We thank the reviewers for their detailed and thoughtful comments. We are encouraged that all reviewers (R1 - R4) find our double over-parameterization approach for robust recovery problems to be novel and appreciate our theoretical analysis of the gradient flow dynamics associated with the proposed formulation. The reviewers think our work help advance the understanding of over-parameterization and implicit bias of gradient descent (R2) which may bear insight into other related works (R1). Moreover, the reviewers find the experiments to be illustrative (R2) and have demonstrated the relative strength of our approach over other unsupervised learning methods (R4).
All minor comments and corrections will be addressed in the final version. Implementation details can be found at line 249 and our submitted code. In the following, we address each reviewer's comments in detail one by one.

Response to Reviewer 1.

- *Q1: Natural images may not have low-rank structures.* A1: We did not model natural images by low-rank structures. Rather, we modeled natural images by untrained deep networks following the work of DIP [34] (see line 46).
- *Q2: Denoising performance for other types of noise.* A2: This paper aims to provide a new framework for dealing with overfitting and parameter tuning for robust learning in over-parameterized models, and our exposition adopts sparse noise modeling *only* as a proof of concept. This opens new ways to handle other types of noise by redesigning the over-parameterization term for noise accordingly, which is certainly a topic of interest for future work.
- *Q3: Needs practical guidance on the learning rate selection.* A3: One strength of our method is precisely that it does *not* require a case-by-case selection of learning rate (see lines 11, 95, 218, 264).
- *Q4: Needs a brief discussion of the impact of sampling rate.* A4: If by "sampling rate" the reviewer meant the sparsity level of the corruption term $\boldsymbol{s}_\star$, its effect is demonstrated in our experiments (see line 241, 264). Or, if the reviewer meant the sampling rate for the robust matrix recovery problem, we proved that it is the same as that of the convex optimization approach which is information-theoretically optimal (see line 168).
- *Q5: More numerical comparisons with SOTA should be included.* A5: The main purpose of the paper is to address the issue of overfitting. Nonetheless, our experiment already demonstrates superior performance when compared with DIP - the SOTA unsupervised method. We will add more comparisons in the full version.
- *Q6: Computational time comparison and convergence behavior discussions are missing.* A6: The running time of our method is comparable to DIP. The convergence behavior is discussed in line 258 and illustrated in Fig. 1 & 5.

Response to Reviewer 2.

- *Q1: Questions on commuting measurements.* A1: We adopt the commutative assumption to simplify the analysis. One example of commuting measurements is when $\{\mathbf{A}_i\}_{i=1}^n$ are symmetric and jointly diagonalizable. Nevertheless, there are both empirical [5,7] and theoretical [6] evidence showing that this assumption is not necessary. We leave the study of recovery under more generic assumptions as interesting future work.
- *Q2: Commonalities and differences of the proof to [7].* A2: The work of [7] only handles low-rank matrix recovery while our work handles both sparse and low-rank recovery. Although our proof follows a similar procedure as that in [7], our contribution is on handling additional sparse terms and characterizing the discrepant learning rates in the verification of the dual certificate. We will clarify this in the final version.
- *Q3: Uniqueness of solution.* A3: Thanks for the reference. We will cite it and add a discussion in the final version.

Response to Reviewer 3.

- *Q1: Conflicting assumptions on the learning rate τ between theory and practice.* A1: To simplify the analysis, we worked in a asymptotic setting where $\tau \to 0$. In experiments (see Sec. 4.1), we showed that our method converges non-asymptotically with a reasonably small learning rate (e.g., $\tau = 10^{-4}$). While we agree that there is a gap between theory and practice, we believe that it can be addressed (e.g., by adapting the proof in [6]) and leave it to future work.
- *Q2: Correctness of proof to Theorem 1.* A2: We appreciate the reviewer's efforts in reading into the proofs and pointing out a typo. Indeed, we intended to use the KKT as a *sufficient* condition (which holds by Slater's condition).

Response to Reviewer 4.

- *Q1: Novelty and significance.* A1: Our method is *not* a trivial or naive combination of low-rank and sparse parameterizations. As explained in Sec. 2.3, it is crucial to design a proper learning algorithm, by means of implicit bias of discrepant learning rates, to obtain correct recovery. With such a learning framework not only did we provide theoretical justification but also demonstrated its good performance. We believe that this framework could have broad implications beyond the robust matrix and image recovery problems: for many other problems in (deep) learning (such as with label noise), it could help us to design scalable and principled ways to deal with epochwise overfitting.
- *Q2: Experimental comparisons.* A2: 1) Reason for using DIP as a baseline. Our method is an unsupervised approach that performs single image denoising, and DIP is the best performing method in this category. We believe by "deep learning approaches" the reviewer refers to supervised methods such as [32], which require extra training data and cannot serve as a fair baseline. We will clarify this in the final version. 2) Fairness of comparison with DIP. Our comparison is more than being fair: we granted DIP the privilege of using the ground truth clean image to determine the best early termination, while only took the results for our method at its convergence that requires no access to the ground truth. We have provided the code for the interested readers to verify our results.