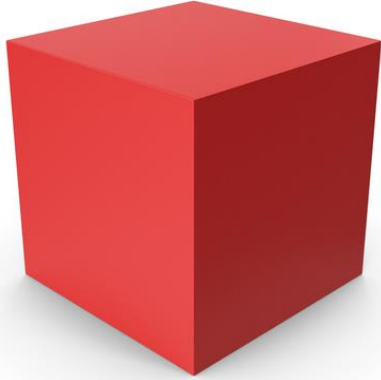


Sequential Representation Learning via Static-Dynamic Conditional Disentanglement

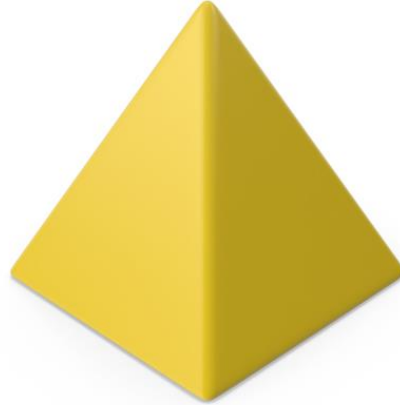
Mathieu Cyrille Simon, Pascal Frossard, and Christophe De Vleeschouwer



Motivation



Red cube



Yellow pyramid

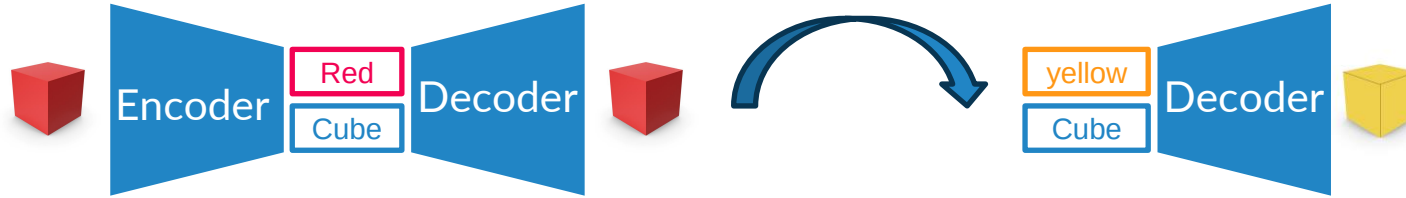
Could I have a
yellow cube?

For a human it is obvious that we have two different separate concepts!



The success of machine learning algorithms generally depends on data representation.

Would like learned representations to have the same flexibility



However, this disentanglement does not happen in practice!

Disentanglement representation learning: How to learn a low-dimensional representation that factorizes the underlying factors of variations

Asking for
disentanglement leads to
very deep questions



- Hair color/length
- Eyes
- Skin
- Nose shape
- Facial expression
- Face orientation

Define

Problem

Method

Unsupervised disentanglement:
presupposes intrinsic definition
of the factors

How to be sure
problem well-posed

How to be sure will
extract the factors

Identifiability

Here we focus on DRL for **sequential data**:

- **Static** factors : **time independent** information, what is common
- **Dynamic** factors : **time varying** information, what is specific

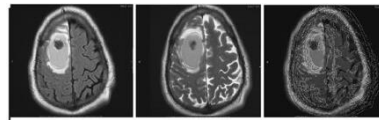
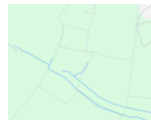


Dynamic (motion- facial expression)



Static (appearance)

As soon as we have multiple views, we might want to separate what is shared between the views and what is specific



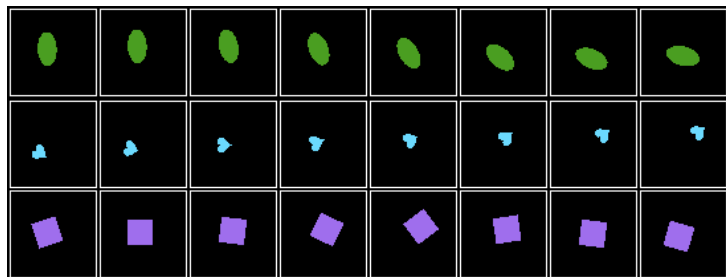
Several works related to this topic...

- Zhu, Y., Min, M.R., Kadav, A., Graf, H.P.: *S3vae: Self-supervised sequential vae for representation disentanglement and data generation.*
- Tonekaboni, S., Li, C.L., Arik, S.O., Goldenberg, A., Pfister, T.: *Decoupling local and global representations of time series.*
- Naiman, I., Berman, N., Azencot, O.: *Sample and predict your latent: Modalityfree sequential disentanglement via contrastive estimation.*
- Luo, Y.J., Ewert, S., Dixon, S.: *Towards robust unsupervised disentanglement of sequential data—a case study using music audio.*
- Li, Y., Mandt, S.: *Disentangled sequential autoencoder.*
- Hsu, W.N., Zhang, Y., Glass, J.: *Unsupervised learning of disentangled and interpretable representations from sequential data.*
- Han, J., Min, M.R., Han, L., Li, L.E., Zhang, X.: *Disentangled recurrent wasserstein autoencoder.*
- Haga, T., Kera, H., Kawamoto, K.: *Sequential variational autoencoder with adversarial classifier for video disentanglement.*
- Denton, E.L., et al.: *Unsupervised learning of disentangled representations from video.*
- Berman, N., Naiman, I., Azencot, O.: *Multifactor sequential disentanglement via structured koopman autoencoders.*
- Bai, J., Wang, W., Gomes, C.P.: *Contrastively disentangled sequential variational autoencoder.*
- Albarracin, J.F.H., Rivera, A.R.: *Video reenactment as inductive bias for contentmotion disentanglement.*
- And many more..

... But we had some problems with these current approaches

- Lack of a proper, non ambiguous, definition of the factors
- Require a lot of supervision
- Assume factors to be mutually independent

Moving shapes



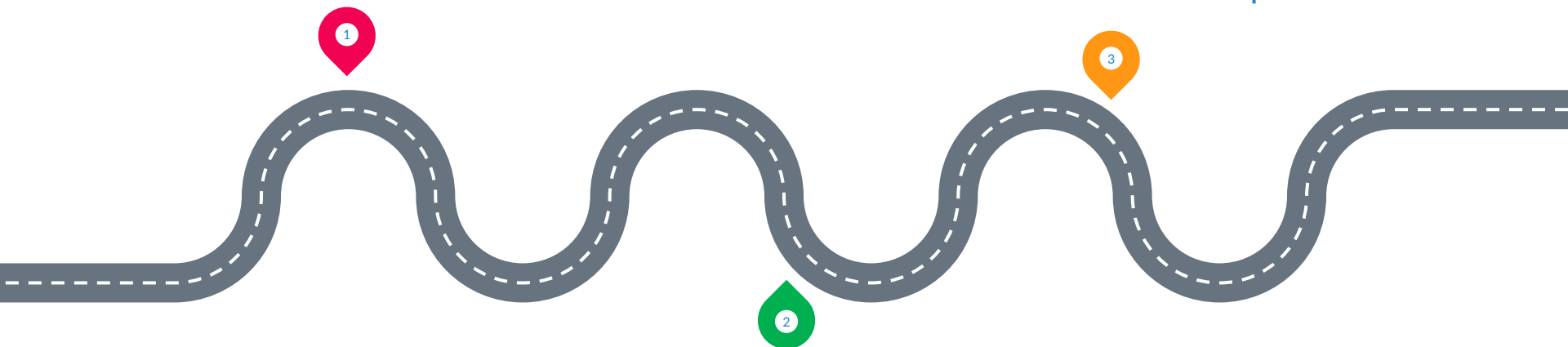
Sample
dynamic



fails to provide a disentangled representation

Formalize the problem and the
definition of the factors

Build a model that directly fulfils
those conditions without
additional losses or supervision



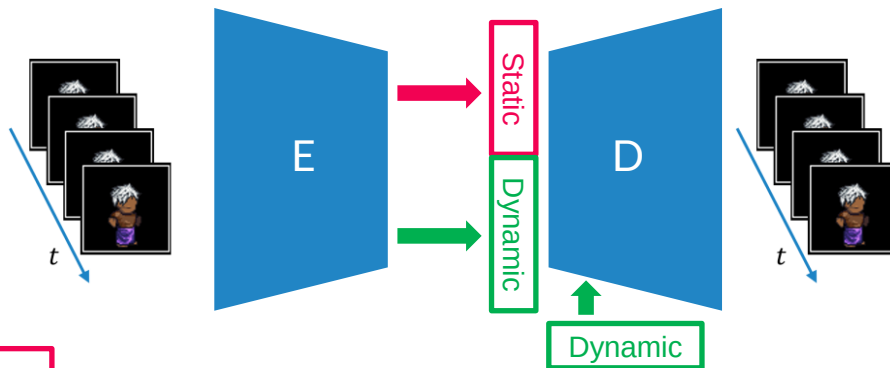
Derive sufficient conditions
for the ground truth factors
to be identifiable

- Static factors s shared across all time steps
- Dynamic factors $d_{1:T}$ varying with time, possibly not in every sequence

Objective :

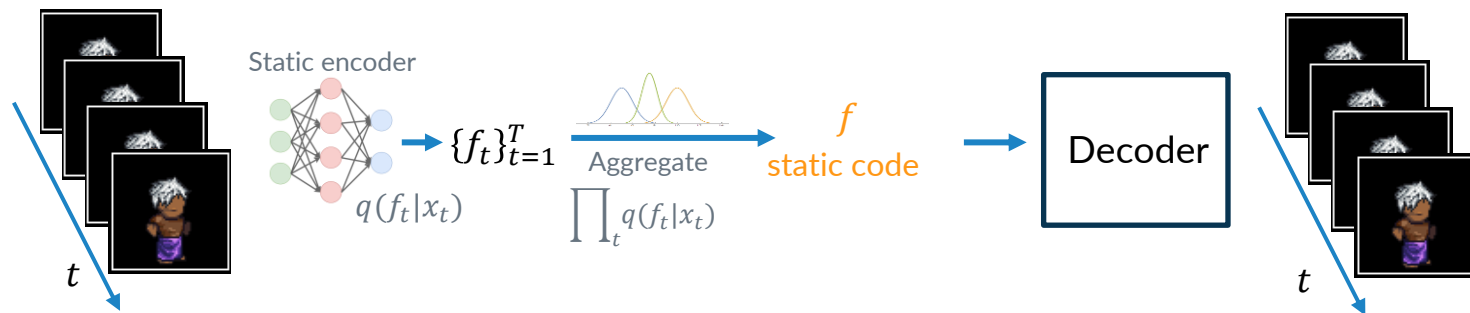
- Learn a latent vector f that captures exactly the static factors s
- Learn latent vectors $\lambda_{1:T}$ that capture the $d_{1:T}|s$

Classical approach :



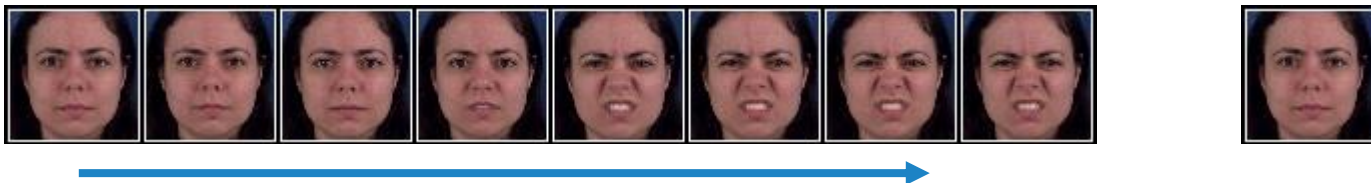
Change paradigm !

Build an autoencoder with a single fixed latent code for the static part



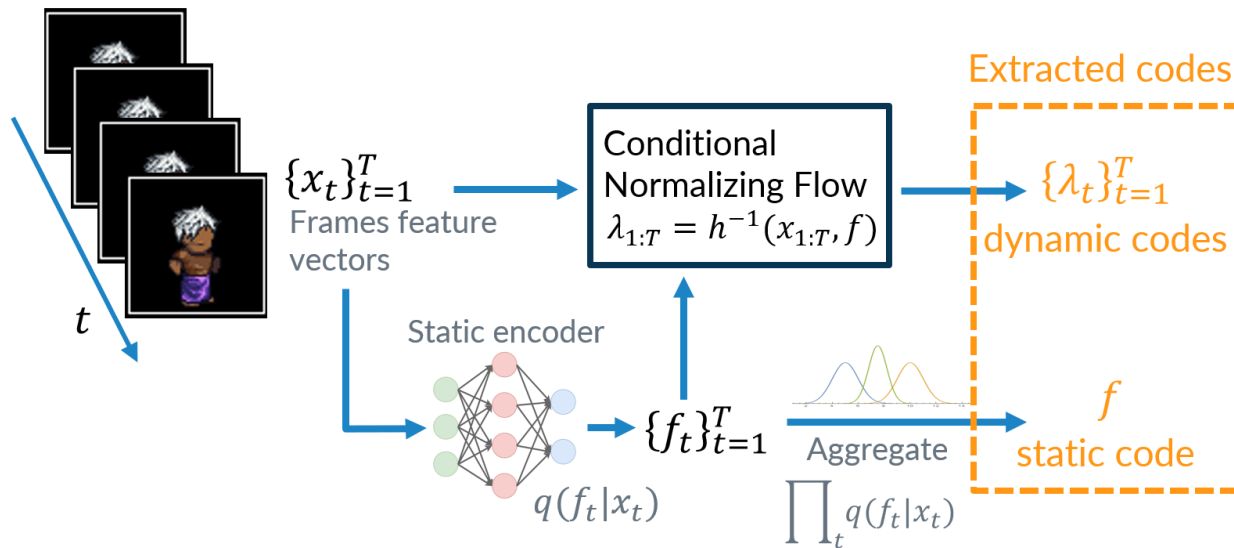
If f disentangled the dynamic $\approx$ residual

Intuitively:



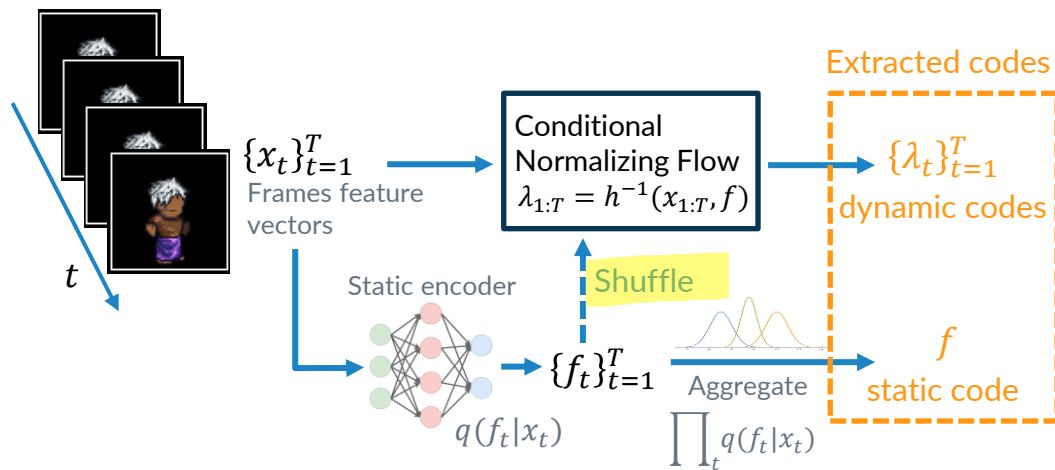
a conditional **Normalizing Flow** to model the frames as decoder !

$$\text{ELBO: } \max_{p,q} \mathbb{E}_{x_{1:T}} \mathbb{E}_{q(f|x_{1:T})} \log p(\lambda_{1:T}) + \log \left| \det \frac{\delta h^{-1}(x_{1:T}, f)}{\delta x_{1:T}} \right| - KL(q(f|x_{1:T}) || p(f))$$



If f disentangled

$$\text{ELBO: } \max_{p,q} \mathbb{E}_{x_{1:T}} \mathbb{E}_{q(f|x_{1:T})} \log p(\lambda_{1:T}) + \log \left| \det \frac{\delta h^{-1}(x_{1:T}, f_{\pi(t)})}{\delta x_{1:T}} \right| - KL(q(f|x_{1:T}) || p(f))$$



Only need to shuffle !

The ground truth factors are identified by the learned codes up to an element-wise invertible transformation

Definition 1. The ground truth factors $\mathbf{s}, \epsilon_{d,1:T}$ are said to be identified by learned vectors $\mathbf{f}, \lambda_{1:T}$ if there exists invertible transformations $\mathbf{m}$ and $\mathbf{a}$ s.t. $\mathbf{f} = \mathbf{m}(\mathbf{s}), \lambda_{1:T} = \mathbf{a}(\epsilon_{d,1:T})$ i.e., $\mathbf{f}, \lambda_{1:T}$ are reparametrizations of $\mathbf{s}, \epsilon_{d,1:T}$ respectively.

Proposition 1. Consider the generative process and conditions presented in Sec.3. Further assume that $\dim(\mathbf{f}) = n_f \geq n_s$. Let $\mathbf{x}$ and $\mathbf{x}^*$ denote two frames (or disjoint sets of frames) belonging to the same sequence $\mathbf{x}_{1:T}$ and $\mathbb{D}$ be a divergence between two distributions. Given unlimited data from $p(\mathbf{x}_{1:T})$, any candidate posterior distribution $q(\mathbf{f}|\mathbf{x}_{1:T})$ that minimizes

$$L_{prop1} = \mathbb{E}_{p(\mathbf{x}_{1:T})} \mathbb{D}(q(\mathbf{f}|\mathbf{x}), q(\mathbf{f}|\mathbf{x}^*)) - I(\mathbf{f}, \mathbf{s}) + I(\mathbf{s}, \mathbf{s}), \quad (8)$$

is disentangled in the sense that $q(\mathbf{f}) = \int q(\mathbf{f}|\mathbf{x}_{1:T})p(\mathbf{x}_{1:T})d\mathbf{x}_{1:T}$, the aggregated posterior, is a reparametrization of $p(\mathbf{s})$.

Proposition 2. Consider the same framework as in Prop.1 and the static code $\mathbf{f}$ to be disentangled as described above. Further assume that $\dim(\lambda) = n_\lambda \geq n_d$. Any smooth and invertible candidate function $\mathbf{h}$ s.t. $\mathbf{x}_{1:T} = \mathbf{h}(\mathbf{f}, \lambda_{1:T})$ which minimizes

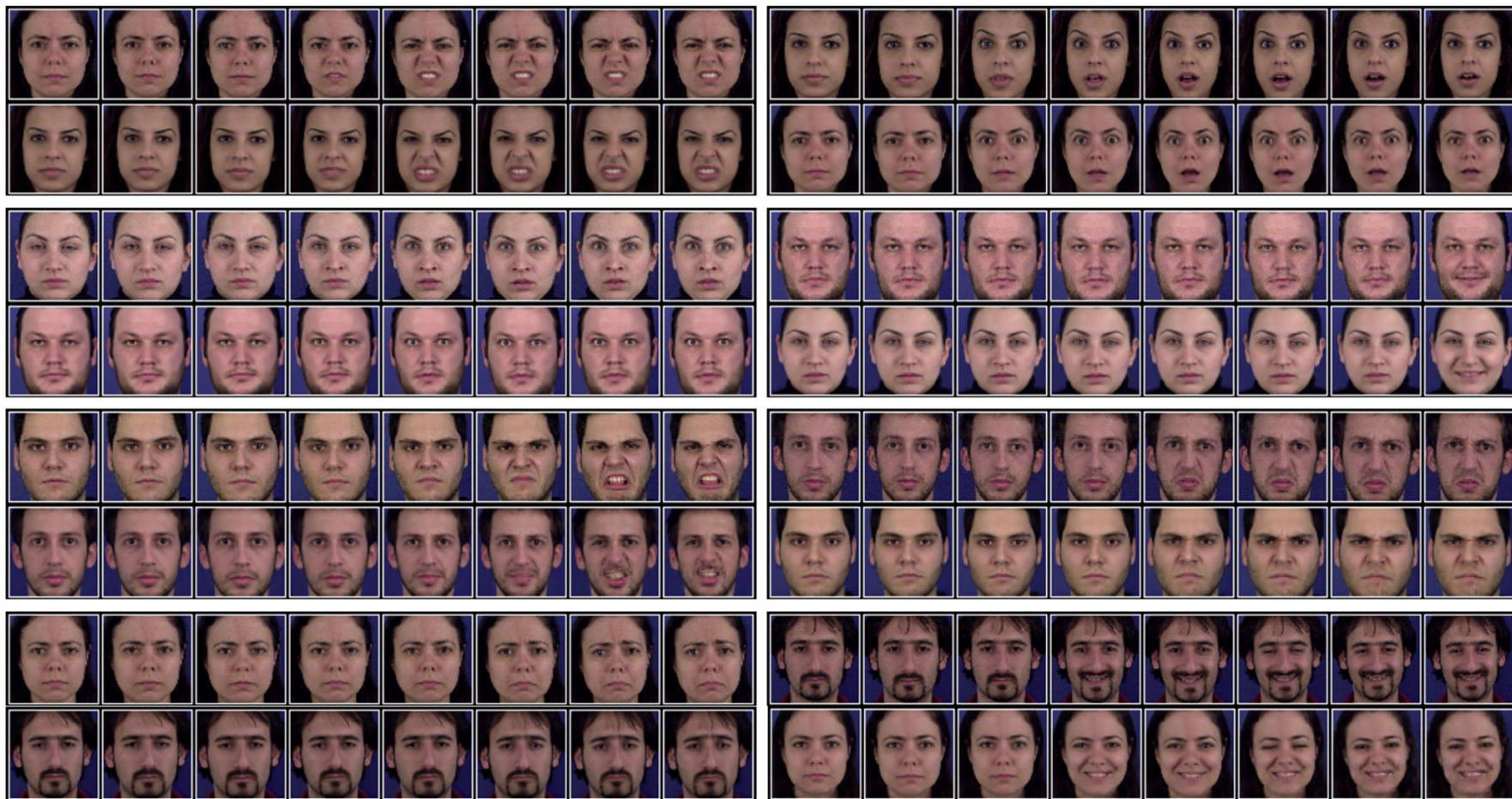
$$L_{prop2} = I(\mathbf{f}, \lambda_{1:T}), \quad (9)$$

is disentangled in the sense that $q(\lambda_{1:T})$ is a reparametrization of $p(\epsilon_{d,1:T})$.

Proposition 3. Consider the model proposed above and further assume $\alpha \gg \beta$. Minimizing $\mathcal{L}$ is equivalent to minimize both L_{prop1} and L_{prop2} . The model satisfies both Prop.1 and 2.

Intuitively :

- shuffling the static codes $f_{1:T}$ and using them to reconstruct the frames, enforces the static code to be time invariant, while simultaneously being informative due to the reconstruction objective $\Rightarrow$ capture static factors
- Thanks to the cNF the dynamic codes $d_{1:T}$ are encouraged to encode the residuals of the frames given the static factors $\Rightarrow$ capture dynamic factors

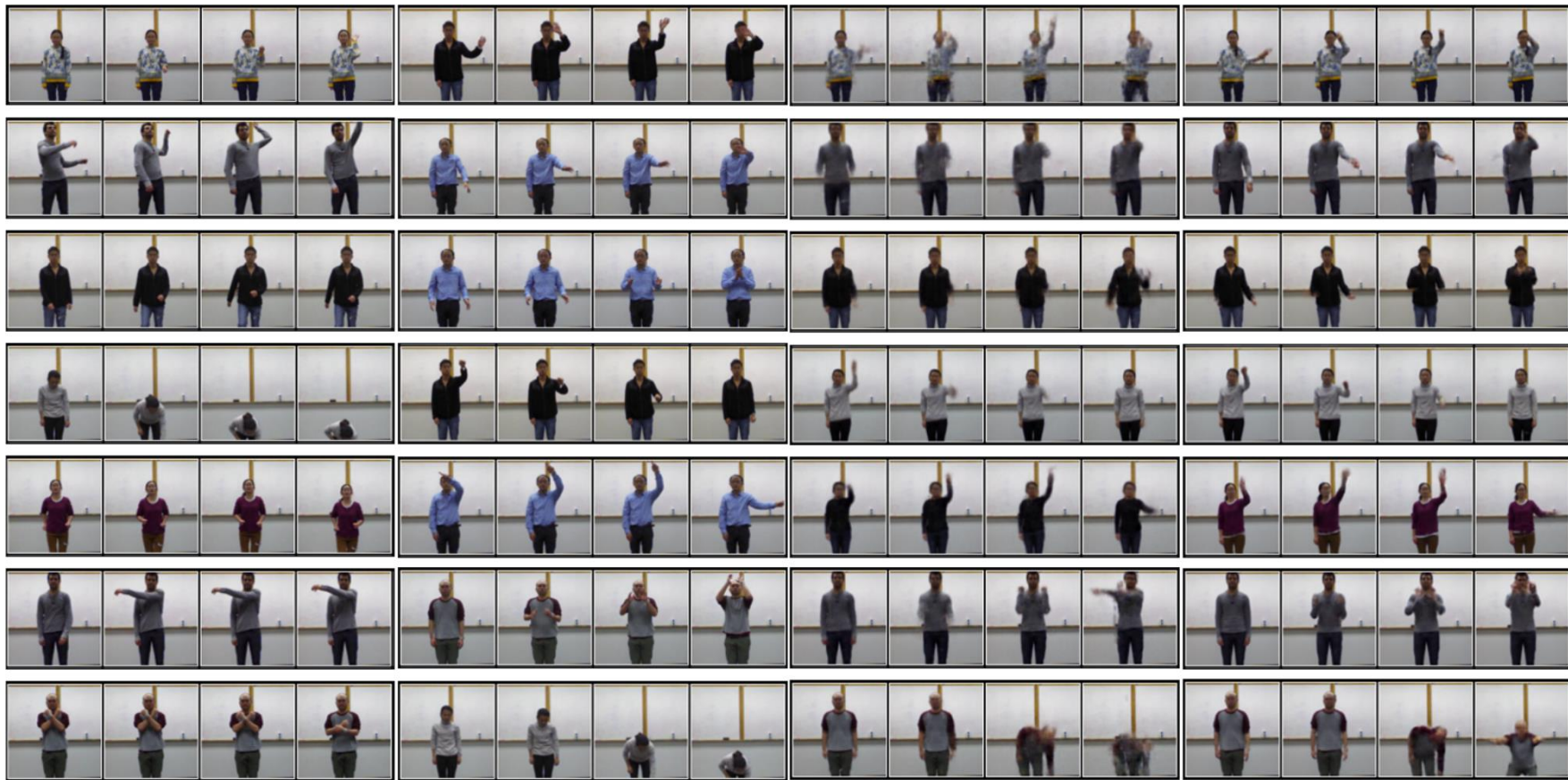


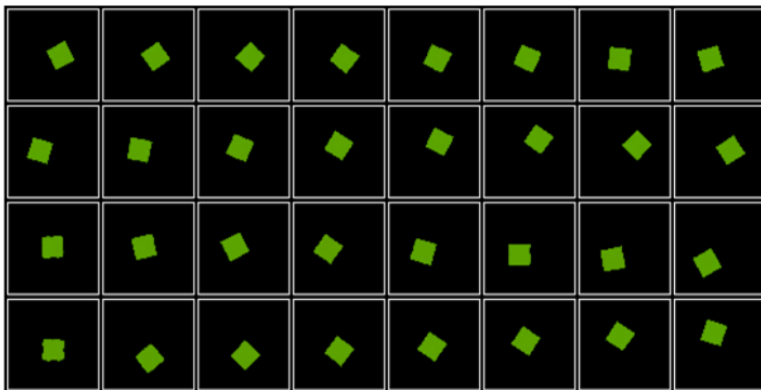
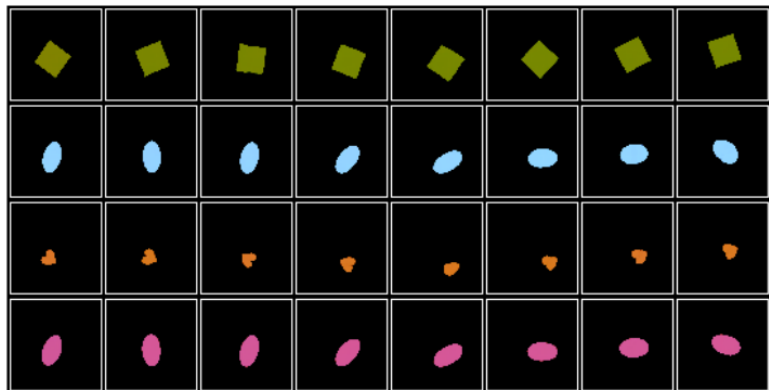
Static

Dynamic

CDSVAE

Proposed





Conclusion

Novel **formal** approach
toward sequential
disentanglement



Extend VAE framework
with a **cNF** and
**generalize to both
dependent and
independent factors**



Proposed a **novel
evidence lower bound**,
provably disentangled



Through a **simple shuffle**,
**disentangle without
requiring any additional
complex losses**, modality
free approach.

Thanks!

Any questions?

Sequential Representation Learning via Static-Dynamic Conditional Disentanglement

Mathieu Cyrille Simon, Pascal Frossard, and Christophe De Vleeschouwer

