A. Formulations for Adversarial Attack

Adversarial attack is often formalized as an optimization problem under some certain constraints, which vary among different attack settings. Here, we give a brief introduction to the settings mentioned in the paper and formalize their objectives following the notations in Sec. 3.

ℓ_∞ Adversarial Perturbations. To corrupt the inputs of one frame with ℓ_∞ adversarial perturbations for untargeted attack, the problem is formalized as

$$\max_{x'} \mathcal{L}(f_\theta(x'), \hat{y}), \text{ s.t. } ||x' - x||_\infty < \epsilon, \tag{5}$$

where ϵ is an allowed perturbation budget.

Instance-specific Adversarial Patches. Formally, the instance-specific patch attack is described as

$$\max_{\delta} \mathcal{L}(f_\theta((1 - m) \odot x + m \odot \delta), \hat{y}), \tag{6}$$

where δ is in the same space as image input x and $m \in \{0, 1\}^{N_c \times 1 \times H \times W}$ represents binary masking matrix to appoint the location of patches with element-wise multiplication of pixels denoted by $\odot$. The masking matrix m is defined according to the ground-truth 3D bounding boxes.

Category-specific Adversarial Patches. The formulation of this problem is

$$\max_{\delta_1, \cdots, \delta_C} \mathbb{E}_{(x,\hat{y})\sim\mathcal{D}}[\mathcal{L}(f_\theta((1-\sum_{j=1}^{C} m_j^x)\odot x+\sum_{j=1}^{C} m_j^x\odot\delta_j), \hat{y})], \tag{7}$$

where δ_j is for objects of j-th category in the dataset $\mathcal{D}$ which has C categories in total, while m_j^x denotes the binary mask for objects of the j-th category in the sample x. Similarly, the masking matrix m_j^x is defined according to the ground-truth 3D bounding box coordinates.

B. Additional Details for ℓ_∞ Adversarial Perturbations

B.1. Raw Data for ℓ_∞ Attack

The raw data for ℓ_∞ attack, including FGSM and PGD10, is shown in Tab. 7 and Tab. 8.

B.2. Details for Transfer Attack

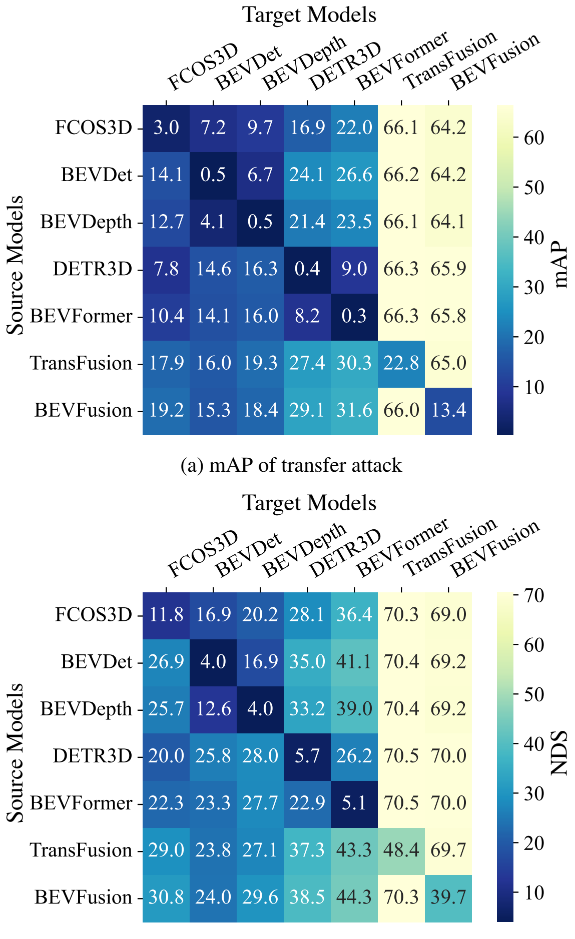
The details of transfer attack are shown in Fig. 6.

B.3. mAP data for ℓ_∞ Perturbations for Fusion Models

The mAP data for ℓ_∞ perturbations for two fusion models, TransFusion [4] and BEVFusion [35], is shown in Fig. 7.

C. mAP Data for Partial Cameras

The mAP data is shown in Fig. 8



(a) mAP of transfer attack

(b) NDS of transfer attack

Figure 6. **Transfer attack.** mAP, NDS of Camera-LiDAR fusion models under different adversarial perturbations added to images and point clouds.

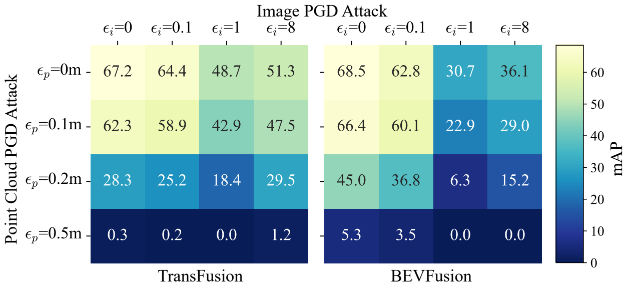


Figure 7. ℓ_∞ **Perturbations for Fusion Models.** mAP of Camera-LiDAR fusion models under different adversarial perturbations added to images and point clouds.

	Clean	FGSM(e=0.1)	FGSM(e=0.2)	FGSM(e=0.5)	FGSM(e=1)	FGSM(e=2)	FGSM(e=4)	FGSM(e=8)
FCOS3D	29.8/37.7	25.9/34.3	23.4/32.5	19.6/29.3	16.9/27	14.9/25.4	14.2/24.6	13.9/24.3
BEVDet	29.2/37.2	17.9/26.9	13.7/23.1	8.7/17.9	6.1/14	4.7/10.7	3.9/10	3.4/9.2
BEVDepth	33.2/40.4	21.6/30.4	16.7/27	10.9/21.8	7.9/17.7	6.2/14.4	5.3/13.7	4.6/9.1
DETR3D	34.7/42.2	21.5/31.1	17/27.6	12.4/24.1	10.4/22.7	9.7/22.1	10.4/22.5	12.3/24.3
BEVFormer	37.0/47.9	16.2/26.9	12.4/23.9	8.2/19.9	6.3/17	5.4/16.1	5.6/16.3	6.6/17.3
TransFusion	67.2/70.9	64.6/69.5	63.3/68.9	61.7/68.1	60.9/67.6	60.4/67.4	60.6/67.5	61.4/68
BEVFusion	68.5/71.4	62.7/68.1	58.5/65.7	52/62.1	48.1/59.8	46.2/58.8	46.9/59.2	50.6/61.4

Table 7. ℓ_∞ **FGSM attack.** mAP/NDS of FGSM attack at different ϵ settings

	Clean	PGD10(e=0.1)	PGD10(e=0.2)	PGD10(e=0.5)	PGD10(e=1)	PGD10(e=2)	PGD10(e=4)	PGD10(e=8)
FCOS3D	29.8/37.7	25.4/33.6	21.8/30.3	14.6/24	9.3/19.1	5.2/13.4	2.6/9	1.2/6.4
BEVDet	29.2/37.2	17.2/26.2	11.6/19.9	5.7/11.9	3.3/6.8	1.8/5.1	0.8/4.3	0.2/0.7
BEVDepth	33.2/40.4	21/29.8	14.6/25.5	7/12.7	3.7/7	1.8/5.2	0.8/4.4	0.3/0.7
DETR3D	34.7/42.2	19.5/29.6	12.2/24.4	4.7/19.2	1.8/15.9	0.8/8.9	0.4/5.6	0.3/5.1
BEVFormer	37.0/47.9	15.7/26.6	10.3/22.7	4.1/16	1.6/11.7	0.7/8.4	0.3/5.3	0.2/3.6
TransFusion	67.2/70.9	64.4/69.5	62.5/68.4	57.8/65.9	48.7/61.4	36.2/55.2	28.2/51.4	28.3/51.3
BEVFusion	68.5/71.4	62.8/68.2	57.3/65	44.1/57.7	30.7/50.1	19.1/43.3	11.6/38.6	7.5/36.1

Table 8. ℓ_∞ **PGD10 attack.** mAP/NDS of PGD10 attack at different ϵ settings

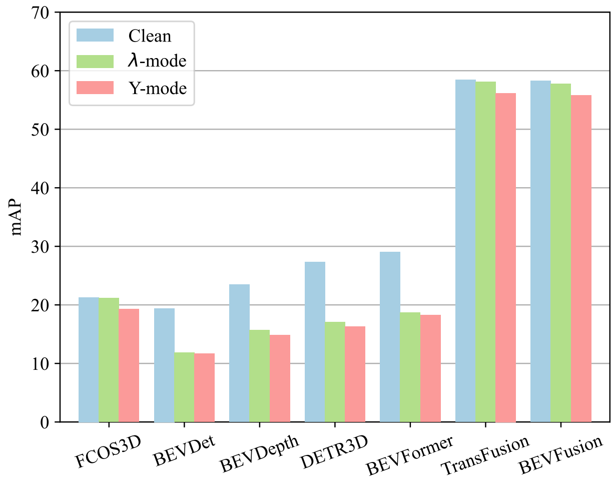


Figure 8. **Partial Cameras.** Performance with partial cameras in terms of mAP

D. Additional Details for 3D Consistent Patch Attack

Full results including mAP and NDS is show in Tab. 9 and Tab. 10.

E. Training strategies

We note that different training strategies could influence the model robustness to some extents and also take them into account. Though prior works on benchmarking robustness in classification [17] and detection [16] usually treat training strategies as part of models instead of dissociating them, we investigate the training strategies of each model including data augmentation, learning rate, optimizer, and find that there is an insignificant difference within the three comparing sub-groups of models. The detailed training schemes are summarized in Tab. 11.

Patch Size	0%	5%	10%
FCOS3D	21.3/32.5	**12.9/22.2**	**8.5/17.5**
BEVDet	19.4/33.8	3.3/11.1	1.4/6.2
BEVDepth	23.5/37.7	4.4/11.8	1.9/8.8
DETR3D	27.3/39.1	3.5/15.2	1.3/10.3
BEVFormer	**29.0/45.0**	4.1/14.8	1.8/9.6
TransFusion	**58.4/66.8**	**52.6/63.7**	**50.7/62.8**
BEVFusion	58.3/66.4	36.4/54.2	29.5/50.8

Table 9. **Multi-view Patch Attack.** mAP/NDS of vision-dependent models with 3D consistent patches in the cases of Multi-view Patch Attack. 0% for clean images.

Patch Size	0%	5%	10%
FCOS3D	29.8/37.7	11.9/20.6	6.0/15.3
BEVDet	29.2/37.2	3.8/12.7	1.7/5.9
BEVDepth	33.2/40.4	6.3/17.7	2.5/8.7
DETR3D	34.7/42.2	16.3/28.3	9.5/23.2
BEVFormer	**37.0/47.9**	**18.8/35.0**	**11.7/29.0**
TransFusion	67.2/70.9	**61.9/68.0**	**58.9/66.4**
BEVFusion	**68.5/71.4**	54.9/63.9	49.1/60.7

Table 10. **Temporally Universal Patch Attack.** mAP/NDS of vision-dependent models with 3D consistent patches in the cases of Temporally Universal Patch Attack. 0% for clean images.

	Optim.	lr	b.s.	epoch	image aug	3D aug	GT aug
FCOS3D [54]	SGD	2e-3	2*8	12	-	RandomFlip3D	-
BEVDet [26]	AdamW	2e-4	8*8	24	ResizeRotFilp	GlobalRotScaleTrans, RandomFlip3D	-
BEVDepth [31]	AdamW	2e-4	8*8	24	ResizeRotFilp	GlobalRotScaleTrans, RandomFlip3D	-
DETR3D [55]	AdamW	2e-4	1*8	24	PhotoMetricDistortion	-	-
BEVFormer [32]	AdamW	2e-4	1*8	24	PhotoMetricDistortion	-	-
TransFusion [4]	AdamW	1e-4	2*8	20+6	-	GlobalRotScaleTrans, RandomFlip3D	GTPaste
BEVFusion [35]	AdamW	1e-4	4*8	20+6	ResizeRotFilp	GlobalRotScaleTrans, RandomFlip3D	GTPaste

Table 11. **Different training strategies.** Training strategy information of different models evaluated in this paper, including optimizer (Optim.), learning rate (lr), batch size (b.s.), training epoch and different types of data augmentation strategies.

F. Different truncation levels for partial cameras

Following KITTI [19], we get the truncation ratio of objects at image boundaries and examine the performance of models at different levels. Results of NDS between DETR3D and BEVFormer are shown in Tab. 12. BEV model outperforms non-BEV model at all truncation levels for partial cameras. This is consistent with our finding in Sec. 4.3.

NDS	λ-mode				Y-mode			
	All	Easy	Moderate	Hard	All	Easy	Moderate	Hard
DETR3D	32.0	37.7	30.0	29.2	30.6	35.1	29.0	27.5
BEVFormer	**37.0**	**43.3**	**39.8**	**36.1**	**35.5**	**40.0**	**34.9**	**32.3**

Table 12. Partial results at different truncation levels for partial cameras.

G. BEVDet/BEVDepth with high resolution

Considering that the input resolution may influence the robustness of detectors, we test this influence on BEVDet and BEVDepth. We increase the resolution from 704x256 to 1408x512 and retrain the models by ourselves due to the lack of official models.

From Tab. 13 of PGD attack, we see that higher resolution improves the robustness of BEVDet and BEVDepth, but the performance is still inferior to FCOS3D. We also test them on temporally universal patch attack, and find similar trends, *e.g.*, under 5% patch size, the NDS of high resolution BEVDet and BEVDepth are 18.9 and 20.5, still lower than 20.6 NDS of FCOS3D. The conclusion is that the original resolution improves the robustness, but cannot substantially overturn our findings.

H. Similar schemes when discussing temporal information

To further investigate the effectiveness of temporal information on improving robustness, we conduct additional experiments among BEVDet & BEVDet4D and BEVFormer-Static (re-implemented) & BEVFormer. The results under two time-related attack settings are shown in Tab. 14. The results confirm that temporal information fortifies robustness to universal attack along time.

ϵ	0	0.1	0.2	0.5	1	2	4	8
BEVDet	37.2	26.2	19.9	11.9	6.8	5.1	4.3	0.7
BEVDet-HighRes	41.0	26.7	20.1	15.0	10.8	9.5	7.6	5.2
BEVDepth	40.4	29.8	25.5	12.7	7.0	5.2	4.4	0.7
BEVDepth-HighRes	**43.9**	31.2	26.4	18.2	11.0	9.1	7.0	4.4
FCOS3D	37.7	**33.6**	**30.3**	**24.0**	**19.1**	**13.4**	**9.0**	**6.4**

Table 13. NDS under ℓ_∞ adversarial perturbations generated by PDG10.

Category-Specific Patch (5%)				Temporally Universal Patch (5%)			
BEVD	BEVD4D	BEVF-S	BEVF	BEVD	BEVD4D	BEVF-S	BEVF
15.4	21.3	32.9	35.9	12.7	17.5	30.3	35.0

Table 14. Partial results (NDS) among models with temporal information.