CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon
Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, Jérôme Revaud
NAVER LABS Europe
https://github.com/naver/CroCo

NEURAL INFORMATION PROCESSING SYSTEMS

NAVER LABS Europe

Monocular Downstream Tasks

- Finetuning the pre-trained encoder
- Comparison with DINO, MAE, MultiMAE pre-trainings

Results

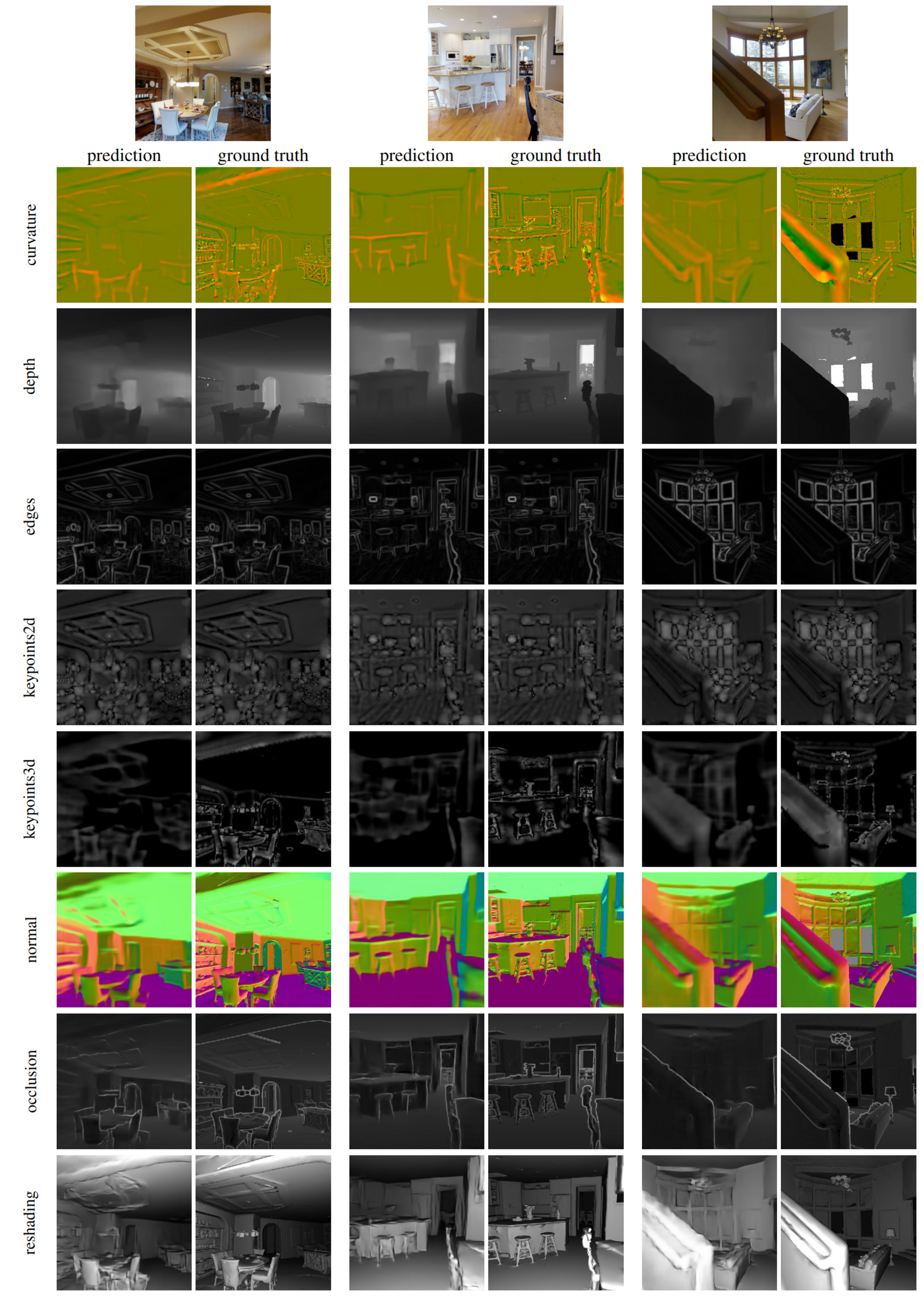
pre-training method (data)	NYUv2 ↑	Taskonomy ↓									
	depth	curv.	depth	edges	kpts2d	kpts3d	normal	occl.	reshad.	avg.	rank.
DINO (IN1K)	81.3	43.04	38.42	3.80	0.16	45.85	65.71	0.57	115.02	39.07	5.00
MAE (IN1K)	85.1	41.59	35.83	**1.19**	0.08	44.18	59.20	**0.55**	106.08	36.09	2.13
MultiMAE (IN1K)	86.4	41.42	35.38	2.17	**0.07**	44.03	60.35	0.56	105.25	36.17	2.75
MAE (Habitat)	84.0	42.06	33.63	1.79	0.08	44.81	59.76	0.56	102.54	35.65	2.88
CroCo (Habitat)	**87.8**	**40.91**	**31.34**	1.74	0.08	**41.69**	**54.13**	**0.55**	**93.58**	**33.00**	**1.25**

→ compares favorably on geometric tasks

pre-training method (data)	IN1K ↑	ADE ↑
	lin.	segm.
DINO (IN1K)	**78.2**	44.7
MAE (IN1K)	68.0	46.1
MultiMAE (IN1K)	60.2	**46.4**
MAE (Habitat)	32.5	40.3
CroCo (Habitat)	37.0	40.6

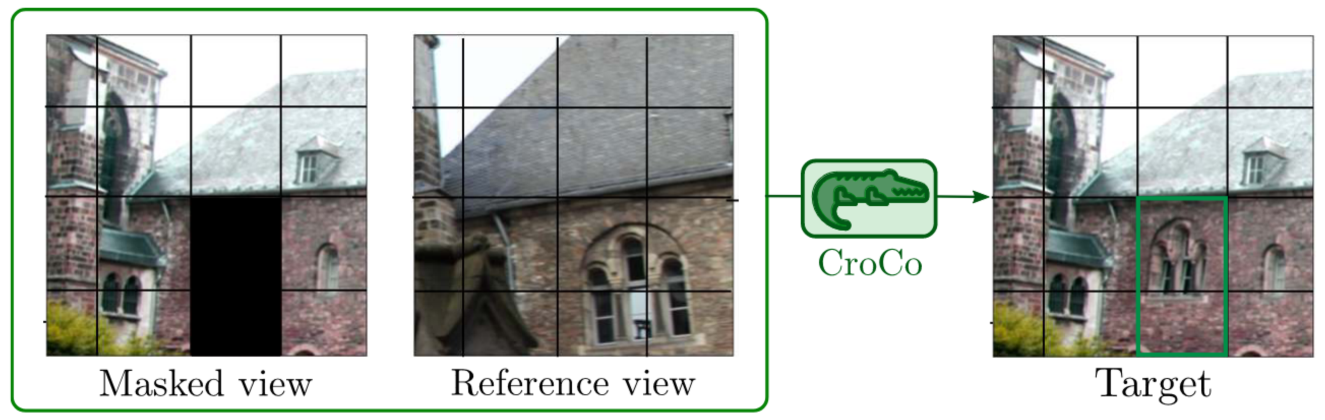
→ small performance drop on semantic tasks

Qualitative visualizations



A Novel Pretext Task for 3D vision

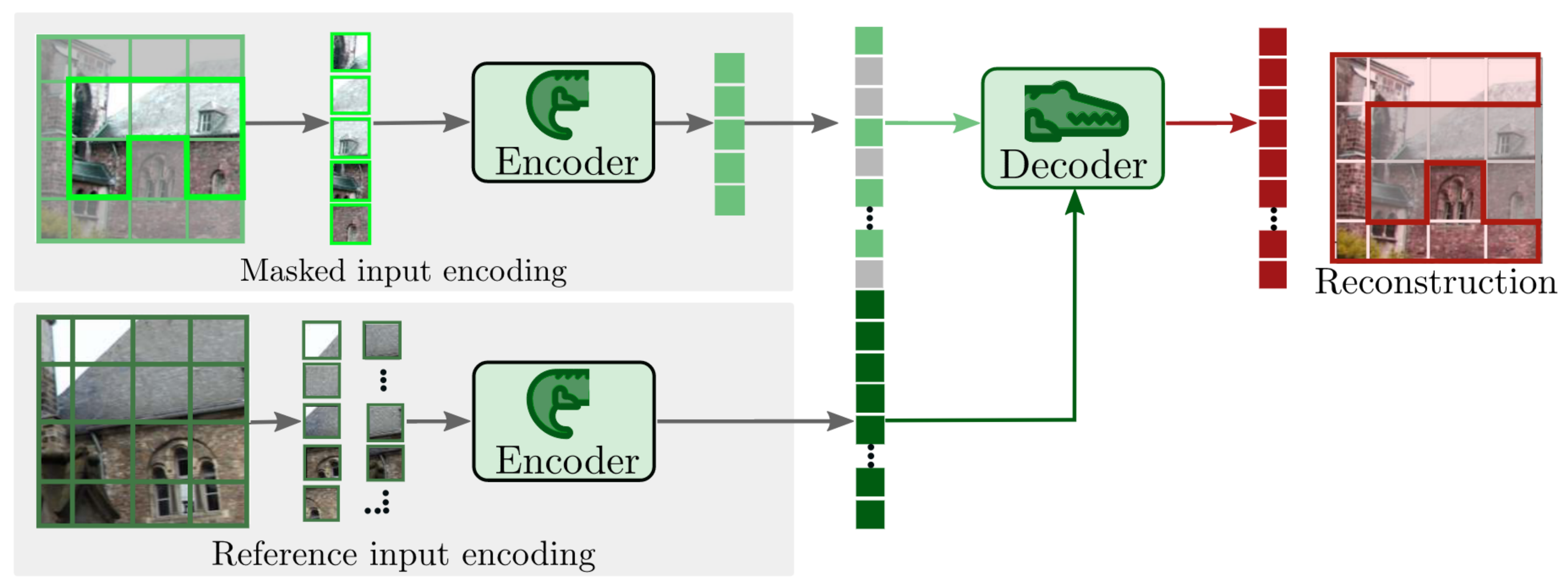
Cross-view Completion (CroCo)



- Masked image modeling (like MAE), but conditioned on a reference view
- **Implicitly learns 3D geometry to solve the task**

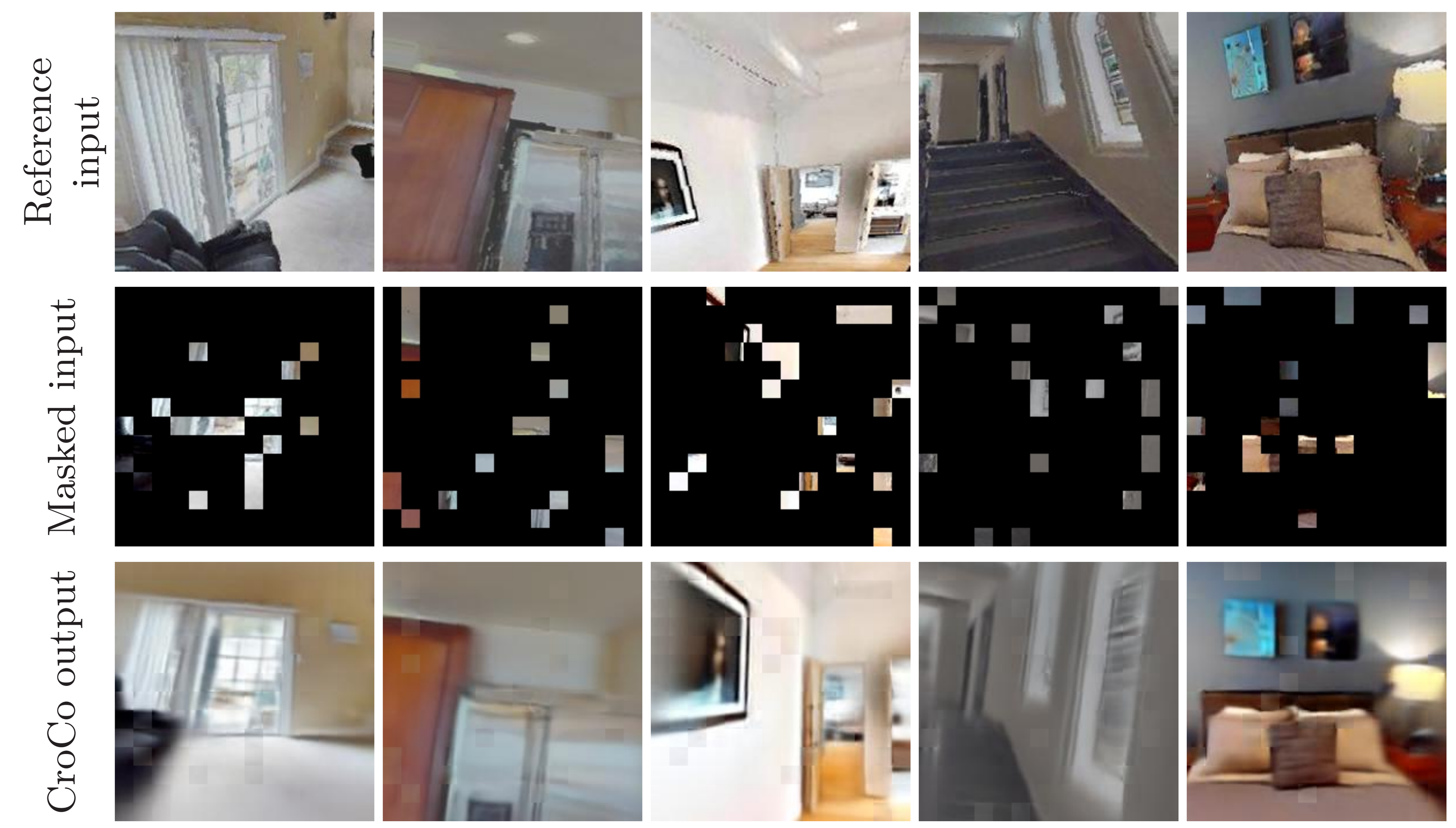
CroCo Architecture and Pre-training

Two views encoded separately and decoded together with cross-attention



pre-training data: 2M synthetic pairs of indoor scenes generated using Habitat-Sim

Cross-view Completion on validation scenes



Binocular Downstream Tasks

- Finetuning both the pre-trained encoder and decoder
- On par with state of the art methods without task-specific design

Stereo matching estimation on KITTI and ETH3D

KITTI 2015

Method	D1-bg↓	D1-fg↓	D1-all↓
AdaStereo	2.59	5.55	3.08
HITNet	1.74	3.20	1.98
PCWNet	**1.37**	3.16	1.67
GMStereo	1.49	3.14	1.77
ACVNet	**1.37**	3.07	**1.65**
LEAStereo	1.40	2.91	**1.65**
CREStreo	1.45	2.86	1.69
CroCo	1.54	**2.58**	2.03

ETH3D

Method	bad@0.5 (%)↓		bad@1.0 (%)↓		avg err (px)↓	
	noc	all	noc	all	noc	all
AdaStereo	10.22	10.85	3.09	3.34	0.24	0.25
HITNet	7.89	8.41	2.79	3.11	0.20	0.22
RAFT-Stereo	7.04	7.33	2.44	2.60	0.18	0.19
DIP-Stereo	6.74	6.99	1.97	2.12	0.18	0.20
GMStereo	5.94	6.44	1.83	2.07	0.19	0.21
CREStereo	3.58	3.75	**0.98**	**1.09**	**0.13**	**0.14**
CroCo	**3.27**	**3.51**	0.99	1.14	0.14	0.15

See [Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow, Weinzaepfel et al., arXiv'22] for details

Optical flow on MPI Sintel

End-Point-Error (↓)

Method	clean	final
PWC-Net+	3.45	4.60
RAFT	1.61	2.86
CRAFT	1.44	2.42
FlowFormer	1.20	**2.12**
SKFlow	1.30	2.26
GMFlow+	**1.03**	**2.12**
CroCo	1.22	2.58

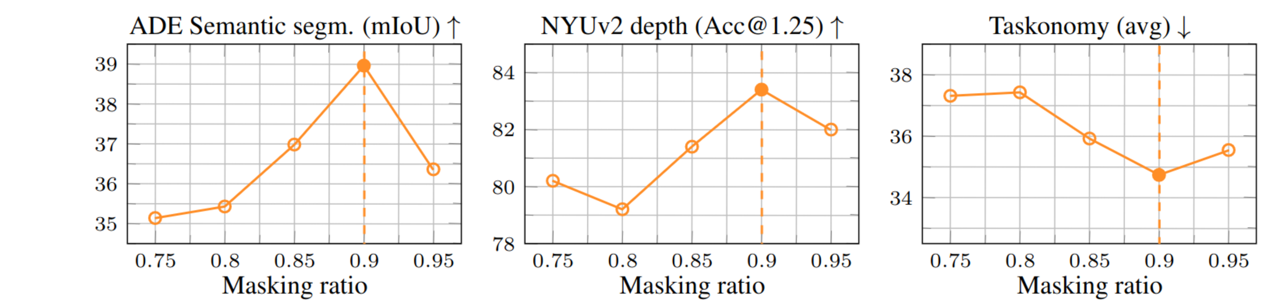
Relative pose estimation on 7-scenes

Method / pre-training	Average
RelocNet*	21cm, 6.74°
NC-EssNet*	21cm, 7.50°
CamNet*†	**4cm, 1.69°**
top1 AP-GeM-18	36cm, 14.2°
MAE (Habitat)	24.8cm, 13.09°
CroCo (Habitat)	5.0cm, 3.46°

*: fuse multiple pose predictions
†: exploit temporal information and multi-step retrieval

Ablations

Masking ratio: 90% performs best on all downstream tasks



Viewpoint change between images is important

image pairs from	ADE ↑	NYUv2 ↑	Taskonomy ↓	
	segm.	depth	avg.	rank.
two viewpoints	**38.8**	**86.8**	**33.56**	**1.00**
geometric transformations of one image	26.1	65.0	48.33	2.00