We thank the reviewers for helpful comments and suggestions. We will address the concerns raised by the reviewers.

(R1:Q1) On using our framework for learning from label proportions (LLP).

(R1:A1) Our proposed framework is applicable for tackling learning from label proportions, even for the multiclass case, by using class proportions as aggregated labels. However, due to space limitation and the fact that LLP has been explored extensively, we would like to focus on other problem settings to expand the usage of learning from aggregate observations and provide the theoretical foundation of a more general case. Nevertheless, we agree that it is interesting to see the performance of this framework compared with other LLP methods to see the competency of our framework. We will add more explanations of LLP, potentially in Appendix due to lack of space.

(R1:Q2) Baselines are quite weak in the experiments, however I note that there might not be too much related work.

(R1:A2) As pointed out, related work that can be used as baselines for our experiments are quite limited. For example, we are not aware of any methods for multiclass classification from triplet comparison data. We tried to come up with several baselines and found that a representation learning method is reasonable and its performance is quite reasonable. It worked quite well in the pairwise comparison case but failed to work well in the triplet case which might be because more data are needed. For regression via mean observation, [1] is the most related as suggested. The difference is that [1] used Gaussian processes and variational inference. We believe both frameworks have different advantages/disadvantages such as the variety of model choices, scalability, or the uncertainty measure. We will add such discussion in the final version.

(R1:Q3) Why linear regression is used as one of the methods in the experiments?

(R1:A3) As R1 suggested, we can use any differentiable model. Linear regression was used because it is one of the standard models for regression on these datasets. Moreover, it is insightful to see the difference in performance between a linear model and a more complex model with the same objective function. Thus, we implemented the proposed objective and the baseline objective on both the linear model and a gradient boosting machine (GBM). We will add more discussion on the choice of models.

(R2:Q4) How does this setup differ from the weakly labeled setting?

(R2:A4) Our problem setting can be regarded as a weakly-supervised learning problem, where only a group-level label is observed although we want to predict a label for an individual instance. It is different from many weakly-labeled settings in the literature (e.g., partial labels, complementary labels, positive-unlabeled learning, noisy labels) in the sense that individual labels are given in those settings although they are weak (i.e., not clean fully-supervised).

(R2:Q5) On how to improve paper's presentation

(R2:A5) Thank you. We will provide explanations to give key ideas how to interpret our results and why they are useful.

(R3:Q6) On the practicality of Assumption 2

(R3:A6) We admit that it is possible that Assumption 2 is violated in real-world problems. Thus, it is interesting to relax Assumption 2 and investigate the situation when this assumption does not hold. For example, we may try to explore a new framework that relies on another assumption that is more practical in some settings. Then, a practitioner can select an appropriate method depending on their problem of interests. We believe there are many issues to be discussed when going beyond this assumption and it is a good future direction. we will discuss these issues as a future work in the final version.

(R4:Q7) Why the proposed method is much better in classification from triplet comparisons?

(R4:A7) One explanation is we might need much more data to learn a reasonable representation compared with simply learning a probabilistic classifier to separate between classes. In Appendix E, we also showed the performance in the binary classification task and found that the baseline can be quite competitive for some datasets. But in the multiclass cases especially when the number of classes is quite high, baselines become much weaker than the proposed method. We will add more discussion in the final version.

(R4:Q8) What are some specific/practical examples where the triplet comparison is given?

(R4:A8) Examples include the sensor network problem and search engine query logs, which were discussed in [12]. Triplets has been used a lot for representation learning but not classification (maybe due to lack of methods). We will include this issue in Introduction.

(R4:Q9) How did you select multiple samples that are aggregated in the experiments?

(R4:A9) It was randomly selected. We are aware that this way may not be ideal. To make up for that, we did experiments on many datasets (59 datasets including Appendix E). We will add more explanations in the final version.