Rethinking Gradient Step Denoiser: Towards Truly Pseudo-Contractive Operator

Shuchang Zhang, Yaoyun Zeng, Kangkang Deng,
and Hongxia Wang

College of Science, National University of Defense Technology
zhangshuchang19@163.com

October 31, 2025

Here, we recall that an operator $T : \mathcal{H} \to \mathcal{H}$ is d-pseudo-contractive (d-PC) if there exists a constant $d \in (-\infty, 1]$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{H}$, it holds that

$$\|T(\mathbf{x}) - T(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 + d\|(\mathbf{x} - T(\mathbf{x})) - (\mathbf{y} - T(\mathbf{y}))\|^2, \qquad (1)$$

- If $d = -1$, T is called firmly nonexpansive (FNE) operator.
- If $d = 0$, T is called nonexpansive (NE) operator.
- If $d < 1$, T is called d-strictly PC (d-SPC) operator.

Theoretical assumption on denoisers

- The FNE and nonexpansive assumptions: **Theorem III.1** (Sreehari et al., IEEE TCI, 2016); **Lemma 5** (Edward T. Reehorst and Philip Schniter, IEEE TCI, 2019); **Assumption 2** (Sun et al., IEEE TCI, 2021).
- **Demicontractive assumption** (Definition 1) in RED-PRO[1], which is defined by:

$$\min_{x \in \mathrm{Fix}(T)} \frac{1}{2\sigma^2} \left\| \mathbf{A}\mathbf{x} - \mathbf{y} \right\|^2 .$$

Definition 1

The mapping $T : \mathcal{H} \to \mathcal{H}$ is d-demicontractive with $d < 1$, if for any $\mathbf{x} \in \mathcal{H}$ and $\mathbf{z} \in \mathrm{Fix}(T)$ it holds that

$$\left\| T(\mathbf{x}) - \mathbf{z} \right\|^2 \leq \left\| \mathbf{x} - \mathbf{z} \right\|^2 + d \left\| T(\mathbf{x}) - \mathbf{x} \right\|^2 .$$

[1] Regev Cohen et al., Regularization by Denoising via Fixed-Point Projection (RED-PRO), SIAM Journal on Imaging Sciences, 2021, 14(3): 1374-1406.

Spectral methods

- Pesquet et al. [2]:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})} \left\| T_\theta(\mathbf{x}+\mathbf{n}) - \mathbf{x} \right\|^2 + \lambda \max\{ \left\| J_{Q_\theta}(\tilde{\mathbf{x}}) \right\|_*^2, 1-\varepsilon \}.$$

.

- Hurault et al. [3]:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})} \left\| T_\theta(\mathbf{x}+\mathbf{n}) - \mathbf{x} \right\|^2 + \lambda \max\{ \left\| J_{\nabla\psi_\theta}(\mathbf{x}+\mathbf{n}) \right\|_*^2, 1-\varepsilon \}.$$

- Wei et al. [4]:

$$\begin{aligned}\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\mathbf{n}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})} & \left\| T_\theta(\mathbf{x}+\mathbf{n}) - \mathbf{x} \right\|^2 \\ & + \lambda \max\{ \left\| J_{dI+(1-d)T_\theta}(\mathbf{x}+\mathbf{n}) \right\|_*^2, 1-\varepsilon \}.\end{aligned}$$

---

[2] Pesquet J C, Repetti A, Terris M, et al. Learning maximally monotone operators for image recovery[J]. SIAM Journal on Imaging Sciences, 2021, 14(3): 1206-1237.

[3] Hurault S, Leclaire A, Papadakis N. Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization[C]//International Conference on Machine Learning. PMLR, 2022: 9483-9505.

[4] Wei D, Chen P, Li F. Learning Pseudo-Contractive Denoisers for Inverse Problems[C]//Forty-first International Conference on Machine Learning, 2024.

Motivation

- How to verify that the denoiser satisfies the demicontractive property? [1].
- Challenges:
  - Spectral methods constrain the spectral norm using limited training samples instead of the entire space.
  - Power iteration methods are computationally expensive, non-deterministic, and lack strict spectral norm bounds during intermediate steps [5].

[5] Blaise Delattre, Quentin Barthélemy, Alexandre Araujo, and Alexandre Allauzen. Efficient Bound of Lipschitz Constant for Convolutional Layers by Gram Iteration. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 7513-7532. PMLR, 23-29 Jul 2023.

Let $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ denote the classes of all operators $T : \mathcal{H} \to \mathcal{H}$ satisfying the assumptions of demicontractive, SPC, NE, and FNE, respectively.

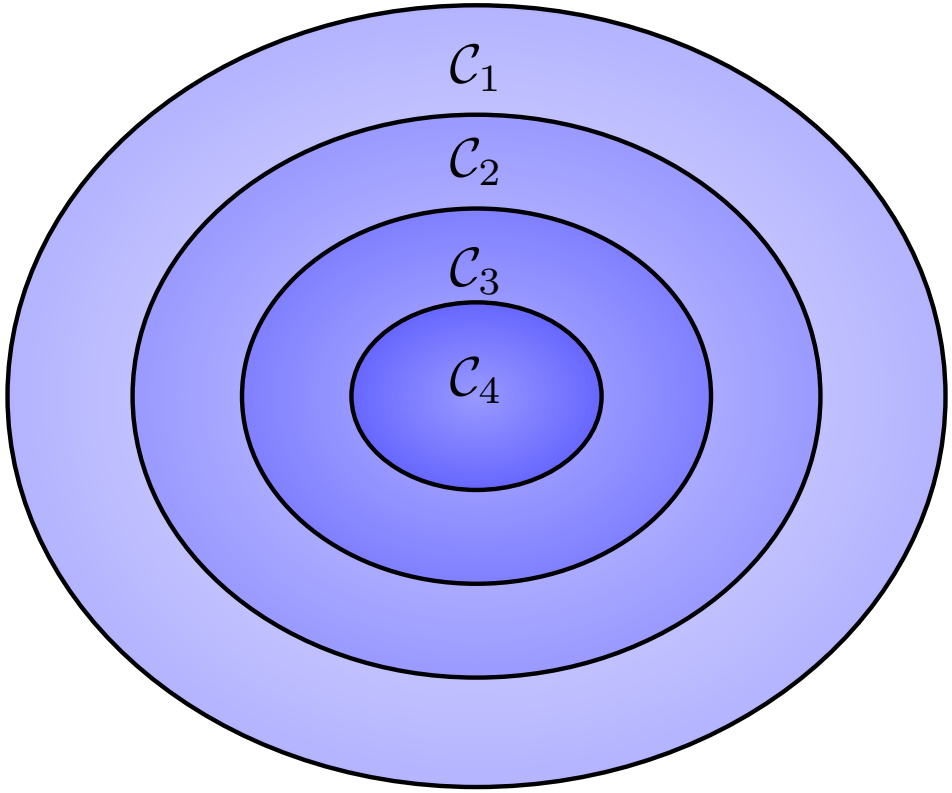


Fig. 1: Relationship between different classes of operators.

- The operator $T : \mathcal{H} \to \mathcal{H}$ is called conically λ-averaged for $\lambda > 0$ if there exists a NE operator U such that $T = (1-\lambda)I + \lambda U$. In particular, when $\lambda \in (0, 1)$ the operator is λ-averaged.
- If U is FNE, then $T = (1-\lambda)I + \lambda U$ is λ-relaxed FNE (λ-RFNE)
- For any $\mathbf{x}, \mathbf{y} \in \mathcal{H}$, it holds

$$\langle \mathbf{x} - \mathbf{y}, R(\mathbf{x}) - R(\mathbf{y}) \rangle \geq \frac{1}{\mu} \|R(\mathbf{x}) - R(\mathbf{y})\|^2, \tag{2}$$

then R is called $\frac{1}{\mu}$-cocoercive.

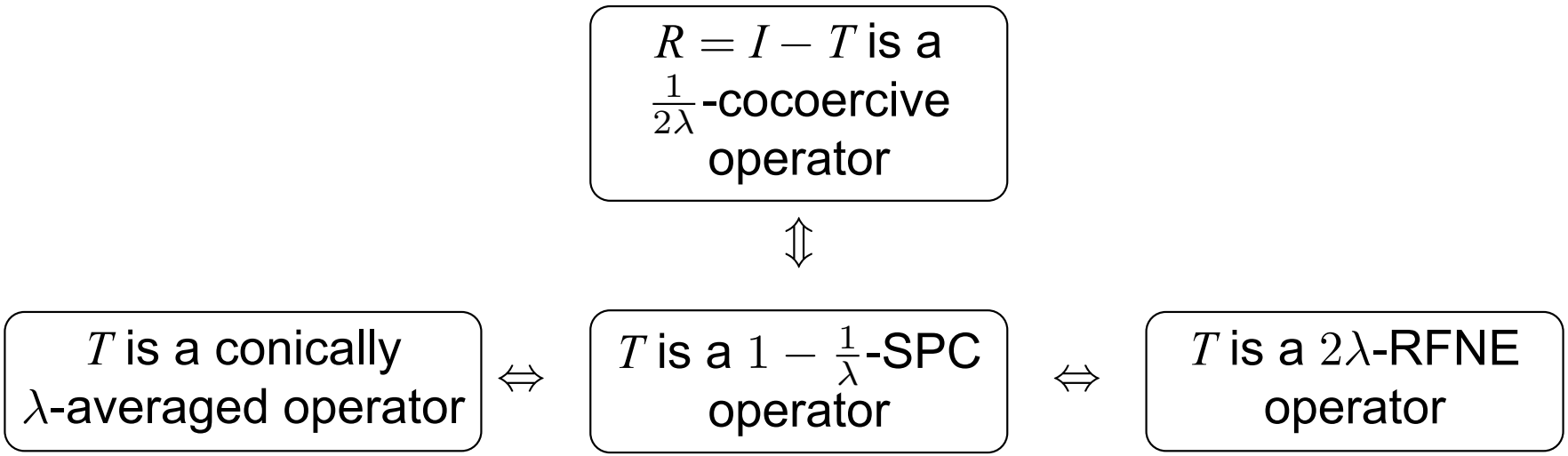


Fig. 2: Equivalent relationships.

Proposition 2 [6]

Let $R = I - T$ and $\mu > 0$, then $T : \mathcal{H} \to \mathcal{H}$ is μ-RFNE if and only if for all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$,

$$\langle \mathbf{x} - \mathbf{y}, R(\mathbf{x}) - R(\mathbf{y}) \rangle \geq \frac{1}{\mu} \|R(\mathbf{x}) - R(\mathbf{y})\|^2. \tag{3}$$

Theorem 3 [7]

Let $h : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be a convex function, differentiable over $\mathcal{H}$, and let $L > 0$. Then the following claims are equivalent:

(i) *h is L-smooth.*

(ii) *For all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$,*

$$\langle \nabla h(\mathbf{x}) - \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\|^2. \tag{4}$$

[6] Andrzej Cegielski. Iterative methods for fixed point problems in Hilbert spaces. Springer, 2012.
[7] Amir Beck. First-order methods in optimization. SIAM, 2017.

Proposition 4

Consider a scalar-valued $(K+1)$-layered neural network $\psi_\theta : \mathbb{R}^n \to \mathbb{R}$ defined by $\psi_\theta(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{z}_K + b$ and the recursion

$$\mathbf{z}_1 = \phi(\mathbf{H}_1\mathbf{x} + \mathbf{b}_1), \quad \mathbf{z}_k = \phi(\mathbf{W}_k\mathbf{z}_{k-1} + \mathbf{H}_k\mathbf{x} + \mathbf{b}_k), \quad k = 2, 3, \ldots, K,$$

where $\Theta = \{\mathbf{w}, b, \{\mathbf{W}_k\}_{k=2}^K, \{\mathbf{H}_k\}_{k=1}^K, \{\mathbf{b}_k\}_{k=1}^K\}$ are learnable parameters, $\phi : \mathbb{R} \to \mathbb{R}$ is a convex, non-decreasing and continuously differentiable scalar function, which operates pointwise. Assume that all entries of $\mathbf{W}_k$ and $\mathbf{w}$ are non-negative, and let ψ_θ be L_θ-smooth, then the GS denoiser $T_\theta = I - \nabla\psi_\theta$ is $\frac{L_\theta - 2}{L_\theta}$-SPC operator.

Algorithm 1 RED-PRO with the learned truly SPC denoiser

Input: initialization $\mathbf{x}^0 \in \mathbb{R}^n, \mu_k = \frac{c}{(1+k)^\alpha}, w \in (0, \frac{2}{L_\theta})$, and the GS denoiser $T_\theta = I - \nabla\psi_\theta$.

1: **for** $k = 1, 2, \cdots, K$ **do**
2: $\quad \mathbf{y}^k = (1-w)\mathbf{x}^{k-1} + wT_\theta(\mathbf{x}^{k-1})$
3: $\quad \mathbf{x}^k = \mathbf{y}^k - \mu_k \nabla f(\mathbf{y}^k)$
4: **end for**

Output: $\mathbf{x}^K$.

Theorem 5

Let $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ and $\{\mathbf{y}^k\}_{k\in\mathbb{N}}$ be sequences generated by Algorithm 1. Assume that $\mathcal{S} = \arg\min_{\mathbf{x}\in \mathrm{Fix}(T_\theta)} f(\mathbf{x})$ is the solution set of the RED-PRO model and the sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ is bounded, then

(i) *For any $\mathbf{x}' \in \mathcal{S}$ and $k \geq 1$, there exist $D_1, D_2 > 0$ such that*

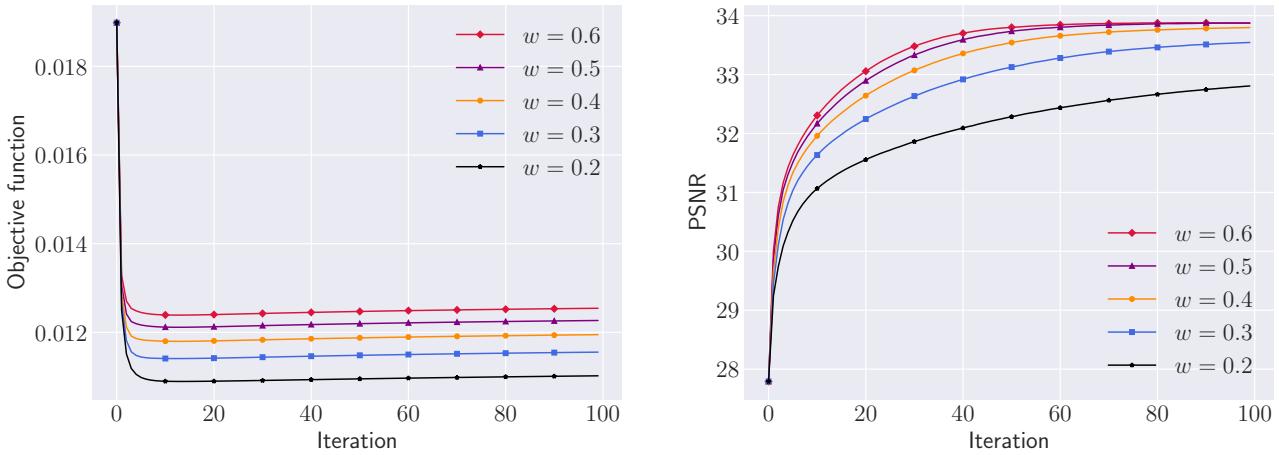
$$u_k \leq \frac{D_1^2}{ck^{1-\alpha}} + \frac{cD_2^2}{k^\alpha},$$

where $u_k = \min\{\langle \nabla f(\mathbf{y}^j), \mathbf{y}^j - \mathbf{x}' \rangle : k \leq j \leq 2k\}$.

(ii) *For any $\mathbf{x}' \in \mathcal{S}$ and $k \geq 1$, we have*

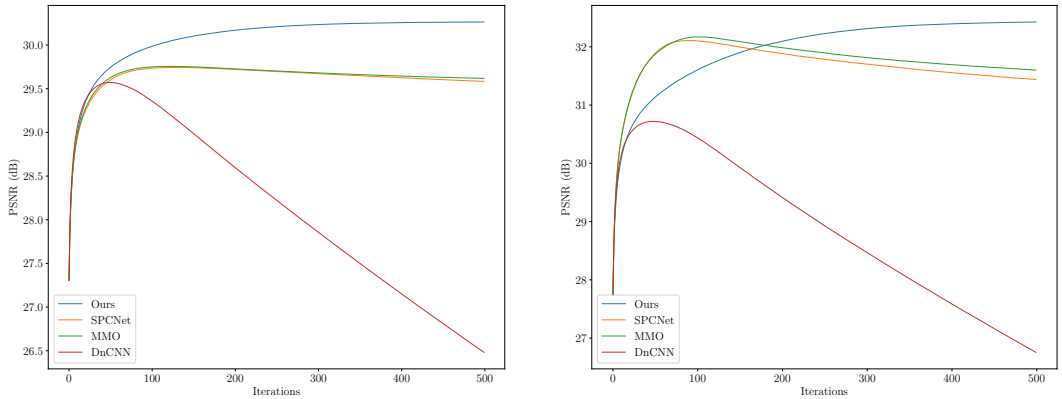
$$f(\mathbf{y}^k_{best}) - f(\mathbf{x}') \leq \frac{D_1^2}{ck^{1-\alpha}} + \frac{cD_2^2}{k^\alpha},$$

where $k_{best} = \arg\min_{k\leq j\leq 2k} f(\mathbf{y}^j)$.



(a) (b)

Fig. 3: Convergence of Algorithm 1 on one image in the CelebA dataset. (a) Objective function. (b) PSNR.

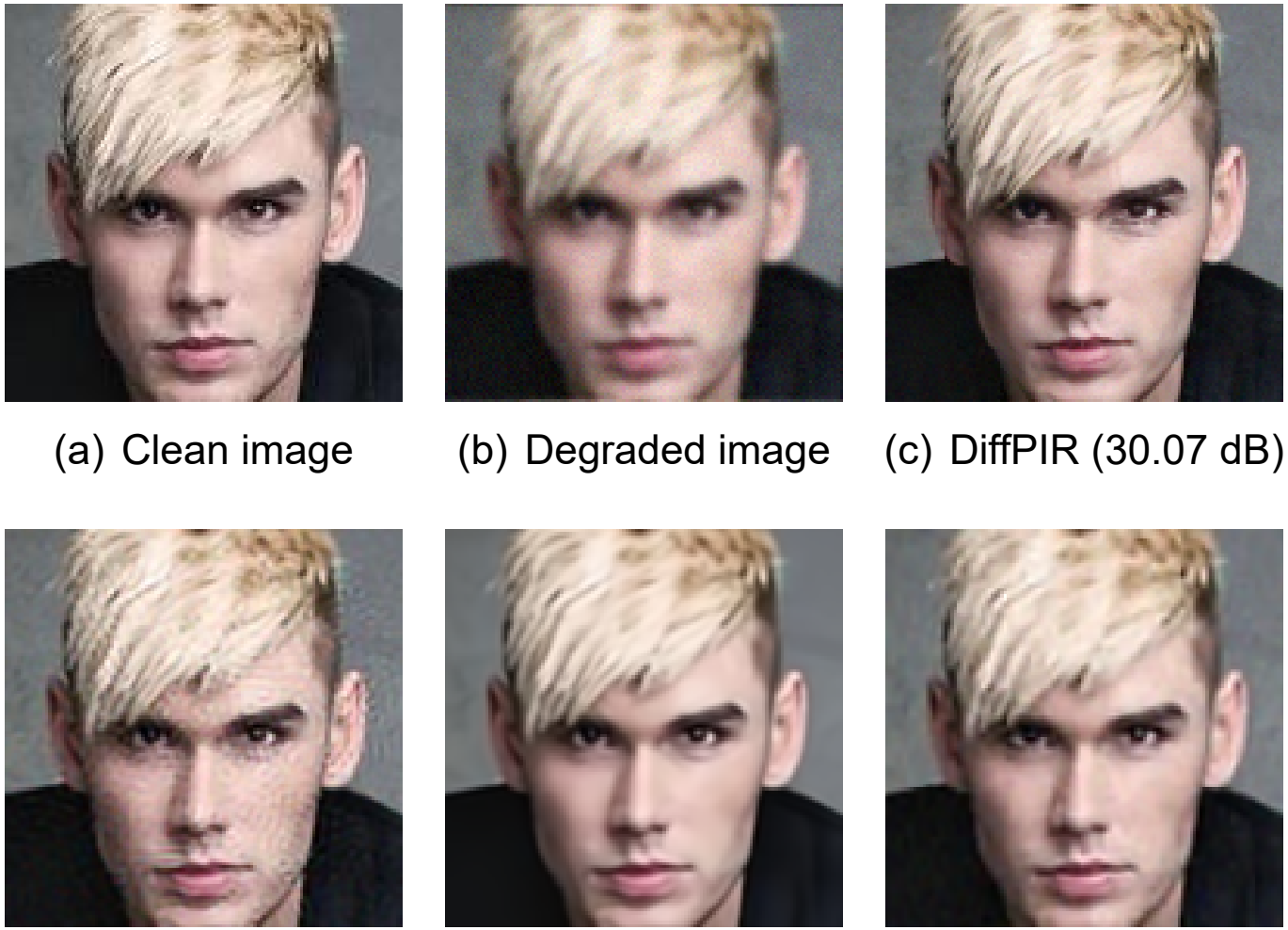


(a) Man (b) House

Fig. 4: PSNR curves of RED-PRO with the learned truly SPC denoiser and non-SPC methods on two images in the Gaussian deblurring task.

Table 1: Deblurring results on CelebA over 20 samples.

METHOD	$\sigma_{blur} = 1$, $\sigma_{noise} = .02$		$\sigma_{blur} = 1$, $\sigma_{noise} = .04$		$\sigma_{blur} = 2$, $\sigma_{noise} = .02$		$\sigma_{blur} = 2$, $\sigma_{noise} = .04$	
	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)
DiffPIR	30.8 ± 2.0	.86 ± .03	29.5 ± 1.8	.82 ± .03	28.6 ± 2.0	.80 ± .05	27.6 ± 1.8	.77 ± .05
PnP-PGD	31.4 ± 1.9	.87 ± .02	27.6 ± 0.9	.71 ± .05	29.9 ± 2.3	.85 ± .05	28.8 ± 2.0	.81 ± .05
DPIR	**33.2 ± 3.0**	**.92 ± .03**	**31.8 ± 2.6**	**.89 ± .04**	**30.1 ± 2.5**	**.86 ± .05**	**29.1 ± 2.2**	**.83 ± .05**
RED-PRO	32.4 ± 2.8	.92 ± .03	30.8 ± 2.3	.88 ± .03	29.3 ± 2.3	.86 ± .04	28.4 ± 2.0	.83 ± .04



(a) Clean image (b) Degraded image (c) DiffPIR (30.07 dB)

(d) PnP-PGD (30.53 dB) (e) DPIR (31.95 dB) (f) Ours (30.93 dB)

Fig. 5: Visual comparison on CelebA for Gaussian deblurring with $\sigma_{blur} = 1, \sigma_{noise} = 0.02$.