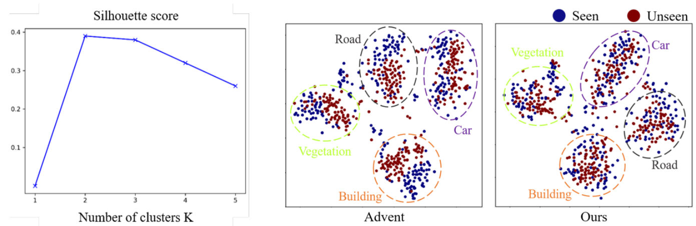
We thank all the reviewers for their constructive reviews. We answer each question below.



(a) Clustering Evaluation (b) T-SNE Visualization

Figure 1: Additional analysis.

Novelty & Contribution We carefully design a unified OCDA framework for semantic segmentation. While some components adopt existing methods, these are well combined in a novel (task-specific) way as also noted by [R3]. We provide extensive ablation studies to verify the individual contribution of three complementary principles. We finally achieved new state-of-the-art OCDA performance. We believe our findings and results can benefit the communities and practitioners.

R1: Additional cost of adopting multi-stage training and multiple discriminators Compared to the baselines, our DHA framework slightly increases the memory usage and computation at the training time. Specifically, the total training time of [31], [33], [*], and ours are 33.8hr, 34.1hr, 61.2hr, and 64.7hr, respectively. The according final performances are 28.8, 29.1, 29.5, and 32.0. This implies that our framework brings significant performance improvement with the moderate computational cost increase during training. We note that the test time costs are all the same, as we only utilize an identical segmentation model.

R1: Why multiple discriminators? To explicitly capture the underlying multi-mode structures in the data, we adopt using multiple discriminators. Our strong empirical results backs our design choice.

R1,R3: Effectiveness of the hallucination step (Table 1-(b)) We apologize for the incorrect notations in the main paper Table 1-(b). As noted in the main paper (see section 3.3 Framework Design), we learn target-to-source alignment using multiple discriminators in Method-(1). Thus, the '+trad' must be replaced with /check. In fact, to see the effect of hallucination step, we should compare the result of the source only and Method-(2) or traditional UDA and Method-(3). The clear improvement demonstrates its efficacy.

R1: Baselines with longer training scheme The followings are the results: [Ours 32.0 / ADVENT 29.1 / Adaptseg 28.8 / CRST 26.9 / CBST 26.7 / Source-only 25.7]. Our framework acheives the best result. We note that the result of [19] is not included, since the official code (for semantic segmentation) is not available currently.

R1,R4: end-to-end training The end-to-end training causes the model to diverge.

R1,R2,R4: Issues in the hyperparameter K If K value is much less than the optimal, the target distribution might be oversimplified, and some latent domains could be ignored. On the other hand, the images of similar styles might be divided into different clusters, and also each cluster may contain only a few images. In this work, we have set the value of K empirically. Instead, we see one can set the value using existing cluster evaluation metrics such as silhouette score. It evaluates the resulting clusters by considering the intra-cluster variation and inter-cluster distance at the same time. As shown in the Fig. 1-(a), K=2 and 3 are the strong candidates, and the quality of clusters drops after K=3.

R1,R2: Applying DHA framework on UDA setting (GTA5 to Cityscapes) The followings are the results: [Ours 36.7 / ADVENT 36.1 / Cycada 35.4 / Adaptseg 35.0 / CBST 30.9]. Our framework achieves the best result.

R2: DHA framework with the ResNet backbone The followings are the results: [Ours 37.2 / ADVENT 36.0 / Adaptseg 36.2 / CRST 36.4/ CBST 35.8 / Source-only 35.7]. We achieve state-of-the-art again.

R2,R4: T-SNE visualization We analyze the feature space learned with our proposed framework and the advent baseline in the Fig. 1-(b). In appears that our framework yields more generalized features. More specifically, the feature distributions of seen and unseen domains are indistinguishable in our framework while not in advent.

R3: Hallucinate in opposite direction We rather observe degraded performance due to the undesirable translation. It is mainly because the semantic labels do not exist and the styles are diverse in the target domain.

R3: Quantitative analysis on style consistency loss In the main paper Table 2-(a), we already provided the quantitative ablation results (ours vs. TGCF-DA).

R3: Figure 1-(c) modification As suggested, we will modify the figure in the final version.

R1,R3,R4: Missing implementation details For the fair comparison, we use same discriminator, learning rate, optimizer, and train/test-time image resolutions with [31,33]; To compute the feature statistics in the "Discover" step, we use vgg-16 relu1_2.

R4: How the DHA framework can deal with the open domain? Our framework aims to learn domain-invariant representations that are robust on multiple latent target domains. As a result, the learned representations can well generalize on the unseen target domains by construction (please also refer to the Fig. 1-(b)). The similar learning protocols can be found in recent domain generalization studies as well.

R4: Comparison with the another strong baseline ([A-B] + [C-D]) As mentioned above, hallucinating the images in the opposite direction produces undesirable images. Therefore, even with the strong translation technique of [A], we get inferior result (27.3) compared to the source only (35.7). Not surprisingly, applying the recent adaptation method of [D] on top of this translation result did not improve up to ours (A+D 30.3 / Ours 37.2).