Pre-Trained Binocular ViTs for Image-Goal Navigation

Guillaume Bono[1] **Leonid Antsfeld**[1] **Boris Chidlovskii**[1]

Philippe Weinzaepfel[1] **Christian Wolf**[1]

[1]**Naver Labs Europe**
Meylan, France
`<firstname>.<lastname>@naverlabs.com`

Abstract

Most recent work in visual goal-oriented navigation resorts to large-scale machine learning in simulated environments. The main challenge lies in learning compact map-like representations that generalize to unseen environments and high-capacity perception modules capable of reasoning on high-dimensional input. The latter is particularly difficult when the goal is given as an exemplar image (*Image Goal*), as the perception module needs to learn a comparison strategy requiring to solve an underlying visual correspondence problem. This has been shown to be difficult from reward alone or with standard auxiliary tasks. We address this problem using two pretext tasks, which serve as a prior for what we argue is one of the main bottleneck in perception: wide-baseline relative pose estimation and visibility prediction in complex scenes. Our first pretext task, cross-view completion, is a proxy for the underlying visual correspondence problem, while the second task addresses goal detection and localization directly. We propose a new dual encoder making use of a binocular ViT model. Experiments show significant improvements on *Image Goal* navigation performance.

1 Introduction

Visual goal-oriented navigation (*ImageNav*) is usually addressed through large-scale training in simulation. While decision taking has not yet been solved either, recent research provides evidence that perception faces major challenges: learning representations required for planning, extracting 3D information without reliable depth measurements, and generalizing to unseen environments.

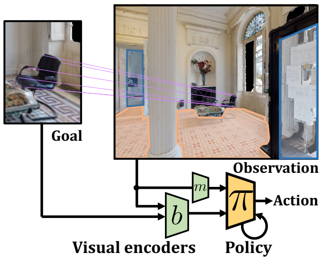

Figure 1: Navigation requires detecting navigable space, exits, and the goal. The correspondence solutions required by pose estimation emerge from pretraining.

As shown on Figure 1, the perception module of the agent needs to learn several skills, including the detection of obstacles, navigable areas, exits, and goals. The detection of visual goals given by exemplar requires to solve a partial matching task, which in essence is a classical *wide-baseline visual correspondence problem*, at the heart of many methods in visual localization and relative pose estimation [11, 22, 23]. However, in navigation, robot

NeurIPS 2023 Workshop on Robot Learning: Pretraining, Fine-Tuning, and Generalization with Large Scale Models, New Orleans, United States of America

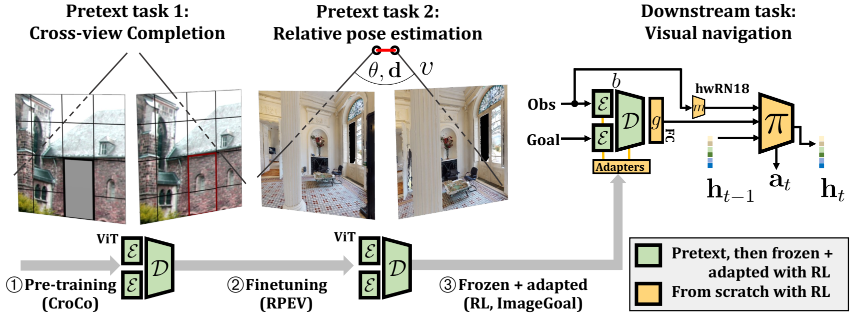

Figure 2: Two pretext tasks: ① CROss-view COmpletion [26], and ② Relative Pose Estimation with Visibility. They are learned by a binocular ViT b which is then combined with a monocular encoder m taking only observations, forming the dual encoder *DEBiT*. The combined embeddings are provided to a recurrent policy π, maintaining memory $\mathbf{h}_t$ and predicting actions a_t. Monocular encoder m and policy are trained with RL ③, the high-capacity model b is frozen, except for –optional– adapters.

perception is often addressed through scene reconstruction, eg. with SLAM [3, 15, 25] or end-to-end reinforcement [12, 30] or imitation [8] learning of a visual encoder. The former delegates goal detection to an external component. The latter attempts to solve the problem implicitly without direct supervision, through weak learning signals.

We propose a new method to address *ImageNav* through end-to-end training which takes advantage of CROss-view COmpletion [26] (*CroCo*), a powerful unsupervised pre-training for low-level scene understanding, followed by wide-baseline Relative Pose Estimation with Visibility (*RPEV*). We show that the underlying correspondence problem solved by *CroCo*, and the representations learned for *RPEV* are particularly relevant to the *ImageNav* task. We obtain SotA performance on two standard benchmarks of *ImageNav* and *Instance-ImageNav* tasks.

2 Learning perception for Visual goal-oriented navigation

In the *ImageNav* task, an agent is asked to navigate in an unknown scene to a goal only described as a randomly-oriented view $\mathbf{x}^* \in \Omega_{\text{rgb}} = [0, 255]^{3\times112\times112}$, taken from an unknown target location (Figure 1). The agent observes the scene through a RGB camera providing at each time-step t a single image $\mathbf{x}_t \in \Omega_{\text{rgb}}$, and selects actions $a_t \in \mathcal{A} = \{\texttt{STOP}, \texttt{FORWARD 25cm}, \texttt{TURN LEFT 10°}, \texttt{TURN RIGHT 10°}\}$. Navigation is successful if the agent calls STOP within $1m$ (geodesic) of the goal location.

Instance-ImageNav differs in the fact that goal views $\mathbf{x}^* \in \Omega_{\text{goal}}$ now target specific objects, and are taken from a different camera (eg. field of view, resolution, tilt, height). It also allows the agent to tilt its camera up and down using two new actions.

Our objective is to learn a perception module to predict a latent representation from an observation and goal. We conjecture that this requires the following three perception skills: **(S1)** Low-level geometric perception (eg. nav. area, exits, . . .); **(S2)** Perception of semantic categories (eg. floor, walls, . . .); **(S3)** Specific object detection and localization (goal).

In end-to-end approaches, these skills have been traditionally learned directly from reward or with expert demonstrations, potentially supported by auxiliary tasks [7, 6, 20, 16]. We propose a dual visual encoder combined with two phases of pre-training. Dubbed "*DEBiT*" for *Dual Encoder Binocular Transformer*, it consists in a binocular ViT model $b(\mathbf{x}_t, \mathbf{x}^*)$, which targets skill **(S3)**, goal detection and localization, and a monocular model $m(\mathbf{x}_t)$, implemented as a half-width ResNet18 (*hwRN18*) which targets skills **(S1)** and **(S2)** not related to the goal $\mathbf{x}^*$ — see Figure 1. The two encoders produce embeddings $\mathbf{e}^b_t$ and $\mathbf{e}^m_t$, respectively, which are integrated into a recurrent policy:

$$\langle a_t, \mathbf{h}_t \rangle = \pi(\mathbf{e}^b_t, \mathbf{e}^m_t, \text{emb}(a_{t-1}), \mathbf{h}_{t-1}) \tag{1}$$

Training the large-capacity binocular encoder b entirely from scratch through reward in navigation is difficult. This weak learning signal cannot handle the underlying geometric correspondence problem.

Training perception separately through losses highly correlated to the perception skills we identified above, in particular **(S3)**, proved to be a key design choice — see Figure 2.

2.1 CROss-view COmpletion

Introduced in [26], *CroCo* is a method inspired by masked image modeling [10], extended to a binocular configuration. It learns to capture low-level geometry cues to extract information from one input image $\mathbf{x}^*$ guided by a loss on reconstruction $\hat{\mathcal{P}}$ of masked patches $\bar{\mathcal{P}} \subset \mathcal{P}$ in a second image $\mathbf{x}_t = \bigcup \mathcal{P}$ taken from a slightly different point of view:

$$\hat{\mathcal{P}} = d(g(b(\mathcal{P} \backslash \bar{\mathcal{P}}, \mathbf{x}^*))) \tag{2}$$

$$l_{\text{croco}} = \sum_{\hat{\mathbf{p}}, \mathbf{p} \in \hat{\mathcal{P}}, \bar{\mathcal{P}}} \text{mse}(\hat{\mathbf{p}}, \mathbf{p}) \tag{3}$$

where b is the binocular encoder we want to pre-train, g is a patch-wise linear projection, and d is a deconvolution head reconstructing patches from embeddings.

2.2 Relative Pose Estimation with Visibility

Once the binocular encoder b has been pre-trained by *CroCo*, we finetune it on a second pretext task: Relative Pose Estimation with Visibility (*RPEV*). Given a pair of images $\langle \mathbf{x}_t, \mathbf{x}^* \rangle$, the network needs to predict the position $\mathbf{t} \in \mathbb{R}^3$, rotation $\mathbf{R} \in SO(3)$ of the goal relative to the observation. In addition, we make the network predict an additional scalar called *visibility* $v \in [0, 1]$ representing the fraction of goal pixels visible from current observation. We use this visibility metric as a threshold on the translation and rotation loss to ensure feasibility of the prediction, but also learn a very relevant signal for goal detection during navigation:

$$\langle \hat{\mathbf{t}}, \hat{\mathbf{R}}, \hat{v} \rangle = f(g(b(\mathbf{x}_t, \mathbf{x}^*))) \tag{4}$$

$$l_{\text{rpev}} = |\hat{v} - v| + \mathbb{1}_{v > \tau}(|\hat{\mathbf{t}} - \mathbf{t}| + |\hat{\mathbf{R}} - \mathbf{R}|) \tag{5}$$

where f is a prediction head, implemented as three MLPs, $\mathbb{1}$ is the indicator function, and $\tau = 0.2$ is a parametrizable threshold to toggle relative pose supervision.

We collect our own synthetic dataset of image pairs annotated with ground truth relative pose and visibility in simulated scenes from the MP3D [2], Gibson [27] and HM3D [21] datasets. We control for the distribution of geodesic distances between observations and goal by uniformly sampling observations along the path from start to goal locations among 4 categories: "in reach" $d \leq 1$m, "very close" $d \leq 1.5$m, "close" $d \leq 2$m, "approaching" $d \leq 4$m, and "far" $d > 4$m.

2.3 Navigation

We train the parameters of the recurrent policy π and the monocular encoder m jointly from scratch with PPO [24] with a reward definition in the lines of the one proposed by [4] for *PointGoal*:

$$r_t = 10 \cdot \mathbb{1}_{\text{success}} - 1 \cdot \Delta d - 0.01 \tag{6}$$

where a large sparse reward is granted on success, agent is densely guided to reduce geodesic distance to goal d, and a small slack cost encourages shorter path.

The backbone of the binocular encoder b is kept frozen from the *RPEV* pre-training. Optionally, we augment it with small *adapter* layers [5] which are trained with RL, alongside m and π.

3 Experimental results

3.1 Experimental setup

We evaluate *DEBiT* on the standard validation episodes for both *ImageNav* [19] and *Instance-ImageNav* [13] tasks. We train our models for 200M steps on a single A100 GPU. RPE is evaluated over the pairs with visibility $> \tau$ as percent of correct predictions for a given tolerance. Visibility is evaluated over all pairs in the same way. Finally for navigation, we report both Success Rate (**SR**) and Success weighted by Path Length (**SPL**) [1].

3.2 Impact of our pre-training strategy

Table 1 gives *RPEV* and *ImageNav* results after 100M steps of PPO training comparing different pre-training strategies for two variants of *DEBiT*, DEBiT-L (130M params) and DEBIT-B (66M params). Directly training the binocular encoder b from scratch did not lead to exploitable results, reward as a learning signal is too weak. *CroCo* pre-training is essential, directly training on RPEV led to low performance, but it is not optimal, and *RPEV* adds a significant boost to the gain provided by self-supervision alone.

Table 1: Impact of pre-training (100M nav. steps)

Variant	Pre-train		% corr. poses		Vis-acc	ImageNav	
	CroCo	RPEV	1m&10°	2m&20°	5%	SR	SPL
DEBiT-L	✗	✗	n/a	n/a	n/a	7.0	4.4
DEBiT-L	✓	✗	n/a	n/a	n/a	60.2	33.1
DEBiT-L	✗	✓	40.1	66.7	58.3	11.8	9.9
DEBiT-L	✓	✓	**97.5**	**98.9**	**94.0**	**82.0**	**54.8**
DEBiT-B	✗	✗	n/a	n/a	n/a	6.8	4.0
DEBiT-B	✓	✗	n/a	n/a	n/a	65.7	37.3
DEBiT-B	✗	✓	39.7	66.4	58.8	23.6	17.4
DEBiT-B	✓	✓	**92.5**	**96.8**	**89.3**	**81.2**	**53.0**

3.3 Comparison with prior work

Table 2 compares the proposed model with prior work. *DEBiT* largely outperforms the competing methods, including ones based on large-capacity ViTs like *VC1* [17] and *OVRL-v2* [29]. Both have been pre-trained with monocular masked image encoding, but perform late fusion of observation and goal, which does not ease learning geometric comparisons.

We also compare with the state-of-the-art in the *Instance-ImageNav* task. As both pre-training phases have been conducted in *ImageNav* settings, adapting *DEBiT* was a key design choice. Without adapters performance was actually unexploitable. We outperform the current SotA method *Mod-IIN*[14] and show that this task can also be addressed without feature matching.

Table 2: Comparison with SotA

Task	Method	#steps	SR(%)	SPL(%)
ImageNav	Siam. hwRN18	180M	10.1	9.6
	Siam. hwRN18 [2]	500M	-	8.0[1]
	Mem. Aug. [19][3]	500M	-	9.0[1]
	ZSEL [9]	500M	29.2[1]	21.6[1]
	ZSON [18]	500M	36.9[1]	28.0[1]
	VC1-ViT-L [17]	500M	81.6[1]	-
	OVRL [28]	500M	54.2[1]	27.0[1]
	OVRL-v2 [29]	500M	82.0[1]	58.7[1]
	Ours (DEBiT-B)	200M	83.0	55.6
	Ours (DEBiT-L)	200M	82.0	59.6
	Ours (adapted DEBiT-L)	200M	**94.0**	**71.7**
Instance ImageNav	IIN RL base[13]	3500M	5.5[1]	2.3[1]
	Mod-IIN[14]	n/a	56.1[1]	23.3[1]
	Ours (adapted DEBiT-L) - max	200M	**61.1**	**33.5**
	Ours (adapted DEBiT-L) - avg	200M	**59.3**	**32.4**

[1] Perf. from orig. papers [2] Mono-view ablation of baseline in Table III of [19]

[3] Retrained in mono-view settings, see Table I of [9]

For *ImageNav*, adding adapters to *DEBiT* also brings large improvements. We conjecture, that the adapters allow to pass richer information to the policy π through $\mathbf{e}_t^b$ from the otherwise entirely frozen binocular encoder b of *DEBiT*.

3.4 Emergence of correspondences

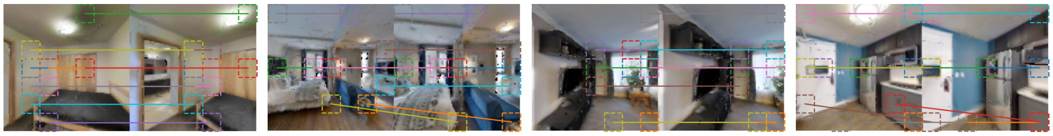

Figure 3: Correspondences emerging from last cross-attention activations in decoder

In Figure 3, we visualize top-k attention over patches, averaged over heads, in the last cross-attention layer of a DEBiT-L model. This gives qualitative results on the relevance of our pre-training strategy for goal detection and localization during navigation. Correspondence solutions naturally emerge without explicit supervision. A video available at this link also shows how the visibility prediction seems to be correlated with the exploration/navigation policy of the agent.

4 Conclusion

Our proposed method enables training a perception module consisting in our dual visual encoder *DEBiT*. It decomposes the problem into three training phases: two pretext tasks, *CroCo* and our *RPEV*, that enable us to address the challenging *ImageNav* and *Instance-ImageNav* tasks, through end-to-end reinforcement learning. We establish new state-of-the-art results on both tasks. We show that this makes solutions of correspondence problem emerge without explicit supervision.

References

[1] Peter Anderson et al. "On Evaluation of Embodied Navigation Agents". In: *arXiv preprint* (2018).

[2] Angel Chang et al. "Matterport3D: Learning from RGB-D data in indoor environments". In: *3DV*. 2018.

[3] Devendra Singh Chaplot et al. "Learning To Explore Using Active Neural SLAM". In: *ICLR*. 2020.

[4] Prithvijit Chattopadhyay et al. "RobustNav: Towards Benchmarking Robustness in Embodied Navigation". In: *CoRR* 2106.04531 (2021).

[5] Shoufa Chen et al. "AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition". In: *NeurIPS*. 2022.

[6] A. Das et al. "Neural Modular Control for Embodied Question Answering". In: *CORL*. 2018.

[7] Abhishek Das et al. "Embodied Question Answering". In: *CVPR*. 2018.

[8] Yiming Ding et al. "Goal-conditioned Imitation Learning". In: *NeurIPS*. 2019.

[9] Ziad Al-Halah, Santhosh Kumar Ramakrishnan, and Kristen Grauman. "Zero experience required: Plug & play modular transfer learning for semantic visual navigation". In: *CVPR*. 2022.

[10] Kaiming He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *CVPR*. 2022.

[11] Martin Humenberger et al. "Investigating the Role of Image Retrieval for Visual Localization". In: *International Journal of Computer Vision* (2022).

[12] Max Jaderberg et al. "Reinforcement Learning with Unsupervised Auxiliary Tasks". In: *ICLR*. 2017.

[13] Jacob Krantz et al. "Instance-Specific Image Goal Navigation: Training Embodied Agents to Find Object Instances". In: *2211.15876*. 2022.

[14] Jacob Krantz et al. "Navigating to Objects Specified by Images". In: *ICCV*. 2023.

[15] Iker Lluvia, Elena Lazkano, and Ander Ansuategi. "Active Mapping and Robot Exploration: A Survey". In: *Sensors* 21.7 (2021), p. 2445.

[16] Arjun Majumdar et al. "SSL Enables Learning from Sparse Rewards in Image-Goal Navigation". In: *ICML*. 2022.

[17] Arjun Majumdar et al. "Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence?" In: *arXiv:2303.18240*. 2023.

[18] Arjun Majumdar et al. "ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings". In: *NeurIPS*. 2022.

[19] L. Mezghani et al. "Memory-augmented reinforcement learning for image-goal navigation". In: *IROS*. 2022.

[20] Piotr Mirowski et al. "Learning to Navigate in Complex Environments". In: *ICLR*. 2017.

[21] Santhosh Kumar Ramakrishnan et al. "Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2021.

[22] Jerome Revaud et al. "R2D2: Reliable and Repeatable Detector and Descriptor". In: *NeurIPS*. 2019.

[23] Paul-Edouard Sarlin et al. "SuperGlue: Learning Feature Matching with Graph Neural Networks". In: *CVPR*. 2020.

[24] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint* (2017).

[25] Sebastian Thrun, Wolfram Burgard, Dieter Fox, et al. *Probabilistic robotics, vol. 1*. MIT Press Cambridge, 2005.

[26] P. Weinzaepfel et al. "CroCo: Self-Supervised Pretraining for 3D Vision Tasks by Cross-View Completion". In: *NeurIPS*. 2022.

[27] Fei Xia et al. "Gibson env: Real-world perception for embodied agents". In: *CVPR*. 2018.

[28] Karmesh Yadav et al. "Offline Visual Representation Learning for Embodied Navigation". In: *arXiv:2204.13226*. 2022.

[29] Karmesh Yadav et al. "OVRL-V2: A simple state-of-art baseline for ImageNav and ObjectNav". In: *arXiv:2303.07798*. 2023.

[30] Yuke Zhu et al. "Target-driven visual navigation in indoor scenes using deep reinforcement learning". In: *ICRA*. 2017.