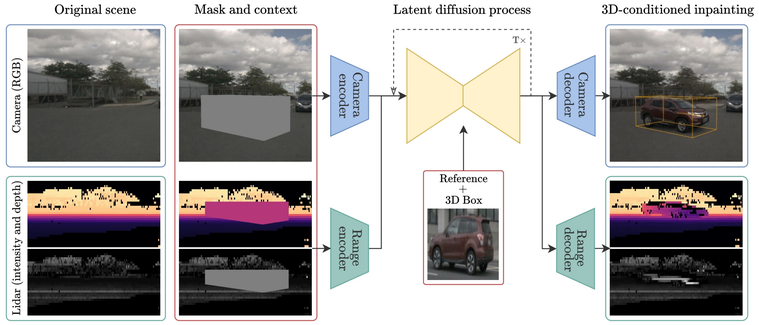
MObI: Multimodal Object Inpainting Using Diffusion Models

Alexandru Buburuzan[1,2] Anuj Sharma[1] John Redford[1] Puneet K. Dokania[1,3] Romain Mueller[1]
[1] Five AI [2] The University of Manchester [3] University of Oxford

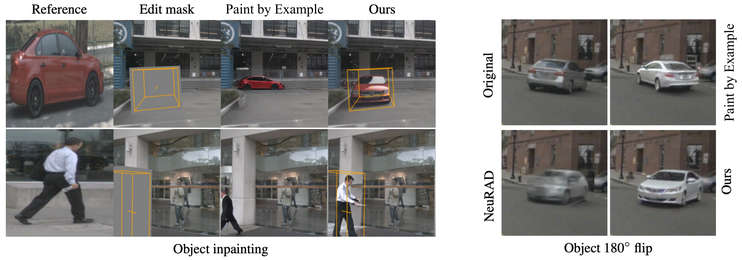
Overview



Extensive, realistic, and controllable multimodal data is critical for rigorous testing of safety-critical applications like autonomous driving, as real-world data collection is costly and complex. We introduce ***MObI***, a framework for Multimodal Object Inpainting that uses a diffusion model to insert realistic objects into driving scenes across camera and lidar, jointly.

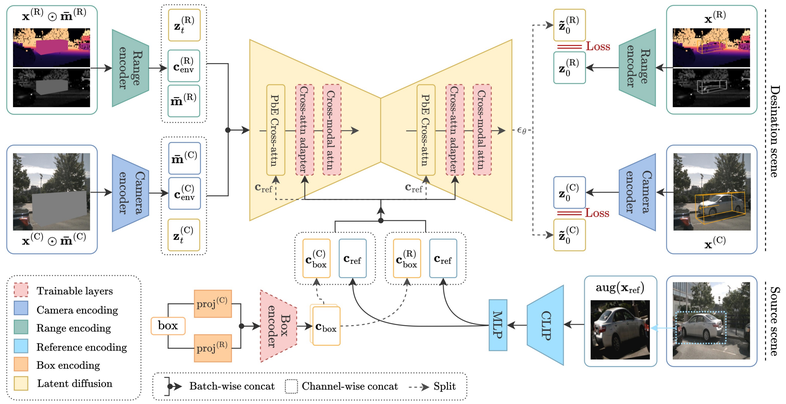
Conditioned on a single reference image and a 3D bounding box, MObI achieves semantic consistency, realistic spatial integration, and multimodal coherence. Our approach supports flexible, high-fidelity object insertion, offering a practical tool for ***generating counterfactuals*** and testing perception models.

Motivation



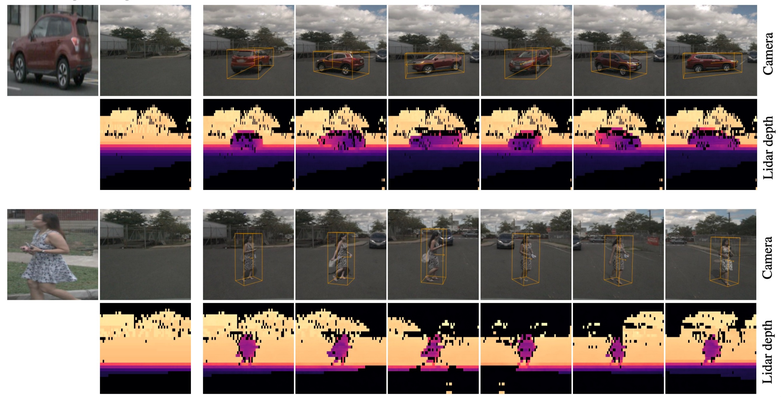
Inpainting methods based on edit masks alone achieve high realism but can lead to surprising results since there are multiple semantically consistent ways to inpaint an object; ***3D reconstruction methods*** are controllable but may lack realism for unobserved viewpoints.

Method



- We extend Paint-by-Example a reference-guided image inpainting diffusion model, to include ***3D bounding box conditioning*** and to ***jointly generate camera and lidar*** by finetuning sandwiched attention layers.
- We adapt the ***image autoencoder of Stable Diffusion*** to the range view modality.

Controllable Object Insertion



Experiments

Model	3D Box	Adapter	Reinsertion: Camera Realism FID↓	LPIPS↓	CLIP-I↑	Reinsertion: Lidar Realism D-LPIPS↓	I-LPIPS↓	Replacement: Camera Realism FID↓	LPIPS↓	CLIP-I↑	Replacement: Lidar Realism D-LPIPS↓	I-LPIPS↓
copy&paste		n/a		n/a		n/a		15.29	0.205	n/a	n/a	
PbE [65]		n/a	7.46	0.133	83.91	n/a		10.08	0.149	**77.25**	n/a	
MObI (256)	✗	✓	8.18	0.123	82.56	0.195	0.231	10.31	0.140	77.22	0.198	0.236
	✓	✗	8.31	0.120	82.88	0.188	0.231	10.43	0.134	76.03	0.191	0.237
	✓	✓	7.74	0.119	83.03	0.192	0.230	9.87	0.133	76.75	0.195	0.236
MObI (512)	✓	✓	**6.60**	**0.115**	**84.22**	**0.129**	**0.148**	**9.00**	**0.129**	76.75	**0.132**	**0.153**

Realism performance for camera and lidar demonstrates strong results across diverse insertion (using the same reference and temporal tracking) and replacement (in-domain and cross-domain reference) settings.

	Scene-level mAP car	mAP ped.	Restricted to reinserted objects ATE car	ATE ped.	ASE car	ASE ped.	AOE car	AOE ped.
Original	0.885	0.873	0.145	0.103	0.138	0.278	0.024	0.462
Reinsertions	0.878	0.863	0.299	0.140	0.145	0.303	0.161	0.754

Camera-lidar detection performance of an off-the-shelf BEVFusion [ICRA'23] object detector on objects reinserted using our method.

Discussion

- We introduce MObI, a method for realistic and controllable multimodal object inpainting across camera and lidar views.
- Results show strong spatial coherence, yet limitations remain in handling open-world references, extreme placements, and overlap with existing objects.
- Despite this, we think our approach offers an interesting, novel avenue to edit multimodal scenes in a realistic and controllable manner.

https://alexbubu.com/mobi