Flow Jepa: Experiments in Bijective Isometry

Decoupling Semantics from Information Destruction

Adam Hibble

Independent Researcher
No Institutional Affiliation

@algomancer

ABSTRACT
This report presents a preliminary exploration into **Flow-JEPA**, a lossless adaptation of Joint Embedding Predictive Architectures. Unlike traditional methods that may achieve semantic invariance by discarding information, this framework enforces a bijective mapping between observation and latent space, preserving exact data density. By integrating a masked normalizing flow with a joint predictive objective, the model learns to sequester aleatoric noise into a specific submanifold rather than destroying it. This yields a representation that is simultaneously semantically structured for perception and topologically complete for precise, differentiable simulation, offering properties that appear mechanistically and computationally desirable for model-predictive control based reasoning.

1 Context

Recent work by Balestriero et al. [1] highlights a fundamental tension: "Learning by Reconstruction Produces Uninformative Features for Perception." This is intuitively sound. In high-dimensional data, "noise" (texture, exact pixel values) often dominates variance, while "signal" (semantics) is low-variance. Naive reconstruction objectives force models to prioritize the former.

However, for a control system, this noise is not always irrelevant. To ensure a generated plan is feasible, the system must be able to constrain the state being optimized to the manifold of valid, realizable configurations. If the representation discards too much information, the optimization process loses the ability to distinguish between a feasible state and a hallucination that satisfies high-level semantics but violates low-level constraints.

This creates a gap. On one hand, we need high-level semantics for planning (to ignore the noise). On the other, we need low-level fidelity for simulation (to respect the noise). The motivating question for this work is whether we can bridge this gap: Is it possible to construct a representation that is lossless yet semantically organized?

2 Hypothesis

To explore this, I am focusing on **Normalizing Flows**. Flows offer a unique mechanical property: they are diffeomorphisms. They map the input space $\mathcal{X}$ to the latent space $\mathcal{Z}$ without destroying topological information or discarding mass.

2.1 Motivation: Planning as Optimization

A primary motivation for selecting a bijective, differentiable map is the operational flexibility it offers for planning.

If a representation space of a world model is mapped via a differentiable and invertible function, we can reframe planning as an optimization problem. Instead of searching through a discrete tree of actions, we can define a high-level goal in the latent space and backpropagate the error to the input space:

$$x^* = x - \alpha \nabla_x \mathcal{L}(f(x), z_{goal}) \tag{1}$$

This setup allows us to find optimal actions by optimizing directly against the model's density function and bijective map, ensuring that the optimized trajectory remains grounded in the model's learned manifold.

2.2 The Challenge: Semantics

However, standard flows typically prioritize local correlation over global structure. A flow might perfectly reconstruct an image, but the "concept" of the object might be distributed across high-frequency dimensions rather than concentrated in a compact manifold.

3 Methodology

The core engine of this experiment is a flow layer I refer to as **Stochastic Masking Coupling**.

3.1 Variational Augmentation

Before the data enters the flow, I introduce a **Variational Augmentation** layer.

Continuous bijections are topologically restricted; they cannot map a disconnected manifold (like a mixture of Gaussians) to a connected one (like a standard Normal) without severing the topology. To relax this constraint, Huang et al. [2] proposed Augmented Normalizing Flows.

In my implementation, I augment the input $x \in \mathbb{R}^{D_{in}}$ with a latent vector $u \in \mathbb{R}^{D_{aug}}$. This effectively embeds the data onto a probabilistic manifold in a higher-dimensional space. The primary role of this augmentation is **Topological Shaping**: it gives the flow the "slack" dimensionality required to untangle complex, crossed topologies (like knots) that would be computationally intractable to resolve in the original pixel space, allowing the flow to map disjoint manifolds to a connected support.

```
// FIELD NOTE: The Kernel Intuition
```

Beyond topology, there is a simple geometric benefit. By mapping $x \to [x, u]$, we increase the ambient dimensionality of the manifold. Just as in SVMs or kernel methods, projecting complex data into a higher-dimensional space often renders it linearly separable. This extra "headroom" gives the flow the geometric freedom to untangle class boundaries that would be jammed tight in the original pixel space, aiding the subsequent linear separation in the VAE.

3.2 Stochastic Masking Coupling

Motivated by the observation that learning from partial information encourages the emergence of semantic features [4], I adopt a randomized masking approach.

Formally, an affine coupling layer splits the input dimension D into two parts. In architectures like RealNVP [3], this split is deterministic. While mathematically sufficient, these fixed patterns reduce the density estimation problem to a static task.

In this architecture, let $x \in \mathbb{R}^D$ be the input tensor at layer l. We sample a random binary mask $m_l \sim \text{Bernoulli}(0.5)^D$. The transformation $f_l : x \to y$ is defined as an affine coupling where the scale s and translation t parameters are functions of the masked input. This formulation ensures the Jacobian remains triangular regardless of the mask topology.

3.3 The Linear VAE Prior: Resolving Tension

A fundamental tension exists in flow-based modeling: as a flow approaches a perfect map to an isotropic Gaussian $\mathcal{N}(0, I)$, the latent space $\mathcal{Z}$ becomes increasingly entropic. In a perfectly optimized flow, all clusters merge, and the marginal distribution of any dimension becomes independent of class labels. This tension creates a significant hurdle for discrimination tasks.

To resolve this, I utilize a **Conditional VAE Prior** with a strictly **Linear Encoder**.

Instead of mapping x directly to $\mathcal{N}(0, I)$, the flow maps $x \to z$, and the VAE models the density $p(z)$. By restricting the VAE encoder to a linear map, we enforce a strong constraint: the flow must untangle the data manifold such that the semantic structure is **linearly accessible**.

- If the VAE were non-linear, the flow could remain "lazy," leaving the data entangled for the VAE to solve.
- With a linear encoder, the flow is forced to perform the non-linear "straightening" of the manifold.

This setup allows the latent space z to retain discriminative structure (facilitating linear separation) while still permitting a valid density estimation via the VAE's generative path.

3.4 Complementary Masking Strategy

If I were to sample masks independently at random for every layer, there would be a non-zero probability that some dimensions remain masked (and thus untransformed) for many consecutive layers.

To prevent these "stale" pixels, I enforce a **Complementary Masking Strategy**. For every pair of layers $(l, l+k)$, I generate a base mask m_{base} for layer l, and strictly enforce that the mask for a subsequent layer is its inverse, $m_{l+k} = \neg m_{base}$.

This simple heuristic guarantees that every single dimension in the input vector is updated exactly once every two steps. Furthermore, because these masks are generated stochastically at each evaluation, I avoid **depth-wise spatial bias**. No specific spatial location is tied to a specific layer index in the processing hierarchy. Consequently, the processing of any given feature is not localized to a specific depth but rather distributed dynamically throughout the stack. Unlike fixed checkerboard patterns, which risk creating "blind spots," this approach ensures that over the course of training, every pixel is conditioned on every other pixel at every depth, preventing the model from overfitting to static connectivity graphs.

3.5 The Projection Head & SIGReg

One of the central tensions in this work is the desire for the backbone to be lossless (bijective) while the downstream tasks often require invariance (lossy). As noted by Gupta et al. [5], the projection head in contrastive learning acts as a filter: it absorbs the specific invariances required by the loss function, allowing the backbone to maintain a higher-rank, more generalizable representation.

In this architecture, I apply the invariance objectives *only* to the output of a projection head, not the

flow's latent variable directly. This allows the flow to remain a valid density model (preserving all information) while the projector learns to extract the invariant semantic manifold.

We define the invariance task over multiple views (augmentations) of the same input. In the projected space z_{inv}, I enforce a variance-minimization objective. We compute an anchor embedding as the weighted mean of the views and minimize the squared distance of each view from this anchor. This effectively pulls semantically related views together into tight clusters.

To structure this projected space and prevent the variance minimization from collapsing all points to a singularity, I utilize **SIGReg** (Sketched Isotropic Gaussian Regularization), as introduced in LeJEPA [6]. Rather than using heuristics like stop-gradients or negative pairs to prevent collapse, SIGReg enforces that the projected embeddings follow an isotropic Gaussian distribution. It achieves this via random 1D projections and characteristic function matching (specifically the Epps-Pulley test).

4 Observations

These experiments were conducted on a TinyBox, a dedicated deep learning workstation equipped with 6x NVIDIA 4090 GPUs (total 144GB VRAM). All results are preliminary and based on training runs using the Imagenette dataset.

Preliminary online linear probing on Imagenette yields a top-1 accuracy of $\approx 88.6\%$. While not state-of-the-art, this places the method in the ballpark of established SSL baselines, which is encouraging given the resource-constrained environment and lack of extensive hyperparameter tuning. We have not yet performed ablations to isolate the effects of the VAE prior or capacity allocation, so these results should be interpreted as a signal of viability rather than a performance ceiling.

5 Discussion

This is a preliminary exploration into the viability of lossless representation learning for perception and control. The goal is not to beat large-scale benchmarks immediately, but to validate whether we can decouple the learning of semantics from the destruction of information.

The primary constraint on this research is computational scale. Normalizing Flows are notoriously compute-intensive, and training high-resolution, deep bijections requires significant memory and throughput. The results presented here, achieved on a single TinyBox node, suggest that the architecture is sound. The immediate next phase involves constructing the transition dynamics model (the World Model) within this frozen latent space to facilitate planning. To fully investigate these mechanics requires scaling to larger datasets and deeper flow stacks. I am seeking compute sponsorship to continue this line of inquiry.

References

[1] R. Balestriero, Y. LeCun. *Learning by Reconstruction Produces Uninformative Features For Perception*. arXiv:2402.11337, 2024.

[2] C. Huang, L. Dinh, A. Courville. *Augmented Normalizing Flows: Bridging the Gap Between Generative Flows and Latent Variable Models*. arXiv:2002.07101, 2020.

[3] L. Dinh, J. Sohl-Dickstein, S. Bengio. *Density estimation using Real NVP*. arXiv:1605.08803, 2016.

[4] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick. *Masked Autoencoders Are Scalable Vision Learners*. CVPR, 2022.

[5] K. Gupta, T. Ajanthan, A. van den Hengel, S. Gould. *Understanding and Improving the Role of Projection Head in Self-Supervised Learning*. arXiv:2212.11491, 2022.

[6] R. Balestriero, Y. LeCun. *LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics*. arXiv:2511.08544, 2025.

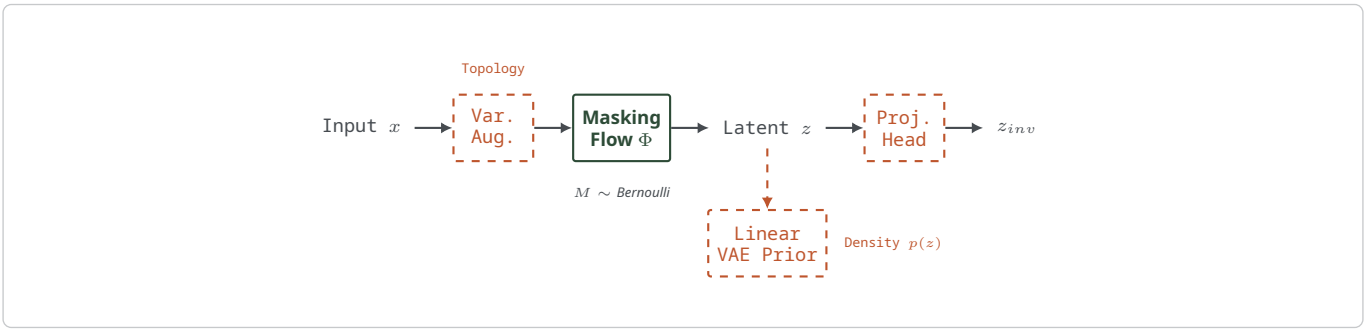

Figure 1: **Proposed Flow.** The input x is first lifted via Variational Augmentation (to untangle topology), then transformed losslessly into z. A Linear VAE Prior models $p(z)$, allowing z to retain structure instead of collapsing to noise. The projection head extracts invariant features z_{inv}.