\sum_{c,b} \varepsilon_{c,b} = \sum_{c,b} \frac{\sqrt{W_{c,b}}}{\sqrt{\lambda}} = \varepsilon_{\text{total}},$$

Solving for $1/\sqrt{\lambda}$:

$$\frac{1}{\sqrt{\lambda}} = \frac{\varepsilon_{\text{total}}}{\sum_{c',b'} \sqrt{W_{c',b'}}}. \tag{9}$$

Finally, substituting (9) back into (8) gives the final solution

$$\varepsilon_{c,b} = \sqrt{W_{c,b}} \cdot \frac{\varepsilon_{\text{total}}}{\sum_{c',b'} \sqrt{W_{c',b'}}},$$

or equivalently,

$$\boxed{\varepsilon_{c,b} = \varepsilon_{\text{total}} \cdot \frac{\sqrt{W_{c,b}}}{\sum_{i \in \{\text{Y,Cb,Cr}\}} \sum_{j=1}^{8} \sqrt{W_{i,j}}}}. \tag{10}$$

C. Complete proofs

C.1. Proof of pixel-level LDP guarantee

In Theorem 3, we claim that LDP-Slicing guarantees that each individual pixel in the output image satisfies $\varepsilon_{\text{total}}$-LDP with respect to the original pixel value. Let $\mathcal{X}$ denote the domain of possible pixel values after LL pruning, and let

$$\mathcal{M} : \mathcal{X} \to \mathcal{Y}$$

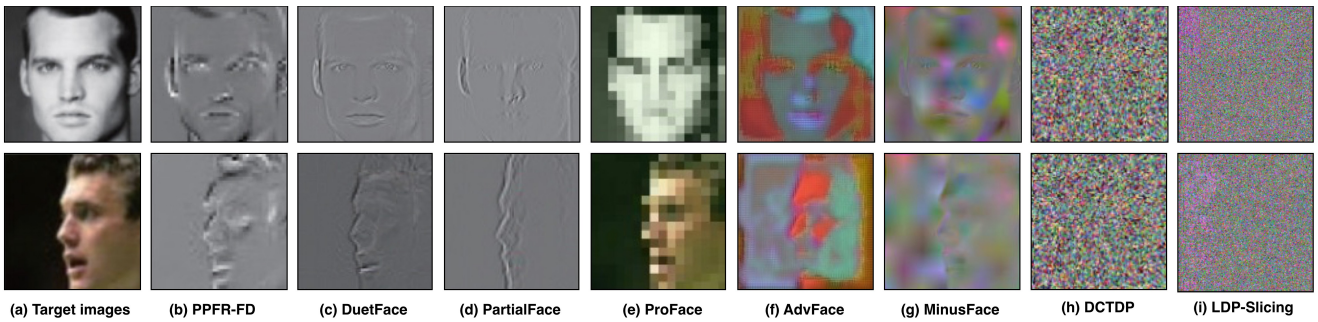

Figure 9. **Qualitative comparison against other SOTAs.** We compare LDP-Slicing (i) against (b) PPFR-FD [46], (c) DuetFace [32], (d) PartialFace [33], (e) ProFace [50], (f) AdvFace [47], (g) MinusFace [34], and DCTDP [23]. DCTDP and LDP-Slicing achieve a higher level of perceptual obfuscation.

be the per-pixel LDP-Slicing mechanism. By Definition 2, this means that for any two possible pixel values $x, \tilde{x}$ and any measurable subset of possible randomized outputs S:

$$\Pr[\mathcal{M}(x) \in S] \;\leq\; e^{\varepsilon_{\text{total}}} \cdot \Pr[\mathcal{M}(\tilde{x}) \in S], \tag{11}$$

where $\varepsilon_{\text{total}}$ is the total privacy budget allocated per pixel. In this section, we provide the complete formal proof.

Setup

Each pixel after LL pruning has Y, Cb, and Cr channels, each with 8 bits. We index the resulting $d = 24$ bits as:

$$x \longmapsto (x_1, \ldots, x_d) \in \{0,1\}^d.$$

Let ε_ℓ denote the privacy budget allocated to the ℓ-th bit, where the total budget is $\varepsilon_{\text{total}} = \sum_{\ell=1}^{d} \varepsilon_\ell$.

Proof. **1. Single bit privacy.**
For any bit x_ℓ in bit-plane B_ℓ, we apply the binary randomized response mechanism $\mathcal{M}_{\text{RR}}$ with budget ε_ℓ. Let p denote the probability of true response:

$$p \;=\; \frac{e^{\varepsilon_\ell}}{e^{\varepsilon_\ell} + 1}$$

By Definition 3, for any output $\tilde{x}_\ell \in \{0,1\}$, we have:

$$\Pr[\mathcal{M}_{\text{RR}}(x_\ell) = \tilde{x}_\ell] = \begin{cases} p, & \text{if } x_\ell = \tilde{x}_\ell, \\ 1-p, & \text{if } x_\ell \neq \tilde{x}_\ell. \end{cases} \tag{12}$$

The worst-case likelihood ratio between two different inputs is then:

$$\frac{\Pr[\mathcal{M}_{\text{RR}}(x_\ell) = \tilde{x}_\ell]}{\Pr[\mathcal{M}_{\text{RR}}(x_\ell) \neq \tilde{x}_\ell]} \;\leq\; \frac{p}{1-p} \;=\; \frac{\dfrac{e^{\varepsilon_\ell}}{e^{\varepsilon_\ell}+1}}{\dfrac{1}{e^{\varepsilon_\ell}+1}} \;=\; e^{\varepsilon_\ell}.$$

Thus, for every ℓ and for all $x_\ell, \tilde{x}_\ell \in \{0,1\}$ and all $S \subseteq \{0,1\}$:

$$\Pr[\mathcal{M}_{\text{RR}}(x_\ell) \in S] \;\leq\; e^{\varepsilon_\ell} \Pr[\mathcal{M}_{\text{RR}}(\tilde{x}_\ell) \in S] \tag{13}$$

2. Basic composition.
For a fixed pixel $x \in \mathcal{X}$, let its d-bit representation after bit-plane slicing be $(x_1, \ldots, x_d) \in \{0,1\}^d$. For each bit-plane index $\ell \in \{1, \ldots, d\}$, we apply the mechanism $\mathcal{M}_{\text{RR}}$ in parallel with budget ε_ℓ:

$$\mathcal{M}_{1:d}(x) := \big(\mathcal{M}_1(x_1), \ldots, \mathcal{M}_d(x_d)\big) \in \{0,1\}^d. \tag{14}$$

Where the mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_d$ are run with mutually independent on each bit. For any output pixel $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_d) \in \{0,1\}^d$ we have:

$$\Pr[\mathcal{M}_{1:d}(x) = \tilde{x}] \;=\; \prod_{\ell=1}^{d} \Pr[\mathcal{M}_\ell(x_\ell) = \tilde{x}_\ell]\,.$$

Fix any two pixel values $x, \tilde{x} \in \mathcal{X}$ we obtain:

$$\begin{aligned} \frac{\Pr[\mathcal{M}_{1:d}(x) = \tilde{x}]}{\Pr[\mathcal{M}_{1:d}(x) \neq \tilde{x}]} &= \frac{\prod_{\ell=1}^{d} \Pr[\mathcal{M}_\ell(x_\ell) = \tilde{x}_\ell]}{\prod_{\ell=1}^{d} \Pr[\mathcal{M}_\ell(x_\ell) \neq \tilde{x}_\ell]} \\ &= \prod_{\ell=1}^{d} \frac{\Pr[\mathcal{M}_\ell(x_\ell) = \tilde{x}_\ell]}{\Pr[\mathcal{M}_\ell(x_\ell) \neq \tilde{x}_\ell]} \\ &\leq \prod_{\ell=1}^{d} e^{\varepsilon_\ell} \;=\; e^{\sum_{\ell=1}^{d} \varepsilon_\ell} \;=\; e^{\varepsilon_{\text{total}}}. \end{aligned}$$

Therefore $\mathcal{M}_{1:d}$ is $\varepsilon_{\text{total}}$-LDP.

3. Immunity to post-processing.
We now prove formally that the reconstruction step preserves the $\varepsilon_{\text{total}}$-LDP guarantee. Let

$$\mathcal{M}_{1:d} : \mathcal{X} \to \{0,1\}^d$$

denote $\varepsilon_{\text{total}}$-LDP the LDP mechanism from Eq. (14). The reconstruction map $f : \{0,1\}^d \to \mathcal{Y}$ is deterministic and given by:

$$f(\tilde{x}_1, \ldots, \tilde{x}_d) = \sum_{\ell=1}^{d} 2^{d-\ell} \tilde{x}_\ell.$$

The per-pixel LDP-Slicing mechanism is therefore:

$$\mathcal{M}(x) = f\big(\mathcal{M}_{1:d}(x)\big).$$

Fix any $x, x' \in \mathcal{X}$ and any measurable subset $S \subseteq \mathcal{Y}$. Define

$$T = \{z \in \{0,1\}^d : f(z) \in S\}.$$

Then

$$\Pr[f(\mathcal{M}_{1:d}(x)) \in S] = \Pr[\mathcal{M}_{1:d}(x) \in T]\,.$$

Since $\mathcal{M}_{1:d}$ is $\varepsilon_{\text{total}}$-LDP, we have:

$$\Pr[\mathcal{M}_{1:d}(x) \in T] \;\le\; e^{\varepsilon_{\text{total}}} \Pr[\mathcal{M}_{1:d}(\tilde{x}) \in T]\,.$$

Hence LDP-Slicing is per-pixel $\varepsilon_{\text{total}}$-LDP. □

C.2. Proof of total variation bound under ε-LDP

Proof. Let P and Q be the output distributions of the LDP-Slicing mechanism $\mathcal{M}$ on pixel x and $\tilde{x}$. By ε-LDP (Definition 2), for all measurable $S \subseteq \mathcal{Y}$,

$$P(S) \le e^{\varepsilon} Q(S)$$

The total variation distance between distribution P and Q is:

$$\mathrm{TV}(P, Q) = \sup_{S \subseteq \mathcal{Y}} |P(S) - Q(S)|.$$

The LDP inequalities imply that the likelihood ratio between P and Q is bounded:

$$e^{-\varepsilon} \;\le\; \frac{dP}{dQ}(y) \;\le\; e^{\varepsilon}$$

then

$$\mathrm{TV}(P, Q) \;\le\; \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1} = \tanh\big(\varepsilon/2\big),$$

which proves (6).

For the identity distinguishing attack in Definition 1, let P and Q be the distributions of the observation conditioned on $b = 1$ and $b = 0$, respectively. Any (possibly randomized) adversary corresponds to some decision region S in which it outputs $b' = 1$. Then

$$\Pr[b' = b] = \tfrac{1}{2}\big(1 + P(S) - Q(S)\big),$$

so:

$$\big|\Pr[b' = b] - \tfrac{1}{2}\big| = \tfrac{1}{2}\,|P(S) - Q(S)| \;\le\; \tfrac{1}{2}\,\mathrm{TV}(P, Q).$$

Combining this with the bound on $\mathrm{TV}(P, Q)$ yields:

$$\mathsf{Adv}_{\mathcal{M}}^{\text{dist}} \;\le\; \tfrac{1}{2}\,\tanh\big(\varepsilon/2\big),$$

□

D. Experiments

D.1. Experimental details

This subsection provides detailed information about our experimental setup and hyperparameters. We conduct all experiments on NVIDIA H100 GPUs using PyTorch.

Image pre-processing. For face recognition, we resize each facial image to 112×112 pixels to align with other SOTAs. For image classification, we apply augmentation like random crop and random horizontal flip and optional cutout.

Privacy-preserving face recognition. We adopt the ResNet-50 architecture with the improved residual [18] backbone. The network is optimized using the ArcFace loss function [9] with a feature scale $s = 64$ and angular margin $m = 0.4$. We use SGD [40] with a momentum of 0.9 and a weight decay of 5×10^{-4}. Training proceeds for 24 epochs with an initial learning rate of 0.1, decayed by a factor of 10 at epochs 10, 18, and 22. We also initialize weights from a pre-trained IR-50 model and apply a 5-epoch warm-up period.

Privacy-preserving image classification. For image classification on CIFAR-10 and CIFAR-100 dataset, we adopt a ResNet-56 architecture trained for 250 epochs with a batch size of 128. We use SGD with a momentum of 0.9 and a reduced weight decay of 1×10^{-4}. We set the initial learning rate to 0.1 and decay by a factor of 10 at epochs 80, 160, and 200. For training with low privacy budgets (*e.g.*, $\varepsilon = 1$), we apply the gradient clipping with a max norm of 0.5.

D.2. Additional experiments

We extend LDP-Slicing to the medical domain using the Chest X-Ray dataset [25] at 224×224 resolution. We adopt ResNet-50 as backbone and remove color weights during the budget optimization process. As shown in Tab. 6, LDP-Slicing retains high utility even under strict privacy budgets (*e.g.* $\varepsilon_{total} = 5.2$). Furthermore, we perform a zero-shot test of our MS1MV2 trained model directly on the VGGFace2 [5] and CelebA [30] datasets. LDP-Slicing maintains a competitive utility under moderate privacy settings (in Tab. 6).

D.3. Ablation on different color weights

We evaluate LDP-Slicing under 2 additional different weights (in Tab. 7). The original weight (4:1:1) consistently outperforms other weights.

D.4. Ablation on LL pruning

We perform an ablation study (in Fig. 10) by removing the perceptual obfuscation stage. Without LL pruning, LDP-Slicing can still suppress identity-defining high-frequency cues, and the inversion recovers at most blurred, non-identifying content. Although perceptual obfuscation is an

Table 6. **Privacy-utility trade-off (%).** We perform additional experiment on another 3 benchmarks

Dataset	ε=1	ε=2.4	ε=5.2	ε=12	ε=20	ε=32	ε=58
Chest X-Ray [25]	76.60	88.94	90.54	91.99	92.15	92.95	93.43
VGGFace2 [5]	48.95	50.31	51.71	66.69	67.15	68.99	70.64
CelebA [30]	49.88	50.97	50.84	75.90	80.44	82.68	84.89

Table 7. **Utility under different color weight (%).**

Color W.	CIFAR10	CIFAR100	AgeDB	LFW	CALFW	CPLFW
4 : 1 : 1 **(Ours)**	**80.36**	**53.55**	**96.68**	**99.75**	**96.02**	**91.08**
2 : 1 : 1	78.44	50.63	95.85	99.63	95.68	90.87
1 : 1 : 1	74.92	45.16	94.68	99.47	94.17	89.85

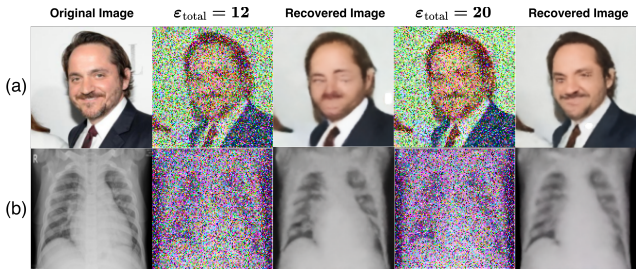

Figure 10. **Black-box reconstruction attack (w/o LL pruning)**

auxiliary step, it allows the same ε noise to be spent more on task-relevant signals.

D.5. Comparison with block-level privacy

In the main paper, we show that LDP-Slicing outperforms DCTDP [23] in face recognition benchmarks. To provide a fair comparison, we must convert the privacy guarantees at same level, as DCTDP defines privacy at the block-level, not pixel-level. Assume the standard configuration, DCTDP upsamples the facial image by 8 and applies a Laplace mechanism with $\varepsilon_{\text{mean}} = 0.5$ to each DCT coefficient within an 8×8 block. It also removes the DC coefficient for each color channel. By the central Differential Private (DP) Sequential Composition Theorem [12], the effective privacy budget bound for a single pixel is:

$$\varepsilon_{\text{pixel}} \leq ((8 \times 8)_{\text{block}} - 1_{\text{DC}}) \times 3 \times 0.5 = 94.5.$$

In contrast, LDP-Slicing operates under a total budget of $\varepsilon_{\text{total}} = 20$. This derivation suggests that our method is approximately $4.7\times$ stricter than DCTDP.

D.6. Qualitative comparison against other SOTAs.

To demonstrate the effectiveness against human observer, we add a visual comparison against other SOTAs in Fig. 9. Heuristic methods like PPFR-FD, DuetFace, and PartialFace retain sensitive facial features visible to the human observer. In contrast, DP/LDP methods conceal sensitive attributes in noise.

E. Ethical considerations.

Our primary experiment is trained on MS1Mv2 dataset [17]. We acknowledge the ethical concerns about the MS1Mv2 dataset. Our usage is strictly limited to privacy defense research and adheres to CVPR's ethical guidelines. We also recognize that our defense mechanism could theoretically be used by malicious actors to evade lawful systems (*e.g.*, hiding illegal content). However, we believe the benefit of allowing end-users local control over data privacy outweighs these risks.