Are NeRFs ready for autonomous driving? Towards closing the real-to-simulation gap

Supplementary Material

Table 5. Hyperparameters for image augmentation during fine-tuning. p's are referring to the probability of an augmentation being applied.

Parameter	Value
Gaussian noise p	0.5
Additive Gaussian noise δ	$\sim \mathcal{N}(0, 10)$
Gaussian blur p	0.5
Gaussian blur kernel size	5×5
Down- and upsampling factor	10
Down- and upsampling method	Bilinear
Photometric, additive brightness δ	$\sim \mathcal{U}(-32, 32)$
Photometric, multiplicative contrast δ	$\sim \mathcal{U}(0.5, 1.5)$
Photometric, multiplicative saturation δ	$\sim \mathcal{U}(0.5, 1.5)$
Photometric, additive hue δ	$\sim \mathcal{U}(-18, 18)$

A. Training details

A.1. Image augmentation hyperparameters

Tab. 5 shows hyperparameter selection for the image augmentation method.

A.2. NeRF augmentation

To generate NeRF-rendered training images for the perception models, we train NeuRAD [37] on a subset of the nuScenes [3] train set. We select all scenes that were collected at daytime and do not have any rain, resulting in 491 out of 750 scenes. Then we overlap this set with the geographical train split proposed in [23], leaving us with 316 scenes. This way, we can use the NeRF rendering both for the 3D object detection and online mapping task. Among the 316 scenes, we randomly select 110 scenes, namely 0402, 0323, 0252, 0048, 0419, 0856, 0949, 0769, 0435, 0812, 0284, 0394, 0673, 0250, 0288, 0006, 0400, 0736, 0264, 0527, 0359, 0290, 0990, 0256, 0234, 0731, 0300, 0439, 0244, 0698, 0525, 0122, 0075, 0254, 0055, 0163, 0740, 0978, 0712, 0544, 0976, 0021, 0292, 0848, 0792, 0066, 0405, 0200, 0675, 0260, 0375, 0542, 0710, 0988, 0242, 0294, 0381, 0165, 0685, 0157, 0053, 0388, 0286, 0304, 0507, 0298, 0706, 0665, 0790, 0218, 0190, 0034, 0687, 0421, 0671, 0032, 0236, 0505, 0854, 0726, 0044, 0351, 0384, 0805, 0539, 0203, 0407, 0373, 0246, 0361, 0767, 0139, 0194, 0701, 0058, 0230, 0228, 0716, 0392, 0437, 0302, 0060, 0192, 0655, 0240, 0128, 0296, 0787, 0206, 0679. The selected sequences result in a total of 26478 images, constituting 15.7% of the original training set.

A.3. Image-to-image training

We use the NeRF renderings outlined in Appendix A.2 to train a pix2pixHD model [39]. The 110 scenes with six cameras result in 26478 training samples for the pix2pixHD model. We use the official implementation[1] and train the base model for 80 epochs, followed by tuning at a higher resolution for 45 epochs.

A.4. Image-to-image augmentation

Using our image-to-image model, trained as outlined in Appendix A.3, we generate images for all 750 scenes in the nuScenes training set.

B. Experiment details

B.1. Evaluation scenes

To validate the perception models, we train NeuRAD [37] on multiple nuScenes validation scenes and generate images for annotated frames. Note that these frames are held out from the NeuRAD training. From the original 150 nuScenes validation scenes, we select all scenes collected at day-time without rainy weather, yielding 111 scenes, namely 0003, 0012, 0013, 0014, 0015, 0016, 0017, 0018, 0035, 0036, 0038, 0039, 0092, 0093, 0094, 0095, 0096, 0097, 0098, 0099, 0100, 0101, 0102, 0103, 0104, 0105, 0106, 0107, 0108, 0109, 0110, 0221, 0268, 0269, 0270, 0271, 0272, 0273, 0274, 0275, 0276, 0277, 0278, 0329, 0330, 0331, 0332, 0344, 0345, 0346, 0519, 0520, 0521, 0522, 0523, 0524, 0552, 0553, 0554, 0555, 0556, 0557, 0558, 0559, 0560, 0561, 0562, 0563, 0564, 0565, 0770, 0771, 0775, 0777, 0778, 0780, 0781, 0782, 0783, 0784, 0794, 0795, 0796, 0797, 0798, 0799, 0800, 0802, 0916,

[1] https://github.com/NVIDIA/pix2pixHD

0917, 0919, 0920, 0921, 0922, 0923, 0924, 0925, 0926, 0927, 0928, 0929, 0930, 0931, 0962, 0963, 0966, 0967, 0968, 0969, 0971, 0972.

For the geographical split [23], we use all scenes in the geographical validation split that are collected at day-time without rain. Further, as NeuRAD requires annotations for training, we remove all scenes without annotations, resulting in 67 scenes. For more rigorous evaluation, we also include all scenes from the geographical test split that have annotations and were collected at day-time without rain. This adds another 87 scenes, totaling in 154 scenes, namely 0002, 0019, 0043, 0046, 0061, 0151, 0158, 0159, 0348, 0355, 0356, 0357, 0358, 0377, 0385, 0945, 0947, 0981, 0982, 0983, 0018, 0036, 0268, 0275, 0276, 0344, 0345, 0411, 0182, 0183, 0315, 0423, 0424, 0425, 0860, 0861, 0862, 0863, 0864, 0925, 0926, 0927, 0928, 0071, 0170, 0171, 0172, 0173, 0174, 0175, 0209, 0210, 0211, 0212, 0500, 0501, 0518, 0660, 0661, 0662, 0663, 0664, 0738, 0821, 0109, 0331, 0523, 0007, 0008, 0009, 0024, 0025, 0026, 0027, 0028, 0029, 0030, 0042, 0050, 0057, 0123, 0124, 0154, 0155, 0364, 0365, 0370, 0379, 0380, 0383, 0952, 0953, 0955, 0956, 0957, 0958, 0959, 0960, 0016, 0966, 0413, 0414, 0415, 0416, 0417, 0184, 0185, 0187, 0188, 0316, 0427, 0428, 0429, 0430, 0858, 0919, 0920, 0921, 0924, 0069, 0073, 0176, 0207, 0208, 0213, 0263, 0396, 0397, 0398, 0509, 0528, 0529, 0530, 0531, 0532, 0533, 0534, 0535, 0536, 0658, 0744, 0746, 0747, 0749, 0750, 0751, 0752, 0757, 0758, 0759, 0760, 0817, 0110, 0330.

For the evaluations on laterally shifted views, we use a smaller subset of scenes from our previously chosen 111 scenes. We select scenes on the criteria that our lateral shifts do not result in the camera ending up inside other road users or structures. This results in 14 scenes, namely 0523, 0924, 0921, 0928, 0268, 0919, 0109, 0926, 0018, 0344, 0345, 0016, 0276, 0925.

B.2. Fine-tuning of 3D object detection models

We start all fine-tunings from model weights pre-trained on nuScenes. For FCOS3D and PETR, we utilize the implementations from the mmdetection3d-framework[2] and use the model weights and corresponding training configurations reported there. For BEVFormer we use the official implementation[3] and model weights corresponding to the small version. See Tab. 6 for details on the hyperparameters used for each fine-tuning.

Table 6. Hyperparameters used to fine-tune the 3D object detection models.

Model	Augmentation	Learning rate	Epochs
FCOS3D	None	$2e-6$	6
	Image aug.	$2e-5$	6
	NeRF	$1e-4$	6
	Img2Img	$1e-4$	6
PETR	None	$1e-8$	12
	Image aug.	$2e-5$	12
	NeRF	$2e-5$	12
	Img2Img	$2e-5$	12
BEVFormer	None	$2e-5$	4
	Image aug.	$4e-5$	4
	NeRF	$4e-5$	4
	Img2Img	$4e-5$	4

Table 7. Hyperparameters used to fine-tune the online mapping method MapTRv2.

Model	Augmentation	Learning rate	Epochs
Original	None	$1e-5$	5
	Image aug.	$1e-4$	5
	NeRF	$1e-4$	5
	Img2Img	$1e-4$	5
Geogr.	None	$1e-4$	5
	Image aug.	$1e-4$	5
	NeRF	$1e-4$	5
	Img2Img	$1e-3$	5

B.3. Fine-tuning of online mapping

Also for the online mapping method MapTRv2 we start from the pre-trained weights for both the original and geographically disjoint splits. Tab. 7 reports the hyperparameters used for the different fine-tunings.

B.4. NeuRAD results

In Tab. 8, we report standard novel view synthesis metrics for the different data splits. We observe NeuRAD to perform similar for all data subsets, hence expecting the artifacts in the images used for augmentation having similar style as the ones used for evaluation.

[2] https://github.com/open-mmlab/mmdetection3d

[3] https://github.com/fundamentalvision/BEVFormer

Table 8. Novel view synthesis performance for NeuRAD on held-out images for the different splits.

Split	PSNR ↑	SSIM ↑	LPIPS ↓
Orig. val	26.50	0.7893	0.2566
Orig. train	26.52	0.7975	0.2456
Geo. val+test	26.88	0.8009	0.2558

C. Additional results

C.1. Real2sim 3DOD

We visualize the real2sim results on 3DOD models, reported in Tab. 1, as bar plots in Fig. 8.

C.2. Correlation to FID for shifted scenes

We illustrate the correlation between detection agreement and FID, isolated for only shifted sequences and divided by the shift amount and direction, in Fig. 9. The detection agreement is computed for BEVFormer fine-tuned with image-to-image translated images.

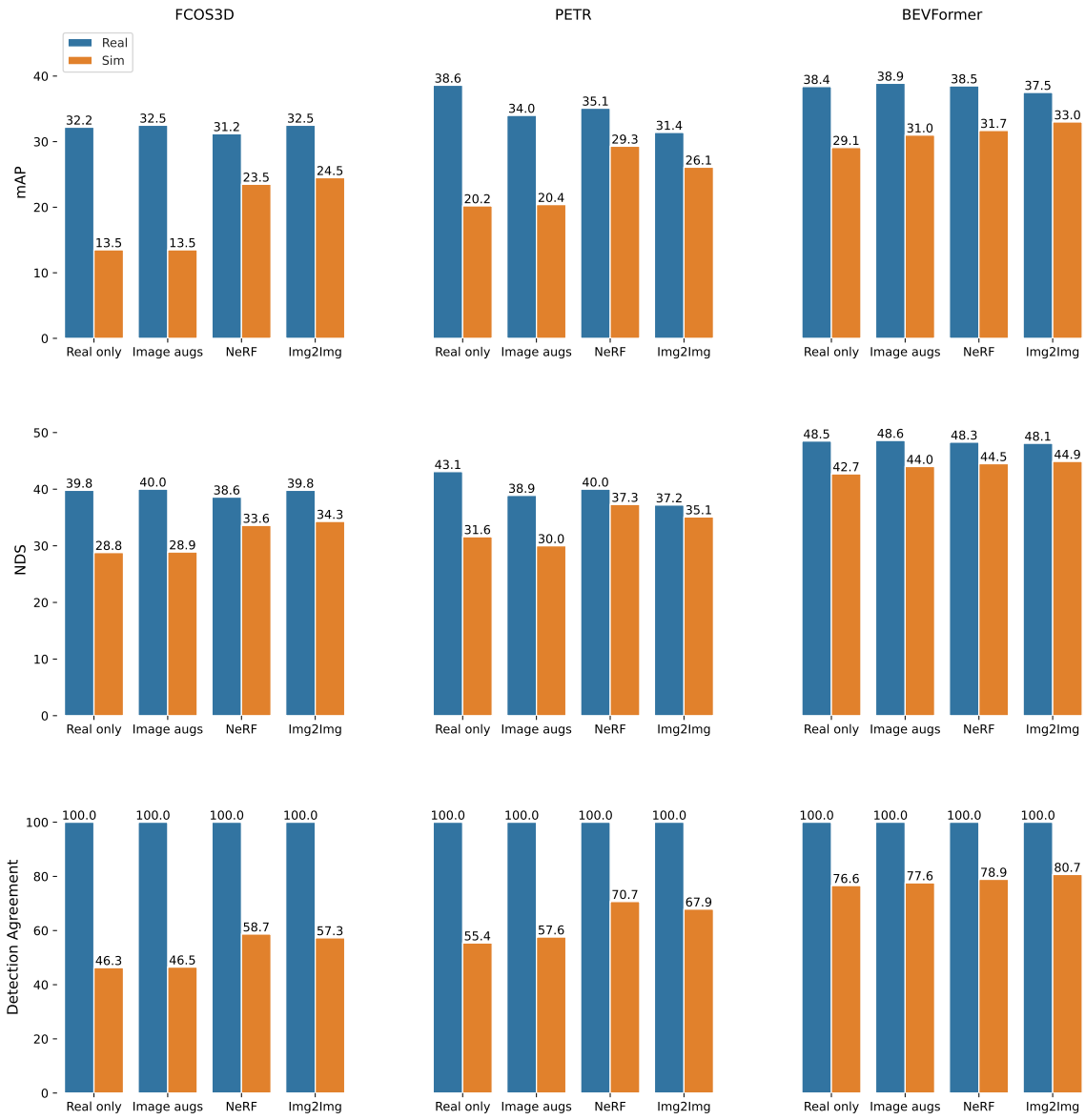

Figure 8. Bar plots of the real2sim gap for the 3DOD-models.

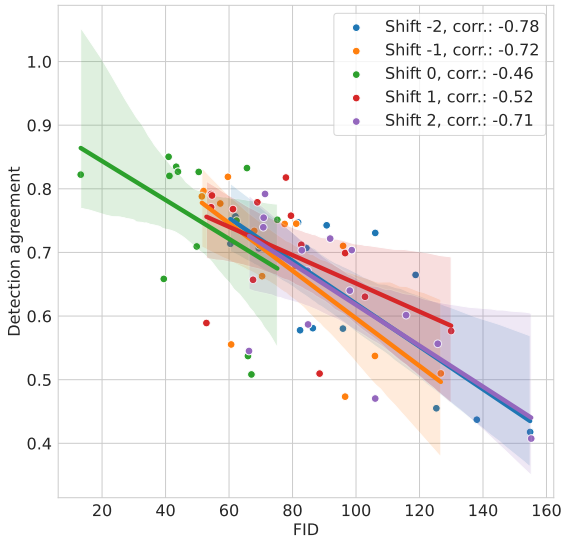

Figure 9. Detection agreement vs. FID scores for BEVFormer fine-tuned with image-to-image translated images, evaluated on the lane shift evaluation set with different shifts.