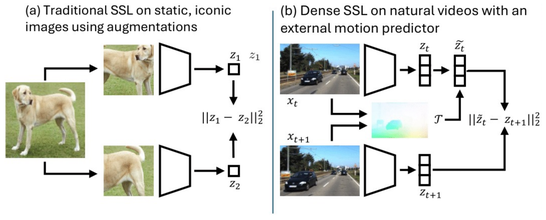
Midway Network: Learning Representations for Recognition and Motion from Latent Dynamics

Chris Hoang, Mengye Ren
https://agenticlearning.ai/midway-network

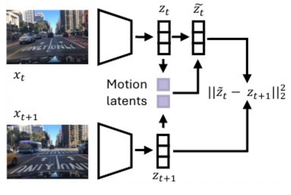
NYU agentic learning ai lab

Components of perception: recognition and motion



Comparison between iconic image SSL (a) and dense video SSL (b) methods

- Image self-supervised learning (SSL), e.g., SimCLR, MoCo, DINO, focuses on learning semantic representations from iconic, human-curated images.
- Video SSL methods (FlowE, DoRA, PooDLe) do not capture object motion while pixel correspondence methods (Uflow, CroCo) result in poor recognition.
- Inspired by works in control (LAPO, DynaMo, LAPA), we propose to leverage latent dynamics to jointly learn representations for recognition *and* motion.

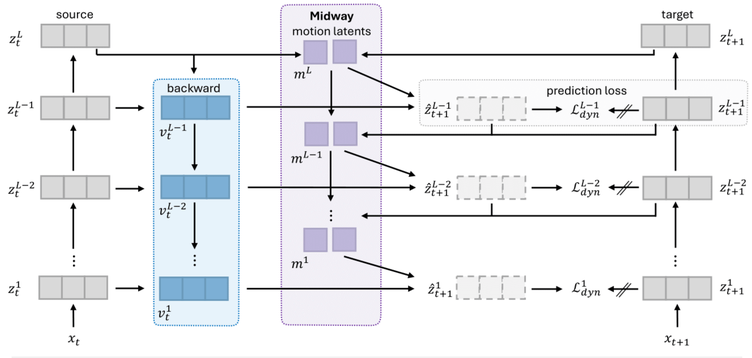


Midway Network infers motions latents between video frames for a latent dynamics SSL objective

Contributions

- Present first SSL architecture to learn rich image-level representations for object recognition and motion understanding solely from natural videos.
- Achieve strong performance on both optical flow and semantic segmentation compared to prior SSL baselines which perform well in only one of the tasks.
- Show Midway Network's ability to capture high-level correspondences over time using novel analysis method based on forwarded feature perturbation.

Hierarchical latent dynamics architecture



- A central *midway* path that infers motion latents, m, to describe transformation between source, z_t, and target, z_{t+1}, features of video frames.
- Forward dynamics model, conditioned on v_t, backward-refined source features, and the motion latents, predicts target features (dense, patch-level).
- Predicted features, $\hat{z}_{t+1}$, from higher feature levels are used instead of source features to refine motion latents at lower feature levels, similar to iterative refinement over feature pyramids in optical flow methods (PWCNet, UFlow).
- The L2 prediction error in latent space, placed at multiple levels, jointly trains all components, and a joint-embedding invariance objective (e.g., DINO) encourages the encoder to learn rich semantic features.

Variant	Latent Dynamics	Backward	Multi-Level	Refinement	Gating	↑mIoU	↓EPE
1 Base model						28.3	6.2
2	✓					30.4	4.4
3	✓	✓				30.0	5.0
4	✓	✓	✓			30.4	5.2
5	✓	✓	✓	✓		31.1	3.9
6 Full model	✓	✓	✓	✓	✓	31.5	4.1
7 No backward	✓		✓	✓	✓	30.4	3.7
8 No multi-level	✓	✓		✓	✓	30.3	5.2
9 No refinement	✓	✓	✓		✓	30.8	5.1

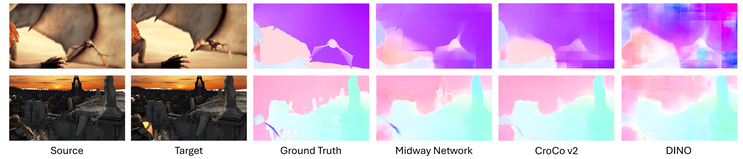
Ablations: latent dynamics loss, backward layers, multi-level loss, iterative refinement, learnable gating on residual connections in forward dynamics transformer. mIoU on BDD100K, EPE on MPI-Sintel

Semantic segmentation and optical flow results

- Midway Network and baselines pretrained on BDD100K and WalkingTours.
- Linear / UperNet readout on sem. seg. and finetuning on optical flow tasks.

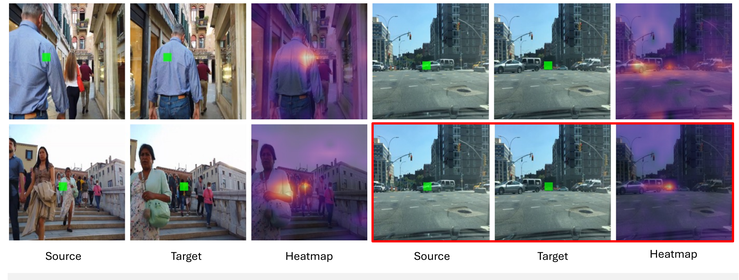
Method	Arch	Ep.	BDD100K Sem. Seg. Linear ↑mIoU	BDD100K Sem. Seg. Linear ↑Acc	BDD100K Sem. Seg. UperNet ↑mIoU	BDD100K Sem. Seg. UperNet ↑Acc	Optical Flow FlyingThings ↓EPE (c)	Optical Flow FlyingThings ↓EPE (f)	Optical Flow MPI-Sintel ↓EPE (c)	Optical Flow MPI-Sintel ↓EPE (f)
PooDLe (Wang et al., 2025)	R50	300	35.1	87.8	47.4	91.0	-	-	-	-
iBOT (Zhou et al., 2021)	ViT-S	800	27.2	85.4	35.5	88.7	18.5	18.0	13.0	13.7
DINO (Caron et al., 2021)	ViT-S	300	36.7	89.3	49.3	92.0	16.8	13.8	11.5	10.8
VideoMAE (Tong et al., 2022)	ViT-S	300	7.8	50.3	10.9	58.6	16.2	16.1	7.2	7.6
CroCo v2 (Weinzaepfel et al., 2023)	ViT-S	300	21.2	80.0	31.9	87.0	9.7	9.4	5.1	5.8
DoRA (Venkataramanan et al., 2024)	ViT-S	300	30.4	87.2	40.8	90.0	16.5	15.1	11.5	11.9
DynaMo† (Cui et al., 2024)	ViT-S	300	36.8	89.4	47.4	91.7	-	-	-	-
Midway (enc. only)	ViT-S	300	-	-	-	-	16.6	13.5	11.7	10.9
Midway	ViT-S	300	**39.7**	**90.3**	**50.4**	**92.4**	**7.3**	**6.8**	**4.1**	**4.9**
DINO (Caron et al., 2021)	ViT-B	300	44.0	90.9	53.8	92.7	17.4	14.8	12.1	14.1
CroCo v2 (Weinzaepfel et al., 2023)	ViT-B	300	16.3	72.4	26.5	84.4	**6.1**	**5.8**	**3.0**	**3.8**
Midway	ViT-B	300	**48.2**	**91.6**	**55.2**	**93.1**	7.0	6.4	4.1	4.8

Experiments pretrained on BDD100K. † DynaMo modified to use ViT-S and DINO objective



MPI-Sintel optical flow finetuning evaluations after WalkingTours pretraining.

Forwarded feature perturbation analysis



- Compute motion latents, perturb source (green), forward predict to propagate perturbation to target, measure perturbation similarity to get heatmap.