Featurising Pixels from Dynamic 3D Scenes with Linear In-Context Learners

– Supplemental Material –

Nikita Araslanov[1 2 3] Martin Sundermeyer[1] Hidenobu Matsuki[1]
David Joseph Tan[1] Federico Tombari[1 2 3]

[1] Google [2] TU Munich [3] Munich Center for Machine Learning

A. Qualitative examples

We provide a video compilation of the LILA representation obtained from the DAVIS dataset. The videos compare LILA to DINOv2 with a ViT-B16 backbone on video object segmentation (VOS) obtained with a linear probe. We observe that LILA reveals spatial detail and tends to emphasise the most prominent objects in the dynamic scenes. At the same time, LILA discriminates well background elements, such as vegetation, pavement and buildings.

B. Backbone generalisation

We train LILA with different backbone families. Concretely, we experiment with a Masked Autoencoder (MAE-B16) [19], a DINOv2 variant equipped with registers (DINO2-B14-Reg) [52] and DINOv3 [37]. Here, we focus on the ViT-B architecture. Tab. 6 reports the results on VOS with linear probing. As expected, LILA provides improved VOS accuracy across the backbone models. For example it improves the MAE-B16 model by $9.4\,\%$ $\mathcal{JF}$.

Table 6. **Generalisation across other backbones.** We report the probing accuracy by pre-training LILA with diverse backbones: Masked Autoencoder, DINOv2 with registers and DINOv3.

Method	Train Data	$\mathcal{JF}$	$\mathcal{J}_m$	$\mathcal{F}_m$
MAE-B16 [30]	LVD*	44.2	41.7	46.7
+ LILA (ours)	+ YT-VOS	**53.6**	**50.4**	**56.8**
DINO2-B14-Reg [30]	LVD*	61.6	59.1	64.2
+ LILA (ours)	+ YT-VOS	**68.4**	**64.7**	**72.1**
DINO3-B16 [30]	LVD*	63.3	60.9	65.8
+ LILA (ours)	+ YT-VOS	**64.8**	**62.0**	**67.6**

C. Practical details

Training LILA is computationally inexpensive, a benefit we achieve through several practical considerations. For example, we downsample the context cue grid, G_{context}, by a factor of 7. This measure significantly improves the efficiency of ridge regression without effecting the downstream performance. We train all models on YouTube-VOS with a batch size of 32 on a single A100 GPU with 40GB of memory. The training typically reaches convergence after 30K iterations, which translates to approximately 24 hours.

PAMR. We apply PAMR [1] in a coarse-to-fine fashion. Specifically, we define an image pyramid with three levels corresponding to the downsampling ratios of 4, 2 and 1 (*i.e.* the original resolution). We define the PAMR parameters identically at each level. The local affinity kernel consists of three 3×3 kernels with dilation ratios of 1, 3 and 5. At the coarsest resolution, the refinement runs for 20 iterations, and we reduce this number by a factor of 2 for each consecutive, finer resolution. This results in a total of 35 refinement iterations. Notably, this coarse-to-fine strategy is approximately four times more efficient than running all 35 iterations at the original image resolution.

D. Zero-shot linear probing

We evaluate LILA under the standard zero-shot semantic segmentation protocol on COCO-Stuff, following the seen/unseen split introduced in prior work [53]. After partitioning the dataset categories into seen and unseen classes, we train the probe using only pixels annotated with seen labels, while evaluating generalization on the 15 held-out unseen classes. During this evaluation, we keep the pre-trained LILA representation frozen and optimise only a lightweight pixel-level probe. Concretely, given frozen encoder and decoder feature maps f^{enc} and f^{dec}, we learn shallow projection heads P_{enc} and P_{dec} that map both branches into a common semantic embedding space. We normalise the projected features with channel-wise ℓ_2-normalisation and classify them with a fixed linear classifier, in which the weights are provided by the class text embeddings (*e.g.*, CLIP text

features) [34]. This yields per-pixel cosine-similarity logits

$$\ell_c^{\text{enc}}(u) = \left\langle \frac{P_{\text{enc}}(f^{\text{enc}}(u))}{\|P_{\text{enc}}(f^{\text{enc}}(u))\|_2}, t_c \right\rangle, \\ \ell_c^{\text{dec}}(u) = \left\langle \frac{P_{\text{dec}}(f^{\text{dec}}(u))}{\|P_{\text{dec}}(f^{\text{dec}}(u))\|_2}, t_c \right\rangle, \tag{7}$$

where t_c denotes the text embedding of class c. The final prediction combines both branches,

$$\ell_c(u) = \exp(\gamma)\big(\alpha_{\text{enc}}\, \ell_c^{\text{enc}}(u) + \alpha_{\text{dec}}\, \ell_c^{\text{dec}}(u)\big), \tag{8}$$

with learnable logit scale γ. We use both encoder and decoder representations because the decoder feature maps are expected to provide dense cues that are complementary to the encoder representation.

The probe is trained with a pixel-wise cross-entropy loss on labeled pixels,

$$\mathcal{L}_{\text{CE}} = \frac{1}{|\Omega_{\text{lab}}|} \sum_{u \in \Omega_{\text{lab}}} \text{CE}(p(u), y(u)), \tag{9}$$

where Ω_{lab} excludes pixels from the unseen categories. In addition, for these ignored pixels Ω_{ign}, we apply a negative regulariser that suppresses probability mass assigned to the seen classes,

$$\mathcal{L}_{\text{neg}} = \frac{1}{|\Omega_{\text{ign}}|} \sum_{u \in \Omega_{\text{ign}}} \left(\sum_{c \in \mathcal{C}_{\text{seen}}} p_c(u) \right)^2. \tag{10}$$

When ignored pixels are present, the training objective is

$$\mathcal{L}_{\text{zero-shot}} = 0.1\, \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{neg}}, \tag{11}$$

and otherwise we optimize only $\mathcal{L}_{\text{CE}}$. This regulariser discourages the probe from collapsing unlabelled regions into seen categories and leaves room for transfer to unseen classes. At test time, we apply the same frozen text-based classifier without further adaptation. As a result, the segmentation accuracy on the held-out classes reflects the zero-shot transfer ability of the frozen LILA features and, in particular, the complementarity of the learned decoder representation (*i.e*. LILA) with that of the encoder.

References

[52] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 1

[53] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *CVPR*, 2019. 1