We thank all reviewers for their constructive feedback and detailed comments. Our responses are provided below.

To Reviewer 1 Thanks for your supportive and helpful comments. We also believe that applying our methods on various fusion settings is an interesting direction for future research (Conclusion).

Q: Other dependencies of input sources, e.g., correlated information.
A: Using multiple camera inputs from various viewpoints can simulate an extremely correlated case. In fact, our experiments deal with correlated information. Object proposals or information about potential surrounding objects can be extracted from both RGB and LIDAR data, and *shared information* is introduced to consider such correlation.

Q: How our methods scale with more than two input sources.
A: Considering its definition, our MaxSSN loss will still try to balance single source robustness for multiple sensory inputs. On computational complexity, our TrainSSN requires $n_s + 1$ forward passes and 1 back propagation. One can further improve on speed by approximating the loss, e.g., mixing TrainSSN and TrainSSNAlt.

To Reviewer 2 We would like to appreciate your valuable and encouraging comments.

Q: Degraded performance of our methods on clean data.
A: Retaining performance on clean data, is an idealistic design goal. In section 3, we solve problem (3) by optimizing over flexible parameters g_1 and g_2. If the parts of input sources contributing to z_3 are known, then indeed we can achieve this goal. In practice however, it is difficult to know which parts of an input source (or latent representation) are related to shared information and which parameters are flexible, and also the design goal becomes a soft rather than a hard constraint. Therefore there is minor degradation in performance, to pay for the added robustness. We will add this discussion in our paper.

Q: More tasks/models to demonstrate the generalization ability of the work
A: AVOD was selected because it is the leading deep fusion model for 3D object detection, and this model has already been favorably compared with prior approaches, so we did not want to take up space with additional comparisons. 3D detection is both an important problem in self-driving cars and one where multiple sensors can contribute fruitfully by providing both complementary and shared information. In contrast, models for 2D object detection heavily rely on RGB data, which typically dominates other modalities. Trying our approaches on other tasks like audio-visual speech recognition is worthwhile, but our paper is already at the page limit covering more basic aspects, e.g., comparison of fusion methods, training algorithms, and corruption methods.
We set the hyper-parameter for ℓ_1 constraint as 0.01, and is available in our source code (omitted for anonymity).

To Reviewer 3 Thanks for your constructive comments and valuable time in understanding our work.

Q: More analyses of Algorithm 1 and 2 and explain why TrainSSNAlt works?
A: We can certainly add more explanation. Essentially the approaches are motivated by findings from the simple linear fusion model, as well as data augmentation. Alternating between clean data and corrupted data aims at increased robustness without much degradation of performance on clean data (strategy is supported by experimental results). We also tried *fine-tuning* only fusion related layers to preserve essential parts for normal data, but Algorithm 1 and 2 are selected for better performance. We discuss the reasons for superior performance of Algorithm 2 in Remark 3. In particular, an element-wise mean operation restricts the features to be optimal for averaging, and hence updating without the maximum loss helps balance the performance. Also note that when AVOD with *concatenation* fusion is trained with TrainSSNAlt, it works a lot worse than TrainSSN.

Q: The proposed latent ensemble layer is not a very novel approach
A: When we devised our LEL, we first set the three objectives (1st sentence of the 2nd paragraph of Section 4.2): (i) error reduction of ensemble methods, (ii) admitting source-specific features to survive, and (iii) allowing different channel depths. Then we wrote the equation in Figure 1 and noted that it can be implemented by using 1×1 convolution. Without applying fancier approaches which could increase computational cost, our LEL showed appealing effectiveness even with simple implementation.

Q: Importance of single source robustness and the generalization ability of the model.
A: Currently there are no formulations even for single source, and our research is the first step. This framework can be extended to robustness against corruption in a subset of input sources, e.g., generalizing to robustness given k of n_s corrupted sources. Self-driving cars using an RGB camera and ranging sensors like LIDAR and radar are exposed to single source corruption. For example, Uber's fatal self-driving crash occurred in nighttime and a camera couldn't detect the victim because of darkness, while LIDAR and radar was not affected. In adversarial settings too, single source attacks seem more feasible. Situations with multiple failures (outside of combat situations) have not been much documented in the automobile industry. Therefore we feel that single source robustness should be studied in depth prior to more general cases.