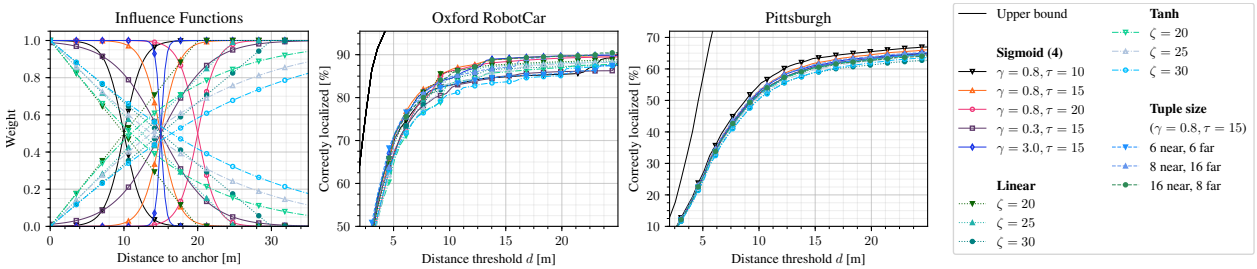

Figure 1: Top-1 localization accuracy on Oxford RobotCar (mean over four conditions) and Pittsburgh as a function of the distance threshold d using our loss (5) for different influence functions and number of near/far samples per tuple.

We thank all reviewers and ACs. We appreciate the many positive and constructive comments. Here, we first address common concerns followed by reviewer-specific answers. We will update our paper and supp. material accordingly.

Sensitivity (R2,R3). Our results for tanh and (4) differ because (4) uses slope γ and offset τ while for tanh we use $y_n = \tanh \frac{x}{\zeta}, y_p = 1 - y_n$. Thank you for pointing out this missing detail. Following your suggestion, Fig.1 shows that our method is not very sensitive with respect to the influence function and the number of near/far samples per tuple.

Difference to MS [28] (R1,R4). In contrast to our method, Multi-Similarity (MS) loss [28] requires hard class assignments. Unfortunately, [28] was lost from the citations in L80, which may have made this unclear. The sum-log-exp in (5) and [28] is a commonly used loss function. Our inspiration to use such a loss is based on [28] being a strong baseline. In [28], the feature similarity kernels of positives and negatives are summed up in separate terms. To avoid such a separation (which our paper argues to be unnatural for the task of localization), our loss weighs each feature similarity with an influence function based on the geometric distance between images, resulting in our positiveness and negativeness scores described in equation (3), which are then used in (5). To summarize, our loss uses geometric distance weighting while [28] requires class labels, making our loss a more natural choice for the task of localization.

R1. Dataset contribution: We will move this from the main list of contributions to the experiments section. Still, we believe that our curated dataset is valuable for the community and will therefore make it publicly available along with the source code. Tab.2: We will add color coding, improve the wording of L267f, and correct the erroneous references.

R2. Mathematical complexity: The proposed method is indeed easily comprehensible, even without the mathematical details. We will guide readers, who are only interested in a general understanding, to skip mathematical details. Additionally, we will add an intuitive explanation similar to yours. Nevertheless, we believe that the multi-objective formulation is important for an in-depth understanding of our interpretations. Evaluation: Tab.2 uses public evaluation benchmarks (i.e. CMU Seasons and Robotcar Seasons). Therefore, these results can be directly compared to the literature. Please note that "RobotCar Seasons" and "Oxford RobotCar" are different (Fig.2 in our paper). Only the much larger latter is cleaned by us. Tables and L267: Please, see R1. Tab.1: The mean of Oxford RobotCar will be added. Arbitrary thresholds: Thresholds based on system requirements are indeed not arbitrary. We will put this into perspective. Localization vs. place recognition: The distinction between (ii) and (iii) on L71 is based on the experimental setups and the underlying theoretical implications. Also, [16] refers to itself as a place recognition method.

R3. Viewpoint: We only consider positives with less than 30° yaw difference to the anchor (L232). This allows us to learn from difficult positives (e.g. night), which are dropped by [16] whenever a tuple has more than one true positive.

R4. Prop.3.3: To avoid discarding images between τ_1 and τ_2 (L122), one may choose $\tau_1 \to \tau_2$. Prop.3.3 shows that this choice contradicts with $\alpha > \epsilon$ in Prob.3.2. Even if there is always a small gap between τ_1 and τ_2 in finite datasets, α may need to become very small for feasibility, which causes the notion of contrastiveness to disappear in Prob.3.2. Sec.3.2 as motivation: The clarification of Prop.3.3 should resolves this issue. L131: The order is not preserved when $\alpha = 0$. Ordinal regression vs. binary classification: We agree. The problem of binary classification directly relates to the hard assignment of images into positive/negative (binary) classes. Therefore, our claim aligns with your analysis. Learning to rank/order preservation: We are not aware of any learning method that *ensures* order preservation for the problem at hand. Most methods attempt to maintain ranking, including ours (L146). Our Prop.4.2 sheds light on this. Note that every feasible solution of Prob.1 is guaranteed to preserve the order. On the other hand, minimizing (5) or similar, in an attempt to minimize the multi-objective function, provides a solution from the pareto optimal front. Since the pareto optimal front passes through the feasible set of Prob.1 (Prop.4.2), it is possible that for some settings of (5), the obtained solution also preserves the desired order (due to being feasible for Prob.1). Please, see L144. Sec.4 in relation to Sec.3.2: We show in Prop.4.2 that our multi-objective formulation may lead to a feasible, order preserving solution of Prob.1. The formulation of Prob.3.2, on the other hand, is problematic in the context of visual localization as shown in Prop.3.3. We will add these high-level connections, to further improve the clarity of our paper. Comparison to [31]: Unlike our method, [31] does not use soft assignments (L77). Their negative visual loss uses hard assignments. Their visual-geometric loss uses geometric distances but is only applied to positives (L61).