Table 1: Hyper-parameters study on Δ_d (GTA5→Cityscapes).

$\lambda_2 = 0.7$, $L = L^s_{CE} + \lambda_2 L^t_{CE}$							
Δ_d	1	1.5	2	2.5	3	3.5	4
mIoU	45.19	45.53	45.93	46.14	46.01	45.96	45.67

Table 2: Hyper-parameters study on λ_1, λ_2 (GTA5→Cityscapes).

$\Delta_d = 2.5$, $L = L^s_{CE} + \lambda_1 (L^s_{dis} + L^t_{dis}) + \lambda_2 (L^t_{CE} + L^{tP}_{CE})$, stage 1										
λ_1	0.3	0.5	0.7	0.9	1.1	0.3	0.3	0.3	0.3	0.3
λ_2	0.7	0.7	0.7	0.7	0.7	0.3	0.5	0.7	0.9	1.1
mIoU	47.13	47.13	47.18	47.21	47.19	47.03	47.16	47.13	47.23	47.20

We appreciate the valuable comments from all reviewers and will carefully take them into account in the final version. In this rebuttal, we focus on major concerns. The source code will be released for verification and reproducibility.

R1.1 Stage-wise training procedure makes training cumbersome. Will the results improve if trained for more stages?

⇒ We tried to train the CAG model in a single stage and update the pseudo-labels at each iteration. However, it is not stable because there are some error-prone pseudo-labels, which may produce incorrect supervision signals, lead to more erroneous pseudo-labels iteratively, and trap the network to a local minimum with poor performance eventually, $i.e.$, less than 30 mIoU. To address this issue, we used the stage-wise training procedure by fixing the anchors at each stage. It is noteworthy that we reduced the training epochs at each stage, $e.g.$, 20 epochs, so that the overall training cost is comparable to previous methods. Nevertheless, we are open to explore more efficient pseudo-labels assignment and training techniques to further accelerate the training procedure. Besides, after the CAG model is trained for an extra stage, it reaches 50.21 mIoU, which is saturated compared with stage 3.

R1.2 The shown pseudo-labels seem relatively smooth (fig. 2). Enforcing local smoothness can benefit the method.

⇒ The pixels in images are spatially coherent, and the features f_D from the penultimate layer are expected to have effective characteristics for clustering. Therefore, the pseudo-labels by the proposed CAs-based assignment are smooth. Enforcing local smoothness can be promising for enhancing the robustness of ATI or PLA and the accuracy of pseudo-labels, $e.g.$, applying CRFs on the CAs-based distance map.

R1.3, **R2.4** It would be clearer to add the explicit definition of the loss L^{tP}_{CE}.

⇒ The explicit definition of L^{tP}_{CE} is given by $L^{tP}_{CE} = -\sum_{i=1}^{M}\sum_{j=1}^{H\times W} a^{tP}_{ij}\sum_{c=1}^{C}\hat{y}^{tP}_{ijc} log\left(p^{tP}_{ijc}\right)$ similar to L^t_{CE} in Eq.(9), where a^{tP}_{ij}, y^{tP}_{ijc} and p^{tP}_{ijc} refer to the active state, assigned pseudo-labels vectors and network output respectively, which will be added in the final version.

R2.1, **R3.1** It would be more convincing if a wider range of validation on the hyper-parameters is provided.

⇒ We investigated the effect of a wider range of hyper-parameters including Δ_d, λ_1 and λ_2. The results are listed in Tables 1 and 2. We can see that the performance peaks at Δ = 2.5, and is not sensitive to the choice of Δ in the range of [2, 3.5]. For λ_1 and λ_2, the performance usually improves when one rises and the other is fixed. The performance is also stable with respect to the changes of hyper-parameters λ_1 and λ_2 in ranges of [0.3, 1.1] and [0.3, 1.1], respectively.

R2.2 Is it more reasonable to assign Δ_d in a normalized setting? The same Δ_d in both domains may not be optimal?

⇒ We used the distance-based threshold in the proposed CAs-based PLA because it is simple and easy to implement. It is noteworthy that we used the same threshold for all categories to reduce the number of free hyper-parameters. It is empirically effective and usually performs well. We did not use the threshold in the source domain, because the ground-truth labels are available. In the final version, we will take this suggestion by comparing the current setting with the normalized threshold.

R2.3 No evidence on "they turn out to be more reliable"; clarification of "they do not depend on the decision boundaries".

⇒ Due to the lack of the target domain labels, the classifier is biased to the source domain and does not generalize well to the target domain, as shown in Fig.1 (c). Consequently, some of the pseudo-labels from predicted probabilities may be erroneous. Based on the observation of the intra-category clustering characteristics, we propose the CAs-based assignment method which is independent of the classifier and the biased probabilities. Under the same setting of the experiments, using the pseudo-labels assigned by anchors achieves better performance than those assigned by predicted probabilities by a large margin. With this regard, we claim that "CAs-based pseudo-labels are more reliable and do not depend on the decision boundary". Here the decision boundary refers to the classification hyperplane formed by the output layer of the network.

R3.2 How important is the warm-up stage? What if you remove it?

⇒ The distance loss and CE loss used in Eq.(11) rely on reliable pseudo-labels to guarantee a correct supervision imposed on the network. Adding the warm-up stage can roughly align both domains (feature distributions) and increase the reliability of the pseudo-labels by the CAs-based PLA. In our experiments, we find that removing the warm-up stage leads to a significant performance drop, $e.g.$, 6.3 mIoU. Nevertheless, we agree that it is valuable (1) to explore more effective PLA and training techniques to further improve the reliability of pseudo-labels and (2) to reduce the side influence from the error-prone ones, $e.g.$, using a progressive PLA based on adaptive thresholds and combining CAG with a style-transfer module suggested by reviewers #1 and #2.