Tripod

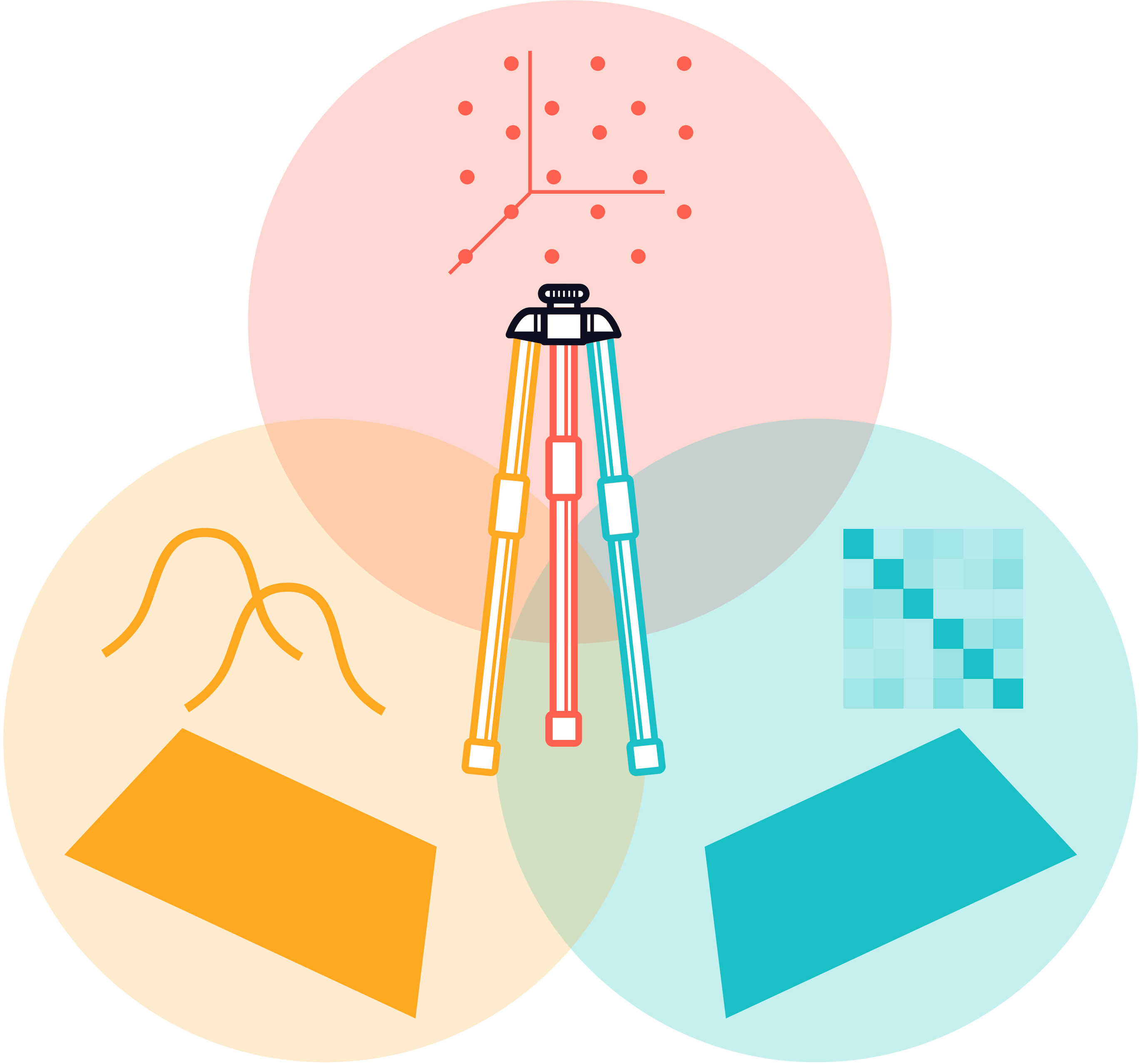
three complementary inductive biases for disentangled representation learning

Kyle Hsu*, Jubayer Ibn Hamid*, Kaylee Burns, Chelsea Finn, Jiajun Wu

Stanford

ICML International Conference On Machine Learning

Inductive biases faciliate disentanglement by paring down the solution space, but prior works largely propose and validate one new approach in isolation from others.



Tripod makes necessary adaptations to three **complementary** inductive biases to meld them into a state-of-the-art disentangling autoencoder.

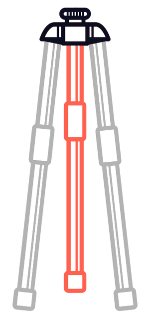
preliminaries

- unlabelled data is generated noiselessly from independent sources

$$p(s) = \prod_{i=1}^{n_s} p(s_i) \qquad x = g(s)$$

- goal: learn autoencoder whose latents recover sources

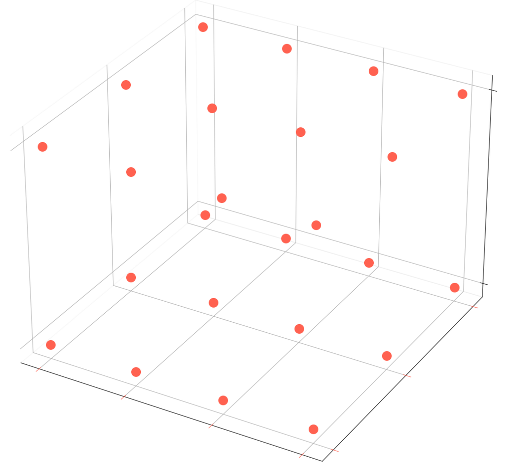
$$z = \hat{g}^{-1}(x) \qquad \hat{x} = \hat{g}(z)$$

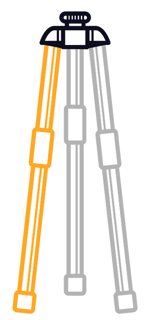


finite scalar latent quantization

- motivation: true sources are highly compressed and organized
- goal: impose quantized, grid-like structure on latent space
- problem: dictionary learning destabilizes other components
- **solution**: use finite scalar quantization (Mentzer et al., 2023)

$$z = \frac{2}{n_q - 1} \operatorname{round}\left(\frac{n_q - 1}{2}\left(\tanh(\hat{g}^{-1}(x)) + 1\right)\right) - 1$$





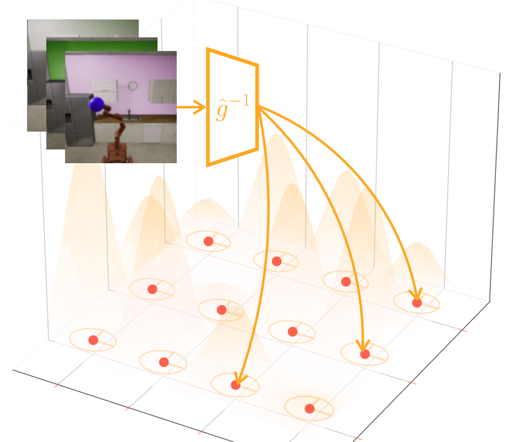
kernel-based latent multiinformation

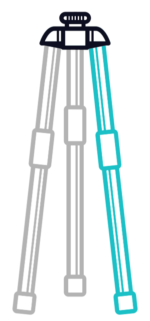
- motivation: true sources are collectively independent
- goal: regularize latent multiinformation

$$D_{\mathrm{KL}}\left(q(z) \,||\, \prod_{j=1}^{n_z} q(z_j)\right)$$

- problem: quantized latents aren't probabilistic
- **solution**: use Gaussian kernel density estimation

$$q(z) \propto \sum_{i=1}^{n_b} \exp\left(-\frac{1}{2}(z - z^{(i)})^\top S^{-1} (z - z^{(i)})\right)$$





normalized Hessian penalty

- motivation: true sources interact minimally to generate data
- goal: regularize off-diagonal entries of decoder Hessians

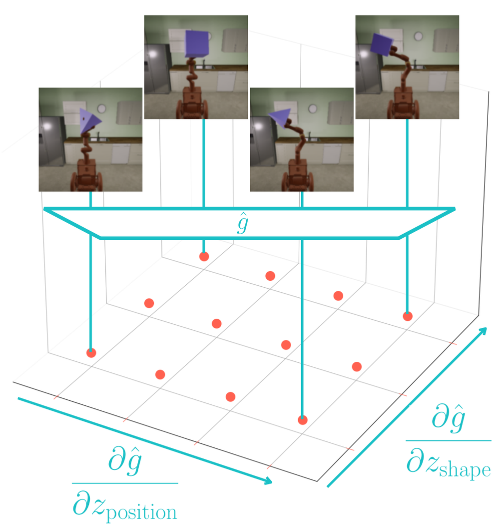
$$\sum_{j_1 \neq j_2} \left(H^{[k]}_{j_1 j_2}\right)^2 \qquad H^{[k]}_{j_1 j_2} = \frac{\partial \hat{g}^{[k]}}{\partial z_{j_1} z_{j_2}}$$

- problem: sensitive to trivial rescalings of latents and activations
- **solution**: replace with a normalized quantity

$$\frac{\sum_{j_1 \neq j_2} \left(H^{[k]}_{j_1 j_2} \sigma_{j_1} \sigma_{j_2}\right)^2}{\sum_{j_1, j_2} \left(H^{[k]}_{j_1 j_2} \sigma_{j_1} \sigma_{j_2}\right)^2}$$

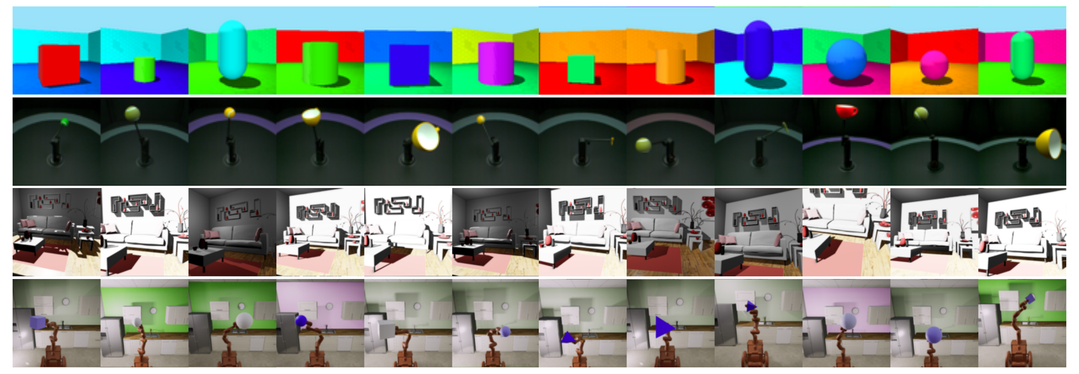
and associated estimator

$$\frac{\operatorname{Var}\left[v^\top H^{[k]} v\right]}{\operatorname{Var}\left[w^\top H^{[k]} w\right]} \qquad \begin{array}{l} v_j \sim \operatorname{Rademacher}(\sigma_j) \\ w_j \sim \mathcal{N}(0, \sigma_j^2) \end{array}$$



experiments

- datasets



Shapes3D	6 sources
MPI3D	7 sources
Falcor3D	7 sources
Isaac3D	9 sources

- **Tripod greatly improves upon prior methods** that use only one of its legs

InfoMEC: (InfoModularity, InfoCompactness, InfoExplicitness)

	aggregated	Shapes3D	MPI3D	Falcor3D	Isaac3D
β-TCVAE	(0.68 0.43 0.88)	(0.86 0.45 **1.00**)	(0.52 0.46 0.75)	(0.62 0.39 **0.82**)	(0.72 0.42 **0.94**)
QLAE	(0.62 **0.57** 0.77)	(0.68 0.55 **0.98**)	(0.45 0.42 0.61)	(**0.71** **0.71** 0.72)	(0.65 0.61 0.78)
Tripod (ours)	(**0.78** **0.59** **0.90**)	(**0.94** **0.59** **1.00**)	(**0.64** **0.53** **0.84**)	(**0.72** 0.56 **0.82**)	(**0.84** **0.68** **0.95**)

- ablating each Tripod leg in turn shows that **all three legs are necessary for best performance**

	aggregated	Shapes3D	MPI3D	Falcor3D	Isaac3D
Tripod (ours)	(**0.78** **0.59** **0.90**)	(**0.94** **0.59** **1.00**)	(**0.64** **0.53** 0.84)	(**0.72** **0.56** **0.82**)	(**0.84** **0.68** **0.95**)
Tripod w/o FSLQ	(0.56 0.46 **0.92**)	(0.69 0.48 **1.00**)	(0.43 0.40 **0.97**)	(0.54 0.41 **0.84**)	(0.57 0.54 0.87)
Tripod w/o KLM	(0.73 0.50 **0.90**)	(0.89 **0.57** **1.00**)	(0.57 0.50 0.80)	(**0.74** 0.54 **0.82**)	(0.72 0.38 **0.96**)
Tripod w/o NHP	(0.70 0.48 0.89)	(0.85 0.46 **1.00**)	(0.60 0.50 0.81)	(0.59 0.40 0.81)	(0.75 0.57 0.93)

- case study: naively combining Tripod's inductive biases fails to disentangle "robot x" and "robot y"

Tripod (ours) | Tripod (naive)