Learning Disentangled Representations of Video with Missing Data

Armand Comas Massagué[1†], Chi Zhang[2‡], Zlatan Feric[3‡], Octavia Camps[4†], Rose Yu[5*]

† College of Electrical and Computer Engineering, ‡ Khoury College of Computer Sciences, Northeastern University, MA, USA, * Computer Science & Engineering, University of California San Diego, CA, USA.
[1] comasmassague.a@northeastern.edu, [2] zhang.chi13@northeastern.edu, [3] feric.z@northeastern.edu [4] camps@coe.neu.edu [5] roseyu@ucsd.edu

Problem

- How to learn representations of video sequences in the presence of missing data?

Our Results

*We propose a deep generative model: Disentangled-Imputed-Video-autoEncoder (**DIVE**).*

- Learns representations factorized into **appearance**, **pose** and **missingness** latent variables;
- **Imputes** missing data by sampling from the learned latent variables;
- Performs unsupervised stochastic **video prediction** using the imputed hidden representation;
- Robustly generates objects even when their appearances are changing by modeling the **static** and **dynamic appearances** separately.
- outperforms the state of the art by a substantial margin on a moving MNIST dataset with various missing scenarios, and on a real-world MOTSChallenge pedestrian dataset.

Code: `https://github.com/Rose-STL-Lab/DIVE`

Disentangled-Imputed-Video-autoEncoder (DIVE)

Deep Generative Model Denote a video sequence with missing data as $(\mathbf{y}^1, \cdots, \mathbf{y}^t)$ where each $\mathbf{y}^t \in \mathbb{R}^d$ is a frame. $\mathbf{x}^t \in \mathbb{R}^d$ are the complete frames in the future.

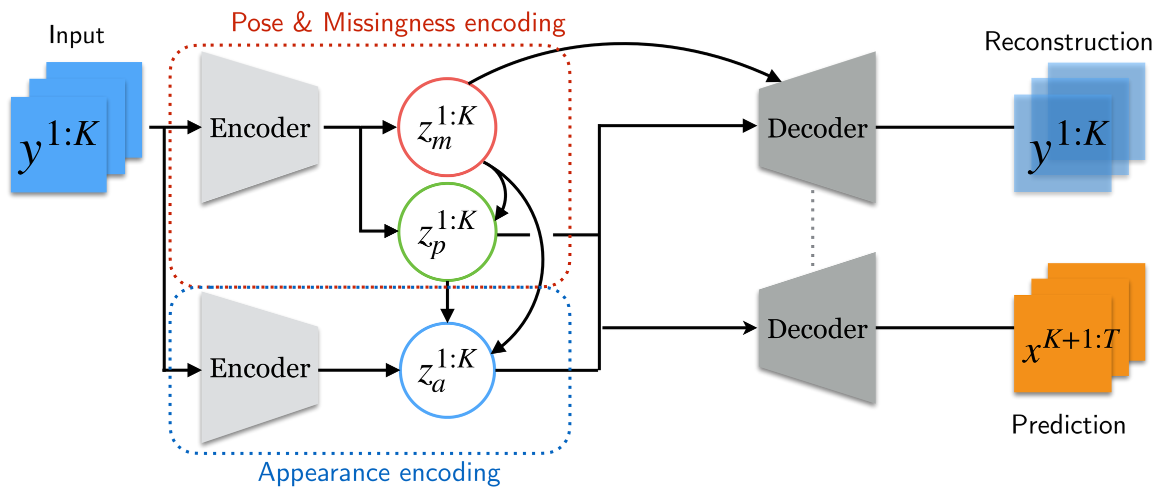
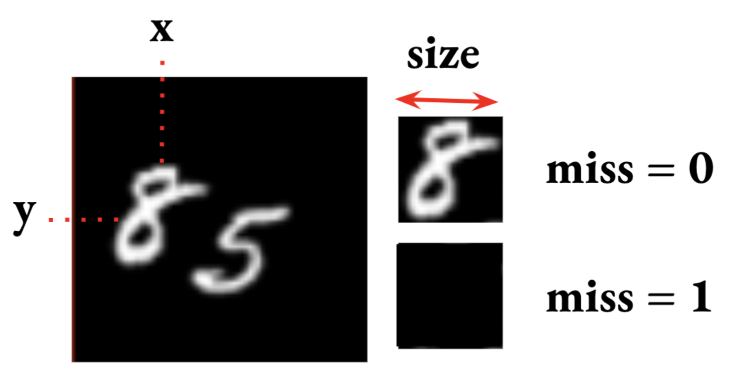


Figure: Overall architecture of DIVE, which takes the input video with missing data, infers the missingness (red), pose (green) and appearance (blue) latent variables. Two separate decoders reconstruct and predict the future sequences.

Disentangled Representation

$$\mathbf{z}_i^t = [\mathbf{z}_{i,a}^t, \mathbf{z}_{i,p}^t, \mathbf{z}_{i,m}^t], \quad \mathbf{z}_{i,a}^t \in \mathbb{R}^h, \mathbf{z}_{i,p}^t \in \mathbb{R}^3, \mathbf{z}_{i,m}^t \in \mathbb{Z} \tag{1}$$



- $\mathbf{z}_{i,a}^t$ **Appearance** vector.
- $\mathbf{z}_{i,p}^t$ **Pose** vector (x, y, size).
- $\mathbf{z}_{i,m}^t$ **Missingness** label.

Generative Model and Learning

- The **generative distribution** is given by:

$$p(\mathbf{y}^{1:K}, \mathbf{x}^{K+1:T} | \mathbf{z}^{1:T}) = \prod_{i=1}^{N} p(\mathbf{y}_i^{1:K} | \mathbf{z}_i^{1:K}) p(\mathbf{x}_i^{K+1:T} | \mathbf{z}_i^{K+1:T}) \tag{2}$$

- We **locate** each decoded object with its estimated pose; **mask** it with the missingness label; and **sum** over all objects.

$$p(\mathbf{y}_i^t | \mathbf{z}_{i,a}^t) = \mathcal{T}(f_{\text{dec}}(\mathbf{z}_{i,a}^t); \mathbf{z}_{i,p}^t) \circ (1 - \mathbf{z}_{i,m}^t) \tag{3}$$

Following the VAE framework, we train the model by maximizing the evidence lower bound (ELBO).

Imputation and Inference Model

Graphical Representation

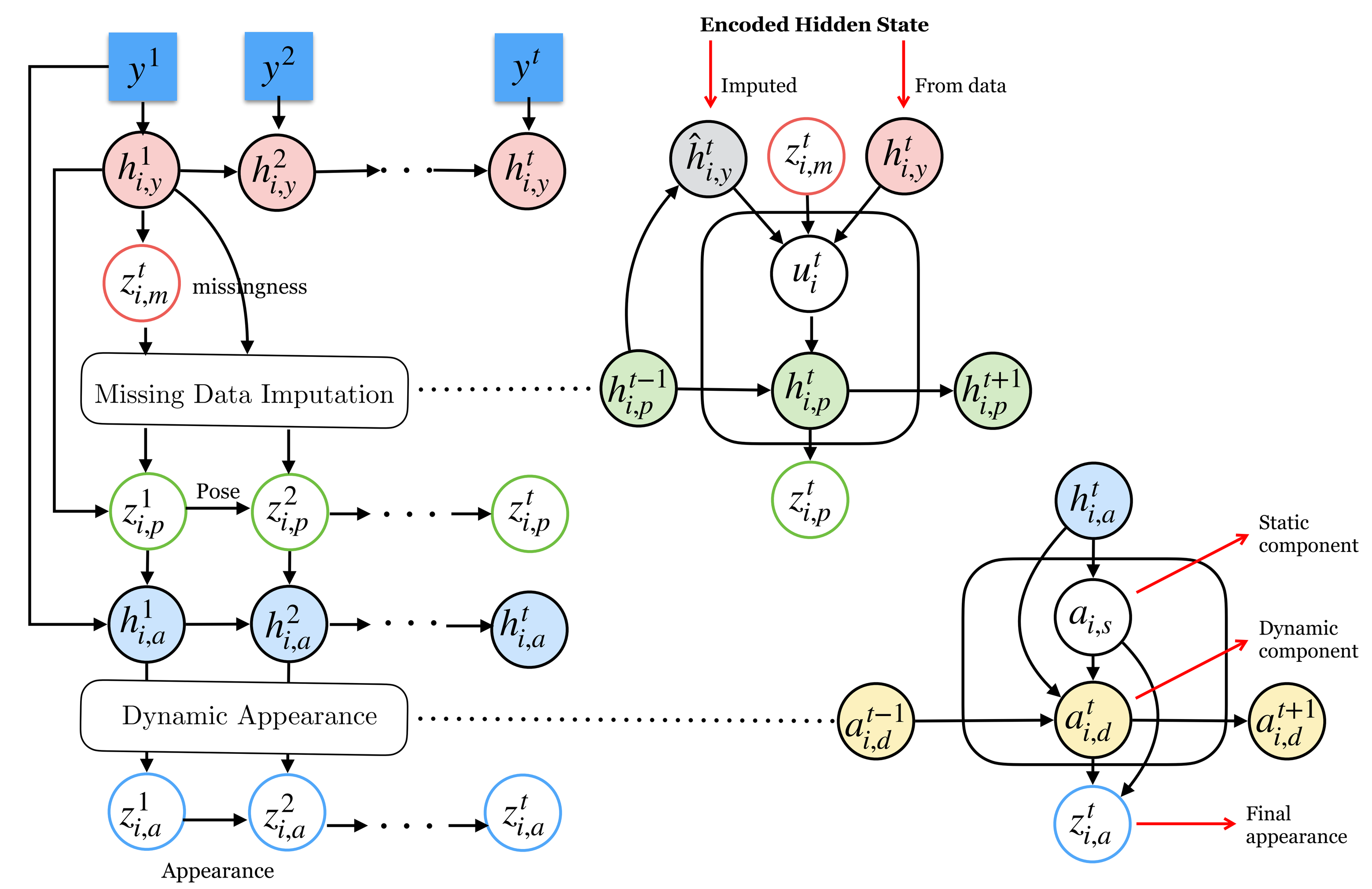


Figure: A graphical representation of DIVE. From top to bottom: inference of the missingness variable $\mathbf{z}_{i,m}^t$, missing data imputations model, inference of the pose vectors $\mathbf{z}_{i,p}^t$ and appearance variable $\mathbf{z}_{i,a}^t$ using dynamic appearance inference.

Variational Inference

Missingness: $\mathbf{z}_{i,m}^t$, we leverage the input encoding and binarize it with a Heavyside step function. $\mathbf{z}_{i,m}^t = \text{Heavyside}(\texttt{FC}(\mathbf{h}_{i,y}^t))$. $\mathbf{h}_{i,y}^t$ is a hidden representation of the input data.

Pose: We obtain $\mathbf{z}_{i,p}^{1:K}$ from the pose hidden representation as $\mathbf{h}_{i,p}^t = \texttt{LSTM}(\mathbf{h}_{i,p}^{t-1}, \mathbf{u}_i^t)$. $\mathbf{u}_i^t = \mathbf{h}_{i,y}^t$ if $\mathbf{z}_{i,m}^t = 0$. In case of missing data, we **impute** $\mathbf{h}_{i,y}^t$ instead of relying on the input data.

Appearance: We decompose it into a **Static component** $\mathbf{a}_{i,s}$: It captures the inherent **semantics**; and a **Dynamic component** $\mathbf{a}_{i,d}^t$: that models the **nuanced variations** in shape. For the **final appearance** vector, we concatenate and mix and the dynamic and static components.

Experiments

MOTS Pedestrian Dataset

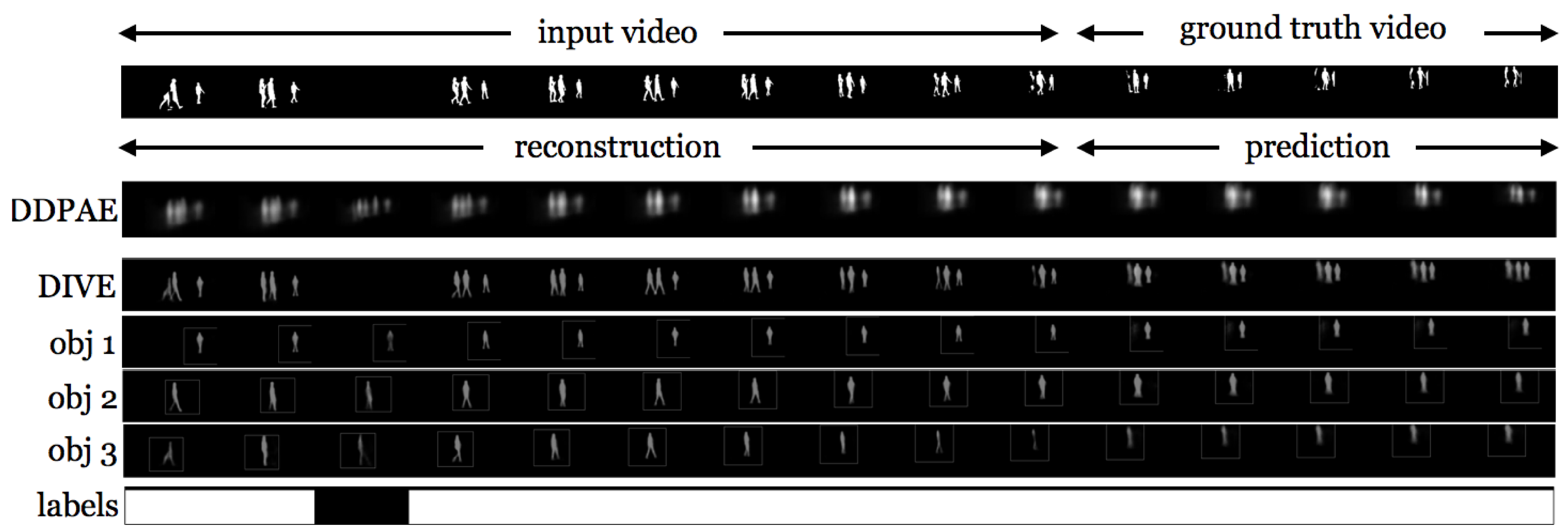


Figure: MOTS dataset qualitative results. Note that our method successfully identifies the missing time step, decomposes the objects and keeps track of the missing pedestrians.

Model	BCE ↓	MSE ↓	PSNR ↑	SSIM ↑	NELBO ↓
DDPAE	2495.08	560.37	22.22	0.90	-0.24
DIVE	**1355.89**	**328.96**	**24.82**	**0.96**	**-0.26**

Table: Quantitative comparison on MOTS pedestrian dataset for DDPAE and DIVE.

Moving MNIST

Scenario 1: Partial occlusion: We can only see 50% of the frame.

Scenario 2: Out of scene:: Objects disappear independently for 2 consecutive time-steps.

Scenario 3: Varying: Out of scene 1 time-step + appearance variation in time.

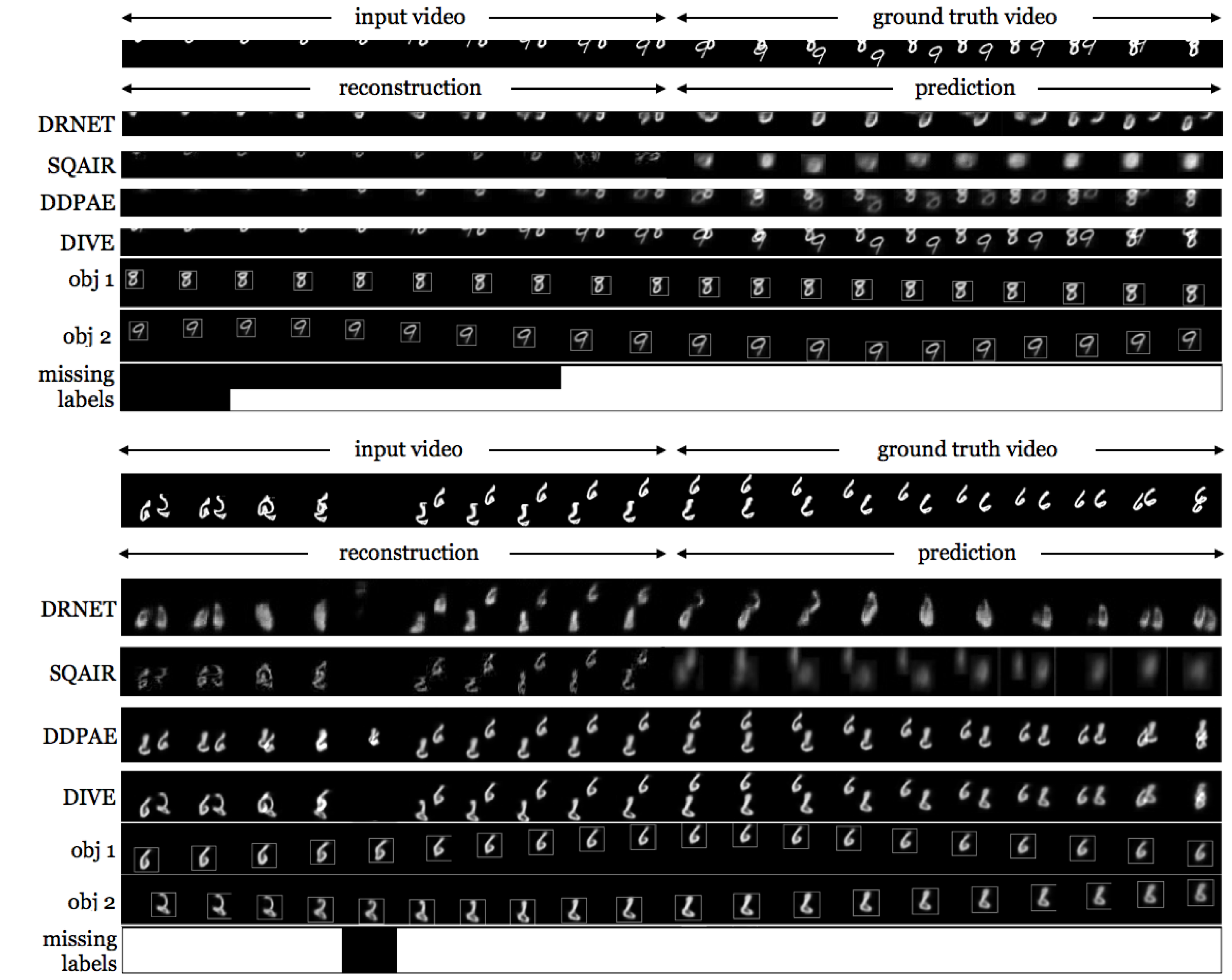


Figure: Qualitative results. *Obj 1* and *Obj 2* show DIVEs individual object generations and *missing labels* indicate whether each object is estimated completely missing in the scene. Note that objects are well decomposed, sharply generated and the labels properly predicted. **Scenario 1** and **Scenario 3** respectively.

Scenario 1	BCE ↓		MSE ↓		PSNR ↑		SSIM ↑		NELBO ↓
Model	Rec	Pred	Rec	Pred	Rec	Pred	Rec	Pred	
DRNET[1]	482.07	852.59	72.21	96.36	7.99	6.89	0.76	0.72	/
SQAIR[3]	178.71	967.20	21.84	84.73	13.19	9.96	**0.90**	0.73	-0.16
DDPAE[2]	182.66	**417.00**	39.09	67.41	17.56	15.49	0.77	0.72	-0.09
DIVE	**119.25**	459.10	**19.73**	**64.49**	**20.64**	**15.85**	**0.90**	**0.78**	**-0.18**
Scenario 2									
DRNET	392.33	1402.45	90.64	187.72	9.59	9.88	0.80	0.67	/
SQAIR	468.22	927.09	73.13	137.04	10.33	8.21	0.84	0.69	-0.17
DDPAE	266.03	409.26	58.37	89.57	18.64	16.94	0.87	0.77	-0.17
DIVE	**165.42**	**321.29**	**27.03**	**64.17**	**22.15**	**18.56**	**0.93**	**0.83**	**-0.21**
Scenario 3									
DRNET	421.72	1304.53	90.46	176.28	9.91	7.33	0.75	0.70	/
SQAIR	560.51	1518.61	74.30	163.25	10.80	7.64	0.83	0.62	-0.16
DDPAE	322.23	403.48	63.63	82.71	18.29	17.22	0.81	**0.78**	-0.18
DIVE	**272.74**	**374.59**	**42.81**	**74.87**	**20.08**	**17.61**	**0.87**	**0.78**	**-0.19**

Table: Quantitative comparison

Acknowledgments This work was supported in part by NSF under Grants IIS#1850349, IIS#1814631, ECCS#1808381 and CMMI#1638234, the U. S. Army Research Office under Grant W911NF-20-1-0334 and the Alert DHS Center of Excellence under Award Number 2013-ST-061-ED0001.

References

[1] E. L. Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423, 2017.

[2] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 517–526, 2018.

[3] A. Kosiorek, H. Kim, Y. W. Teh, and I. Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems*, pages 8606–8616, 2018.