We thank the reviewers for the feedback and will address their concerns in the following.

Motivation and significance of our research direction: (reviewer 1,2,3)
Below we present a result obtained on the realization space, which shows that for sufficiently large architectures all local minima of a regularized neural network optimization problem are almost optimal. We then use inverse stability to translate this result to the practically relevant parametrized problem. As suggested by reviewer 1 we will include this in the camera-ready version.
For the following we fix a depth L as well as input/output dimensions N_0, N_L. We denote by $\mathcal{A}$ the set of all architectures $N = (N_0, N_1, \dots, N_L)$ with this depth and these input/output dimensions, and by $\mathcal{P}$ the set of all parametrizations with architecture in $\mathcal{A}$. Let $(X, \|\cdot\|)$ be a Banach space with $\mathcal{R}(\mathcal{P}) \subseteq X$ and let $\Lambda\colon X \mapsto \mathbb{R}_+$ be a quasi-convex regularizer. Define $S := \{f \in X : \Lambda(f) \le C\}$ and assume that S is compact in the $\|\cdot\|$-closure of $\mathcal{R}(\mathcal{P})$. We denote the sets of regularized parametrizations by $\Omega_N := \{\Phi \in \mathcal{P}_N : \Lambda(\mathcal{R}(\Phi)) \le C\}$ and consider a convex and c-Lipschitz loss function $\mathcal{L}$ on S (note that this is fulfilled for virtually all relevant loss functions).

Theorem 1 (Almost optimality of local realization minima)**.** *For all $\varepsilon, r > 0$ there exists $n(\varepsilon, r) \in \mathcal{A}$ such that for every $N \in \mathcal{A}$ with $N_1 \ge n_1(\varepsilon, r), \dots, N_{L-1} \ge n_{L-1}(\varepsilon, r)$ the following holds:*
Every local minimum h_* with radius at least r of the optimization problem $\min_{h \in \mathcal{R}(\Omega_N)} \mathcal{L}(h)$ satisfies

$$\mathcal{L}(h_*) \le \min_{h \in \mathcal{R}(\Omega_N)} \mathcal{L}(h) + \varepsilon.$$

Proof. Let $\eta := \min\left\{\frac{r\varepsilon}{2c\,\mathrm{diam}(S)}, \frac{r}{2}\right\}$. Due to compactness of S there exists $n(\varepsilon, r) \in \mathcal{A}$ such that for every $N \in \mathcal{A}$ with $N_1 \ge n_1(\varepsilon, r), \dots, N_{L-1} \ge n_{L-1}(\varepsilon, r)$ it holds that $\sup_{f \in S} \inf_{\Phi \in \Omega_N} \|\mathcal{R}(\Phi) - f\| \le \eta$. Let $h \in \mathcal{R}(\Omega_N)$ and define $\lambda := \frac{r}{2\|h - h_*\|}$ and $f := (1-\lambda)h_* + \lambda h \in S$. By the assumptions on h_* and $\mathcal{L}$ it holds that

$$\mathcal{L}(h_*) \le \mathcal{L}(\mathcal{R}(\Phi)) \le \mathcal{L}(f) + c\eta \le (1-\lambda)\mathcal{L}(h_*) + \lambda\mathcal{L}(h) + c\eta.$$

Direct computation establishes the claim. □

Now inverse stability is necessary (see Example A.1 in the paper) and sufficient (see Prop. 1.2 in the paper) in order to get the following corollary for the parametrized problem.

Corollary 2 (Almost optimality of local parameter minima)**.** *Assume that the realization map is (s, α) inverse stable on Ω_N w.r.t $\|\cdot\|$ for every $N \in \mathcal{A}$. Then for all $\varepsilon, r > 0$ there exists $n(\varepsilon, r) \in \mathcal{A}$ such that for every $N \in \mathcal{A}$ with $N_1 \ge n_1(\varepsilon, r), \dots, N_{L-1} \ge n_{L-1}(\varepsilon, r)$ the following holds:*
Every local minimum Γ_* with radius at least r of the optimization problem $\min_{\Gamma \in \Omega_N} \mathcal{L}(\mathcal{R}(\Gamma))$ satisfies

$$\mathcal{L}(\mathcal{R}(\Gamma_*)) \le \min_{\Gamma \in \Omega_N} \mathcal{L}(\mathcal{R}(\Gamma)) + \varepsilon.$$

Proof. $\mathcal{R}(\Gamma_*)$ is a local minimum with radius $(\frac{r}{s})^{1/\alpha}$ (Prop. 1.2 in the paper) and Theorem 1 implies the claim. □

Note that it is important to have an inverse stability result, where the (s, α) does not depend on the size of the architecture, which we achieve in our submission for $L = 2$ and $X = W^{1,\infty}$. Suitable Λ would be Besov norms which constitute a common regularizer in image and signal processing. Moreover note that the required size of the architecture in Theorem 1 and Corollary 2 can be quantified, if one has approximation rates for S. In particular, this approach allows the use of approximation results in order to explain the success of neural network optimization and allows for a combined study of these two aspects, which, to the best of our knowledge, has not been done before. Unlike in recent literature, our result needs no assumptions on the sample set (incorporated in the loss function, as shown in the paper), in particular we do not require 'overparametrization' with respect to the sample size. Here the required size of the architecture only depends on the complexity of S, i.e. the class of functions one wants to approximate, the radius of the local minima of interest, the Lipschitz constant of the loss function, and the parameters of the inverse stability.

Extension to deep architectures and multiple output units: (reviewer 2)
We have the extension to multiple output units worked out (it requires adjusting the notion of balancedness but otherwise follows directly using techniques from the paper) and will include it in the final version. While a full extension to deep networks would exceed the scope of the submission, we will add a discussion of the challenges and possible solutions.

Motivation for the non-degeneracy conditions of Theorem 3.1 and contribution to regularization: (reviewer 3)
We would like to highlight the following additional merit of the study of degenerate parametrizations. Without inverse stability it is possible to get stuck in a local parameter minimum that is not a realization minimum, which could be prevented by regularizations designed to exclude the pathologies we found. In order to make the technical conditions in Theorem 3.1 in the paper more palatable we will link them to practical methods of regularization. Specifically, we will cite and comment on recent NeurIPS resp. ICLR papers where the authors achieved promising empirical results by using a regularization term (e.g. based on cosine similarity) corresponding to conditions C.2 and C.3 of the theorem.

We think this addresses all the points of criticism, regarding motivation, results on neural network optimization, and extensions to (deep) architectures with multiple output units and we will include these improvements in the final version.