End-to-End (Instance)-Image Goal Navigation through Correspondence as an Emergent Phenomenon

NAVER LABS Europe

ICLR Vienna, 2024

Guillaume Bono Leonid Antsfeld Boris Chidlovskii Philippe Weinzaepfel Christian Wolf

Motivation

Task: ImageNav,: navigation to a goal specified by an image.

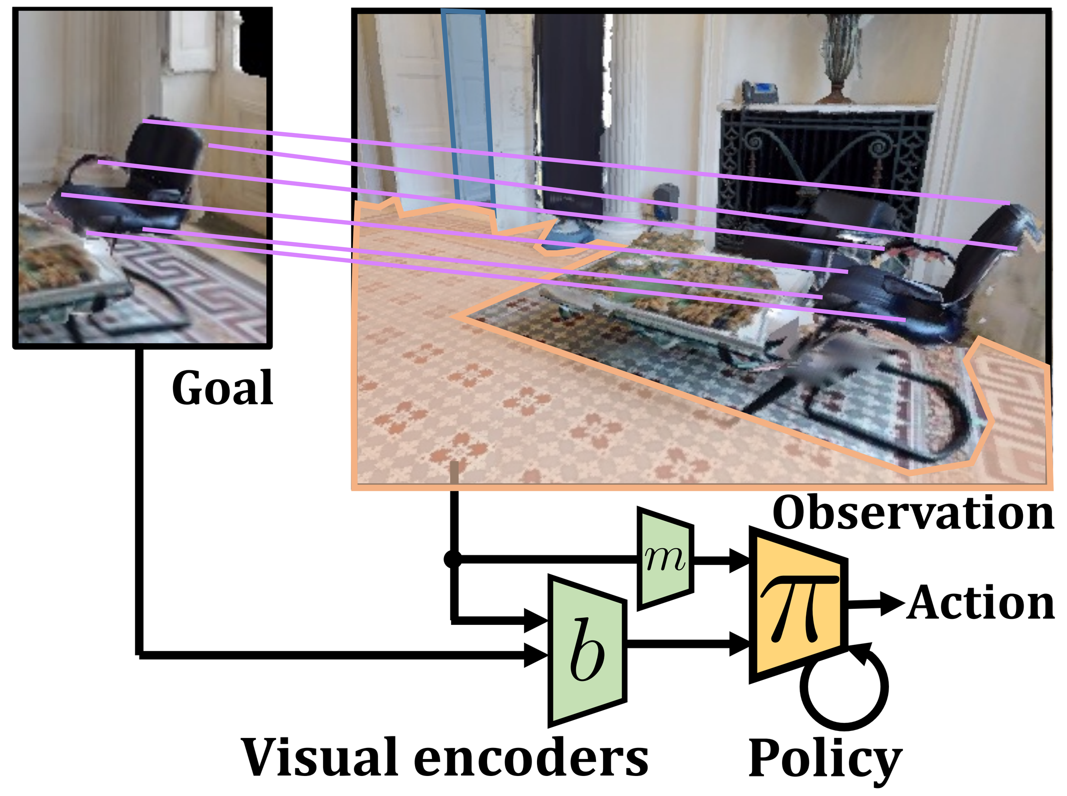
Requires: (1) Nav skills: detection of navigable space, exits etc.
(2) Detection of relative pose wrt the goal

Cur SOTA: Train visual encoder + policy end-to-end, or outsource goal detection to local feature based methods.

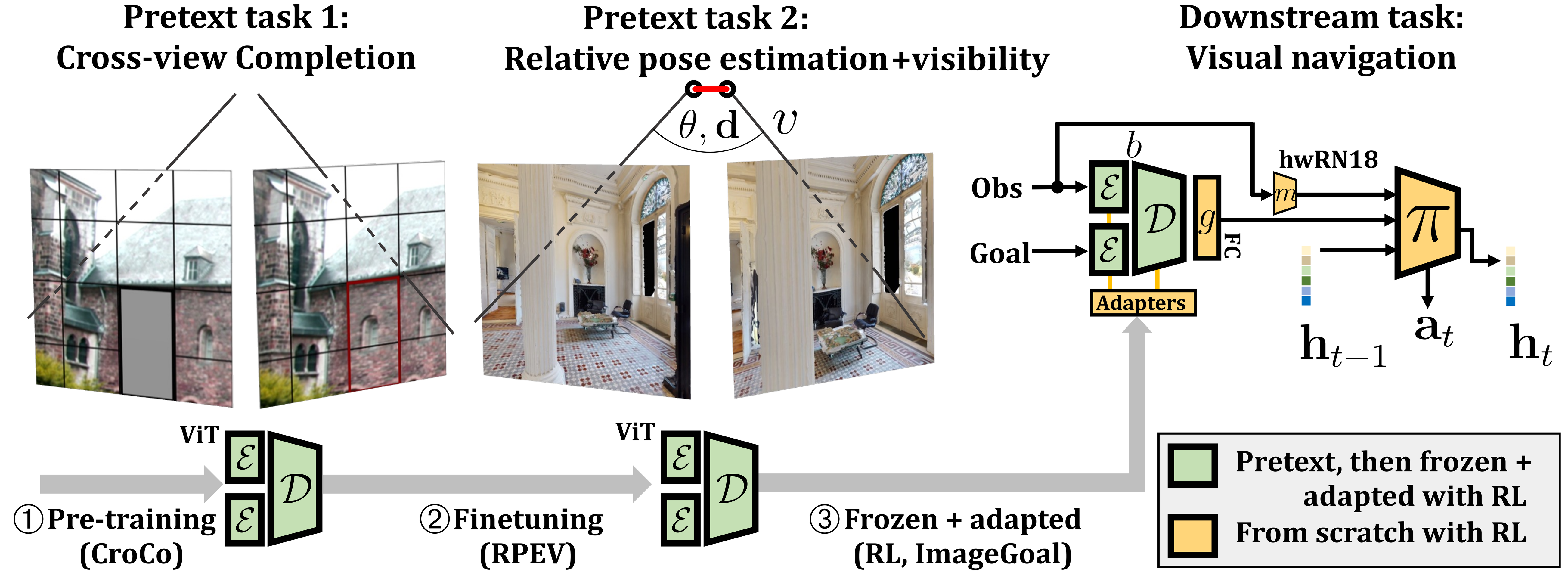
Idea: Pre-train a binocular ViT with cross-attention on different losses:

- Cross-view competition as in Weinzaepfel et al., NeurIPS 2022
- Extremely-wide baseline relative pose + visibility est.

Use as visual encoder in an end-to-end policy trained with RL. Encoder is frozen + adaptors.



Training pipeline



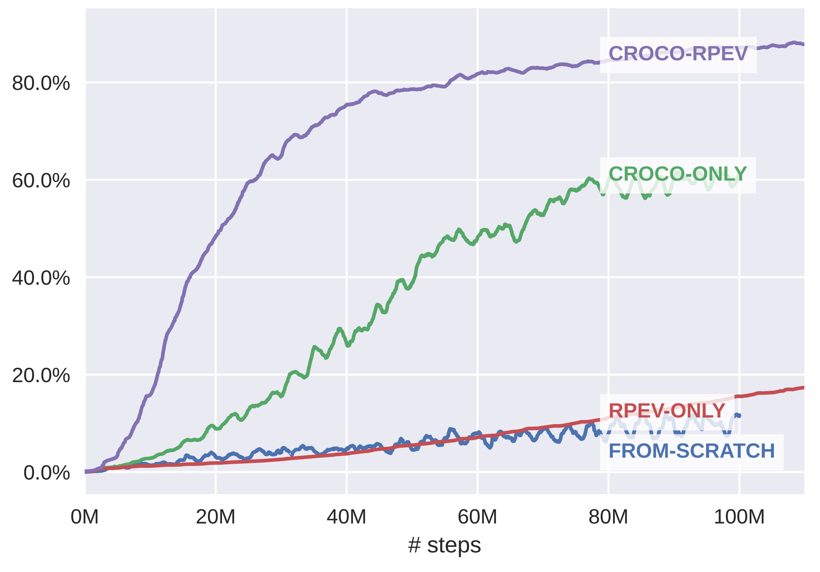
Correspondence solutions emerge after pre-training (CroCo + RPEV)

Shown: cross-attention of the last head average over all heads. Patches selected based on total attention.

Impact of architecture variants and sizes

Variant	Encoder L	Encoder H	Encoder d	Decoder L	Decoder H	Decoder d	#params (binoc)	Monoc	% correct poses 1m&10°	% correct poses 2m&20°	Vis-acc (%)	Nav. perf. SR (%)	Nav. perf. SPL (%)
DEBiT-L ("Large"), no adapters	12	12	768	8	16	512	120M	hwRN18	**97.5**	**98.9**	**94.0**	82.0	**59.6**
DEBiT-B ("Base"), no adapters	12	6	384	8	16	512	55M	hwRN18	92.5	96.8	89.3	**83.0**	55.6
DEBiT-S ("Small"), no adapters	12	6	384	2	8	256	24M	hwRN18	82.7	93.5	81.6	79.6	52.1
DEBiT-T ("Tiny"), no adapters	8	6	384	2	8	256	17M	hwRN18	80.3	92.4	80.6	79.3	50.0

Impact of pre-training strategies



Variant	Pre-train CroCo	Pre-train RPEV	% corr. poses 1m&10°	% corr. poses 2m&20°	Vis-acc (%)	Nav. perf. SR	Nav. perf. SPL
DEBiT-L, no adapters	✗	✗	n/a	n/a	n/a	7.0	4.4
DEBiT-L, no adapters	✓	✗	n/a	n/a	n/a	60.2	33.1
DEBiT-L, no adapters	✗	✓	40.1	66.7	58.3	11.8	9.9
DEBiT-L, no adapters	✓	✓	**97.5**	**98.9**	**94.0**	**82.0**	**54.8**
DEBiT-B, no adapters	✗	✗	n/a	n/a	n/a	6.8	4.0
DEBiT-B, no adapters	✓	✗	n/a	n/a	n/a	65.7	37.3
DEBiT-B, no adapters	✗	✓	39.7	66.4	58.8	23.6	17.4
DEBiT-B, no adapters	✓	✓	**92.5**	**96.8**	**89.3**	**81.2**	**53.0**

Comparison w. SOTA: ImageNav

Method	#steps	SR(%)	SPL(%)	Pretrained weights
Siam. hwRN18	180M	10.1	9.6	None, from scratch
Siam. hwRN18 [2]	500M	-	8.0[1]	None, from scratch
Mem. Aug. (Mezghani et al., 2022)[3]	500M	-	9.0[1]	Finetuned
ZSEL (Al-Halah et al., 2022)	500M	29.2[1]	21.6[1]	Obs.&policy frozen, goal from scratch
ZSON (Majumdar et al., 2022a)	500M	36.9[1]	28.0[1]	Obs. finetuned, goal frozen (CLIP)
VC1-ViT-L (Majumdar et al., 2023)	500M	81.6[1]	-	Finetuned
OVRL (Yadav et al., 2022)	500M	54.2[1]	27.0[1]	Finetuned
OVRL-v2 (Yadav et al., 2023)	500M	82.0[1]	58.7[1]	Finetuned
ANS (Chaplot et al., 2020b) + DEBiT-L		32.0	15.0	Modular architecture + our frozen encoder
Ours (DEBiT-B), no adapters	200M	83.0	55.6	Frozen
Ours (DEBiT-L), no adapters	200M	82.0	59.6	Frozen
Ours (DEBiT-L) + adapters	200M	**94.0**	**71.7**	Frozen + adapted

[1] *Perf. from orig. papers;* [2] *Mono-view ablation of baseline in Table III of (Mezghani et al., 2022);*
[3] *Retrained in mono-view settings, see Table 1 of (Al-Halah et al., 2022)*

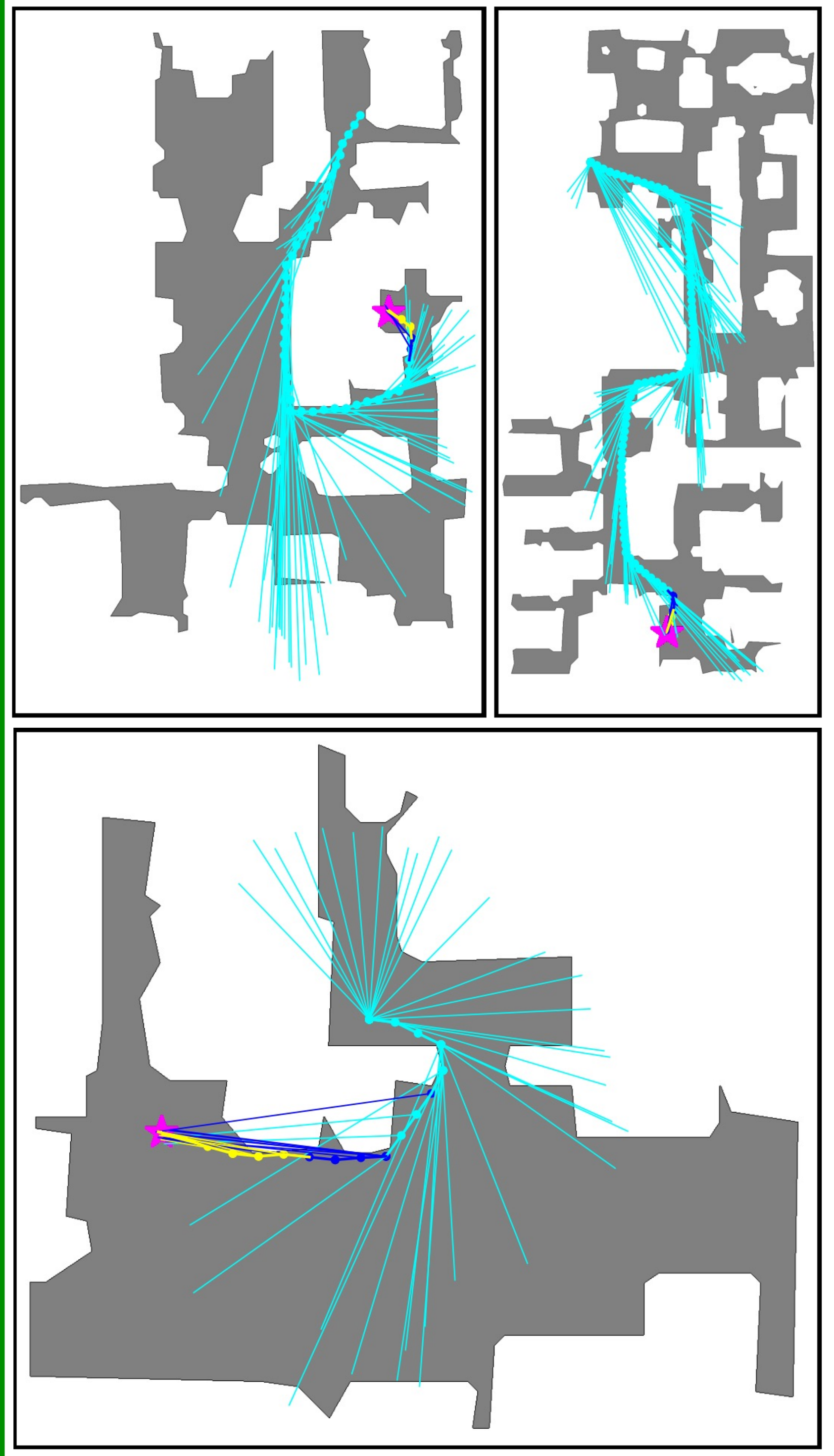
Comparison w. SOTA: Instance ImageNav

Method	#steps	SR (%) max	SR (%) avg	SPL (%) max	SPL (%) avg
(Krantz et al., 2022)	3500M	5.5	n/a	2.3	n/a
(Krantz et al., 2023)	n/a	56.1	n/a	23.3	n/a
Ours(DEBiT-L)+adapters	200M	**61.1**	**59.3**	**33.5**	**32.4**

Task: camera intrinsics and height may differ btw obs and goal.

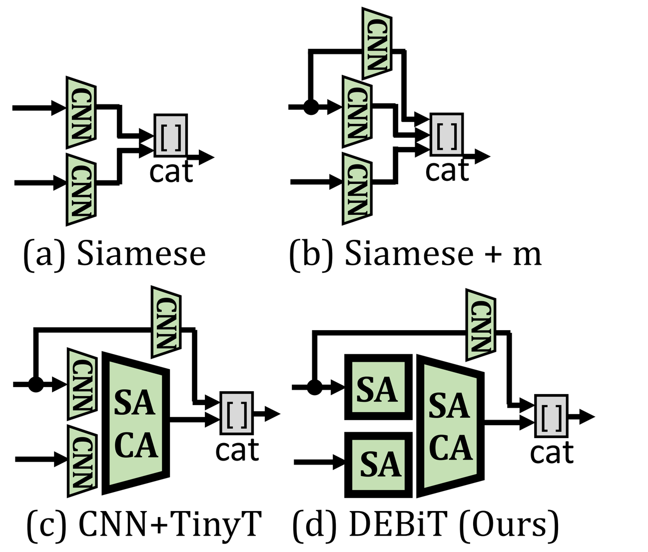
Method: pre-train on ImageNav, finetune policy and adaptors w. RL on Instance-ImageNav

RPEV Results during Nav



TN($v^*<\tau, v<\tau$), TP ($v^*>\tau, v>\tau$)
FP ($v^*<\tau, v>\tau$), FN ($v^*>\tau, v<\tau$)

Alignment architecture / loss



Visual encoder	Pre-train	#parms	SR	SPL
(a) Siamese hwRN18*	No	4.1M	10.1	9.6
(b) Siamese hwRN18*+m	RPEV	8.3M	8.0	7.7
(c) hwRN18+Cross-Att+m	No	10M	7.4	4.7
(c) hwRN18+Cross-Att+m	RPEV	10M	7.4	7.2
(d) DEBiT-B (Ours), no adapters	No	60M	6.8	4.0
(d) DEBiT-B (Ours), no adapters	CroCo+RPEV	60M	**83.0**	**55.6**

* *Baseline in (Mezghani et al., 2022), inspired by (Zhu et al., 2017)*