Learning Proximal Operators to Discover Multiple Optima

Lingxiao Li [1] Noam Aigerman [2] Vladimir G. Kim [2] Jiajin Li [3]
Kristjan Greenewald [4] Mikhail Yurochkin [4] Justin Solomon [1]

[1]MIT CSAIL [2]Adobe Research [3]Stanford University [4]MIT-IBM Watson AI Lab

Multi-solution optimization (MSO)

Finding multiple optima of an optimization problem is a ubiquitous task, either because there are many equally-performant global optima or due to the fact that the optimization objective does not capture user preferences precisely.

We define multi-solution optimization (MSO) as follows:

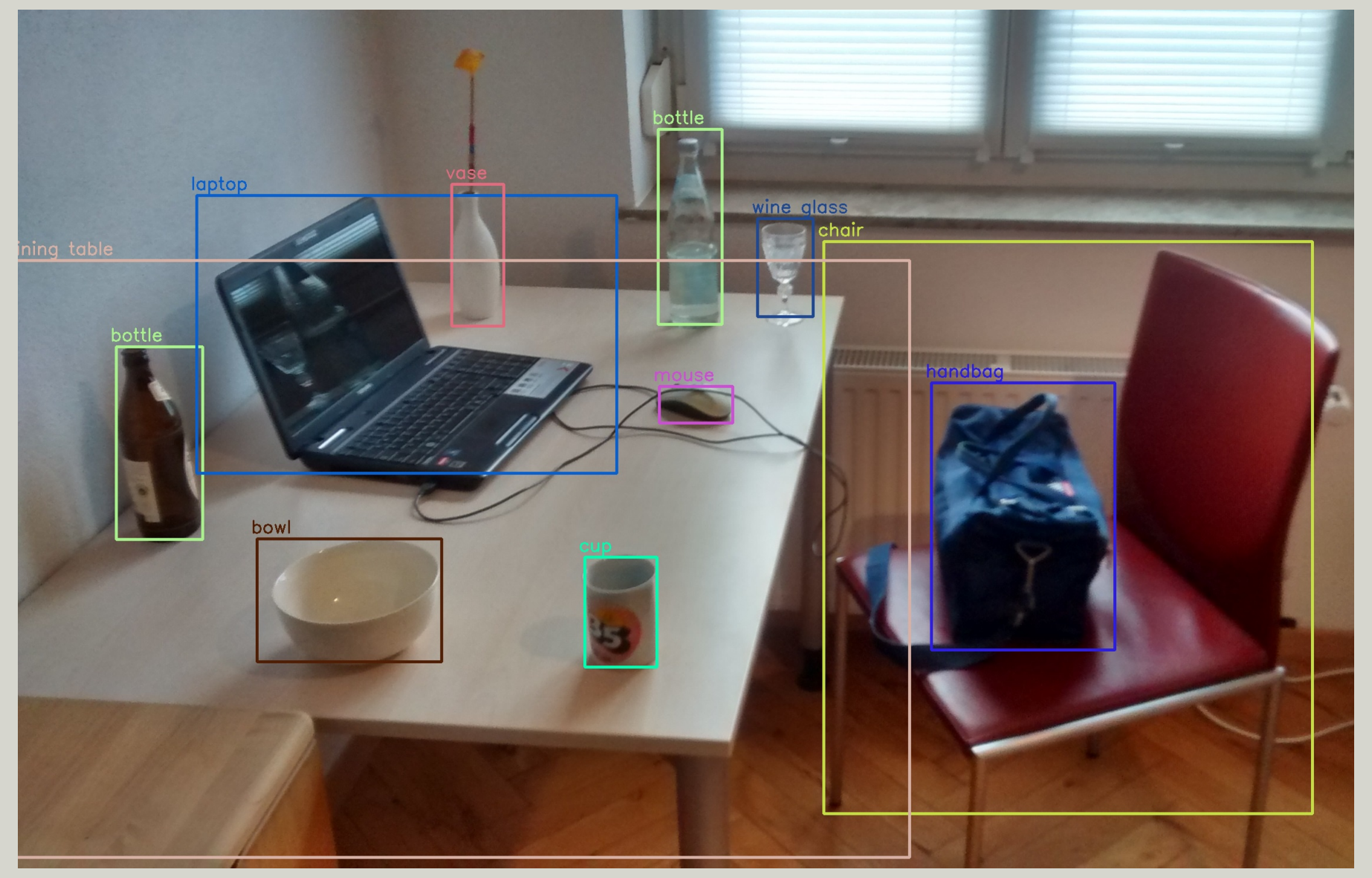
> Given a family of non-convex differentiable $\{f_\tau : \mathcal{X} \to \mathbf{R}\}_{\tau\in\mathcal{T}}$, find all minima of f_τ. Each function f_τ might contain multiple local and global minima.

- *Can we do better than random initial guesses plus gradient-based optimization from each initial guess?*
- *Can we apply machine learning to generalize to an unseen τ' at test time?*

Example: object detection in images

Consider the task of object detection (ignoring labels):

- $\mathcal{X} \subset \mathbf{R}^4$: $x = (w, h, c_x, c_y)$ denotes a bounding box.
- $\mathcal{T}$ is the space of images.
- $f_\tau(x) = \min_{i=1}^{K_\tau} \left\|b_i^\tau - x\right\|_1$, where $\{b_i^\tau\}_{i=1}^{K_\tau}$ are the set of GT bounding boxes in image $\tau \in \mathcal{T}$.
- At test time, we only have access to a test image τ' but not its GT bounding boxes.



Proximal-point algorithm (PPA)

For a fixed $\lambda > 0$, the *proximal operator* of f_τ is defined as

$$\text{prox}(x;\tau) := \arg\min_y \left\{ f_\tau(y) + \frac{\lambda}{2}\|y - x\|_2^2 \right\}.$$

If f_τ is λ-weakly convex ($f_\tau + \lambda/2\|\cdot - x\|_2^2$ is strongly convex), then prox is uniquely defined.

With initial point x^0, the *proximal-point algorithm (PPA)* iterates, for $k \in \mathbf{Z}_{\geq 0}$,

$$x^{k+1} := \text{prox}(x^k;\tau).$$

When f_τ is locally indistinguishable from a convex function, then with reasonable stopping criterion, PPA converges linearly to a local minimum of f_τ, even if prox is approximated [6]. Such convergence can be faster than gradient descent [3].

- *How can we approximate the proximal operator without solving an inner optimization at each evaluation?*
- *If we start at a different initial guess or optimize for a similar but different τ, can we "reused" the approximated proximal operator?*

Learning proximal operators

We propose to learn the proximal operators in an end-to-end fashion for all $x \in \mathcal{X}, \tau \in \mathcal{T}$:

$$\min_{\Phi:\mathcal{X}\times\mathcal{T}\to\mathcal{X}} \mathbf{E}_{\substack{x\sim\mu \\ \tau\sim\nu}} \left[f_\tau(\Phi(x,\tau)) + \frac{\lambda}{2}\|\Phi(x,\tau) - x\|_2^2 \right], \tag{1}$$

where μ is a prior on $\mathcal{X}$ (e.g. a uniform distribution on $\mathcal{X}$) and ν is the training data on $\mathcal{T}$. We parameterize $\Phi : \mathcal{X} \times \mathcal{T} \to \mathcal{X}$ using a residual neural network and train Φ by SGD.

For problems where $\mathcal{T}$ is structured (images or 3D point clouds), we first embed τ using a suitable encoder before passing it to Φ.

- At test time, for an unseen τ' with objective $f_{\tau'}$, we sample a batch of $x \sim \mu$ and run a few steps (≤ 10) of PPA by $x^{k+1} \leftarrow \Phi(x^k;\tau')$ to obtain multiple solutions.
- Can view $\Phi^k(;\tau')_\#\mu$ as a **generative model** from which local minima can be sampled. Hence we can represent *arbitrary number* of solutions even when the set of minima is continuous.

Global convergence of training

How easy is it to train Φ with loss (1)? Based on Kawaguchi and Huang [4], we show theoretically that the proximal term in (1) conveniently elevates the convexity of f_τ to obtain global convergence of training.

Suppose the training dataset is $S = \{(x_i,\tau_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{T}$ of size n. Define the discretized training loss of (1) to be,

$$L(\Phi) := \frac{1}{n}\sum_{i=1}^n \left[f_{\tau_i}(\Phi(x_i,\tau_i)) + \frac{\lambda}{2}\|\Phi(x_i,\tau_i) - x_i\|_2^2 \right].$$

> Suppose for any $\tau \in \mathcal{T}$, the objective $f_\tau \in C^1(\mathcal{X})$ is ξ-weakly convex and ∇f_τ is ζ-Lipschitz with $\xi < \lambda$. Then for any common neural network with $\tilde{\Omega}(n)$ total parameters, with high probability, gradient descent on its weights will eventually reach the minimum loss $\min_\Phi L(\Phi)$.
>
> The number of iterations needed to achieve $\epsilon > 0$ training error is $O((\lambda + \zeta)/\epsilon)$. When this occurs, the mean-squared error of the learned proximal operator compared to the true one is $O(2\epsilon/(\lambda-\xi))$ on training data.

Application: non-convex sparse recovery

Given measurement y generated from $y = Ax^* + e$ where e is noise, sparse recovery aims at recovering a sparse x^* from y. We consider a non-convex sparse recovery formulation that minimizes

$$f_{(\alpha,p)}(x) = \frac{1}{2}\|Ax - y\|_2^2 + \alpha\|x\|_p^p,$$

for $\alpha > 0, p \in (0,1)$. Compared to LASSO ($p = 1$), non-convex ℓ^p norms require milder condtions under which the global optimia of the objective are the desired sparse x^* [1, 2].

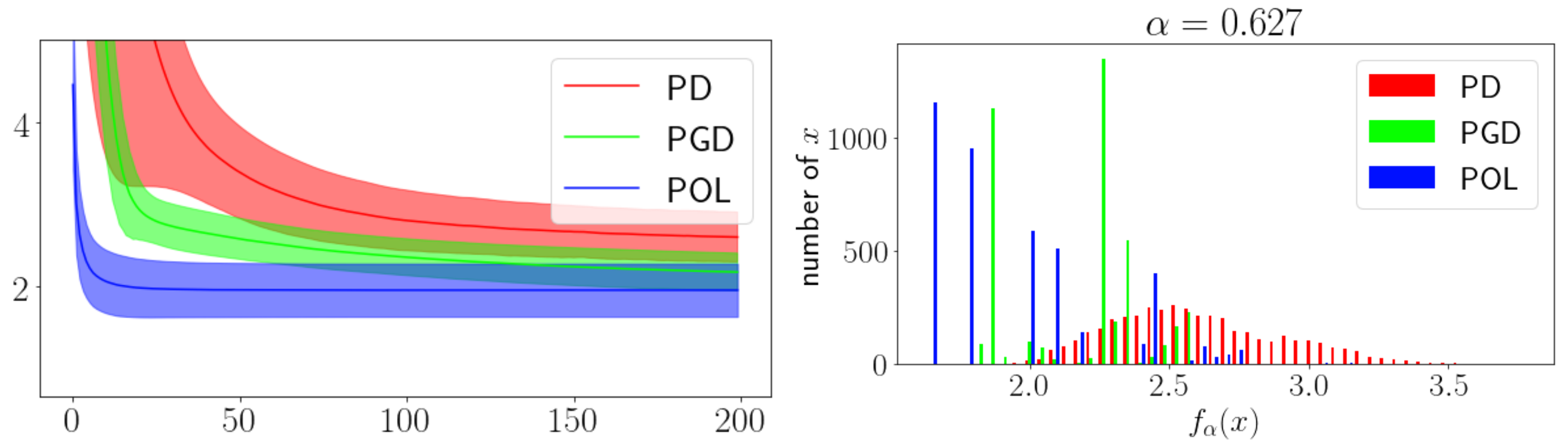


Figure 1. Results for $p = 1/2$. Objective (left) and histogram of found minima (right). PD: gradient descent. PGD: proximal gradient descent. POL: proximal operator learning (proposed).

Application: 3D symmetry detection

Given a surface $\mathcal{M}_\tau$ in $\mathbf{R}^3$, we define, for a reflection plane $x = (n,d)$ with normal n and intercept d,

$$f_\tau(x) = \mathbf{E}_{p\sim\mathcal{M}_\tau}[s_\tau(R_x(p))],$$

where R_x is the reflection transformation corresponding to x, and $s_\tau(p) := \min_{q\in M_\tau} \|p - q\|_2$ is the distance field.

Compared to existing methods that either require ground truth symmetries or detect only a small number of symmetries, our method finds arbitrary number of symmetries including continuous ones and can generalize to unseen shapes.

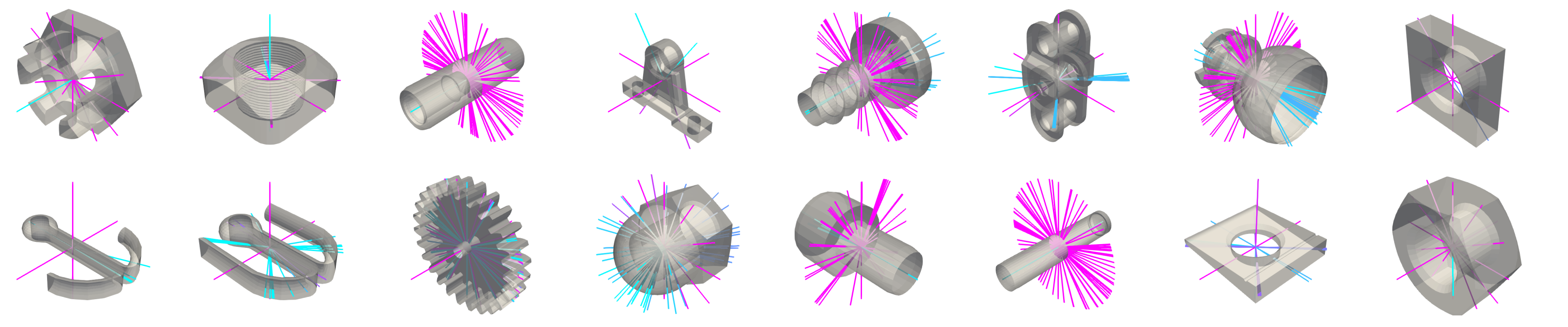


Figure 2. Each reflection is represented as a colored line segment representing the normal of the reflection plane with one endpoint on the plane. Pink indicates better objective values, while blue indicates worse.

Application: object detection in images

Applying the MSO formulation for object detection from the earlier section, we are able to encode the distribution of bounding boxes in the learned proximal operator without needing to predict confidence scores or a fixed number of boxes, unlike existing methods.

METHOD	WD_∞	$\mathrm{WP}_\infty^{0.1}$	PRECISION	RECALL
FRCNN(.80)	**0.140**	**0.624**	0.778	**0.650**
FRCNN(.95)	0.162	0.589	**0.887**	0.515
FN	0.161	0.481	0.139	**0.577**
GOL	0.251	0.243	0.508	0.282
POL (OURS)	**0.149**	**0.590**	**0.817**	0.442

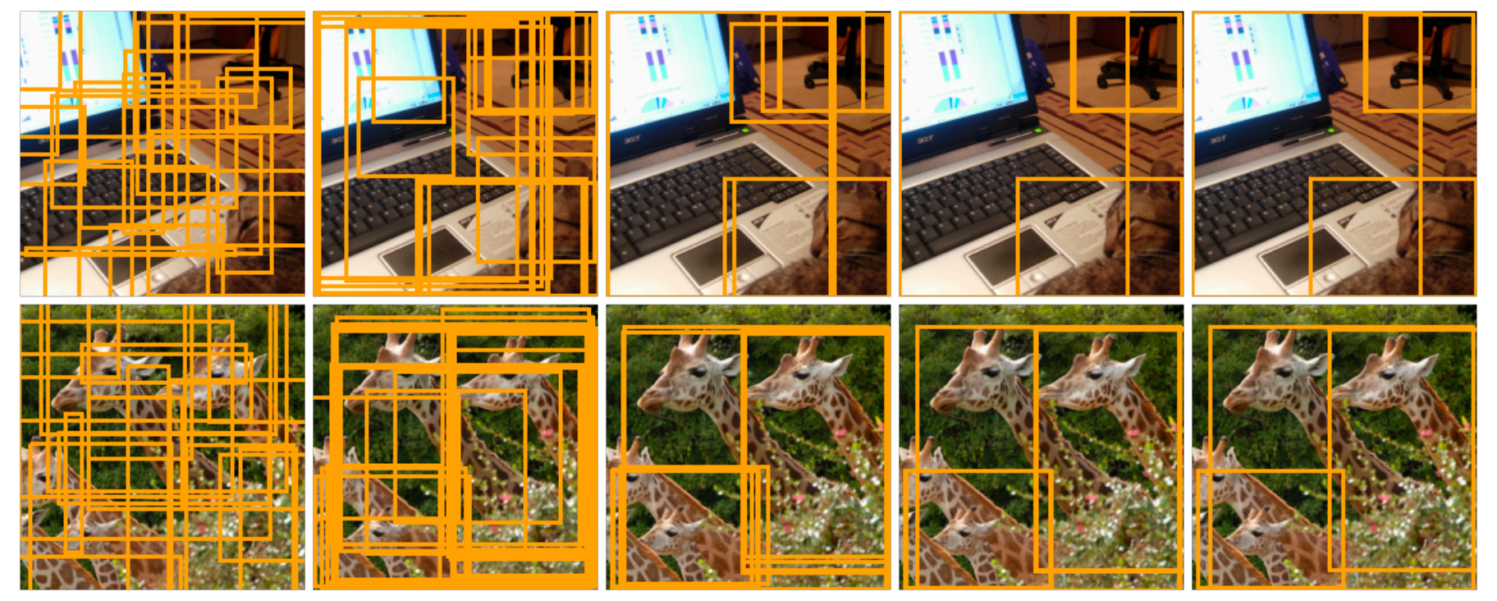


Figure 3. Left: metrics compared to baselines and Faster R-CNN. Right: First 4 iterations of PPA using the learned proximal operator on 20 randomly initialized boxes (leftmost column). Only a few iterations are needed for the boxes to form distinctive clusters.

Compared to the Faster R-CNN [5], we achieve slightly worse results with 39.7% fewer network parameters. While Faster R-CNN contains highly-specialized modeuls, we simply feed the image feature vector output by ResNet-50 to the proximal operator network. Incorporating regional information in our framework is a future direction.

References

[1] Rick Chartrand and Valentina Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(3):035020, 2008.

[2] Laming Chen and Yuantao Gu. The convergence guarantees of a non-convex approach for sparse recovery. *IEEE Transactions on Signal Processing*, 62 (15):3754–3767, 2014.

[3] Tim Hoheisel, Maxime Laborde, and Adam Oberman. A regularization interpretation of the proximal point method for weakly convex functions. *Journal of Dynamics & Games*, 7(1):79, 2020.

[4] Kenji Kawaguchi and Jiaoyang Huang. Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 92–99. IEEE, 2019.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[6] R Tyrrell Rockafellar. Advances in convergence and scope of the proximal point algorithm. *J. Nonlinear and Convex Analysis*, 2021.