We thank the reviewers for their encouraging and constructive comments. We are pleased that they find the paper well written and acknowledge the novelty and originality of the proposed task, which “has a potential to spark interest” (R1) and “may lead to future papers studying it” (R2). Regarding the proposed framework, R1 and R2 not only find it “sound” and “novel” but also stress the “re-implementation ease” from which “practitioners may benefit” (R1). Still, the reviewers raise points of improvement (R1, R3) and suggest a discussion about a related task (R2). We carefully address these comments below. Some of our answers will be included in the paper if accepted.

Reviewer 1

More classical ZSL baselines. Recent works ([42] and [Schonfeld CVPR’19][1]) reported that DeViSe [15] is among the best classic techniques for generalized zero-shot learning (GZSL). Based on this, we believe that other classic methods would perform similarly to DeViSe if used instead in our GZSL semantic segmentation baseline. During the rebuttal period, we nonetheless conducted additional experiments, adapting ALE [1] for zero-shot segmentation on Pascal-VOC with $K = 2$. They yielded 68.1% and 4.6% mIoU for seen and unseen classes (harmonic mean of 8.6%), on a par with the DeViSe-based baseline in the paper. Such poor generalized-setting performance of classical ZSL methods confirm again the conclusion in [9]. Following R1’s suggestion, we will include more of such classical baselines with discussion in the paper, if accepted.

Graph context clarity. We apologize for the lack of details on the graph context (GC). This is in part due to our initial intent to devote lots of attention and space to the experiments. In an attempt to mitigate this unbalance, we had included GC visualization in the supplementary material, which appears insufficient. If accepted, we will use the additional page at best to include more technical details and visualizations on GC.

Reviewer 2

Other datasets. In state-of-the-art semantic segmentation works, Pascal VOC 2012 and Pascal Context still serve as the main benchmarks. The recent COCO-stuff dataset [Caesar CVPR’18], though larger in scale, is very similar to Pascal Context. We thus expect similar performance behaviors on it. While we have not yet completed such experiments as of now, this will be done, and would be reported in our revision of the paper. R2’s suggestion of looking into urban scene datasets like Cityscape is also interesting and worth investigating in the future.

Transductive ZSL. While our central contribution, ZS3Net, is not transductive (no data, even unlabelled, is available at train time for unseen classes), the ZS5Net variant indeed appears related to transductive zero shot learning. We thank the reviewer for bringing this to our attention. Apart from the fact that referred papers concern all image classification, another difference is worth mentioning though: all but [Song CVPR’18] consider purely transductive settings where *all* unseen class samples are already available at training time. By contrast, our ZS5Net learns from a mix of labelled and unlabelled training data, and is evaluated on a different test set (effectively, a form of semi-supervised learning).

Reviewer 3

Novelty. Being the first to address zero-shot semantic segmentation, we naturally built on existing zero-shot learning literature. Yet, as abundantly exemplified in fully supervised learning, moving from image-level categorization to pixel-level recognition is not as direct or straightforward as it might seem. Highly structured prediction remains a challenge, which we revisit in the context of zero shot learning. While previous generative-based ZSL methods like [7] operate on image-level features, our generator operates on pixel-level ones. Moreover, to encode spatial context, we propose a novel graph convolutional generator which, conditioned on context graphs, generates corresponding structured pixel-level features. Also, as we shall clarify, our framework is not solely bound to GMMN as in [7]; it is in fact agnostic to the choice of the generative model. For instance, we experimented a variant of ZS3Net based on GAN [42], which turned out to be on a par with the reported GMMN-based one. In the submission, GMMN was chosen due to its better stability. In the end, ZS3Net achieves promising, quantitative and qualitative results on a never addressed task, and its ZS5Net extension yields performance very close to the full-supervision upper-bound.

Retraining on ImageNet. We acknowledge R3’s suggestion of re-training ResNet-101 only on seen classes images. Actually, this should be the *de facto* protocol for all zero-shot learning works, to avoid supervision leakage from unseen classes. Our main concern is the challenge of such an undertaking: beside mere time and compute requirements, the absence of current reference performance with such a setting might make training from scratch even more challenging; this might also raise fair comparison issues with future works in the field. Anyhow, we will try our best to overcome these challenges and to extend our manuscript accordingly.

Realism of graph context. One who has never seen a ‘zebra’ can still learn from the fact that zebras live in African treeless grasslands. Such a coarse context prior is actually enough to construct a valid context graph in our approach. Indeed, the way we design this graph is in fact very loose, requiring only relative spatial arrangements, not object shapes (as illustrated in Fig. 2 of supplementary). Using segmentation masks is only one possible strategy, which we chose for the sake of convenience. However, any other, less precise contextual descriptions of unseen objects would suffice to build useful graphs. Anyhow, we would argue that even using segmentation masks for that purpose does not amount to full supervision for the unseen classes since images themselves are not accessible.

Baselines. We kindly refer R3 to our first answer for R1 above.

---

[1] Schonfeld *et al.*, Generalized zero-and few-shot learning via aligned variational autoencoders, CVPR 2019