We appreciate the careful readings and constructive comments. We are encouraged by the reviewer's appreciation of the proposed effective CSRL model[R1,R2,R4] designing for the valuable generalized zero-shot semantic segmentation (GZS3) task[R3]. Moreover, all reviewers recognized the *superior* performance of our simple yet effective method.[R1,R2,R3, R4]. Here we emphasize our main contributions. First, CSRL constrains the feature generation of unseen categories by preserving relation consistency between seen and unseen categories(§4.2), which is *not* exploited by [1]. Second, CSRL exploits the class co-existence by feature and relation aggregation(§4.1). Thus we could not only better learn a feature generator, but also implicitly model the category coexistence in a scene (e.g. the 'cow' is usually on the 'grass'). Such inter-class relationship is *not* explored in classification-based zero-shot task.

[R1,R3] **Discussion about the differences w.r.t [1].** Our method implicitly applies constraints to unseen categories by exploring the relations between seen and unseen categories for the feature generation, while [1] purely employs the seen categories to learn the feature generator, leading to a poor representative ability for the generated unseen features. Specifically, beyond the point-wise consistency of seen classes as adopted by [1], our method further exploits the relations between unseen and seen classes by pair-wise and list-wise consistency (§4.2). Compared with [1], our superior performance can well demonstrate the effectiveness of relation modeling.

[R1] **The choice of temperature γ.** We discuss the effect of γ in supplementary §C. In brief γ is chosen by grid searching the highest hIoU. We experimentally find that the model is robust with γ under different unseen splits.

[R1] **Better to conduct extra evaluation on datasets such as ADE20k.** Completely agree. To fairly compare with [1], we conduct experiments on object segmentation dataset (Pascal VOC) and scene parsing dataset (Pascal Context) in this submission. However, due to the limited rebuttal period, we cannot provide the results on large=scale datasets ADE20K and COCO. We promise to include them in the updated version.

[R2] **Whether the pixels of unseen classes are used.** No. We strictly follow the zero-shot settings described in §3, thus the pixels and visual features of unseen classes are *never* used during training. In §4.1, the input nodes are the semantic word embeddings of both seen and unseen classes. The output nodes are the generated visual embeddings. A classifier is fine-tuned on these generated visual features. Thus, we can segment images with both seen and unseen classes. We will further polish the descriptions in §4.1 to alleviate misunderstandings.

[R2, R3] **Discussion about the difference w.r.t other feature generation methods, e.g. zero-shot image classification.** The difference is described in the second contribution given above. To further validate the argument that our method works well for the semantic segmentation task, we run two state-of-the-art zero-shot image classification methods ([A],[B] with publicly available code) on Pascal-VOC benchmark. Our method achieves a clear performance boost over other classification based ones as shown in Table 1.

Table 1: Comparison on VOC dataset.

Method	Seen mIou	Unseen mIoU	hIoU
DGP [A]	72.9	41.7	53.0
GDAN [B]	73.0	39.8	51.5
Ours	73.4	45.7	56.3

[A] Kampffmeyer, Michael, et al. "Rethinking knowledge graph propagation for zero-shot learning." CVPR. 2019. [B] Huang, He, et al. "Generative dual adversarial network for generalized zero-shot learning." CVPR. 2019.

[R3] **A baseline model without relation aggregation.** This baseline (CSRL w/o relation) achieves 73.0%/43.2%/54.3% in terms of Seen mIoU/Unseen mIoU/hIoU, which validates the effectiveness of mutual feature and relation aggregation. Detailed results will be updated.

[R3] **Detailed implementations e.g. word embedding and learning rate.** The implementation details to re-produce our results are given in §B in the supplementary material.

[R4] **More related works should be mentioned.** We have added the missing GAN-based methods and a thorough related works discussion will be updated.

[R4] **Why not dynamically learn the weights of losses.** In this work, we focus on exploring the relation consistency between seen and unseen categories. To maintain simplicity, we do not dynamically adjust the weights of losses. Even this we have already reached a superior performance, and a better result could be achieved by adopting techniques in e.g. Sener, Ozan, and Vladlen Koltun. "Multi-task learning as multi-objective optimization." NeurIPS. 2018.

[R4] **Intuitive discussion about the similar relations in visual and semantic.** There intrinsically exists a similar relation among categories in both visual and semantic spaces due to the class *coexistence* and *correlation*. For example, animals (*e.g. cat, dog*) tend to appear simultaneously or highly correlated in both visual scenes or in text corpora.

[R4] **The reason of failure cases.** The reason of failure cases caused by, (i) similar classes (row 1&2&4): Some unseen classes tend to be classified as similar seen ones; (ii) highly occlusion (row 1&2): the areas which are highly occluded by multiple instances or other objects tend to be mis-segmented; (iii) complex scene (row 3): our model fails to correctly parse the image with a complex scene. However, for these failure cases our method is still visually better than [1].

[R4] **Stronger or weaker relations in Fig.4.** The consistent losses aim to constrain the relation consistency. However, the relation cannot be exactly the same. Moreover, each row in the relation matrix is normalized in Fig.4. Thus, some categories may be a little bit weaker or stronger compared to semantic space.