We thank the reviewers for their feedback, which includes encouraging comments about the task novelty, the superior results quality, and that our method helps advance an important field. They also raise concerns, which we address below.

R1: **Additional baseline with semantics as input to view synthesis networks.** Great suggestion, we implemented it now, on CARLA dataset. As shown in the table below (cols 2-5), feeding RGB+semantics to view synthesis networks (SPADE+SM^{+}) improves over the original baseline (SPADE+SM). Here we only show SPADE+SM, (the most robust baseline), for space constraints. However, our GVS method consistently outperforms these improved baselines.

R1, **R4**: **Novelty of the technical contribution.** No method currently exists that can synthesize photorealistic images of novel views from single semantics. In fact, we had to design our own baselines. We show that, though sensible, these baselines fall short of matching the quality of our results. This speaks to the importance of the choices behind the proposed method. **R4** suggests that the results are "not surprising". We think that's good: it means that our method is principled and it follows logical steps, which is what we strove for and the reason of the quality of the results.

R2, **R1**: **Benefits of GVS.** Our goal is to simplify content creation by allowing the user to create and edit semantic map and produce photorealistic novel-view images. **R2** (2.1) and **R1** point out that this *can* be done with existing i2i methods. One can edit the semantics for each view independently forcing geometric consistency–a tedious operation at best–and apply i2i to each map. Or, they can apply i2i to the semantic map and perform single-image NVS, precisely one of the baselines in our paper, which we outperform. Our method requires a single semantic map and generates geometrically (and photometrically) consistent novel views. This can be seen in the supp. video, 4:18-4:55. We show semantic editing in Sec.4.1 to confirm that a semantic map edit seamlessly propagates to novel views.

R2, **R1**: **On depth from semantics.** To produce consistent views, we study how to best encourage geometric consistency between the output views and propose a novel semantic uplifting network (SUN). To validate that SUN delivers on the promise, we verify that it can learn the scene geometry from just 2D semantics. We compare against MonoDepth to offer context to evaluate the accuracy of SUN, not as an application of our method. It was a poor choice on our part to call the section "Applications," which confused **R2** (2.1) and **R1** (4). We will clarify this in the revision.

R2: **Whether our representation is a proper MPI.** We agree that this can be confusing. However, following Stereo Magnification (SM) by Zhou et al., which introduces MPIs, we use the term to refer to our multi-plane representation of the scene that can be projected in the novel views via homographies. While MPI planes correspond to physical scene layers, this is not always apparent in the learned MPIs, see Fig.5 of the SM paper. Our lifted semantics is a *compact* representation of the 3D scene semantics which is also converted to an MPI representation. These semantics layers are not learned with an explicit constraint on modelling occlusions. While trying to come up with a compact representation of the scene, the network often learns to perform occlusion removal in Layer-3 and dilation of thin objects in Layer-2.

R2: Multi-view consistency. **R2** points out that our method can generate better images than the obvious baselines, but a numerical validation that our novel views (NV) are consistent and plausible is missing (2.5)–a very good point. We agree with **R2** on *Consistency*. We quantify *consistency* by using the GT depth to warp 2 NVs back to the reference camera and computing the per-pixel-per-channel absolute error between them. Our numbers are comparable with SPADE+SM on CARLA/vKITTI (see table). Note: SPADE+SM is already close to the upper bound in terms of consistency as all the NVs come from the same RGB image. However, if we evaluate the NVs' *plausibility* by comparing them with the ground truth views, SPADE+SM significantly underperforms as compared to our methods (cols 2-5).

R3, **R2**: **Limitations and generalization to non-street scenes.** As **R2** correctly identifies (2.3) our focus on "street scenes" is due to dataset availability. There is nothing fundamental that prevents our approach from working on any other type of scenes. This hurdle is common to many SOTA NVS methods [34, 10, 35] and addressing this limitation forms an important future work.

Method	Cls. Acc ↑	IoU ↑	PD↓	FID↓	Consistency ↓
GVSNet	**74.34**	**66.43**	**1.74**	**62.06**	0.015 / **0.033**
SPADE+SM	69.93	60.82	1.95	75.81	**0.013 / 0.033**
SUN+SPADE	72.92	65.52	1.75	68.96	0.025 / 0.051
SPADE+SM^{+}	72.29	63.55	1.75	73.74	0.012/-

R2: **On two-stage system.** Models like SPADE are trained with relatively large batch sizes on multiple GPUs. To achieve similar photo-realism and due to limited GPU resources, we train SUN and LTN networks separately to enable larger batch sizes. Our model allows end-to-end training. Now, we further finetuned the network in an end-to-end manner (with smaller batch size, for 6K iterations) and it leads to slight ($\sim 0.5 - 0.7\%$) improvement in all metrics.

R2: **"Generate semantic map and image for novel views."** We have tried something similar with SUN+SPADE model, where SUN gives novel-view semantics which is fed to i2i (SPADE), SUN+SPADE in Table above. The Table (col 6), Fig. 5 in the paper and supp. video (3:10-4:05), however, show that this model suffers from inconsistencies.

R4: **Formatting/Clarity issues**; **R2**: **Missing references.** Thanks for the suggestions, we will fix all the issues.

R4: **"More elegant solution?"** We propose several novel techniques to tackle a heavily underconstrained problem. We think that elegance can be achieved with principled choices, in addition to simplicity. However, we welcome **R4**'s personal assessment and hope that our work inspires future research to develop more compact solutions.