This supplementary material contains the following three sections. Section A presents hyperparameter analysis on the weights in our formulated losses. Section B shows more qualitative results to compare the performance of single-modality SSD, multi-modality SSD, and our S2M2-SSD. Section C discusses the limitations of our current approach.

A. Hyperparameter Analysis

To determine the weights, we fix the weights of a module once fine-tuned then tune the weights of the next module.

Table 1. Effect of different w_1^r & w_2^r (see Eq.(2) in main paper).

$w_2^r \mid w_1^r = 1$	3	4	5	6	7
NDS	52.4	52.8	**53.0**	52.6	52.3
mAP	43.5	43.7	**43.9**	43.8	43.6
$w_1^r \mid w_2^r = 5$	0.5	0.75	1	1.25	1.5
NDS	52.8	52.7	**53.0**	52.8	52.6
mAP	43.6	43.8	**43.9**	43.6	43.5

Effect of w_1^r and w_2^r. We try different values of w_1^r and w_2^r in the classification response distillation loss (see Eq.(2) in main paper). As Table 1 above shows, setting w_2^r=5 and w_1^r=1 lead to the highest NDS and mAP values, as focusing more on the false predictions by setting a larger w_2^r can improve both the recall rate and precision.

Table 2. Effect of different w_1^v & w_2^v (see Eq.(5) in main paper).

$w_2^v \mid w_1^v = 2$	6	7	8	9	10
NDS	53.4	53.5	**53.8**	53.1	53.2
mAP	43.6	43.7	**44.2**	43.8	43.7
$w_1^v \mid w_2^v = 8$	0.5	1	2	3	4
NDS	53.3	53.4	**53.8**	53.4	53.1
mAP	43.4	43.7	**44.2**	43.4	43.2

Effect of w_1^v and w_2^v. Similarly, we try different values of w_1^v and w_2^v. Results in Table 2 show that the model attains the best performance with w_1^v=2 and w_2^v=8 (see Eq.(5) in main paper). Also, the difference between the two factors should not be further enlarged, since it may hinder the knowledge transfer of the true positive features.

Table 3. Effect of different w_f^p (see Eq.(9) in main paper).

w_f^p	1.0	1.5	2.0	2.5	3.0
NDS	52.1	52.4	**52.7**	52.4	52.3
mAP	43.2	43.4	**43.8**	43.6	43.4

Effect of w_f^p. From Table 3, we analyze the effect of weight w_f^p in Eq.(9) of the main paper, so we decide to set it as 2.

Effect of G. Further, we analyze the effect of G in Eq.(12) of the main paper. As shown in Table 4, we can see that a larger or smaller grid size than 5×5 decreases the detection performance, so we argue that 5×5 is a trade-off between different classes of objects with various sizes and shapes.

Table 4. Effect of grid size G (see Eq.(12) in main paper).

G	3×3	4×4	5×5	6×6	7×7
NDS	51.9	52.3	**52.6**	52.4	52.3
mAP	42.4	42.6	**43.2**	43.0	43.0

B. More Visualization Results

Figure 1 shows more BEV detection results to compare the performance of single-modality SSD, multi-modality SSD, and our S2M2-SSD with four groups of results shown on its top-left, top-right, bottom-left, and bottom-right. Comparing the first and second rows in each group, we can see that multi-modality SSD (orange frame) predicts objects more accurately than the single-modality SSD (black frame). Due to the rich semantics in the RGB images, multi-modality SSD can remove many false positives far away from the LiDAR sensor or near the ground truths, which is significant to improve the detection precision. Comparing the second and third rows in each group, we can see that our S2M2-SSD (blue frame) can predict bounding boxes very close to those of the multi-modality SSD. Also, we can see that our S2M2-SSD can further remove more false predictions near the ground-truth objects, especially for the large-scale ones, *e.g.*, bus and car. Such results are consistent with the evaluated APs shown in the paper. These improvements show that our designed approach can effectively train the single-modality SSD to simulate LiDAR-image features and responses from the multi-modality SSD.

C. Limitations

First, S2M2-SSD employs both point clouds and RGB images in model training that involves two parallel SSDs, so the training time tends to be longer than conventional detectors. To fuse the images and point clouds, it requires a high-performance 2D segmentation network pre-trained on all images of the dataset to segment the input images for the multi-modality SSD. Particularly, these images should be captured under sufficient light and in high resolution to obtain clear textures. Second, we experiment our approach on all ten classes of objects with varying shapes and sizes in nuScenes. Yet, for objects out of these ten classes, our S2M2-SSD may not recognize them well. To reflect the complexity of real scenarios in autonomous driving, detecting objects of more classes is also important. Building or incorporating another dataset to provide more classes beyond nuScenes can be helpful. Last, our S2M2-SSD is trained on point clouds of around 30k points per frame, which are produced by a 32-beam LiDAR sensor. For point clouds collected by a sensor with more beams or covering a larger scan area, the massive input points may make the training of S2M2-SSD infeasible with the current settings.

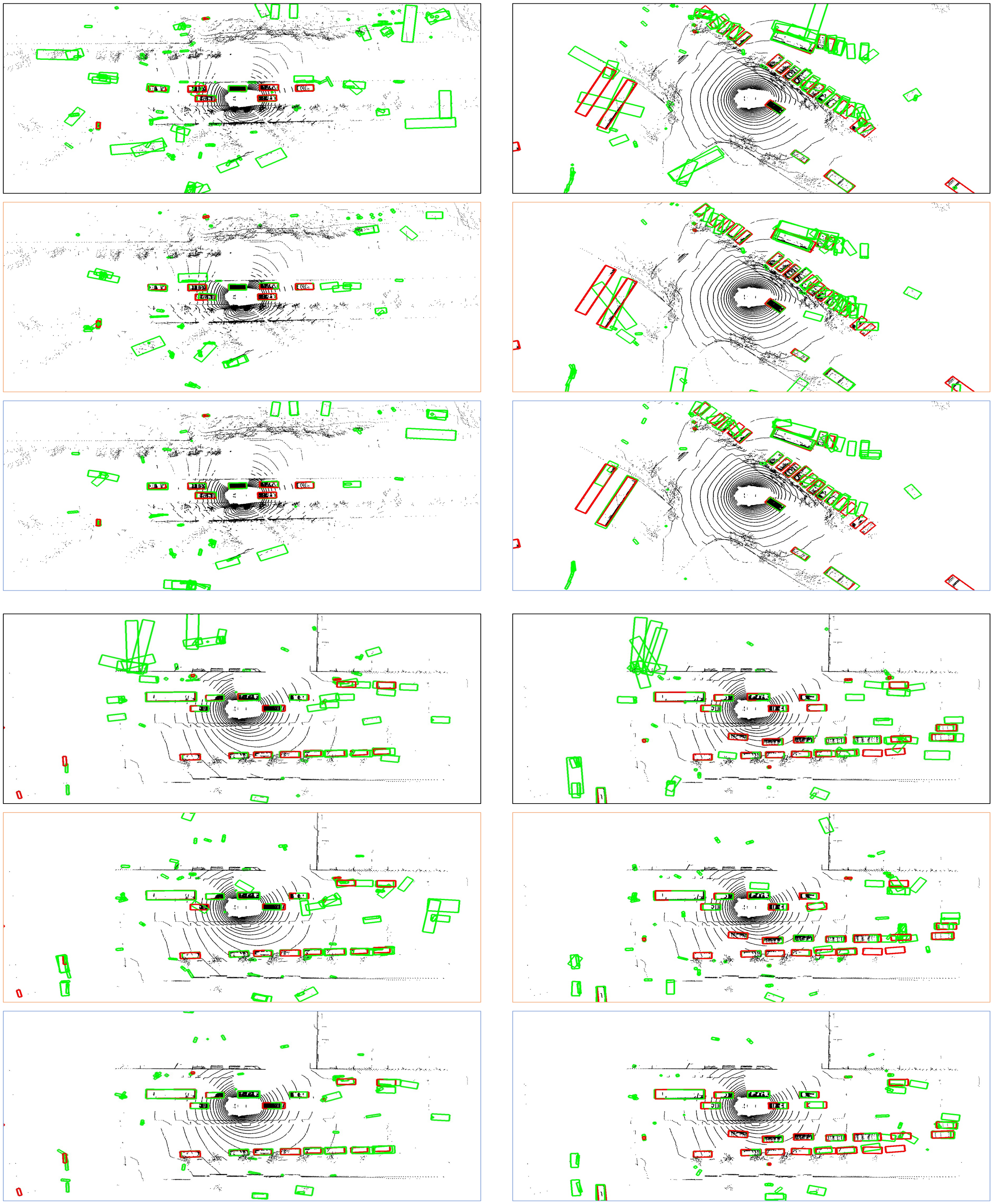

Figure 1. Comparing the BEV detection results produced by single-modality SSD (black frame, 1st & 4th rows), multi-modality SSD (orange frame, 2nd & 5th rows), and our S2M2-SSD (blue frame, 3rd & 6th rows). Exploring the ground-truth bounding boxes (red) and predicted bounding boxes (green) indicates that our S2M2-SSD can remove more false predictions and realize more accurate object localization compared with the single-modality SSD. Our predicted bounding boxes are closer to those of the multi-modality SSD.