Thanks for the helpful feedback. We will address the concerns.

R1. **Q1**. No theoretical proof and explanation. **A1**. Just the lack of proof or theoretical explanation should not be a reason to reject a paper for NeurIPS. Domain adaptation papers without proofs have been accepted by NeurIPS (e.g., Domain Separation Networks, NeurIPS'16 and Transferable Normalization NeurIPS'19). Despite the existence of theoretical papers for domain adaptation, they are not useful for the universal domain adaptation because they do not provide good insight on how to deal with open categories.

R2. **Q1**. What if the known target sample is closer to other targets than to prototypes? **A1**. It depends on how close the target sample is and how many other target samples are nearby. Since we have both ES and NC loss, the target sample can be aligned to the source prototype even if the target is nearer to other targets than the prototype.

R3. **Q1**. Losses are ad-hoc. More intuition for each loss. **A1**. ES is a carefully designed pseudo-labeling loss giving "known" or "unknown" label to target samples. We need to decide whether a sample is "known" or "unknown". Importantly, we do not even know whether we have "unknown" samples in target domain for the universal domain adaptation. Then, even though there are many "unknown" samples or none of them, the objective function for "unknown" needs to work well on both scenarios. The entropy of a classifier output shows the confidence of the prediction. Large entropy implies that the classifier is uncertain about the prediction for the sample and the value intuitively implies a distance from source classes. Such distance should be effective metric for "unknown" score under different proportions of "unknown" samples. We assume that target samples with entropy smaller than a threshold are all known samples while other samples are unknown ones. The entropy of classifier output gets larger in the log scale when the source domain has many more classes. Therefore, the threshold of the entropy (ρ) is set larger with the scale ($\frac{\log(C)}{2}$). We try to select only confident samples using confidence value m. NC forces each target sample to be closer to its neighbors, which results in discriminative features. For example, if the nearest neighbor of sample A is B while that of B is C, all A, B, and C can be put closer. Of course, NC may not form very compact clusters as shown in Fig 3 (d), it does not require to know the number of classes in the target and is suitable for universal domain adaptation. **Q2**. Effectiveness of each module NC, ES. **A2**. Table A, Table 7, and Fig. 3(c) vs (d) show the ablation of NC and ES. Ablation study in Table A (left) corresponds to Table 6 (paper), where we classify unknown samples into their original class given a fixed feature extractor and one labeled target sample per class. To perform well, features have to be well-clustered. We provide DANCE w/o ES and DANCE w/o NC results. The results show that NC extracts well-clustered features for unknown classes. Only with ES (w/o NC), the accuracy is worse or comparable to source-only (SO). Fig. 3(c) vs (d) supports the observation too. Left of Table A shows the ablation of open-set DA for VisDA. From this results, ES is effective for both alignment of known samples and rejection of unknown samples. Combining ES and NC further boosted performance. Table 7 (paper) shows that NC is not enough to ensure the performance since it does not consider the assignment of each sample to source class. **Q3**. Comparison between DANCE and other approaches is not apples-to-apples. **A3**. The main result is summarized in Table 1 (paper), where DANCE performs best in terms of averaged rank. As existing baselines are tailored to specific category-shift, we first perform "universal comparison" where we do not have any prior knowledge on the type of category-shift. This "universal comparison" between ours and SO, DANN, ETN, STA, UAN is fair in that the hyper-parameters and checkpoints are validated in the same way (see "B. Implementation Detail" in the supp. for details). We show that while the category shift settings we evaluate are all different and not directly comparable (and have specially designed algorithms), our method consistently has the best performance (ranked first or close to first), despite not being specifically tuned for each setting. This is a very powerful advantage in real-world settings.

Method	D to A	W to A	R to P	R to C	VisDA	VisDA (OS/UNK)
SO	53.5	54.1	58.2	20.7	72.3	43.3 / 28.5
w/o NC	52.0	51.6	59.5	20.3	73.1	54.7 / 36.0
w/o ES	**57.5**	**58.1**	**64.5**	**23.1**	75.1	60.2 / 52.2
DANCE	**57.5**	57.4	63.5	22.8	**75.2**	**65.2 / 79.8**

Table A: Left: Analysis of ability to cluster "unknown " class samples given one labeled sample per class. Right: Ablation in open-set DA for Visda. UNK shows the accuracy to reject unknown class samples.

R4. **Q1**. Straight forward method. NC is not new. **A1**. We would like to correct the misunderstanding. **NC is a new approach, which is totally different from "Memory-based neighborhood embedding for visual recognition". The mentioned method is basically for supervised learning or few-shot learning and does not have a module to handle unlabeled samples.** They attempted to aggregate the neighbourhood information for the discriminative embedding. By contrast, NC does not have a module to aggregate the information. It tries to make neighboring unlabeled target samples closer and calculates the loss in an unsupervised way. **Q2**. Cost of memory? **A2**. Storing features in memory does not add much cost. All experiments were done with a single 12GB GPU. If we have many more samples, we can limit the number of samples to store. **Q3**. Missing reference in Table 2-6. **A3**. We will add the reference in Table 2-6. **Q4**. Effect of domain-specific batch normalization. **A4**. The technique we used is exactly the same as [3]. The effect is shown in Table 8 of their paper (maybe better to show some results.). [1] "Memory-based neighborhood embedding for visual recognition". [2] "Unsupervised Feature Learning via Non-Parametric Instance Discrimination". [3] "Semi-supervised domain adaptation via minimax entropy"