given variant, are searched as shown in Fig. 4-1. These ranges are based on prior experience of the author. The temporary geo ILM uses the height-threshold-based ground-plane removal since it is the most widely used method in the literature.

$M_F^{(\text{thin})}$	$M_F^{(\text{big})}$	$\varphi_\measuredangle^{(\text{thin})}$	$\varphi_\measuredangle^{(\text{big})}$
$[0.1, 0.2, ..., 0.6]$	$[0.3, 0.4, ..., 0.8]$	$[1°, 1°, ..., 5°]$	$\{10°, 20°, 30°, 40°\}$

M_O	M_D	T
$[0.3, 0.4, ..., 0.8]$	$[0.1, 0.2, ..., 0.5]$	$\{1, 5, 10, 15, 20, 25\}$

Fig. 4-1: Ranges for geo IRM variants' parameter grid search given the parameters are not fixed for the respective variant.

In the second step, the afore best performing geo IRM variant will be kept fixed and used as a reference to analyze the ground-plane removal variants. Here, the two filters under investigation are a purely geometric height threshold-based and a semantic filter. The height thresholds are compared for different heights in the interval $[0, 0.1]$ with step size 0.025 and in the interval $[0.1, 2.0]$ with step size 0.1 in order to have a higher resolution close to zero. For the semantic filters, in the majority of cases the street detections are closest to the ego vehicle in the BEV projection followed by sidewalks and terrain. Since the removal of detections in occluded areas has little to no effect for geo ISMs, a three stepped removal of the semantics is proposed which removes pixels of classes on average increasingly further away from the vehicle. The three removal steps are as follows [no street; no street or sidewalk; no street, sidewalk or terrain].

In order to obtain lidar occupancy maps closer resembling the ground-truth occupancy state, the dynamic state provided by the bounding box labels is propagated to corresponding lidar detections. All detections marked as dynamic are only used to provide boundaries for free space rays but not to define occupied space. Since the sweep information is used for mapping but the labels are only available on a sample level, the bounding box poses of dynamic objects have to be interpolated as described in 3.3.1.

4.1.2 Experimental Results for geo ILM and IRM Parameter Tuning

In order to parameterize the temporary geo ILM, the height-threshold will be defined first. Here, the parameter is increased up to the point where additional structure besides street is being removed. Using the so found height threshold of 0.6 cm, the free and occupied weight M_F, M_O, M_D and the opening angle $\varphi_\measuredangle$ of Alg. 1 are balanced in a way that

- remaining ground points and dynamic object artifacts are being filtered out

- IDM rays don't cut through object boundaries
- object boundaries are as dense as possible
- free and occupied space has maximal assigned evidential mass

leading to the parameters summarized in Tab. 4-4. For an illustration of the ILM tuning, refer to Fig. 4-2.

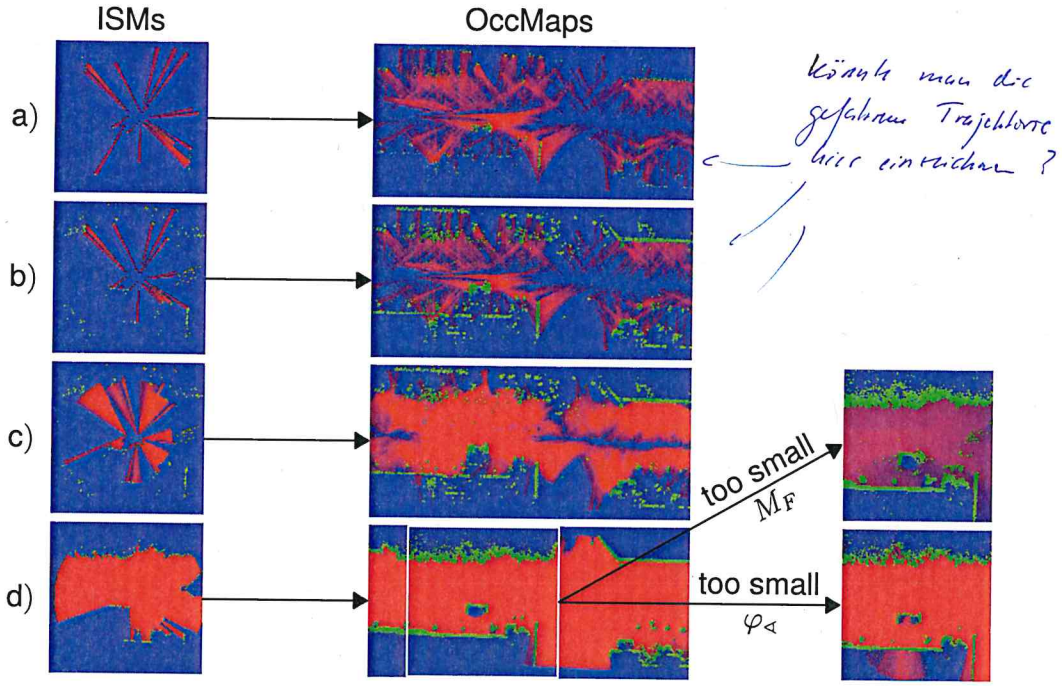

Fig. 4-2: Examples of the chosen ISM parameterization. The first column shows the ISM results and the second one the resulting occupancy maps. The rows show the different investigated ISM variants. Row a) shows the rays casting IRM which is extended in b) with accumulated detections and further extended in c) with additional free space rays. Row d) shows the geo ILM. Additionally, examples of the ILM with sub-optimal free mass M_F and ray opening angle $\varphi_\sphericalangle$ are shown on the right.

After fixing the temporary ILM, it can be used to tune the IRM using the mIoU between their occupancy maps. As mentioned in Section 3.3.3, three IRM variants are separately tuned and compared against each other. As shown in Tab. 4-3, the accumulation of detections results in additional information about occupied space and helps to correct the free space leading to a better overall mIoU. Additionally, enriching the free space with the larger IDM rays keeps the occupied score untouched while moving unknown mass to the free class, leading to the best mIoU score of the considered IRM variants. These variants together with resulting occupancy maps are illustrated in Fig. 4-2.

IRM variants	mIoU			
	free	occupied	unknown	overall
ray casting	67.0	16.8	21.5	35.1
ray casting + acc. detections	70.6	26.6	21.5	39.9
ray casting + acc. detections + free rays	76.2	26.6	22.6	42.1

Fig. 4-3: Comparison of mIoU of occupancy maps generated using three IRM variants and lidar occupancy maps.

ISM	$M_F^{(\text{thin})}$	M_O	M_D	$\varphi_{\measuredangle}^{(\text{thin})}$	$M_F^{(\text{big})}$	$\varphi_{\measuredangle}^{(\text{big})}$	T
ILM	0.025	0.5	0.3	3°	0	0	
IRM #1	0.1	0.8	0.3	5°	0	0	1
IRM #2	0.2	0.3	0.3	5°	0	0	20
IRM #3	0.2	0.3	0.3	5°	0.2	30°	20

Fig. 4-4: Parameters used for geo ILM and IRM (see Section 3.3.4) to produce the qualitative and quantitative results in Fig. 4-2 and 4-3. The parameters are chosen via grid search.

4.1.3 Experimental Results for Lidar Ground Plane Removal Variants

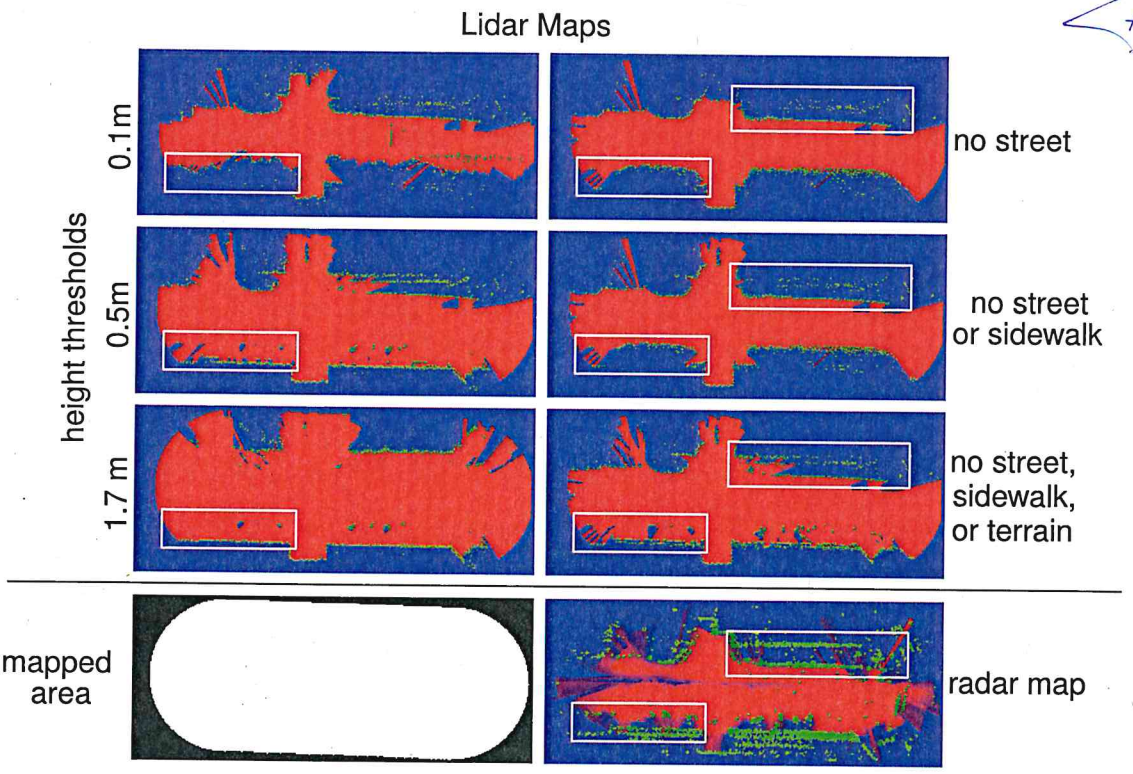

Fig. 4-5: Example of lidar maps created by successively removing ground-plane semantics and three threshold-based filters, together with the mapped area (white) and the radar map.

The quantitative comparison in Fig. 4-6 shows that the successive semantic-based removal up to the terrain level leads to increasingly better overlap up to the best reached mIoU score of 28.18%. On the other hand, the height threshold-based filters show improved performance up to a height threshold of 0.5 m with a score of 26.96% after which the performance starts to decrease. This suggests that for the street and sidewalk level semantics as well as low height threshold large portions of the areas detected by the radar are occluded in the lidar BEV. This is also qualitatively shown in the lower left white boxes in Fig. 4-5. Moreover, when the height threshold is set too high, portions of the areas detected by the radar are increasingly filtered out, as illustrated in the upper right boxes in Fig. 4-5. Thus, a height threshold of about 0.5 m or the semantic-based removal up to the terrain level provide the best compromise of the compared methods. However, since the semantic information is only available for keyframes in the NuScenes dataset and because the 0.5 m height threshold filter rivals the best semantic filter in its performance, it is proposed to use the height threshold-based filter to obtain the labels for further experiments.

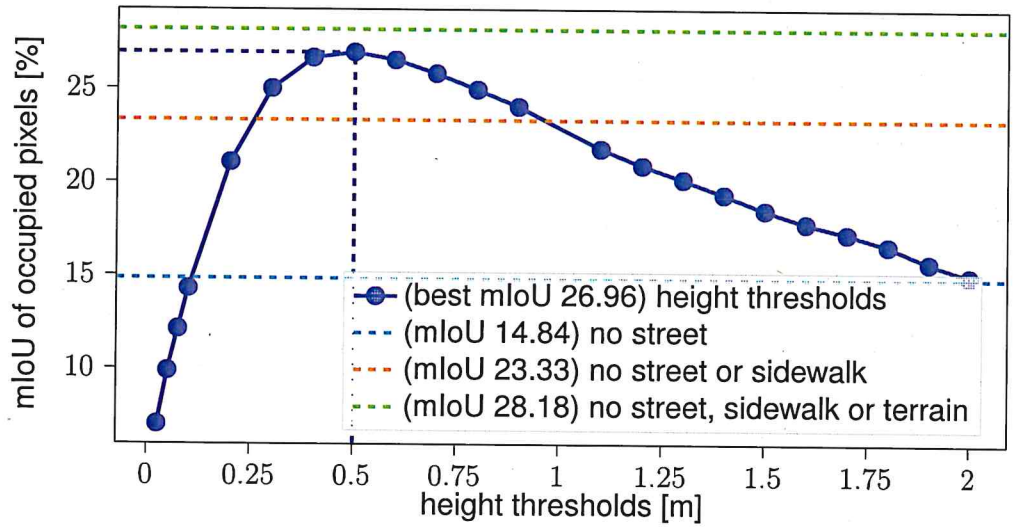

Fig. 4-6: Results of mIoU between different evidential occupancy maps create with variations of geo ILMs and the geo IRM computed in the mapped area.

4.1.4 Discussion

Looking at the results of the geo IRM variant comparison, it becomes clear that the two proposed improvements, namely the accumulation of detections over time and the additional application of the IDM with a wide cone, indeed outperform the baseline by a margin of 7% in mIoU. However, there have been cases in which the accumulation lead to a persistence of outliers increasing their influence. However, those cases are statistically outweighed by the benefit of restricting rays from cutting through occupied regions. It shall also be mentioned that there are still large areas marked as unknown

which are clearly free. An example of which is the vehicle's trajectory. Here, the geo IRM can be changed to set the region beneath the ego vehicle to free. For the other regions further away from detections and ego vehicle further investigations can be done to improve the free space coverage. Overall, this procedure has shown to suffice RQ2, since it strictly improves the mIoU scores of all classes over the baseline model.

With regards to the analysis of ground-plane removal techniques according to RQ1, the comparison of lidar ground plane removal methods demonstrates that the removal of detections up to the terrain level increases the overlap between lidar and radar sensing modalities and provides the best result. Additionally, it can be seen that the application of the simple height threshold removal method can reach similar performance when it comes to aligning the sensing overlap. Thus, since the lidar labels are only provided with low frequency on the sample level and an interpolation down to the higher frequent sweep level is non-trivial, it is proposed to use the height threshold method for the further experiments.

4.2 Choice of UNet Architecture

This section details the experimental determination of hyperparameters of the deep ISM's base architecture, as defined in Section 3.3.6.

4.2.1 Experimental Setup

Since the focus of this work lies on the investigation of radar ISMs, the architecture search is performed based on radar input images with occupancy map patches for targets (see Section 3.3.5). More specifically, radar BEV images based on one sweep's information R_1 are used, since they contain the least information and, thus, provide the most difficult task for the network to handle. Moreover, the considered UNet (see Section 3.3.6) is trained in SoftNet configuration (see Section 3.3.7), since it is the baseline configuration as proposed by the literature.

The hyperparameters searched with the following procedure are the downsample factor D of each ResNet layer in the UNet and the amount of filters for each stage $C_k, k \in [0, 4]$. With regards to the filters, the approach proposed in the literature to double the amount after each encoder and halve it after each decoder stage respectively up to a maximum number of filters is adapted in this work (see Section 2.3.1). Thus, reducing the filter search to the initial amount of filters C_0 and the maximum amount of filter $C_{\max}$. These hyperparameters shall be investigated in a two stepped approach. First, D is set to 0.5 which is half of the most conservative compression rate reported to work without loss in performance (see Section 2.3.1). On the other hand, for C_0, the following variations are evaluated $[4, 8, 16, 24, 32, 40]$ with $C_{\max} = \infty$. Given the results

To fine tune the capacity of the network, C_0 is fixed to 32 while an upper limit on the number of filters is set to $C_{\max} = 128$. Given this setup, D is halved, starting from 0.5 up to the point where the performance collapses. It can be seen that this is the case after the second reduction, as shown by the blue dots in the left plot of Fig. 4-9. The parameter settings for the two experiments labeled "7" and "8" are shown in the lower part of 4-7.

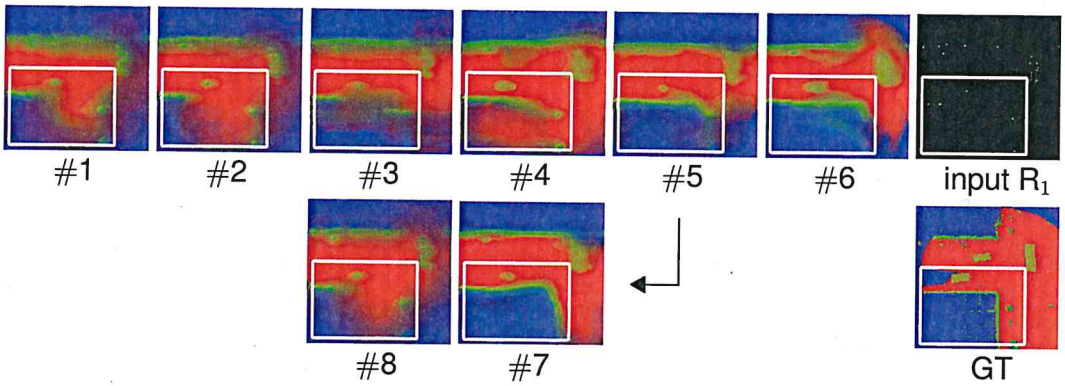

Fig. 4-8: Qualitative results of the different models trained in the experiments as numbered in Fig. 4-7 with the radar input R_1 and the ground-truth GT.

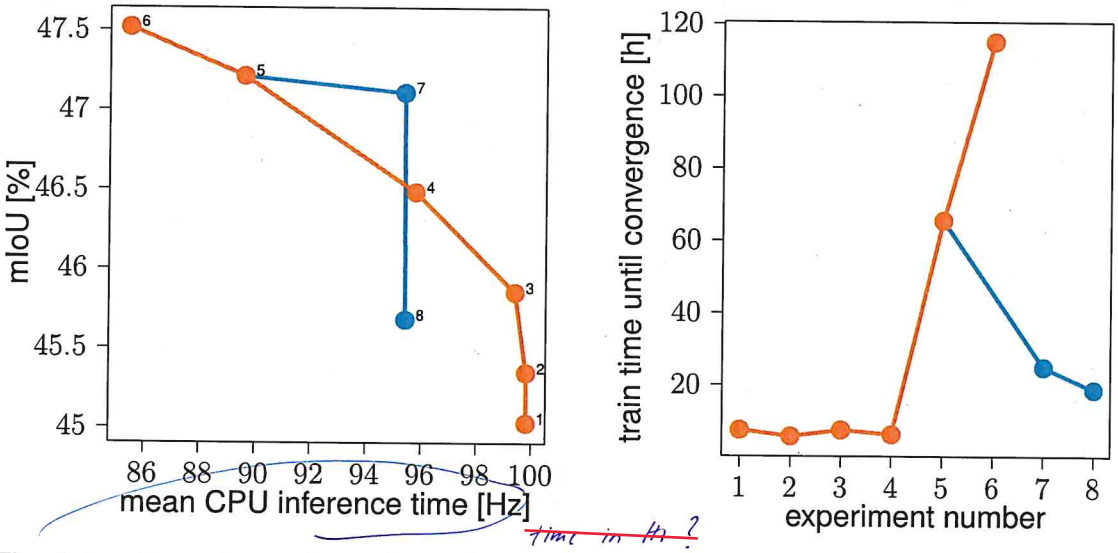

Fig. 4-9: Experimental results of the network tuning experiments. On the left-hand-side, the mIoU over the CPU inference time is shown while on the right-hand-side, the training time is plotted for each experiment. Here, orange marks the first part of the experiments in which the downsamlping rate D is fixed and the initial amount of filters C_0 is searched. On the other hand, blue marks the second stage of the parameter search in which D is finetuned.

The experimental results show that a reduction of D from 0.5 to 0.25 brings the network about 6% closer to the goal of 100 Hz while roughly keeping its performance. Also, the

Here, the right white box in scene B of Fig. 4-13 shows a predicted dynamic object for R_{20} which is already outside the ISM's FoV. This is caused by the trailing dynamic predictions which does not occur for the other deep IRM variants. On the other hand, the overall occupied class performance is slightly but consistently improved by only taking the latest dynamic detection for $R_{20|1}$. One potential cause can be seen in the lower white box in scene B of Fig 4-13. Here, static detections are missing in R_{20} which are present in $R_{20|1}$. This is caused by some detections almost outside of the time horizon that have been falsely identified as dynamic and are being overlayed over the static predictions. Hence, the accumulation of dynamic detections can cause outliers to deteriorate static detections. Eventually, the free scores are mainly the same, showing that the changes only affect the occupied and dynamic class. Finally, comparing $R_{20|1}$ and R_{20} with R_1 it can be seen that the problem of decrease in dynamic prediction performance, indicated by the quantitative results while the qualitative results seem to be sharper, remains.

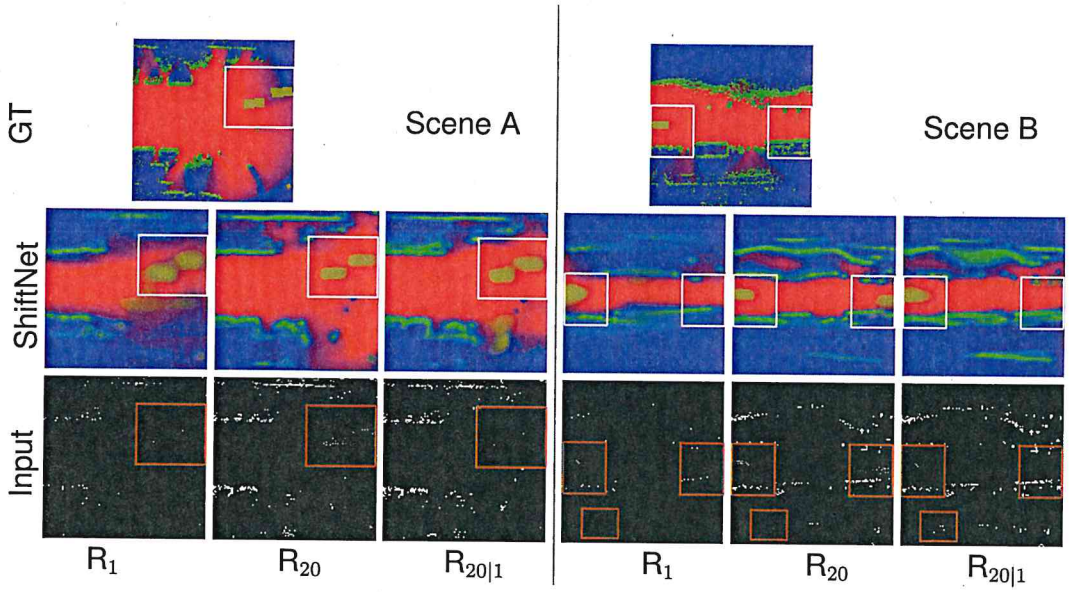

Fig. 4-13: Qualitative comparison of ShiftNet predictions (middle) trained on radar inputs with three different dynamic detection encodings (bottom) shown for two scenes with the respective GT (top). For a detailed explanation of the abbreviations and color format kindly refer to Section 3.3.5.

4.4.3 Discussion

The above results show that the radar encoding $R_{20|1}$ significantly decreases the capability to estimate dynamic objects, slightly increases the occupied area predictions while maintaining about constant for free areas. Here, the decrease for the dynamic class is to be expected, since in some cases only a single detection point is assigned to a dynamic object. Therefore, the $R_{20|1}$, similar to the R_1, based deep ISMs need

4.5.2 Experimental Results Camera and Camera-Radar Inputs

Based on the false rates, the deep IC_DM slightly outperforms the other purely camera-based counterparts with an accumulated overall false rate of 52.7 as compared to 54.1, both for deep $IC_{RGB}M$ and IC_SM. Looking at the true rates, this distinction becomes even clearer with the deep $IC_{RGB}M$ providing the worst true rates in all categories (true rate sum of 103.1), followed by the deep IC_SM (true rate sum 111.7) and, finally, with the deep IC_DM as the best purely camera-based model (true rate sum 126.1). This overall better performance of the deep IC_DM is also reflected in the overall decreased unknown mass compared to the other purely camera-based ISMs. The improved performance of the deep IC_DM is also reflected in the qualitative results in Fig. 4-15 by the more highlighted occupied space and, more importantly, the increased correctness of free and occupied space contours compared to the remaining purely camera-based models.

	$\tilde{k}$	$\tilde{d}$	$\tilde{f}$	$\tilde{o}$	$\tilde{u}$	$\tilde{d}$	$\tilde{f}$	$\tilde{o}$	$\tilde{u}$	$\tilde{d}$	$\tilde{f}$	$\tilde{o}$	
$IC_{RGB}M$	$p(k\|d)$	33.2	23.0	5.3	38.6	28.3	33.0	5.6	33.0	36.3	18.9	4.6	40.2
	$p(\tilde{k}\|f)$	4.0	58.9	3.1	33.9	3.1	70.1	2.7	24.1	6.6	29.8	4.3	59.4
	$p(\tilde{k}\|o)$	4.1	14.6	11.0	70.3	4.3	22.6	12.2	60.8	4.1	12.2	10.6	73.1
	$p(\tilde{k}\|u)$	3.2	4.7	5.0	87.2	-	-	-	-	3.2	4.6	4.9	87.3
IC_SM	$p(k\|d)$	38.1	22.3	4.6	35.0	34.6	30.2	4.8	30.5	40.9	18.6	4.2	36.2
	$p(\tilde{k}\|f)$	4.0	62.4	2.5	31.2	3.2	73.9	1.8	21.1	6.3	32.2	4.2	57.3
	$p(\tilde{k}\|o)$	4.8	15.9	11.2	68.1	5.7	23.7	11.6	58.9	4.6	13.6	11.0	70.8
	$p(\tilde{k}\|u)$	2.6	6.5	5.4	85.5	-	-	-	-	2.6	6.4	5.4	85.6
IC_DM	$p(k\|d)$	40.8	16.8	7.0	35.3	36.1	25.5	6.6	31.7	44.3	12.0	7.1	36.6
	$p(\tilde{k}\|f)$	3.4	69.3	2.4	25.0	2.3	81.4	1.5	14.9	6.4	38.3	4.7	50.5
	$p(\tilde{k}\|o)$	6.9	16.2	16.0	60.9	8.2	25.2	15.5	51.2	6.6	13.5	16.1	63.7
	$p(\tilde{k}\|u)$	3.3	6.5	8.0	82.1	-	-	-	-	3.3	6.4	8.0	82.2
$IC_DR_{20}M$	$p(k\|d)$	42.4	14.3	12.6	30.7	38.0	20.6	13.0	28.5	45.7	11.3	12.3	30.7
	$p(\tilde{k}\|f)$	2.9	69.6	2.5	25.0	2.2	80.0	1.6	16.1	4.9	42.3	4.9	47.9
	$p(\tilde{k}\|o)$	5.0	12.0	28.8	54.1	6.4	17.2	30.0	46.4	4.7	10.5	28.4	56.4
	$p(\tilde{k}\|u)$	2.0	8.1	8.0	81.9	-	-	-	-	1.9	8.0	8.0	82.1
ShiftNet		overall				visible				occluded			

Fig. 4-14: Normed confusion matrix evaluated on the ShiftNet model which was trained on C_{RGB}, C_S, C_D and, finally, the fusion of C_D and R_{20}. For a detailed explanation of the table format kindly refer to Section 3.3.8.

More specifically, the only prominent deficit of the deep IC_DM lies in its increased false rates for $p(d|o)$ and $p(o|d)$. This shows that, compared to the other purely camera-based models, the deep IC_DM is better in estimating whether there is an object present but not in which state (static or dynamic) it is. On the other hand, both homography-based ISMs struggle to distinguish free from dynamic space. But, the incorporation of semantic information leads to a jump in the true positive rate for dynamic objects. This

worsening is observed with regards to false occupied predictions in the dynamic class. Here, the false rate is almost doubled compared to the other camera ISMs. Comparing to the deep $IR_{20}M$, the deep $IC_DR_{20}M$ provides better performance in each of the overall scores without exception.

A qualitative example for the improvement of the deep $IC_DR_{20}M$ over the deep $IR_{20}M$ can be seen in the white box in Scene B of Fig. 4-15. Here, the MonoDepth provides the additional information that the wall proceeds around the corner, leading the model to assign the area behind the wall as unknown rather than free. Also, the shape of occupied space in the deep $IC_DR_{20}M$ predictions is highly improved over the purely camera-based ISM predictions.

4.5.3 Experimental Results Lidar and Lidar-Radar Inputs

When comparing the scores of the deep ILMs in Fig. 4-16 with the ones solely based on camera and radar (see Fig. 4-14 and 4-10) the deep ILM obtains the overall best results. The only exception is with regards to dynamic objects. Here, it obtain the highest false rates of dynamic predictions in the occupied class. Also, the unknown mass is reduced for each category, further illustrating ShiftNet's capability to adapt the unknown mass to account for the higher accuracy in the lidar data. For the geo ILM, the importance of treating the visible and occluded metrics separately becomes more evident than for all other ISMs. Here, the occluded area, as per definition, is mainly assigned to the unknown class. Hence, only the visible area will be considered for the following discussion.

In the visible area, it can be seen in Fig. 4-16 that there are still predictions for dynamic objects remaining. This is due to the fact that not all occupied pixels belonging to the dynamic objects get removed since the bounding box ground-truth in NuScenes often does not fully enclose the objects. The remaining occupied pixels get mixed with the free space rays leading to half occupied half free, and hence, dynamic pixels. This effect can also be seen in Fig. 4-17 in the upper white box in scene A where some contour points of the bus remain as dynamic detections. Regarding the free predictions in the visible area, they provide an almost perfect overlap with the resulting occupancy maps. Thus, the mapping doesn't change the free space much in the visible area. However, looking at the occupied class, the false free rates are among the highest of all models. This is due to the ILM's parameterization allowing free space rays to cut through occupied areas. Here, the ILM is parameterized to correct these errors during mapping resulting in good maps but sub-optimal ILM performance. This effect can be seen qualitatively in Fig. 4-17 looking at the perforated contours of occupied areas.

When comparing the deep with the geo ILM scores in Fig. 4-16, it can be seen that

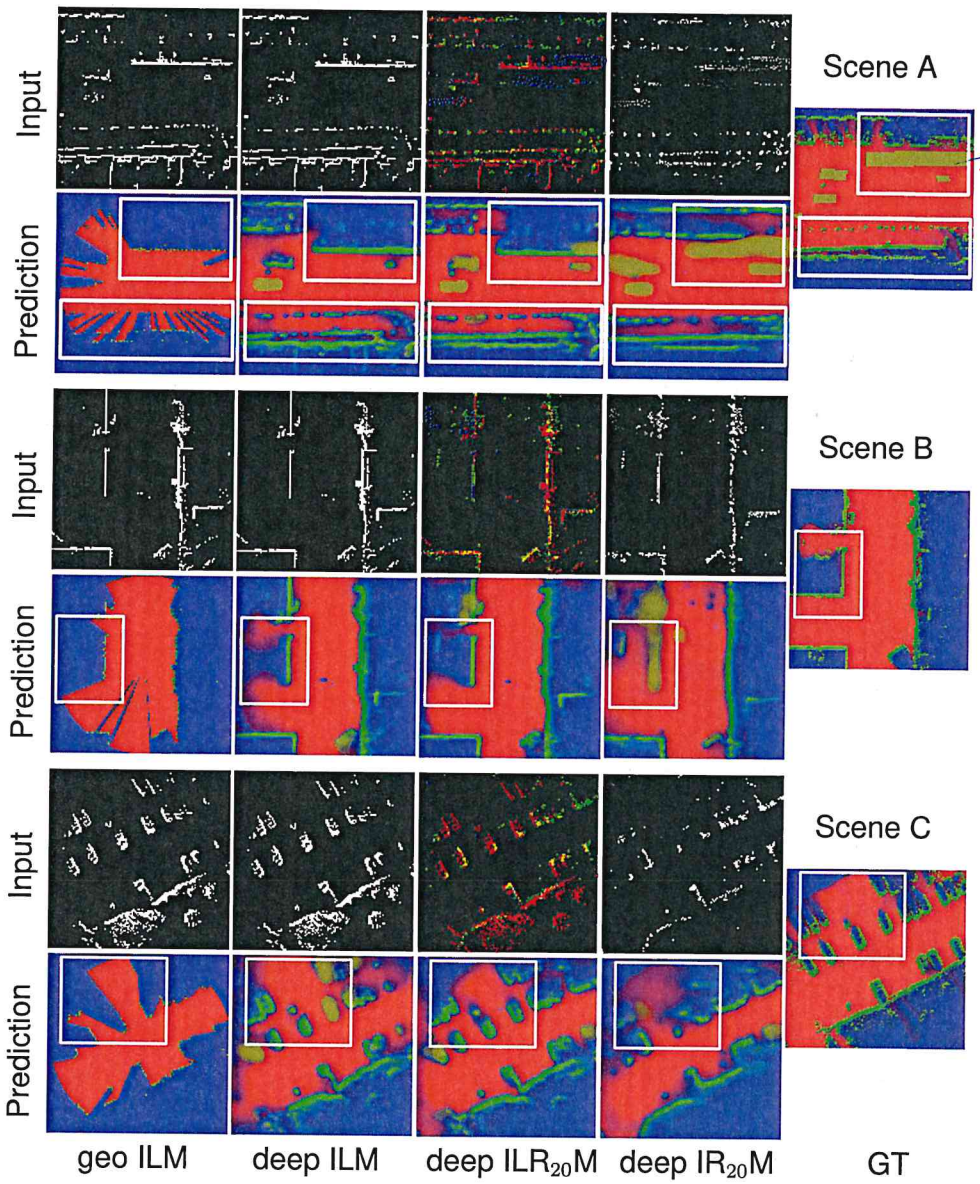

Fig. 4-17: Qualitative comparison of ShiftNets trained on different lidar and radar inputs shown for three scenes with the respective GT to the right. For a detailed explanation of the abbreviations and color format kindly refer to Section 3.3.5.

The deep IRM formerly fully recognized bus as being dynamic while the deep ILM interpreted it as occupied. In the fused result, the bus is only partially predicted as dynamic which, thus, worsened the prediction. However, the static car in the deep ILM is assigned dynamic in the fusion and the dynamic cyclist's contour becomes more accurate in the fusion.