We sincerely thank all the reviewers for their constructive feedback. Please find our answers to the key questions raised by each reviewer (**R**) below.

First and foremost, we would like to emphasize that the focus of our work is to provide a novel regularized OT scheme that provably preserves the class structure. To our knowledge, our work is the first attempt to provide comprehensive theoretical guarantees for such a problem. We also provide a novel optimization algorithm tailored to the complexity of our OT scheme, for which we also provide a theoretical convergence analysis. The experiments in our paper are to support our theoretical results and provide additional insights. Hence in our experiments, we focus on the techniques for which our theoretical insight is relevant. In particular, we do not consider the studies which also consider feature learning, as they fall beyond the scope of our theoretical framework.

Relation to prior work (**R1, R2, R3, R4, R5**): As pointed out by all reviewers, especially **R3, R4** and **R5**, we propose a new stochastic algorithm for solving a *convex* optimization problem and provide rigorous theoretical guarantees for the rate of convergence and uniqueness of the solution. Papers [1] and [3] mentioned by **R1** are mostly experimental, and do not contain theoretical analysis of their algorithms. The "Structured Optimal Transport" paper, which uses the notion of submodularity, does not employ stochastic optimization methods, and thus may not be as scalable as our framework. The performance of our algorithm can be improved by incorporating modern feature learning methods (like in deep DA and Unbalanced OT suggested by **R4**), however, the analysis of such methods in our framework is intractable . On the other hand, the choice of such features is orthogonal to our framework, i.e, they can be combined in a consistent way, which is an interesting subject of a future study.

Misunderstanding of a related work (**R1**): We will revise this sentence in the final version. The complete sentence in our current version is as following: "Note that the Laplacian regularizer in [Courty et al., 2017] acts only indirectly on the transported points and is quadratic, whereas ours acts directly on the transport plan and is of l1 type". Our focus here is on the type of the regularizer used. The l2 norm used in the Laplacian regularizer is quadratic and has a simplifying effect on the terms, while the l1-type regularizer in our framework acts in an element-wise manner.

Choice of kernels (**R1**): The effect of different kernels in our algorithm is not part of the theoretical analysis. In section 8 of the supplement, we analyze our proposed method for general kernel coefficients. However, we agree that different choices of kernels can significantly affect the performance of our method in different applications. Thus, a proper choice is application-dependent. Again, this aspect is beyond the scope of our study.

Stochastic block model (**R1**): The stochastic block model is commonly used for theoretical analysis, as it makes the analysis approachable and is closest to real-world applications.

Experiments (**R1, R2, R4**): The setting used for experiments in section 5.1 induces an imbalanced number of classes in the two domains which is more common in practice. By these experiments, we argue that our method addresses this scenario better than the alternatives. In particular, it prevents the samples from a source class that does not have a target counterpart from being "difussed". This ability of our method is shown in the experiments of Figure 3 in the supplement and we observe its superiority in the respective real-world experiments. To be consistent with the experiments in [Courty et al., 2017a], we use resized images of size 16×16 for the MNIST dataset. Regarding the low accuracy results in our real-world experiments, note that we do not have any training samples in the target domain, and removing classes from target makes the problem more complicated. Therefore, we expect a lower accuracy in this setting compared to the balanced case. The reported accuracies in tables 1 and 2 (MNIST-USPS and Caltech Office) are the *best* results achieved when using different values of regularization parameters (we will clarify this in the final version). Additionally, it is well known that the final result does not change much in a wide range of λ_1 and λ_2. We observe this fact in our experiments as well. Regarding the error bars, please note that as we show in the theoretical analysis, our stochastic algorithm converges to a *unique* solution at the end of optimization and this is indeed observed in our experiments as well. Therefore, there is no variation in the final solution of different runs (which is an advantage of our method).

Resistant optimal point (**R1**): Our definition of a resistant optimal point is consistent with the notion of Stability of Optimal Solutions but it is more general, and in this sense it is novel. Note that in the conventional optimization literature, stability is provided by strong convexity. Our optimization framework lacks strong convexity. Our contribution is to show that a weaker form of stability, which is present in our framework, is sufficient to guarantee uniqueness.

Partial domain adaptation (**R2**): We agree with **R2** that our method is applicable to the PDA problem, but our aim is to solve a more general case without the restrictive assumptions in PDA.

Class structure (**R5**): By class structure we mean the latent random variable that partitions data points into separate clusters. Our framework does not find these clusters but it aims at preserving the clusters in the transported data points in the target domain.

Finally, we appreciate other comments from the reviewers. We will apply them as much as possible to the final version of the paper.