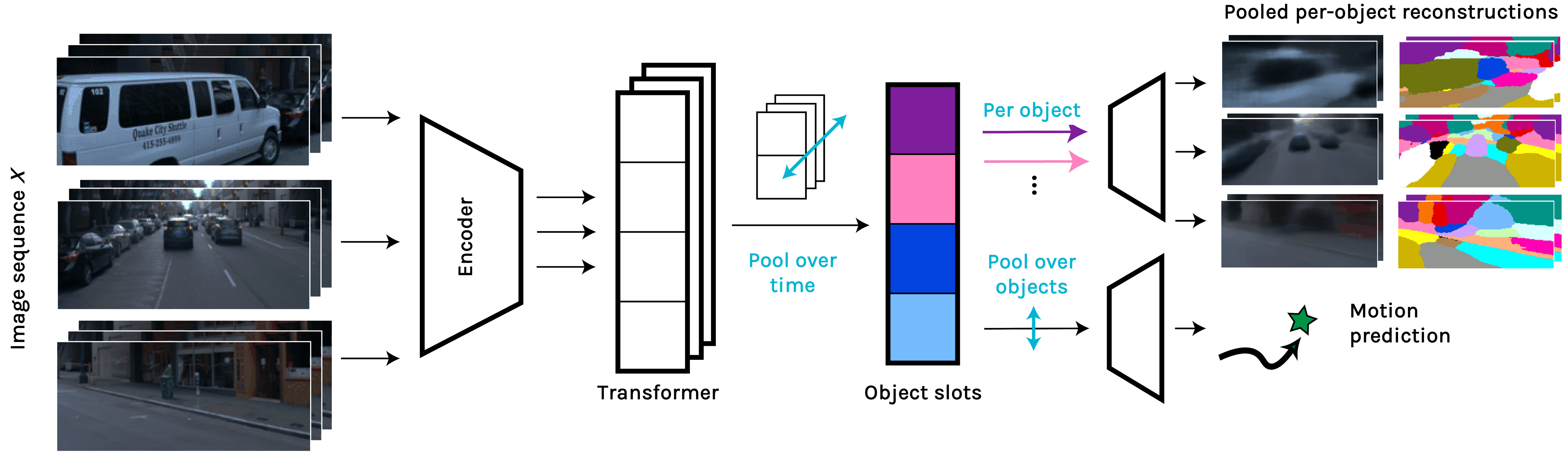
Linking vision and motion for self-supervised object-centric perception

Kaylene Stocking[1,2], Zak Murez[1], Vijay Badrinarayanan[1], Jamie Shotton[1], Alex Kendall[1], Claire Tomlin[2] Christopher P. Burgess[1]
1 Wayve 2 UC Berkeley; Contact: kaylene@berkeley.edu



Object-centric perception for autonomous driving

- Object-centric perception:
  1. Can help with generalization and reasoning about the interactions between multiple independent objects [1]
  2. May be easier to interpret & debug due to similarities with human vision
- Supervised learning can pick out "known" objects but relies on expensive labels & hand-engineered definitions of objects
- Switching to a self-supervised objective unlocks end-to-end learning of representations that are both object-centric and adapted for downstream tasks like autonomous driving

Towards real-world driving images

- Self-supervised methods like SIMONe[2] and SAVi[3] leverage object motion in contiguous frames to segregate the input pixels into object-centric slots
- These methods are typically evaluated on synthetic data, and real images remain challenging without additional labels or privileged information
- In autonomous driving, the vehicle's motion is usually known
- We experiment with leveraging known camera motion in two ways:
  1. As an auxiliary input for image token embeddings and reconstruction queries
  2. As the basis for an auxiliary task of predicting future camera motion, which is equivalent to behavioral cloning for driving actions in the training data
- Our architecture is builds on SIMONe [2]

Training objective

There are three components to the training loss:

$$\mathcal{L}_{total}(X,s) = \mathcal{L}_{recon}(X) + \beta\mathcal{L}_{KL}(X) + \omega_{task}\mathcal{L}_{task}(s)$$

The network outputs a set of K latent vectors, where each latent $q_k(X)$ is a Gaussian distribution containing information about slot (object) k. These latents are used to make independent predictions of the reconstruction distribution for each pixel $x^{(n)}$, a mixture of Gaussians with H modes:

$$p(x^{(n)}|o_k) = \frac{1}{H}\sum_h \hat{\alpha}_k^{(h)}\mathcal{N}(\hat{\mu}_k^{(h)}, \sigma_x)$$

$$o_k \sim q_k(X) \qquad \hat{\alpha}_k, \hat{\mu}_k = f(o_k)$$

The overall reconstruction prediction is a weighted sum over the per-slot predictions:

$$p(x^{(n)}|o_1,...,o_K) = \frac{1}{K}\sum_k \bar{\alpha}_k[p(x^{(n)}|o_k)]$$

$$\mathcal{L}_{recon}(X) = \frac{-1}{N}\sum_n \log p(x^{(n)}|o_1,...,o_K)$$

Adopting a variational autoencoding framework, disentanglement of the slots and latent dimensions is encouraged with a KL-penalty using a unit spherical normal prior:

$$\mathcal{L}_{KL}(X) = \sum_k D_{KL}(q_k(X)||p(\cdot))$$

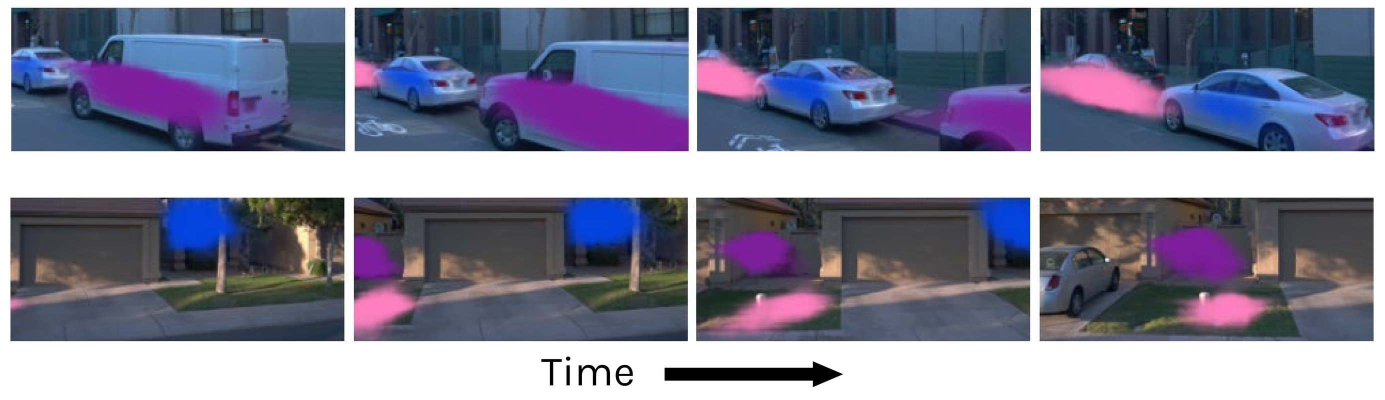
Finally, we use an auxiliary task loss of future motion prediction (behavioral cloning) leveraging the future camera pose:

$$\mathcal{L}_{task}(s) = \sum_{t'} ||s_{t'} - \hat{s}_{t'}||_1$$

Results

Emergent fusion over time & space

Object tracking



Segmentation metrics

- Foreground adjusted Rand index (ARI-F)
- Center-of-mass distance (CoM) with Hungarian mask matching [4]

Method	Privileged information	ARI-F ↑	CoM ↓
SAVi (RGB) [4]	None	-	21.5 ± 1.8
SAVi++ [4]	Bounding boxes, depth	-	4.4 ± 0.2
SAVi++ (unconditioned) [4]	Depth	-	6.9 ± 0.5
SIMONe [4]	Depth	-	7.4 ± 0.2
Ours (no mixture, H=1)	Camera motion	.193 ± .004	10.0 ± 0.3
Ours (no camera motion)	None	.237 ± .003	9.8 ± 0.3
Ours (no motion pred.)	Camera motion	.257 ± .018	9.9 ± 0.7
Ours	Camera motion	.253 ± .009	9.6 ± 0.4

References

[1] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. arXiv:2012.05208 (2020).

[2] Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess. SIMONe: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. NeurIPS (2021).

[3] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. ICLR (2021).

[4] Gamaleldin Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. NeurIPS (2022).