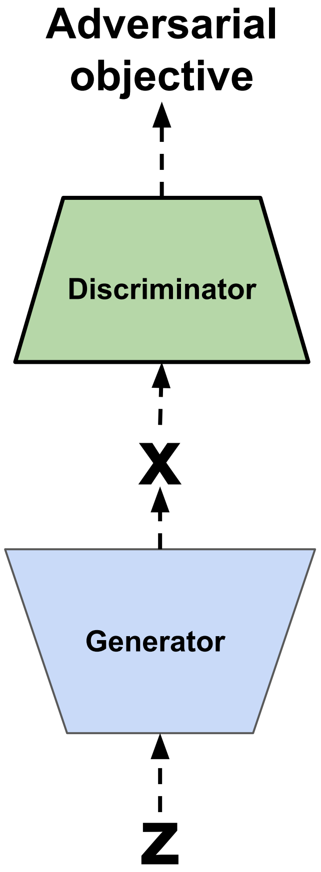
Disentangling Content and Pose with an Adversarial loss

Emily Denton

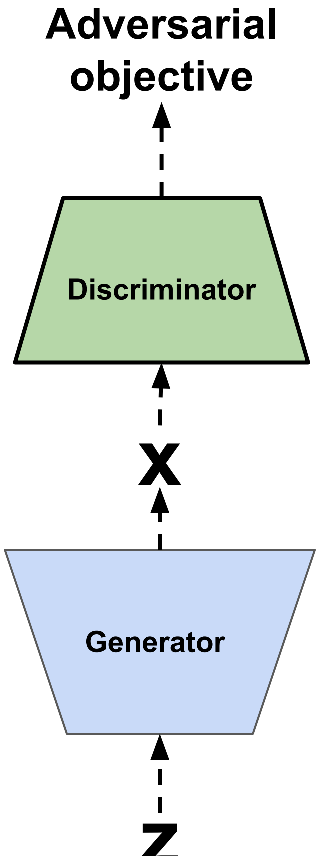
CVPR GAN Tutorial
June 2018

NYU

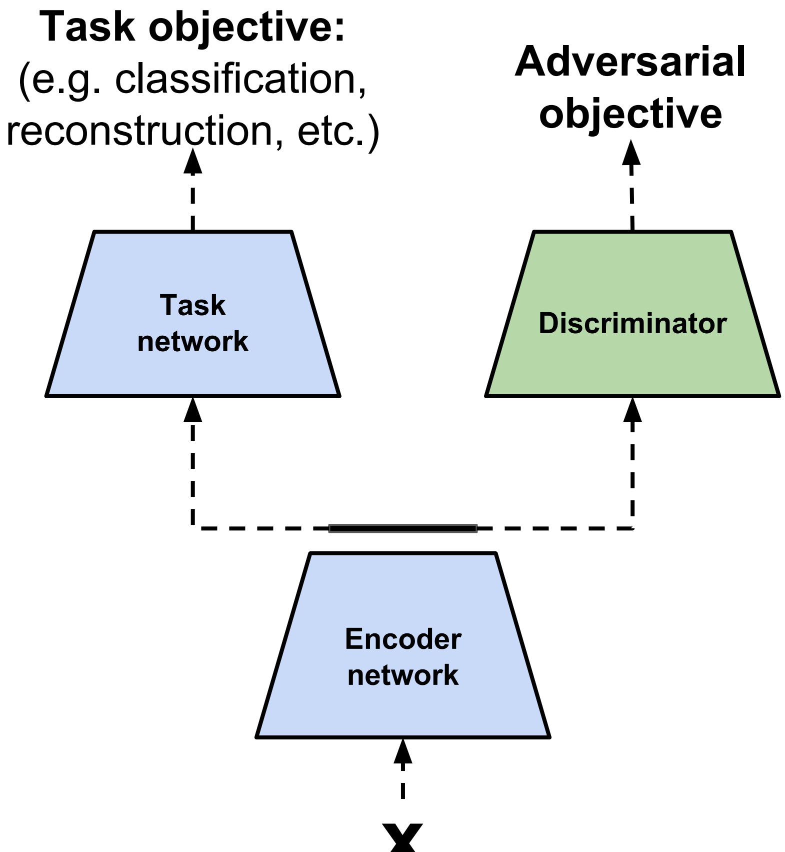
Generative adversarial network framework:



Generative adversarial network framework:



Adversarial losses to shape representations:



Part I: Disentangling content and pose with an adversarial loss

Denton and Birodkar. *Unsupervised Learning of Disentangled Representations from Video*. NIPS, 2017

Part II: Survey of adversarial losses in feature space

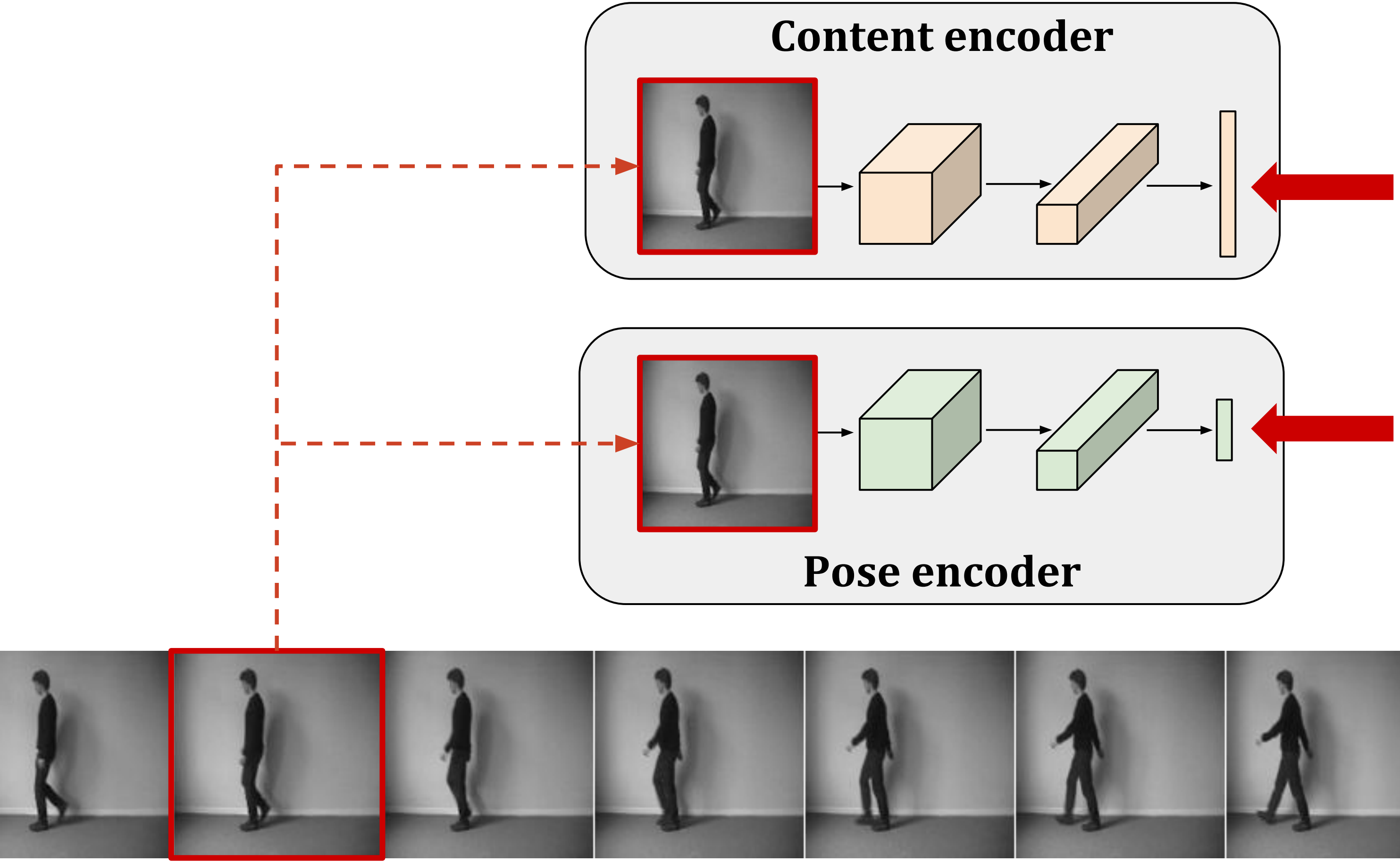
Disentangled Representation Net (DrNet)

Disentangling auto-encoder that factorizes image sequences into **temporally constant (content)** and **temporally varying (pose)** components

Time varying information: Pose of body

Time invariant information: Lighting, background, identity, clothing

DrNet: two seperate encoders



Time invariant information:
Lighting/Background
Identity/clothing

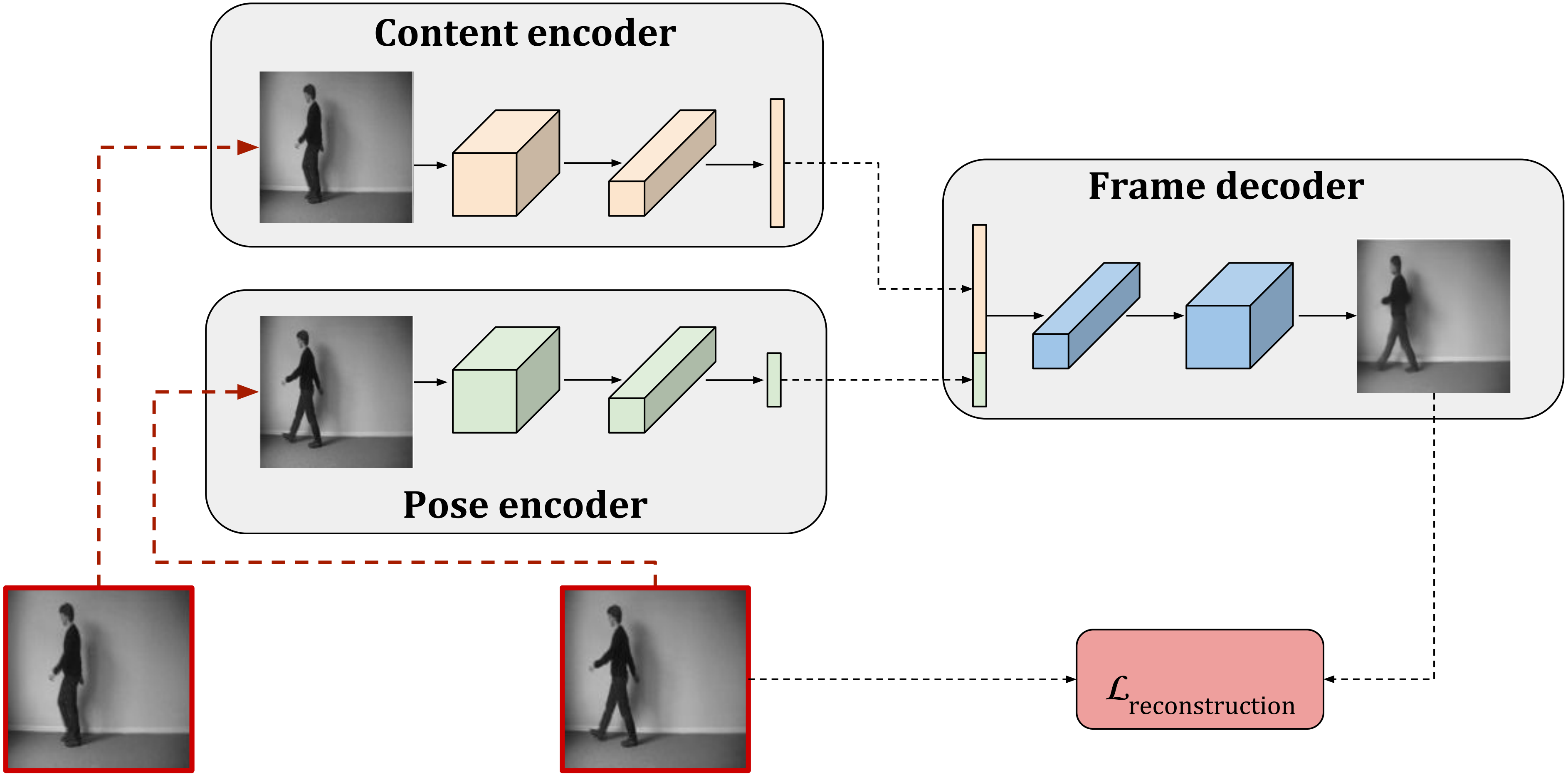
Time varying information:
Pose of body

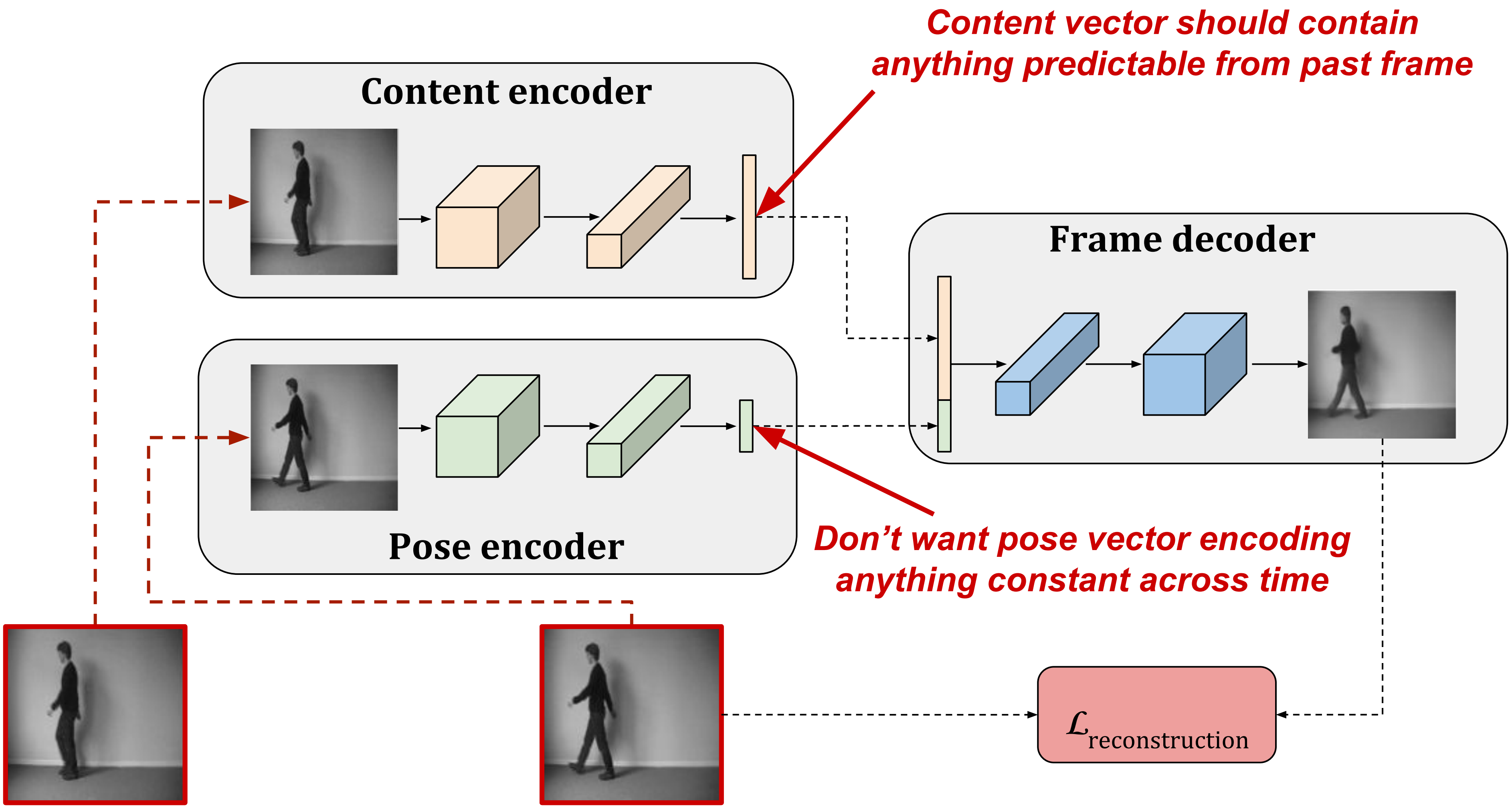
DrNet: training

- **Reconstruction loss** drives training
- **Similarity loss** makes content vectors invariant across time
- **Adversarial loss** enforces pose vectors to only contain info that changes across time

DrNet: training

- **Reconstruction loss** drives training
- **Similarity loss** makes content vectors invariant across time
- **Adversarial loss** enforces pose vectors to only contain info that changes across time

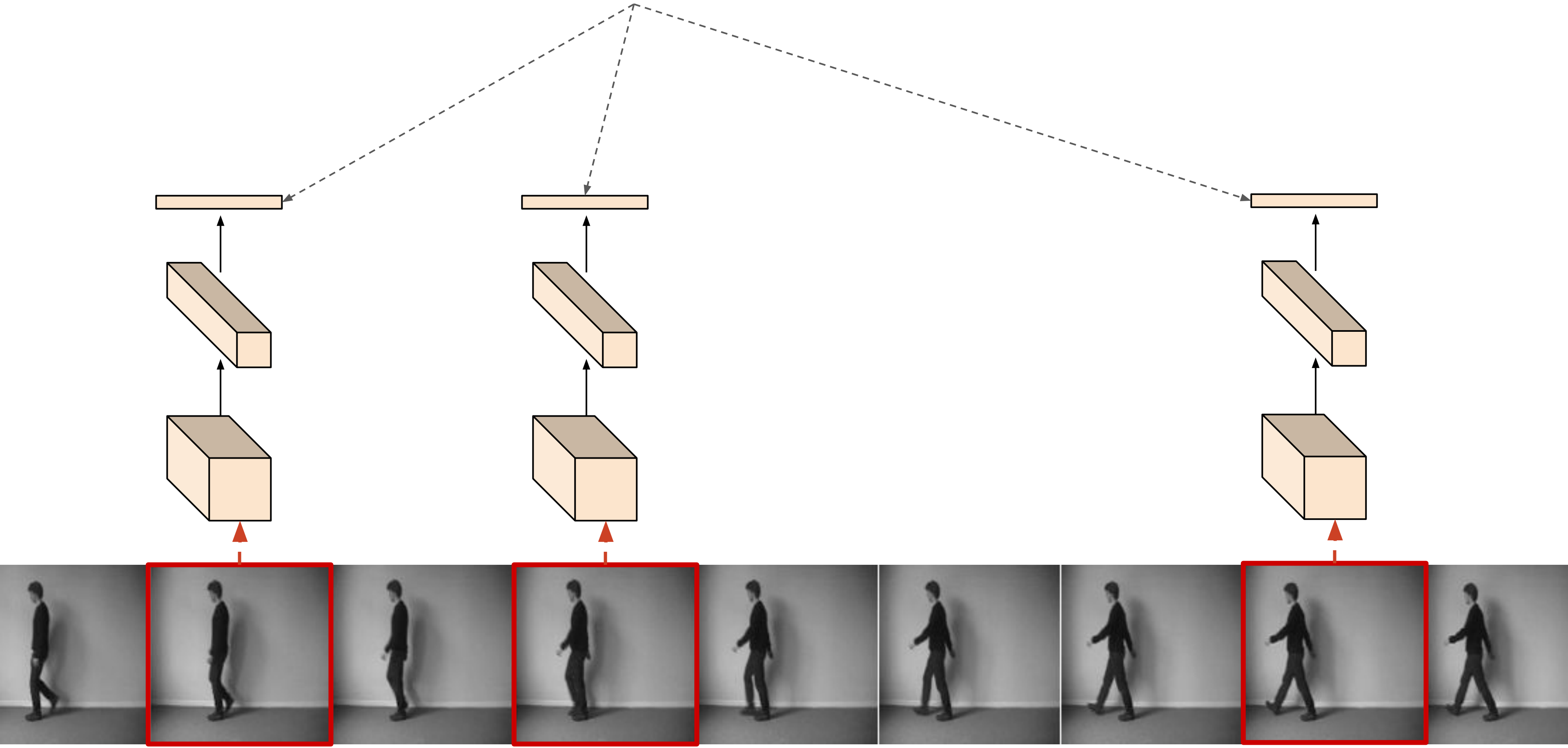

Content encoder
Frame decoder
Pose encoder
$\mathcal{L}_{\text{reconstruction}}$


Content vector should contain
anything predictable from past frame
Content encoder
Frame decoder
Pose encoder
Don't want pose vector encoding
anything constant across time
$\mathcal{L}_{\text{reconstruction}}$

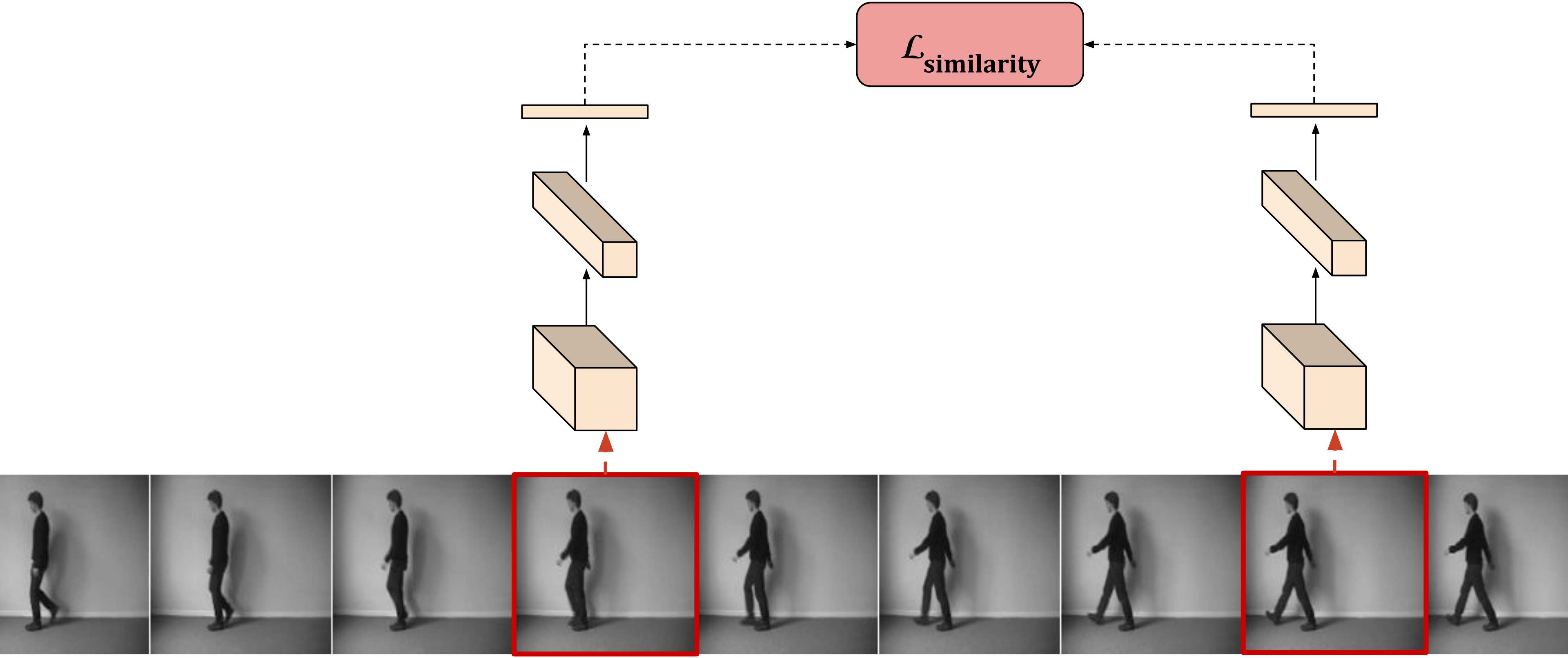
DrNet: training

- **Reconstruction loss** drives training
- **Similarity loss** makes content vectors invariant across time
- **Adversarial loss** enforces pose vectors to only contain info that changes across time

Content vectors should be invariant across time



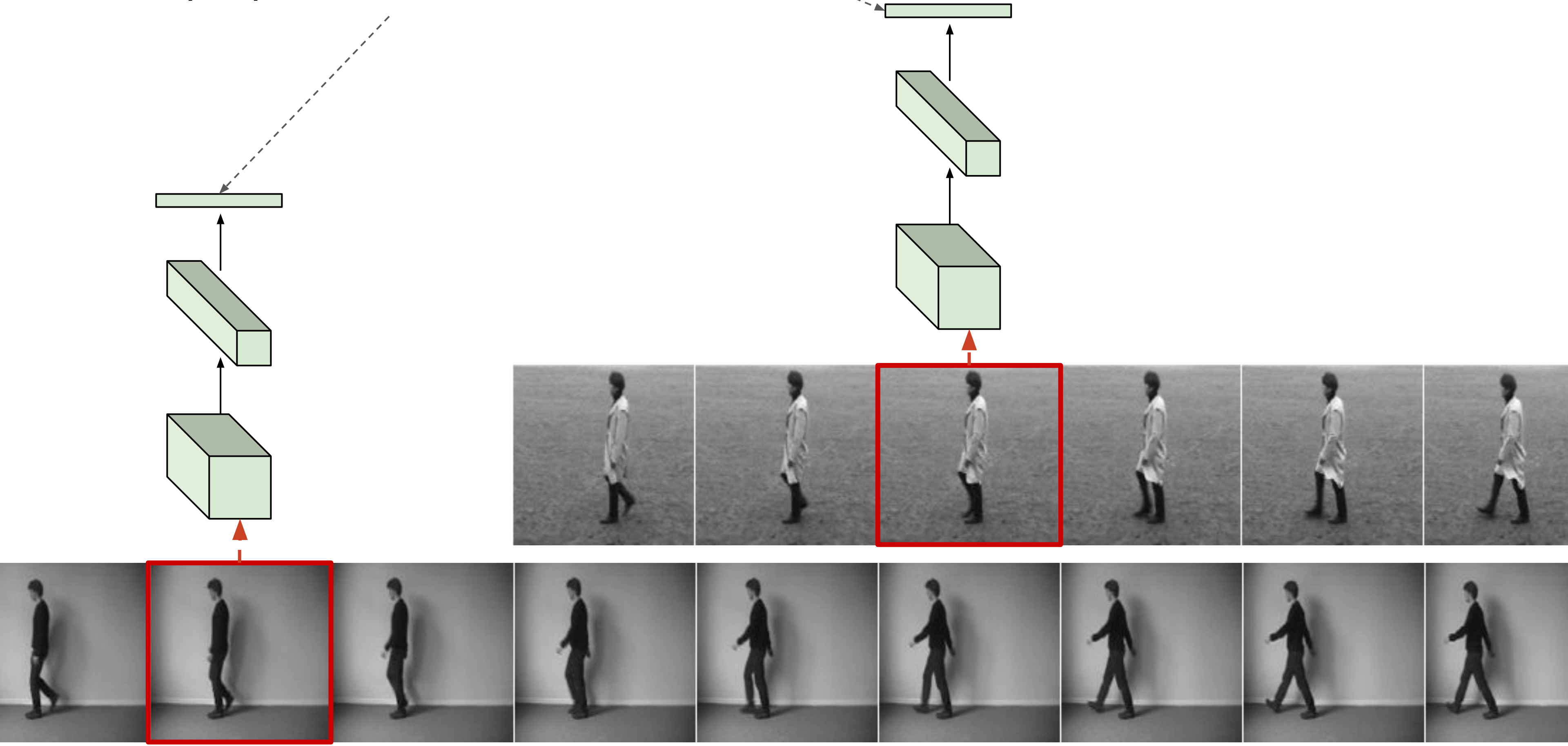
l2 similarity loss on temporally nearby content vectors

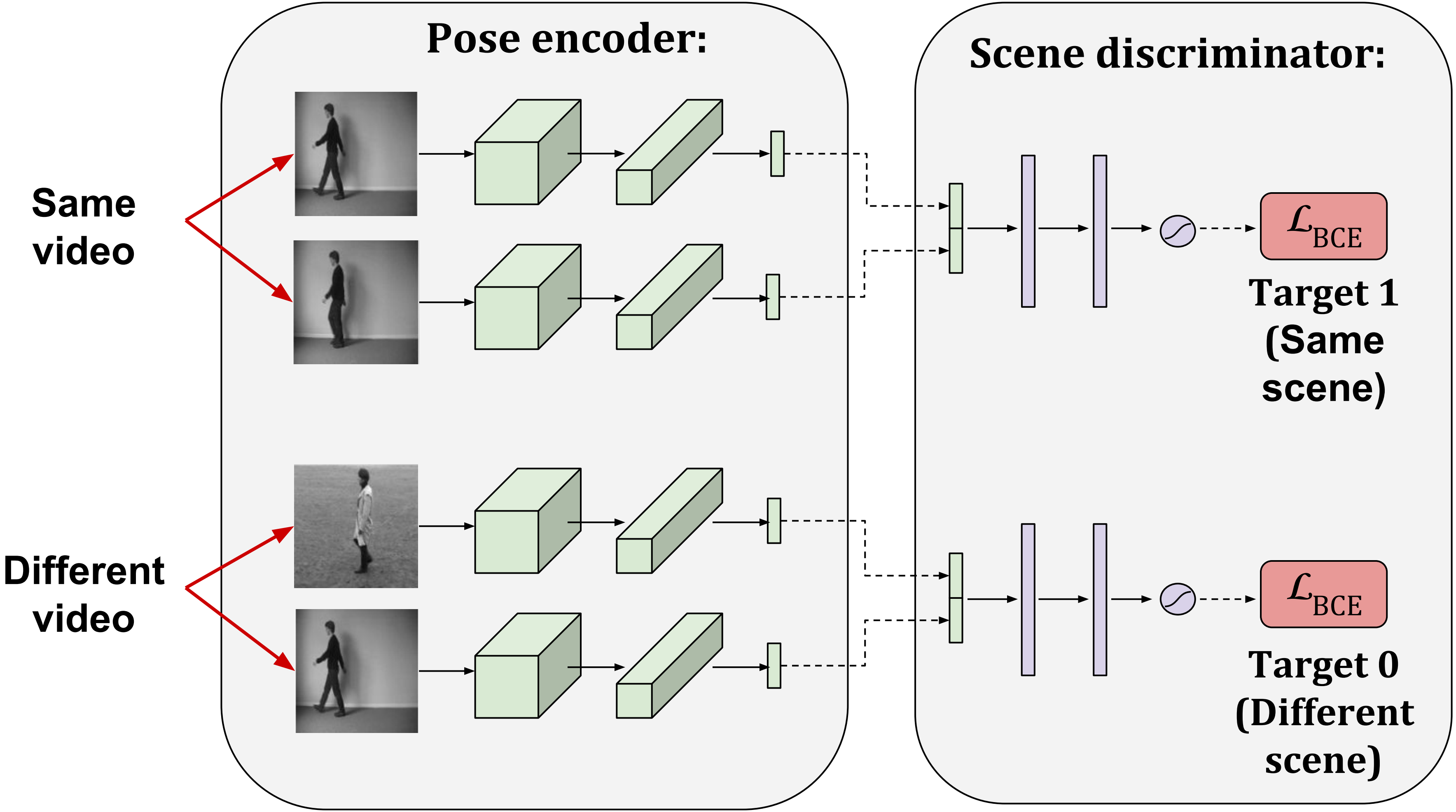


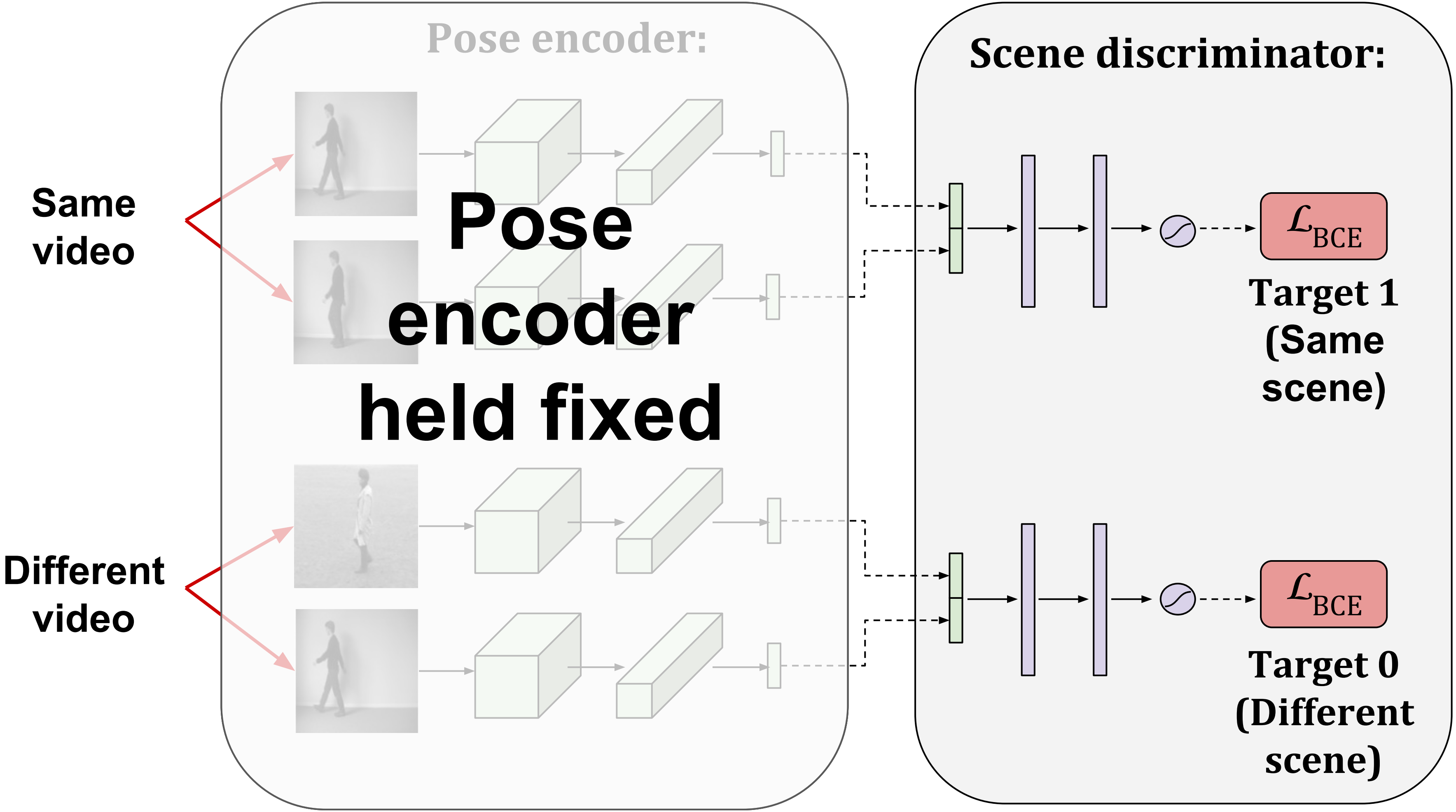
DrNet: training

- **Reconstruction loss** drives training
- **Similarity loss** makes content vectors invariant across time
- **Adversarial loss** enforces pose vectors to only contain info that changes across time

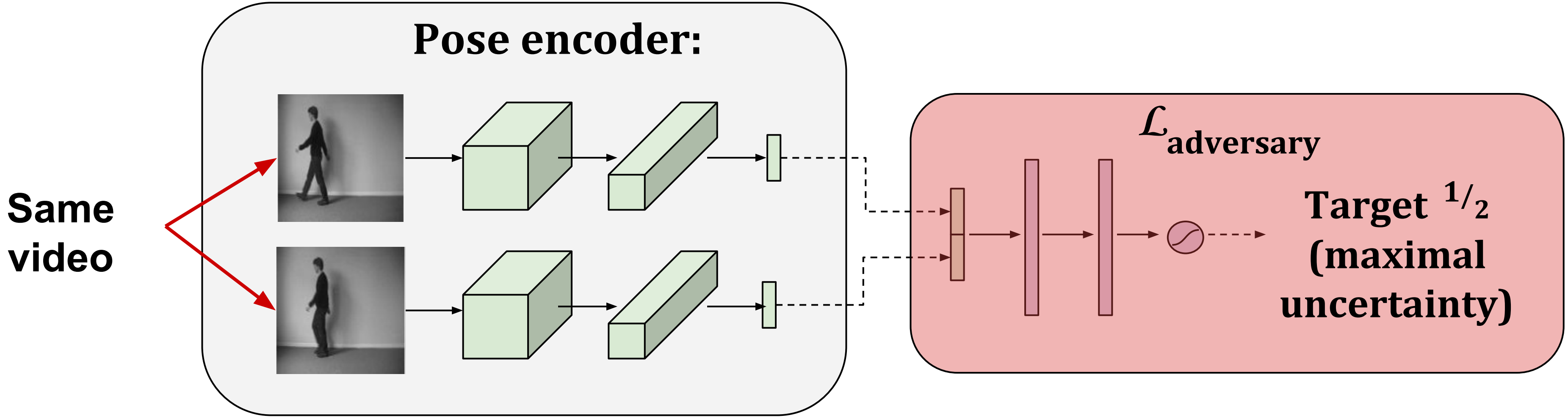
Should not be able to distinguish which video clip a pose vector comes from




Pose encoder:
Scene discriminator:
Same video
Different video
$\mathcal{L}_{\mathrm{BCE}}$
Target 1 (Same scene)
$\mathcal{L}_{\mathrm{BCE}}$
Target 0 (Different scene)

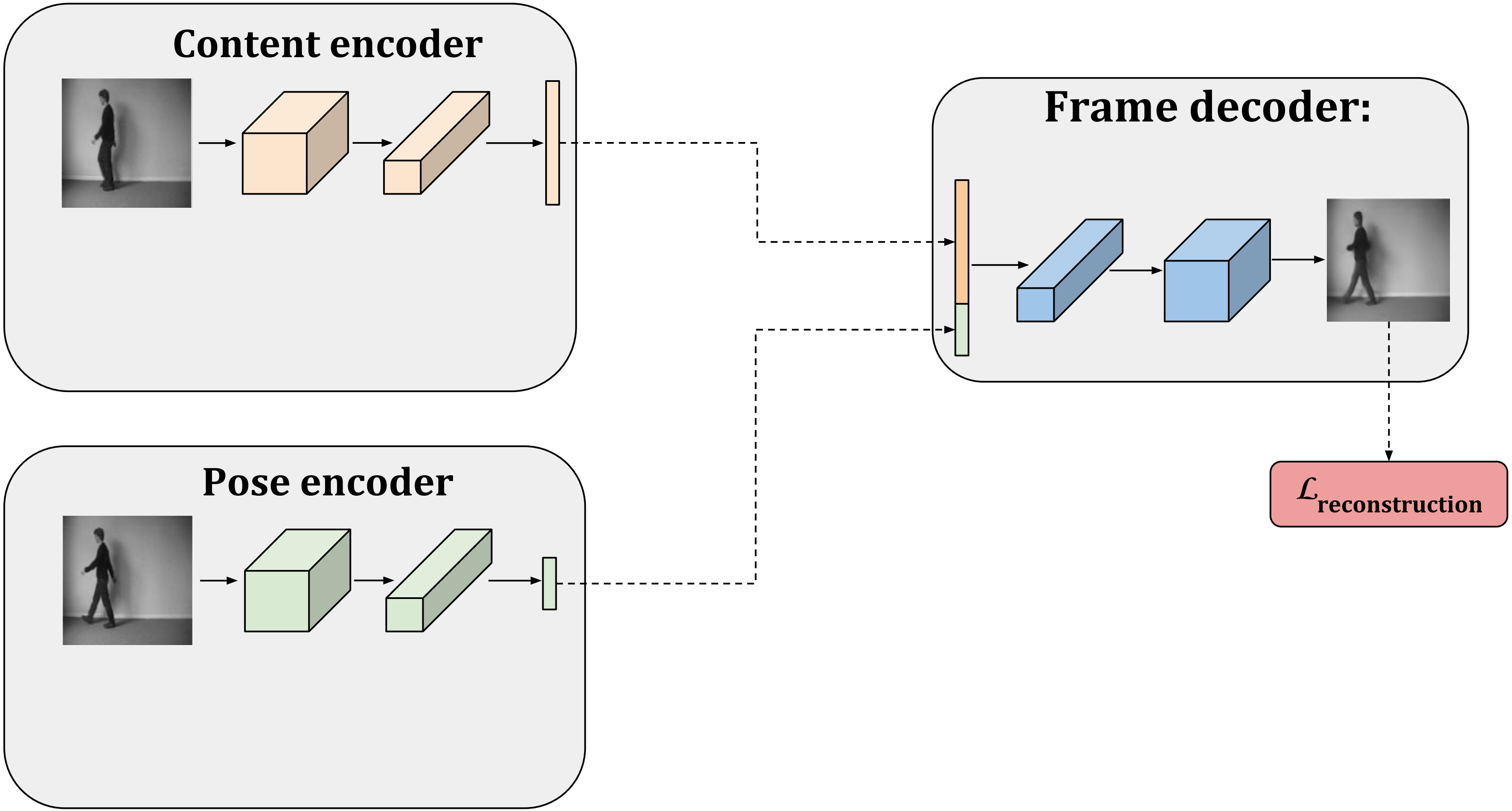


Pose encoder:
Scene discriminator:
Same video
Different video
Pose encoder held fixed
$\mathcal{L}_{\text{BCE}}$
Target 1 (Same scene)
$\mathcal{L}_{\text{BCE}}$
Target 0 (Different scene)

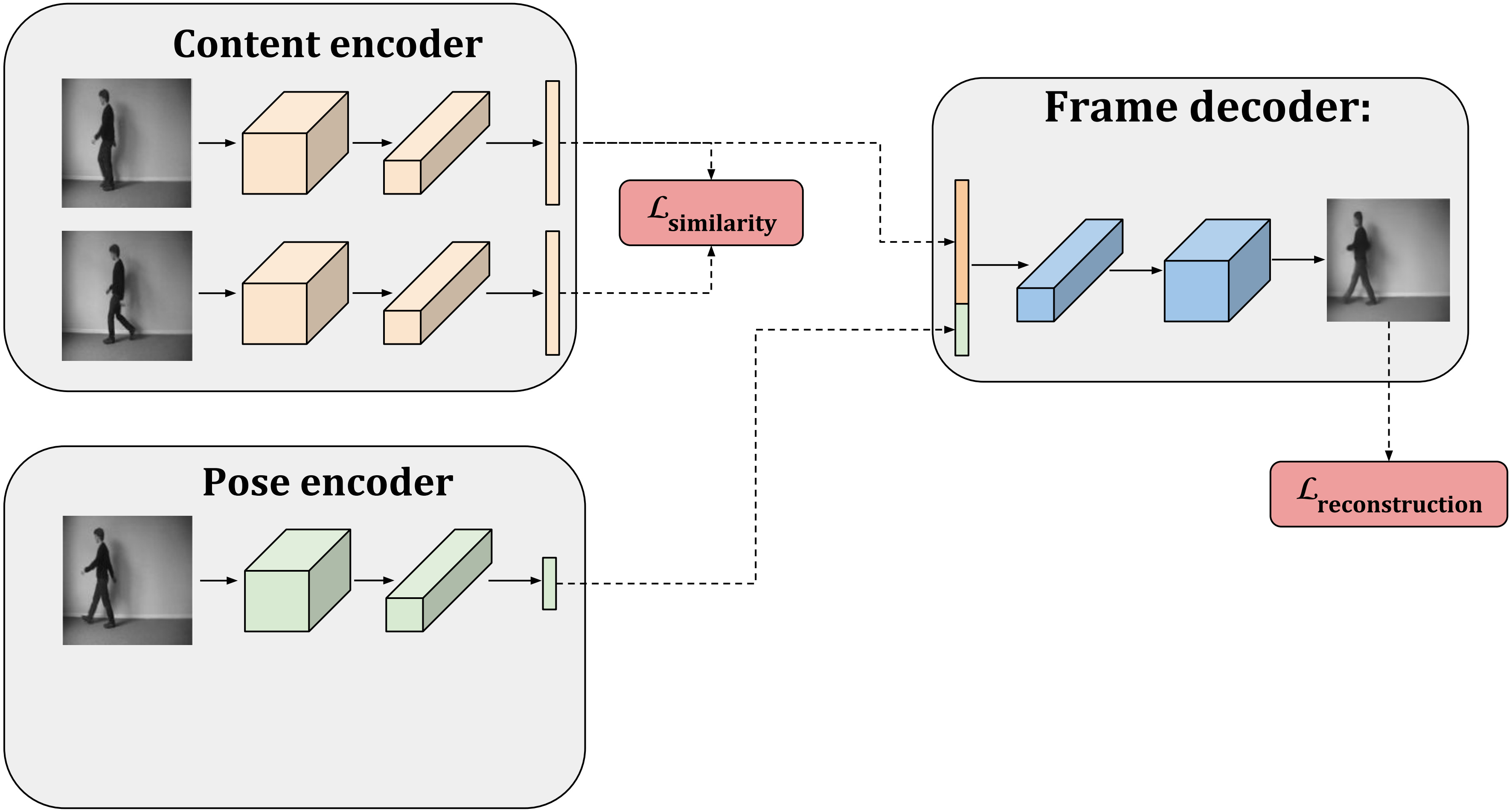


Train pose encoder to produce pose vectors that make the discriminator **maximally uncertain** about the content of the video

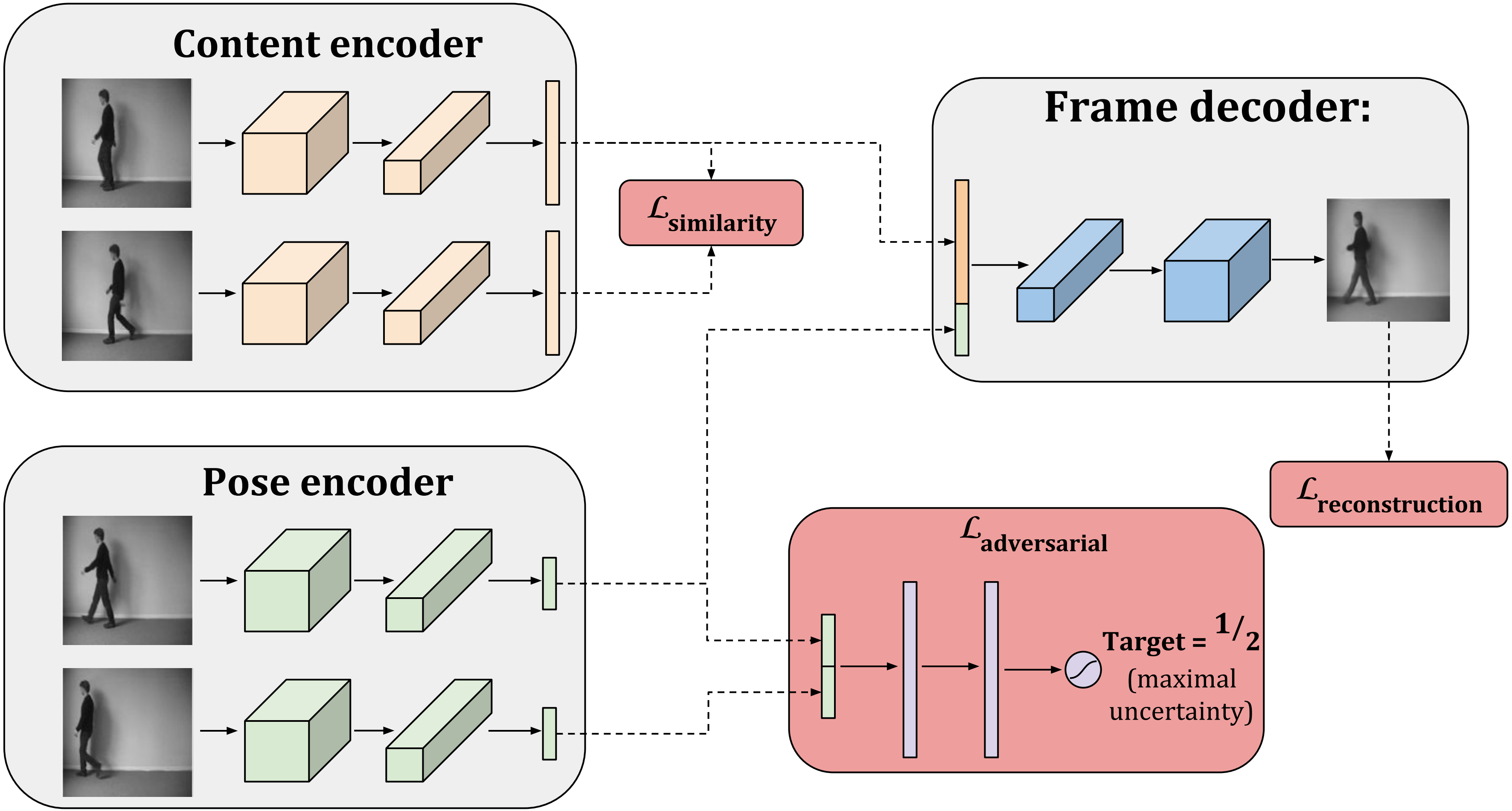
Scene discriminator held fixed, only used to compute gradients for pose encoder



$$\mathcal{L} = \mathcal{L}_{reconstruction}(E_c, E_p, D) + \alpha\mathcal{L}_{similarity}(E_c) + \beta(\mathcal{L}_{adversarial}(E_p) + \mathcal{L}_{adversarial}(C))$$



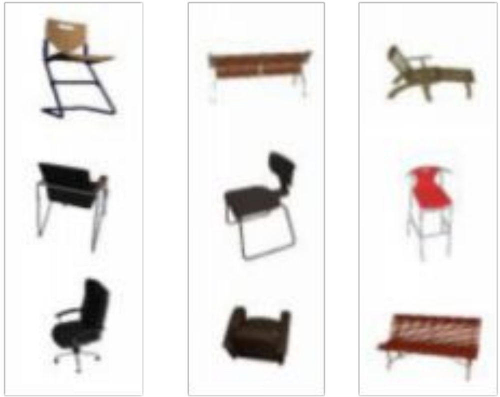
$$\mathcal{L} = \mathcal{L}_{reconstruction}(E_c, E_p, D) + \alpha\mathcal{L}_{similarity}(E_c) + \beta(\mathcal{L}_{adversarial}(E_p) + \mathcal{L}_{adversarial}(C))$$



$$\mathcal{L} = \mathcal{L}_{reconstruction}(E_c, E_p, D) + \alpha\mathcal{L}_{similarity}(E_c) + \beta(\mathcal{L}_{adversarial}(E_p) + \mathcal{L}_{adversarial}(C))$$

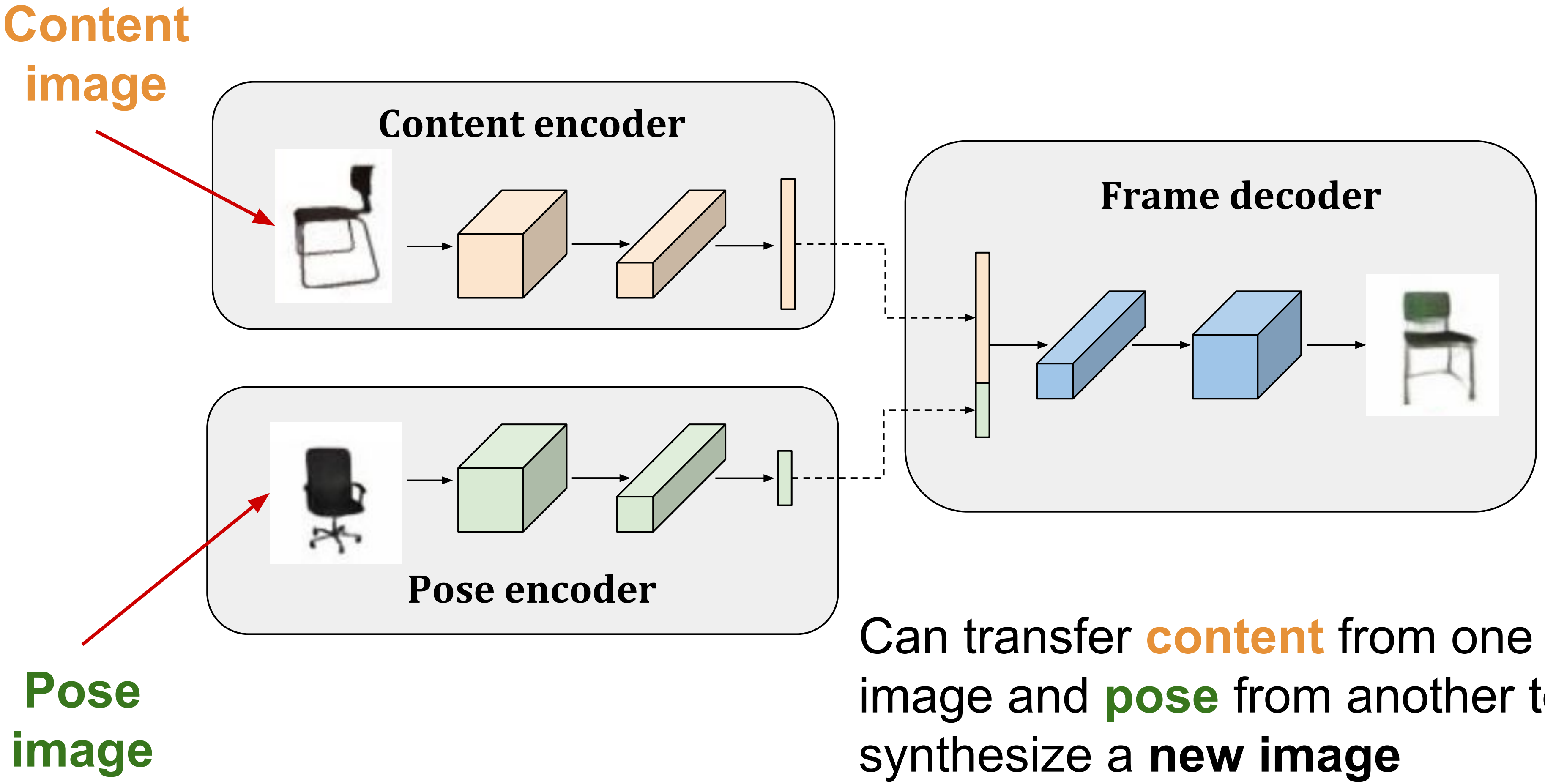
SUNCG dataset: rotating objects

- 280 chair models, 5 elevations, large variability
- Video sequence: camera rotates around chair



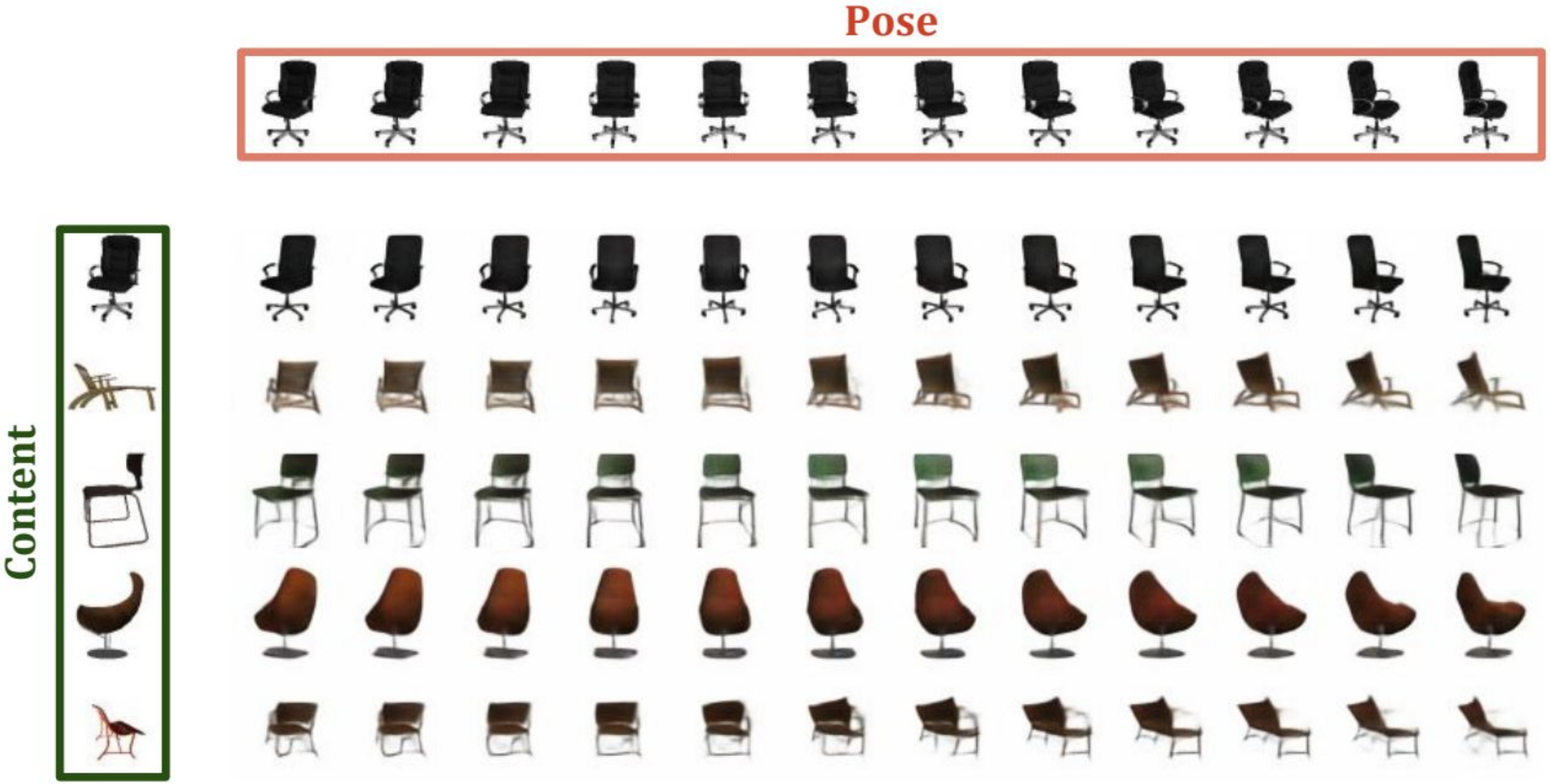
S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene comp

Image synthesis by analogy



Can transfer **content** from one image and **pose** from another to synthesize a **new image**

Image synthesis by analogy



Interpolation in pose space

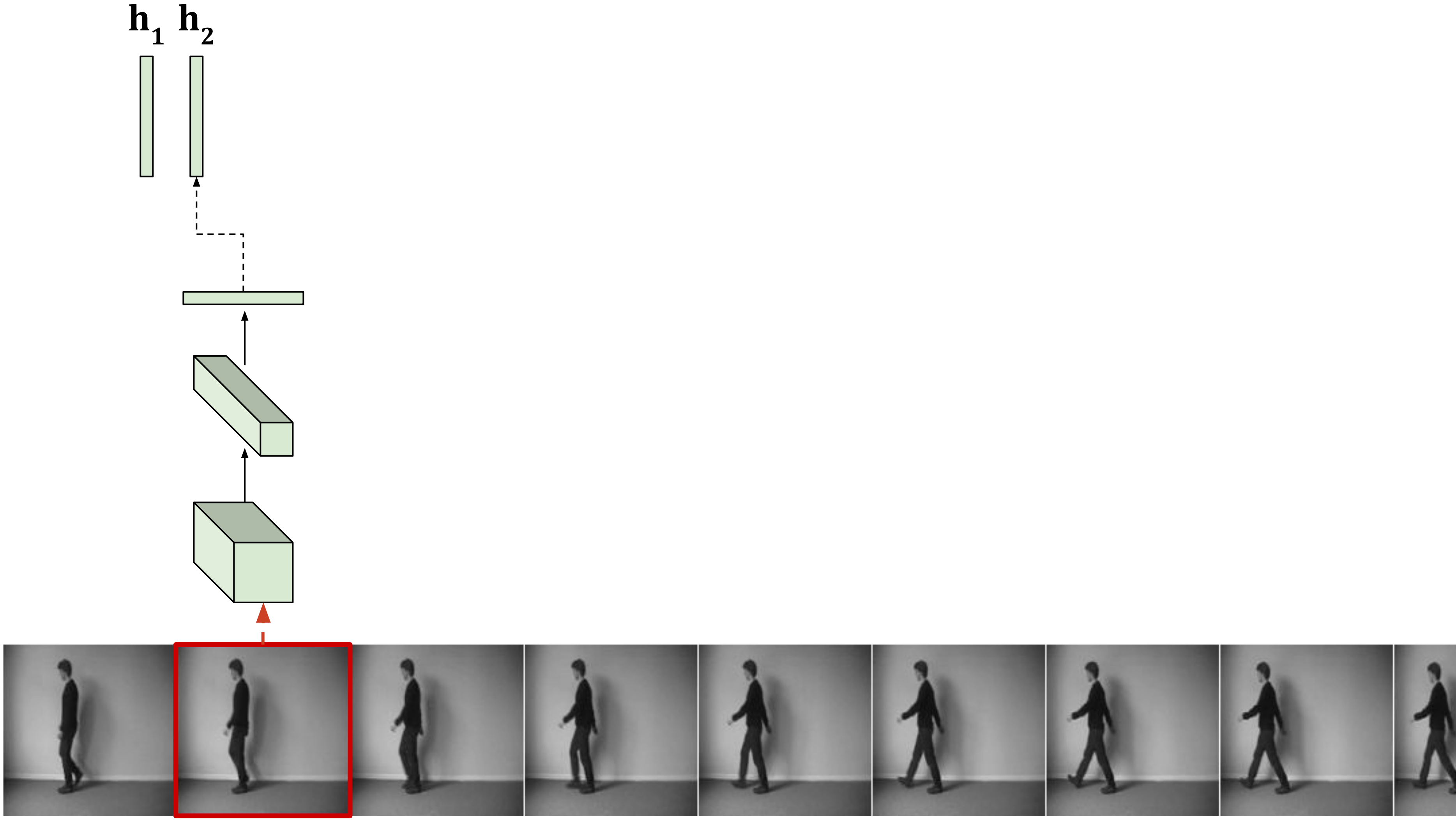
Video prediction

- A representation that factorizes into temporally constant and temporally varying components is particularly useful for video prediction

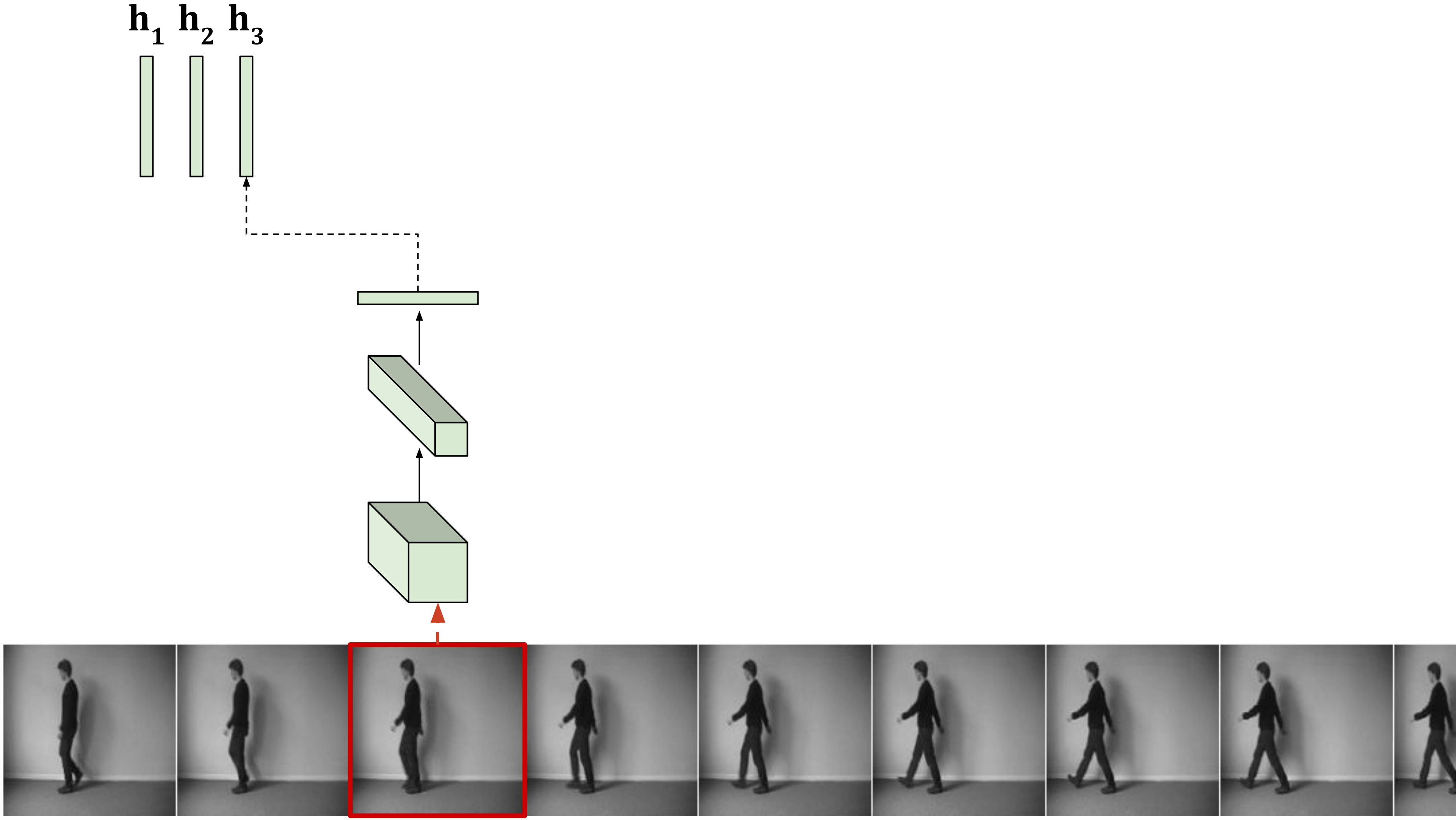
- Instead of modeling how the entire scene changes, **only need to predict the temporally varying component**

- **Prediction** done entirely in latent **pose space**

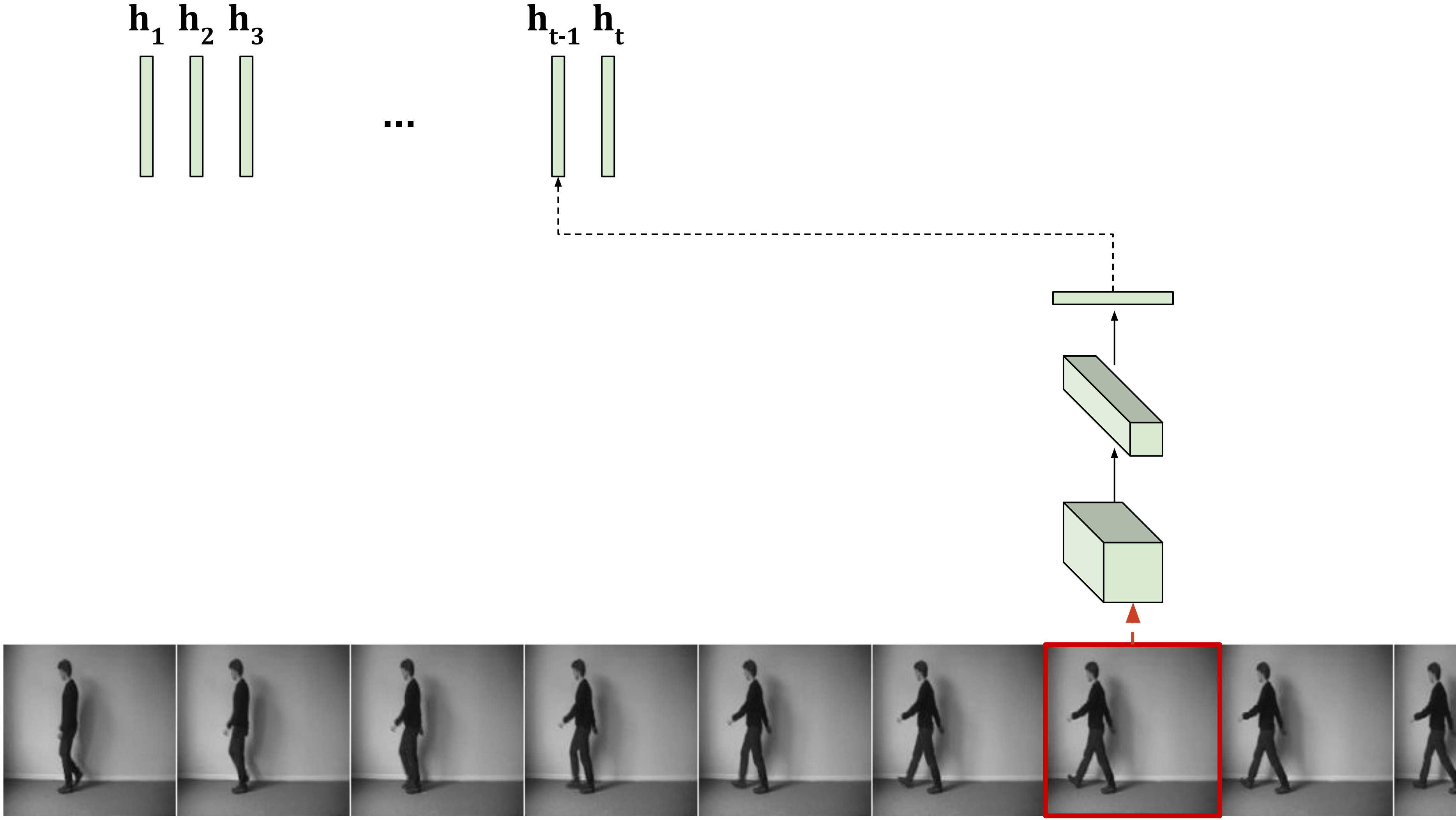

h_1



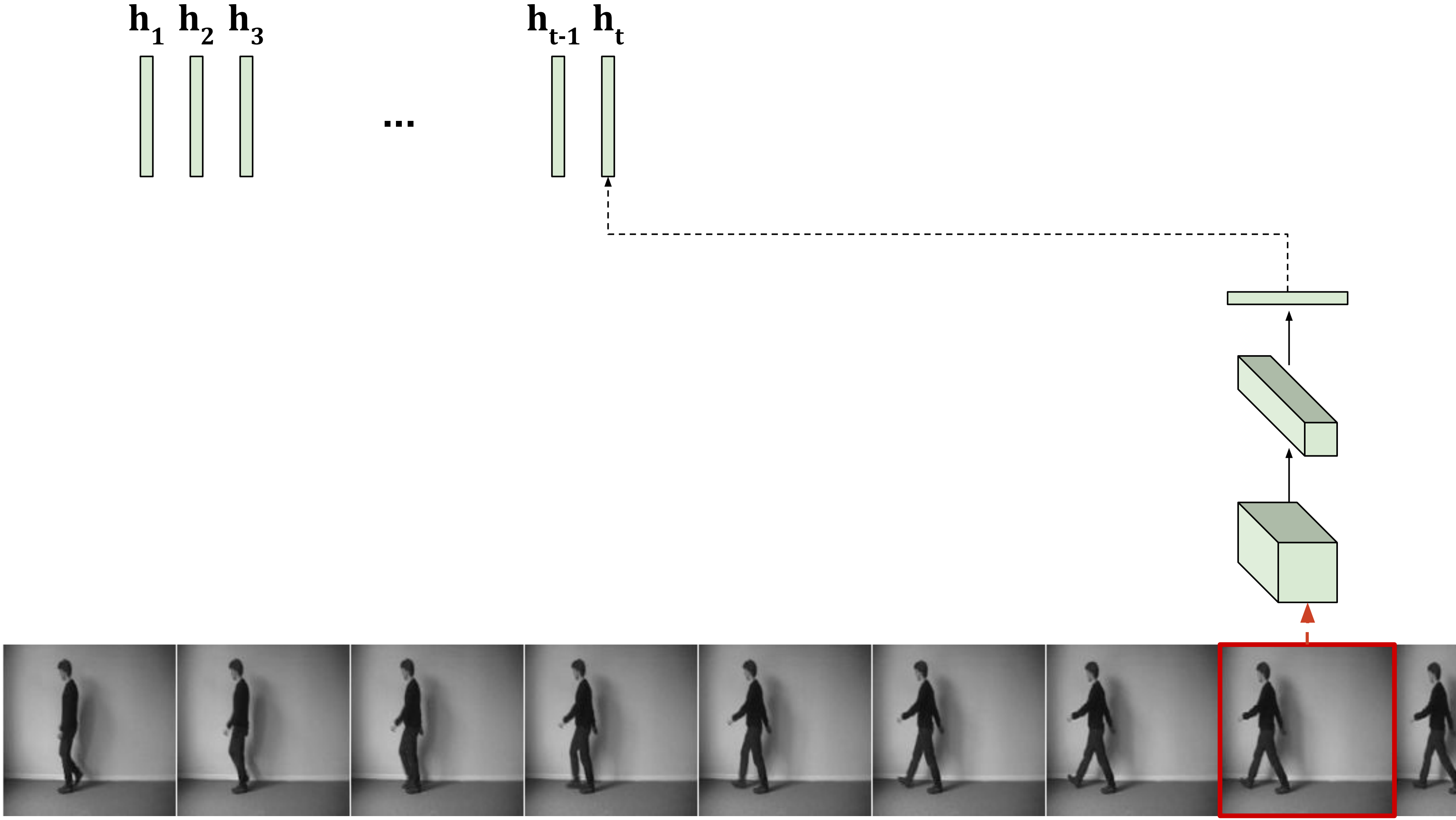
$\mathbf{h}_1$ $\mathbf{h}_2$



h_1 h_2 h_3

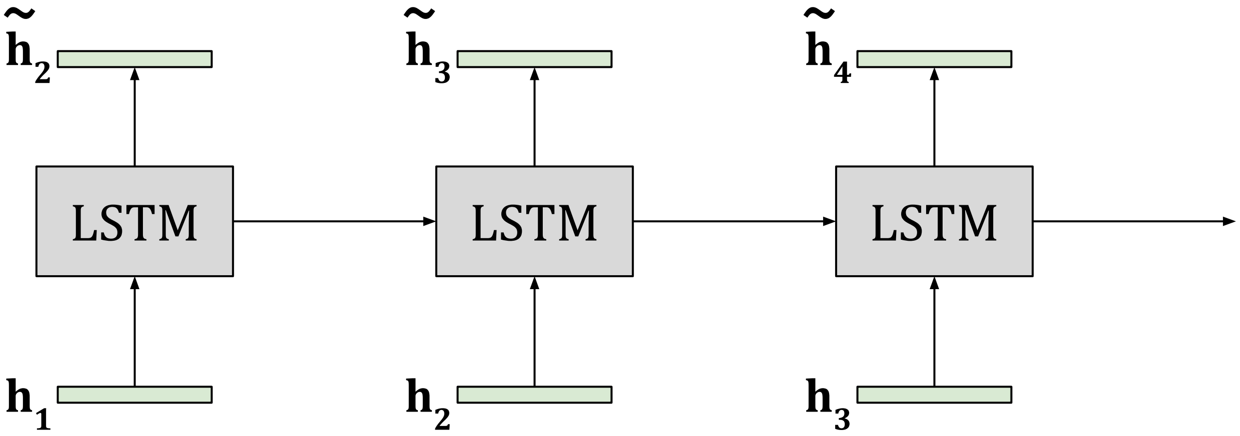


h1 h2 h3
...
ht-1 ht



h_1 h_2 h_3
...
h_{t-1} h_t

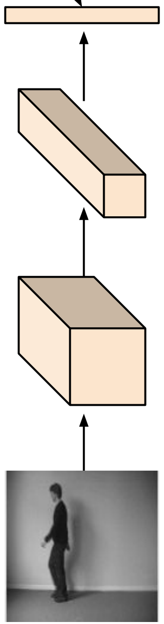
Train LSTM to predict future **pose** vectors



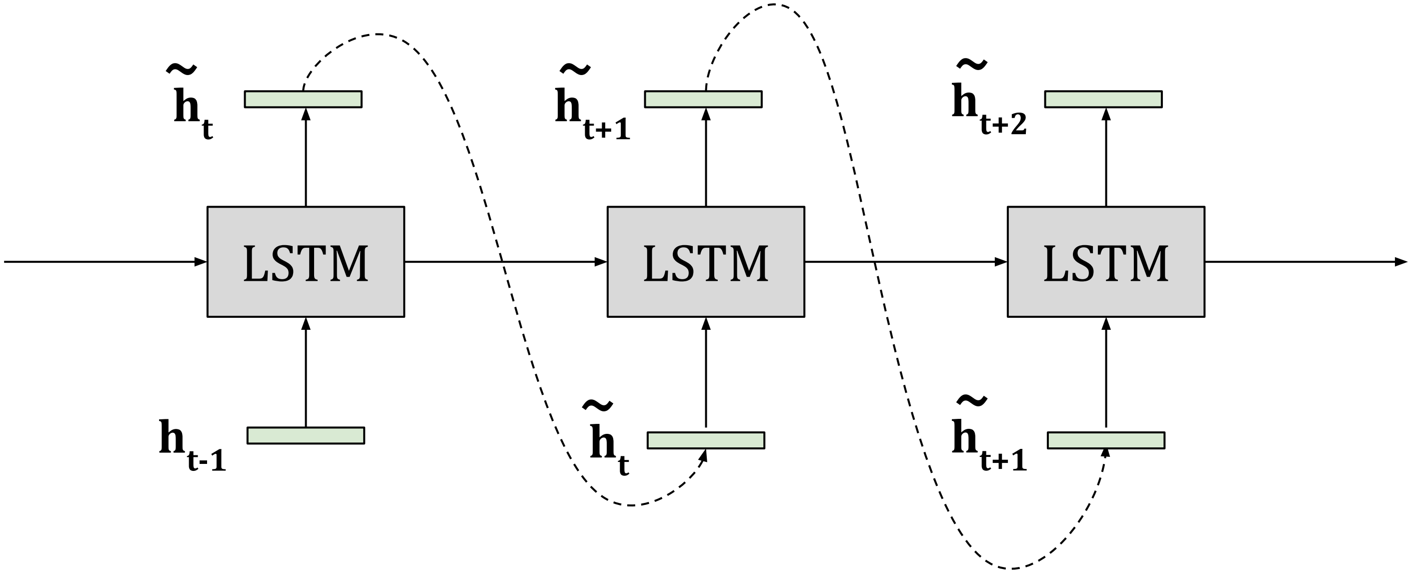
Don’t have to worry about content vectors - they are fixed across time by design

Test time: generating a video sequence

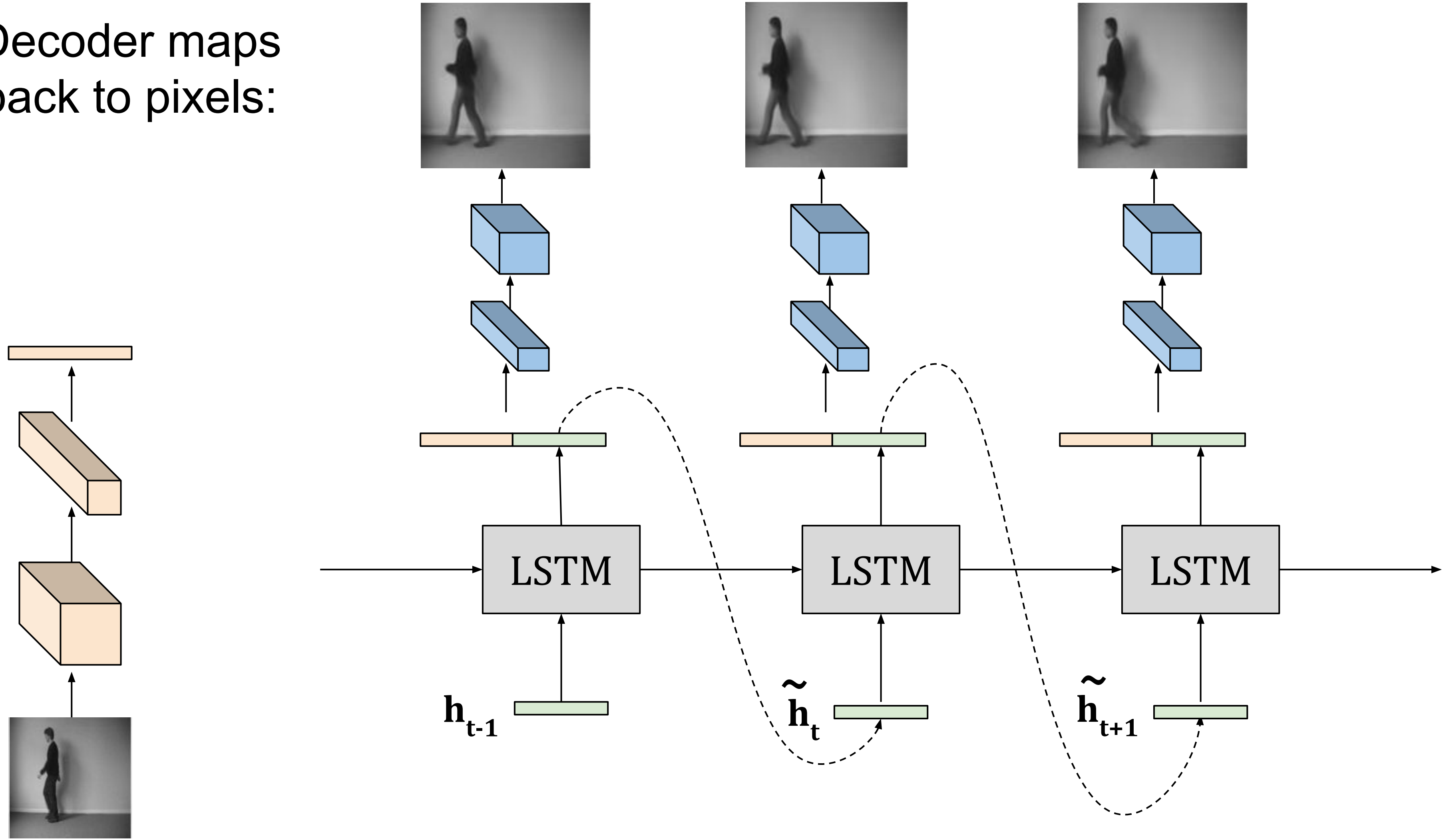
Content vector from any past frame



Feed predicted pose vectors back into model



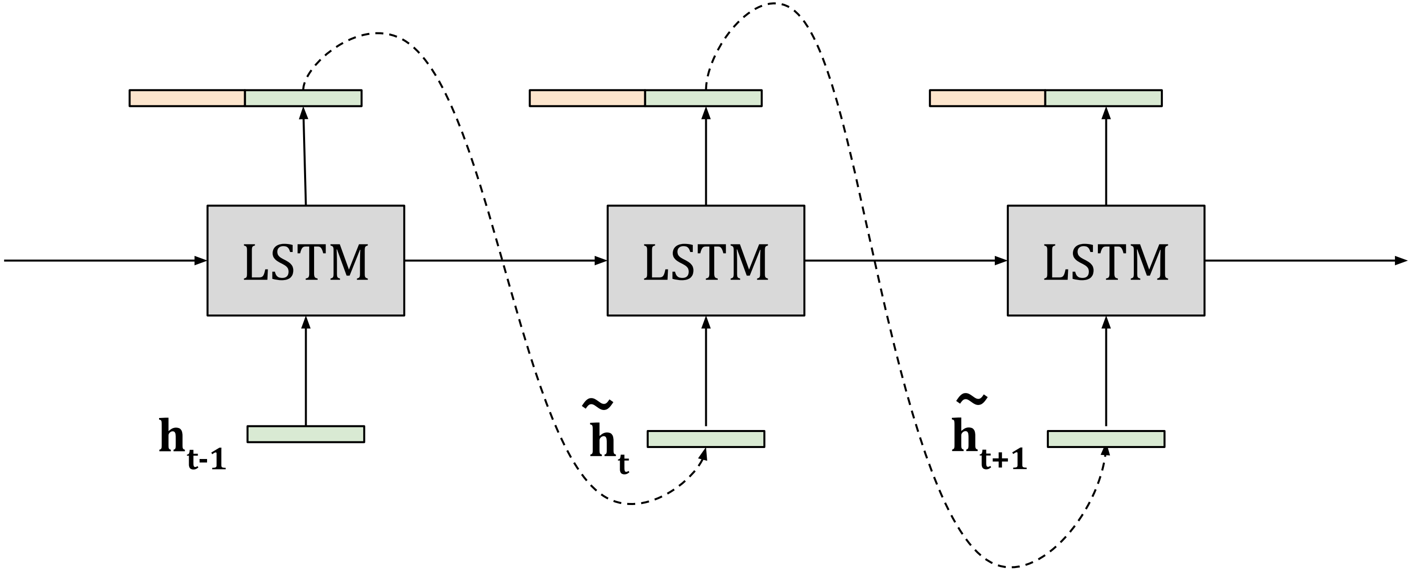
Decoder maps back to pixels:



DrNet video prediction takeaways:

- Prediction done entirely in latent pose space
  - Generated images never fed recursively back into the model

- Small errors in pixel predictions don't propagate through time



Moving MNIST: generating forever...

- Trained model to condition on 5 frames and generate 10 frames into the future

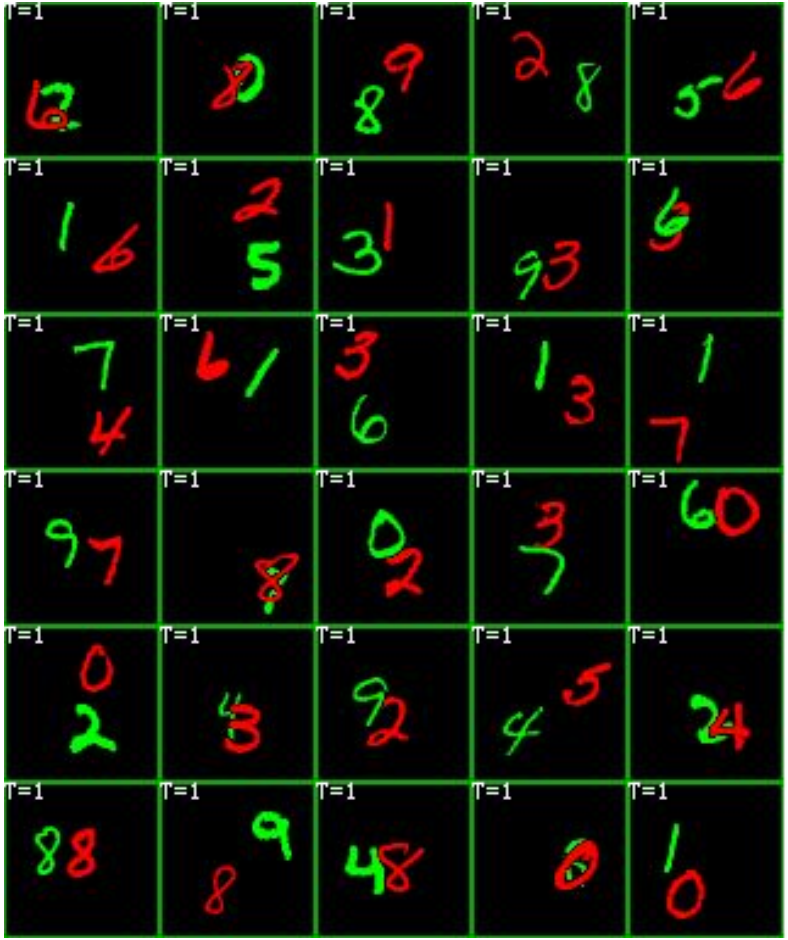
- Can unroll model indefinitely

Green box: Ground truth input (t = 1, ... 5)

Red box: generated frames (t = 6, ..., 500)

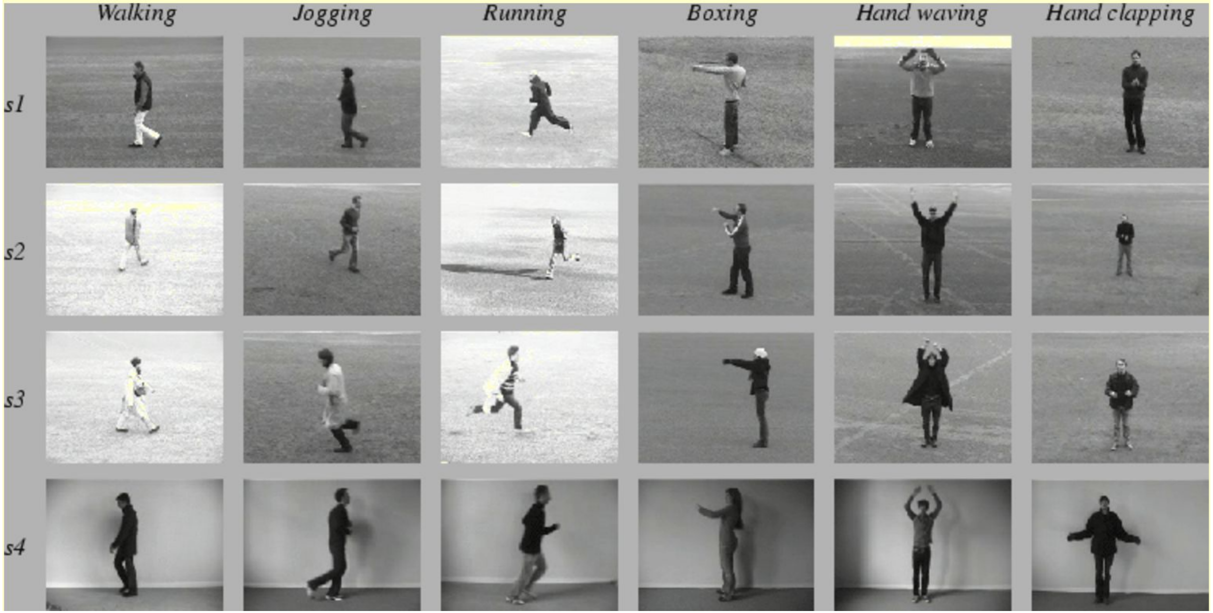
- Content vector fixed across time - helps deal with occlusions

- Digits colored differently so content/pose factorization exists



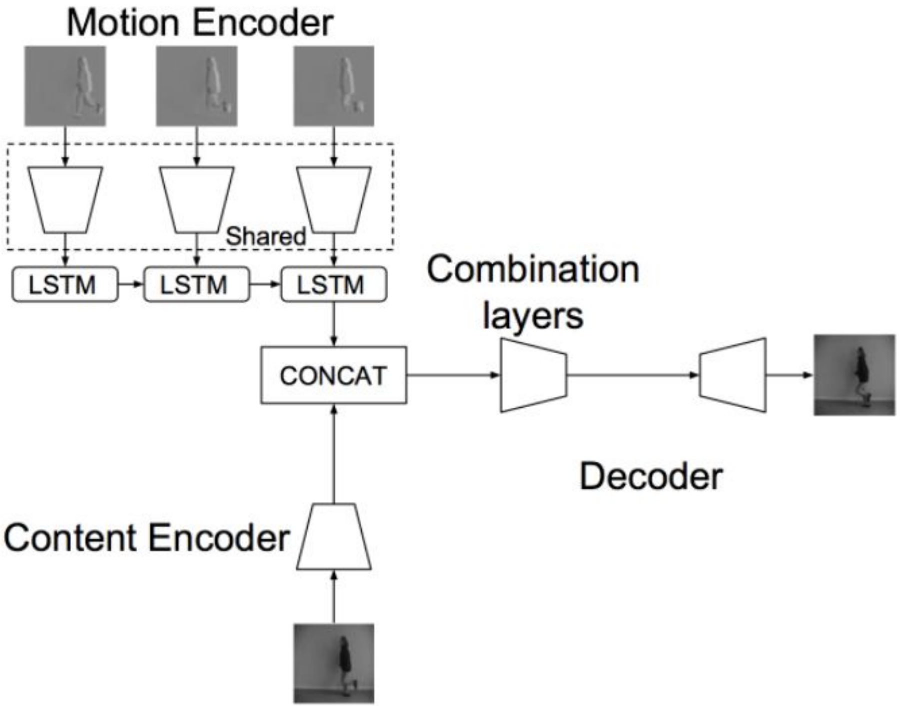
KTH dataset

- Simple dataset of real-world videos
- Six actions
- Fairly uniform backgrounds



C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, volume 3, pages 32–36. IEEE, 2004.

Baseline: MCNet (Villegas et al. 2017)

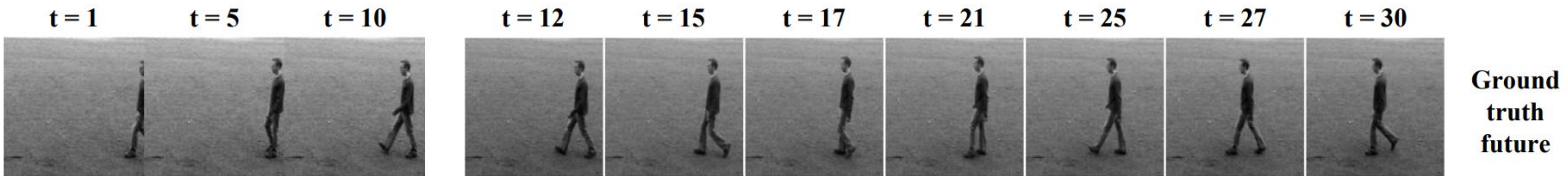


Motion-content net separately models motion and content in video sequences

Trained with combined MSE + GAN loss

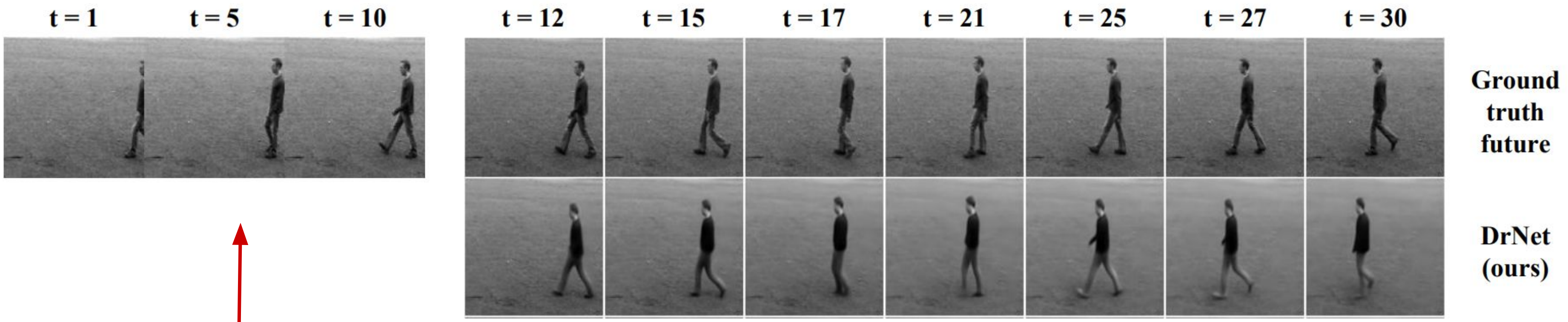
[*Villegas et al. Decomposing motion and content for natural video sequence prediction. In ICLR, 2017.*]

KTH video generation



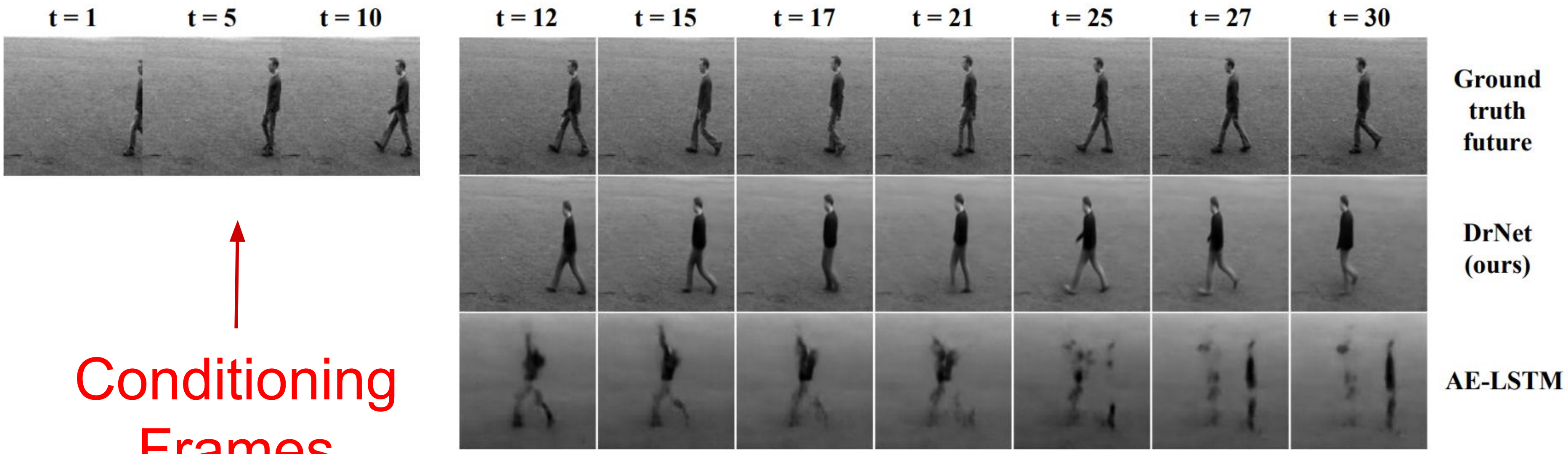
Conditioning Frames

KTH video generation



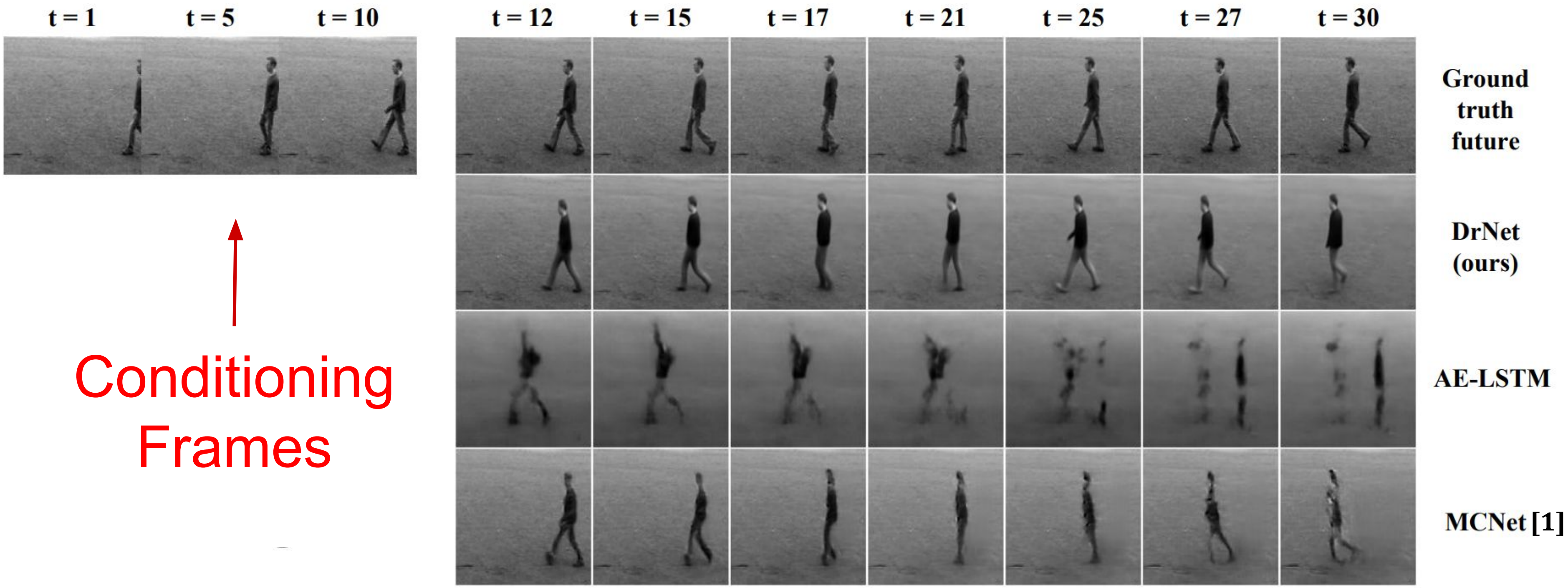
Conditioning Frames

KTH video generation



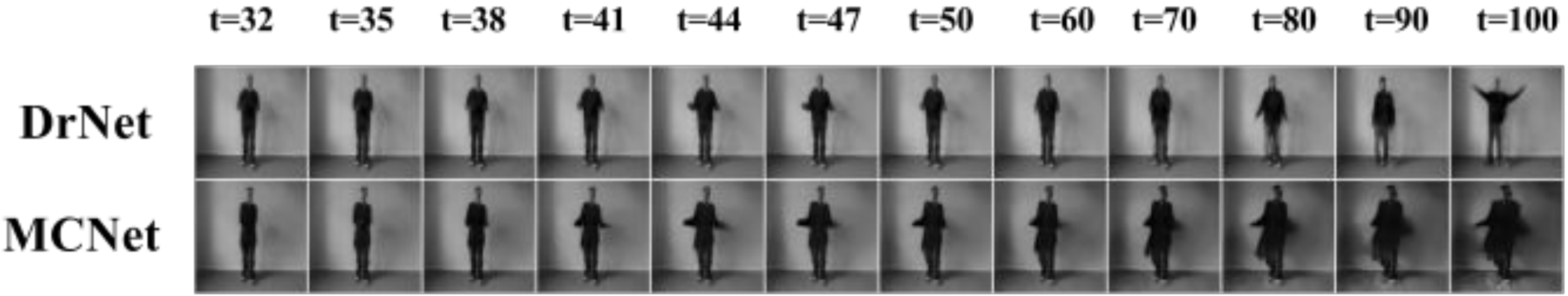
Conditioning Frames

KTH video generation

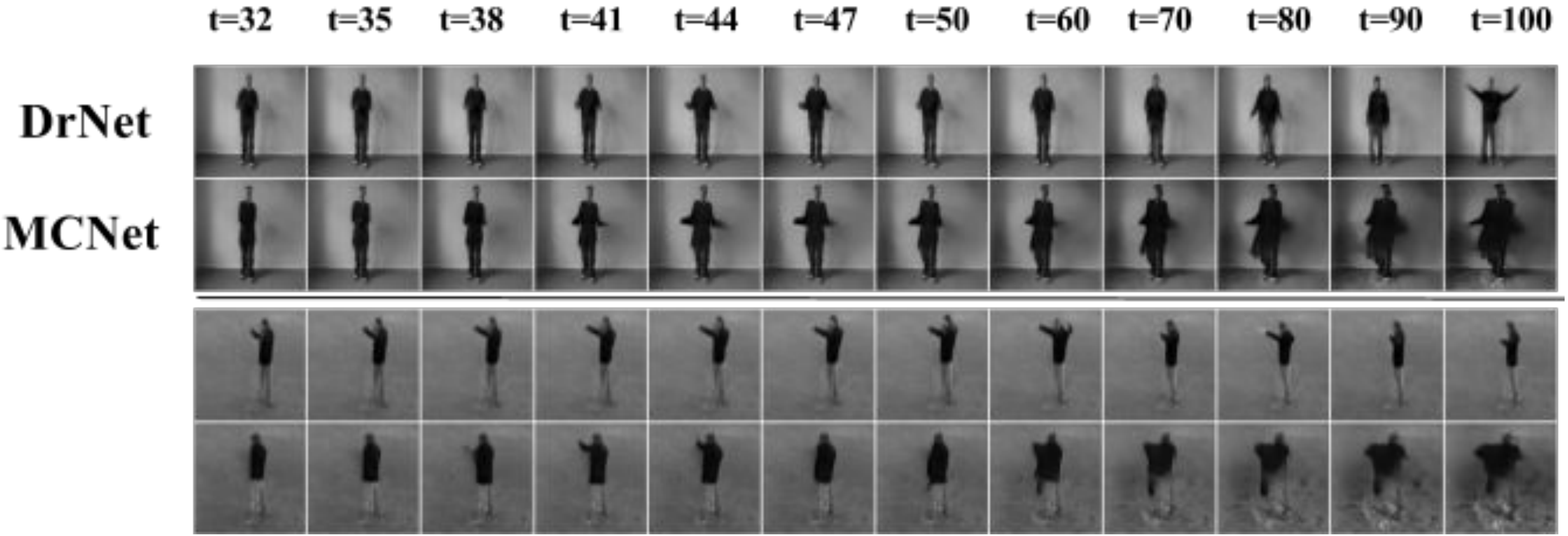


[1] Villegas et al. *Decomposing motion and content for natural video sequence prediction.* In ICLR, 2017.

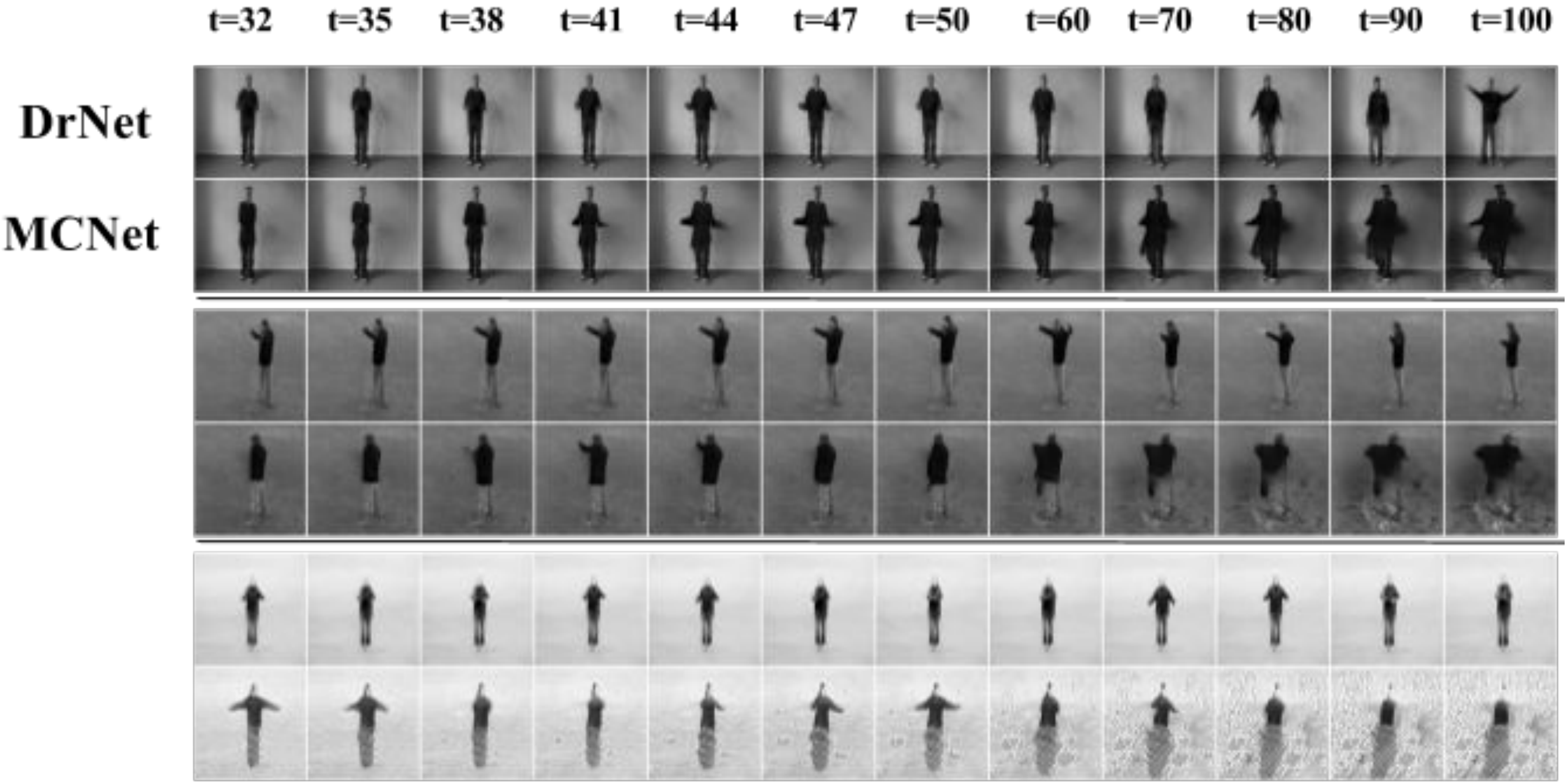
KTH long term video generation



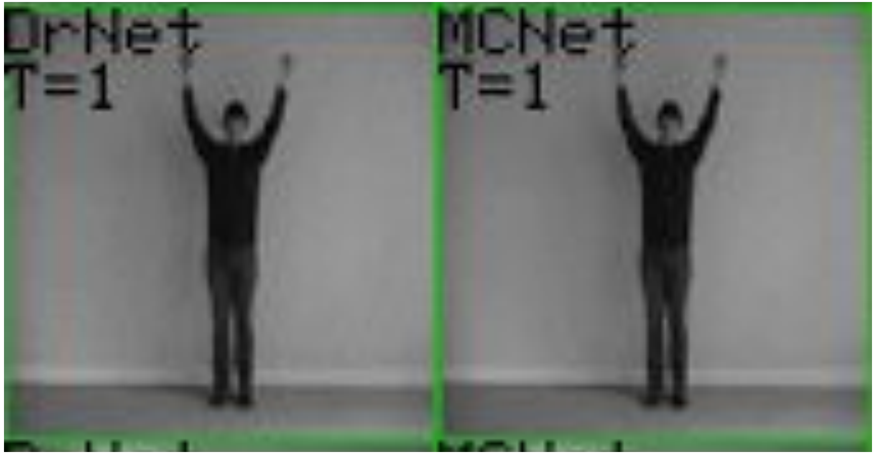
KTH long term video generation

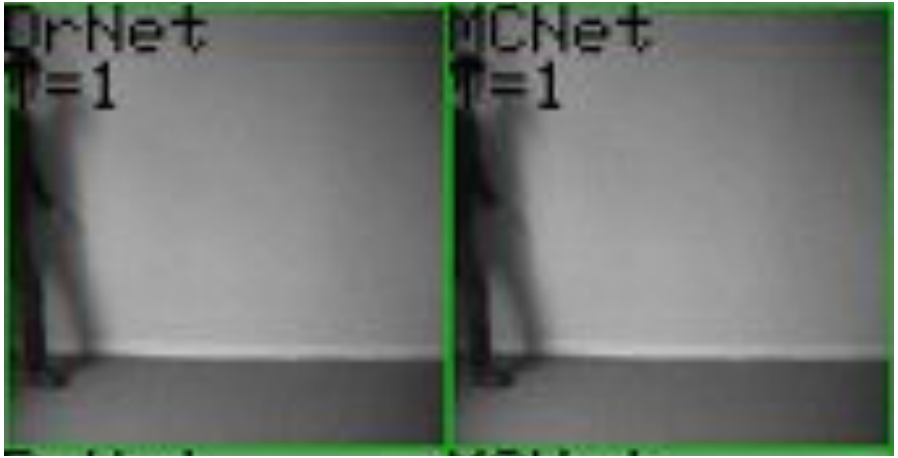


KTH long term video generation

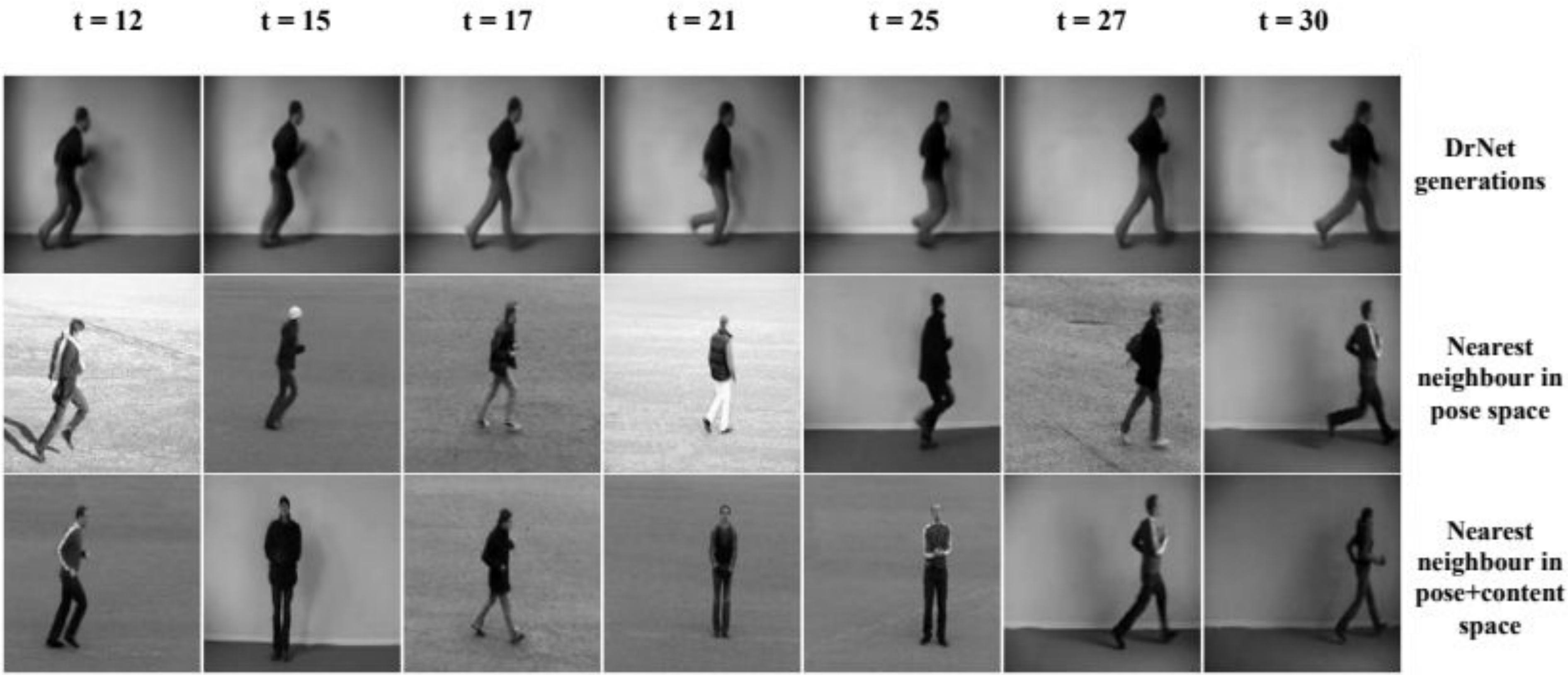


KTH long term video generation

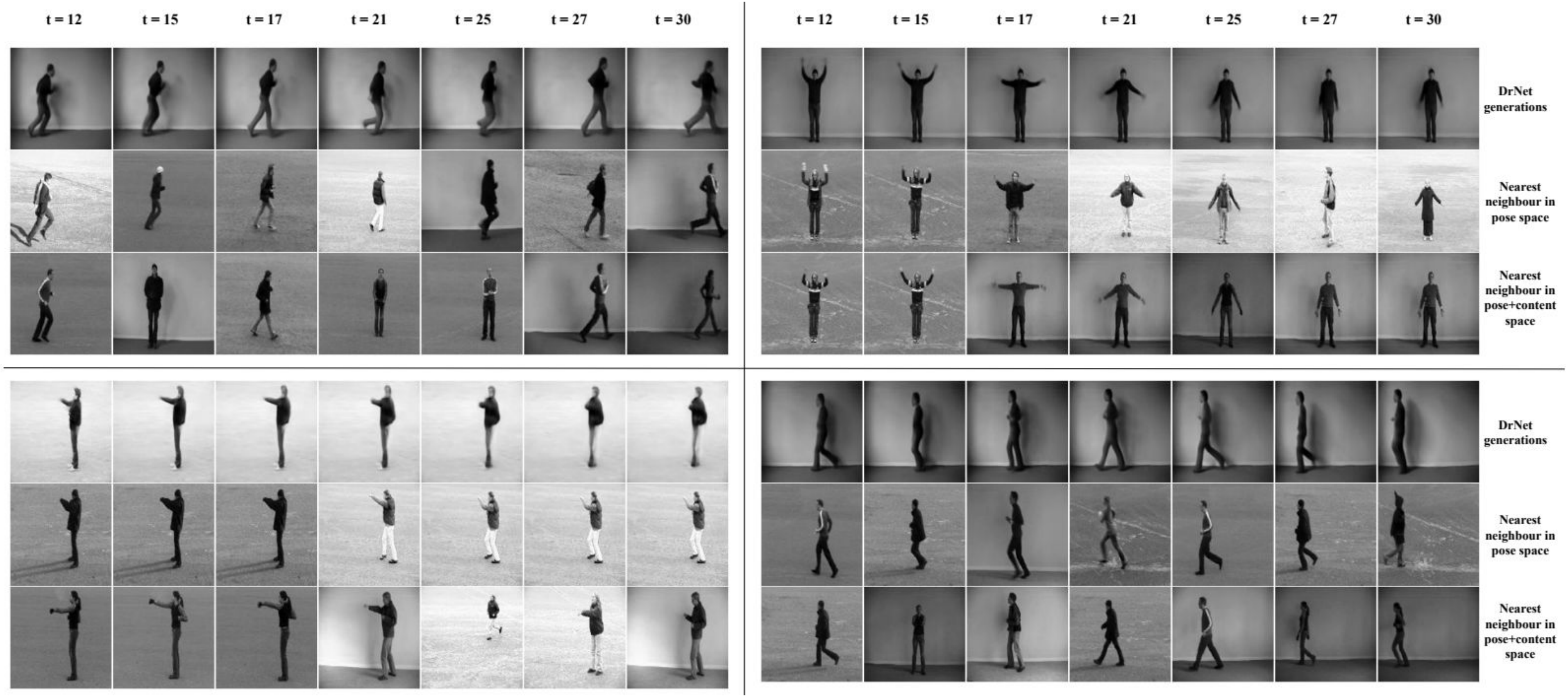




KTH nearest neighbours



KTH nearest neighbours



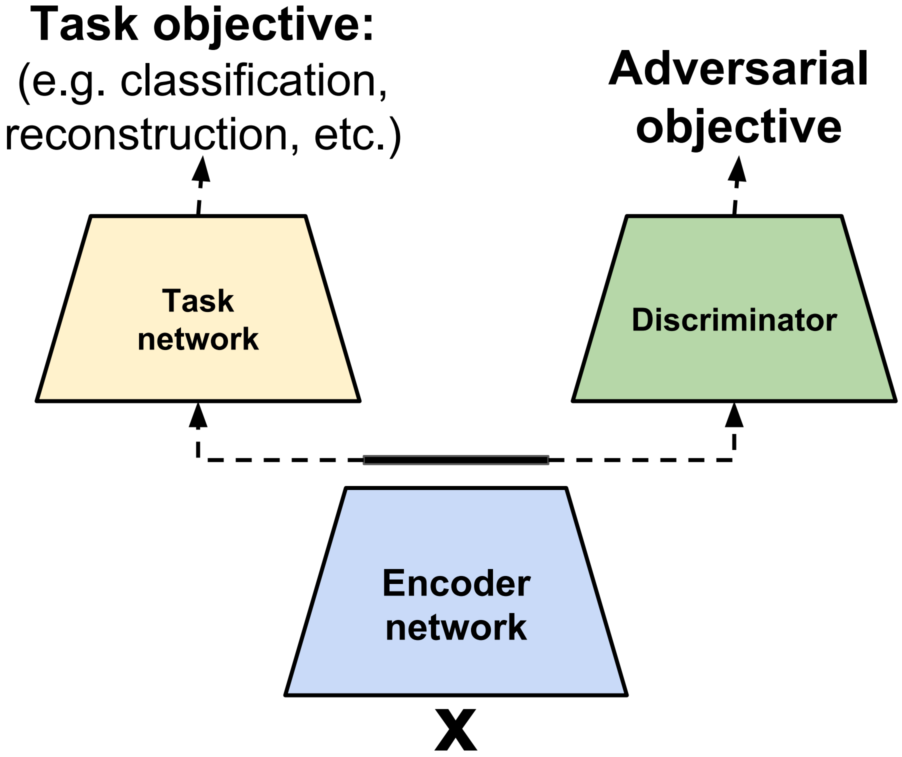
- This adversarial disentangling technique is very general

- Could apply to other datasets where weak labeling is available
  - Only need grouped data - temporal coherence of videos gives us ‘labels’ for free

Part I: Disentangling content and pose with an adversarial loss

Denton and Birodkar. *Unsupervised Learning of Disentangled Representations from Video*. NIPS, 2017

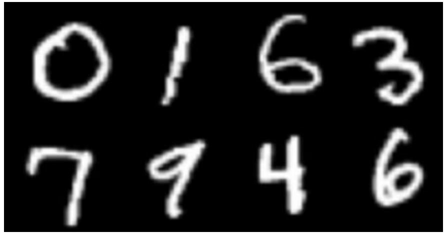
Part II: Survey of adversarial losses in feature space



Domain adaptation

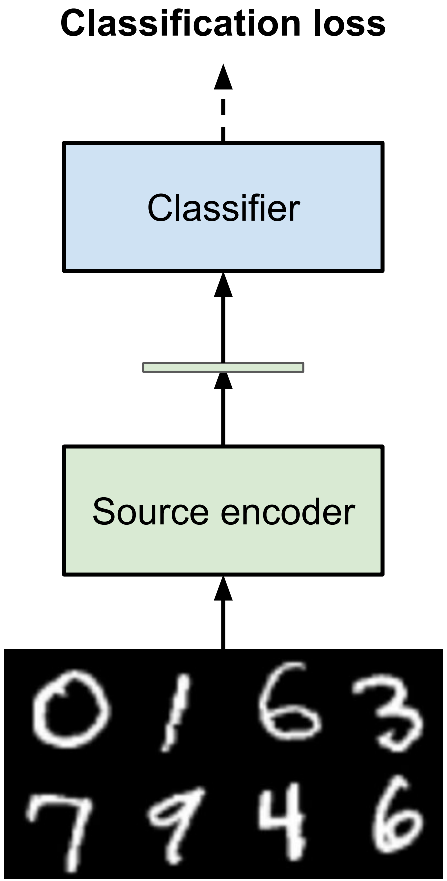
Labelled examples from **source domain**,
few or no labels from **target domain**

Source domain



Target domain

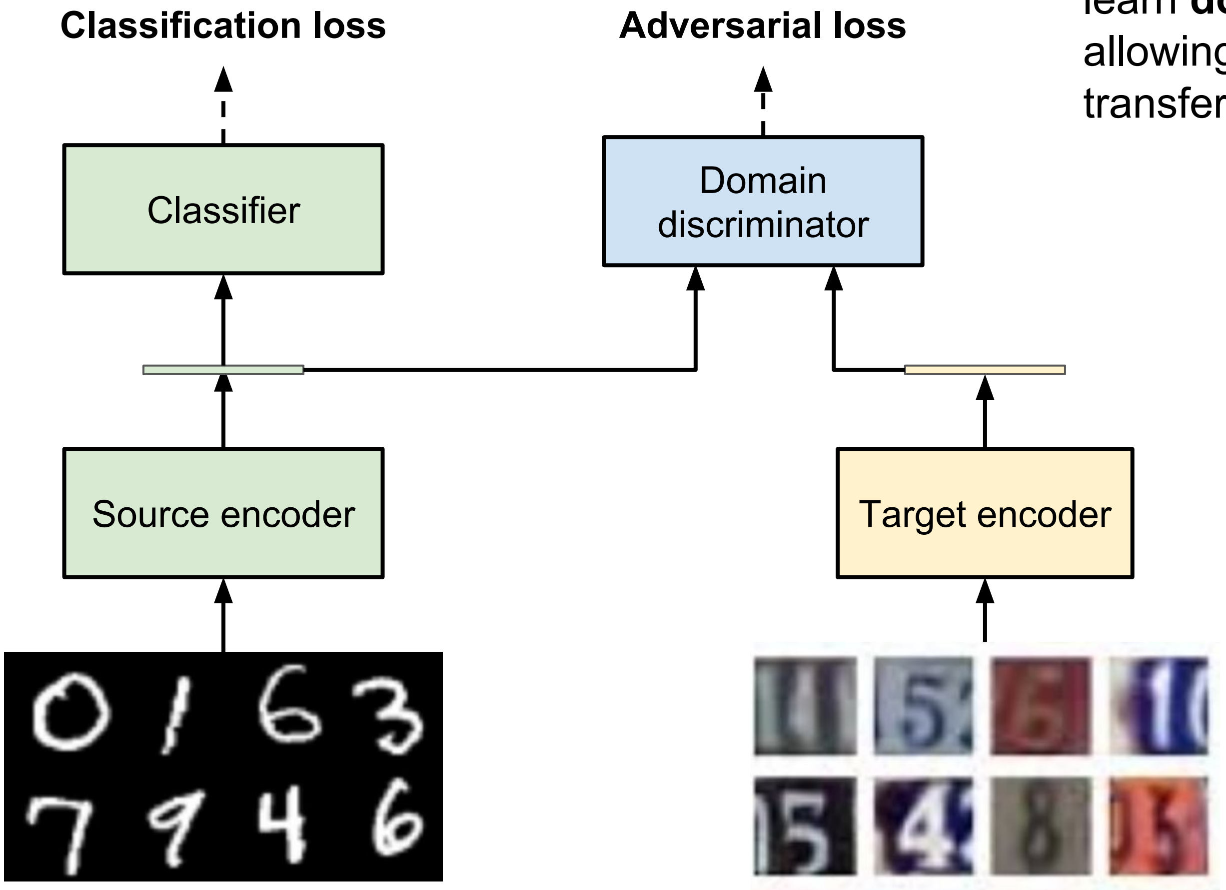
Domain adaptation



Labelled examples from **source domain**,
few or no labels from **target domain**

Target domain

Domain adaptation



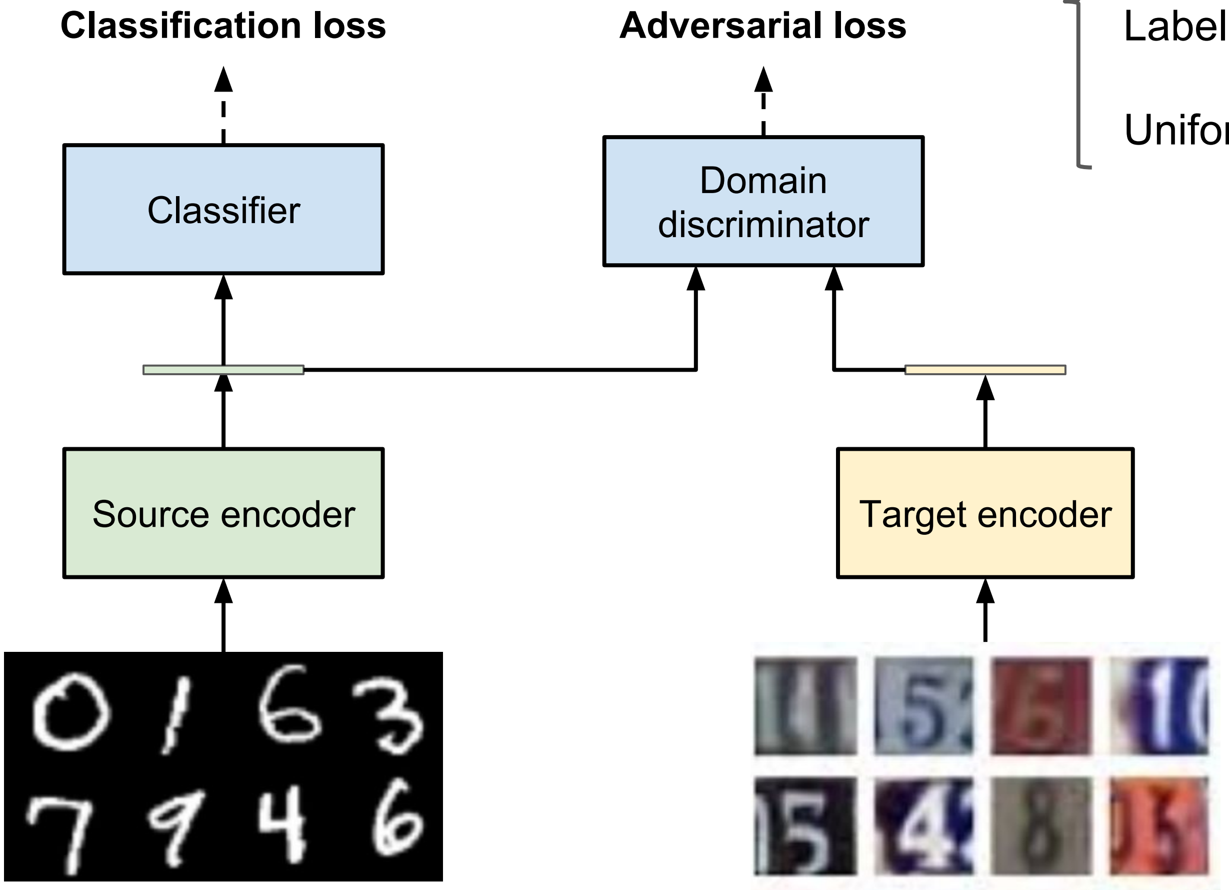
Adversarial loss can be used to learn **domain invariant features,** allowing source classifier to transfer to target domain

Domain adaptation

Gradient reversal [*Ganin and Lempitsky, 2015*]

Label flip [*Tzeng et al. 2017*]

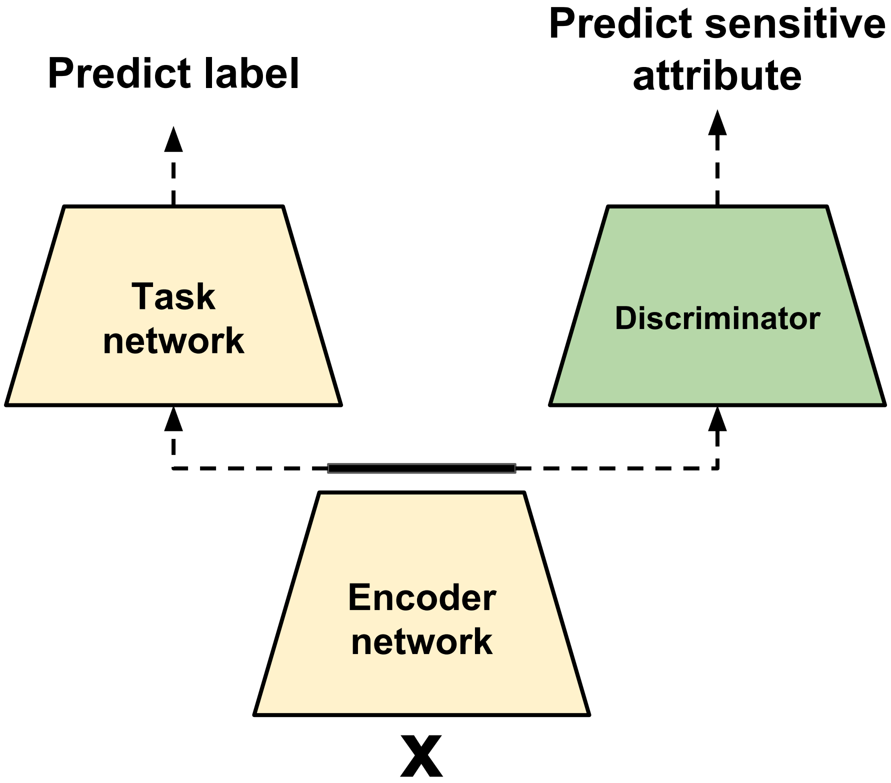
Uniform target [*Tzeng et al. 2015*]



Learning fair representations

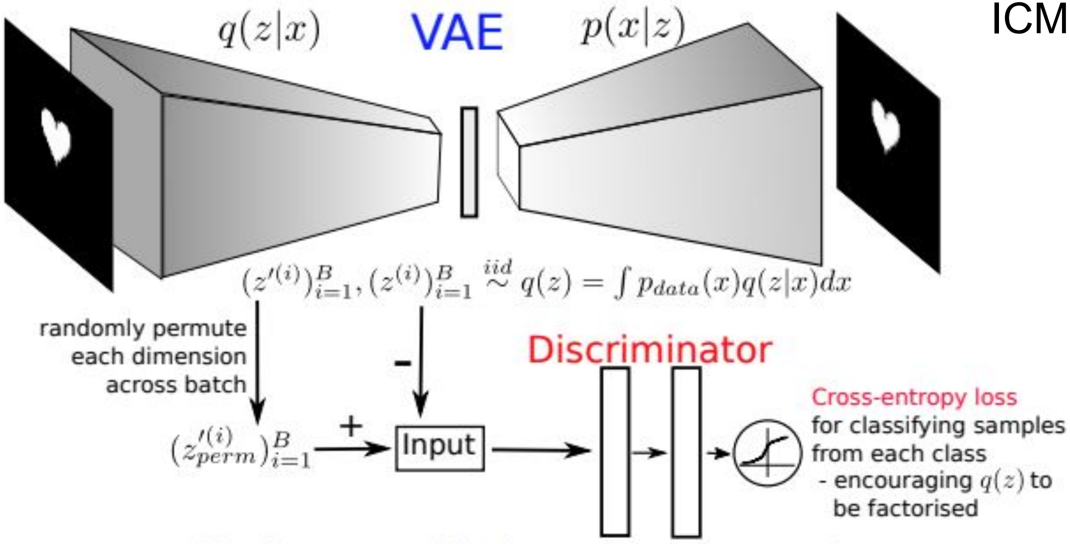
- Closely related to problem of domain adaptation
  - source/transfer domain vs. demographic groups

- Different formulations of adversarial objectives achieve different notions of fairness
  - Edwards & Storkey, 2016
  - Beutel et al. 2017
  - Zhang et al. 2018
  - Madras et al. 2018



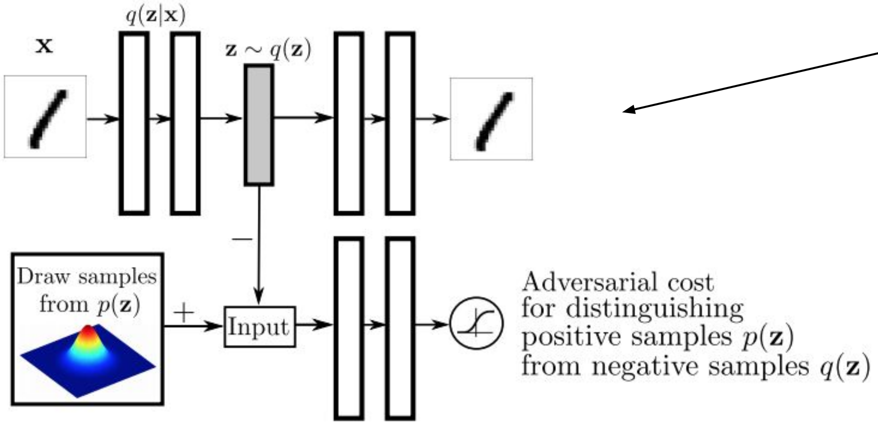
Independent components

Kim and Mnih. Disentangling by Factorising. ICML, 2018



- Discriminate marginal distribution vs. product of marginals: $q(z_1, ..., z_n)$ vs. $\prod q(z_i)$
- Earlier work on discrete code setting by Schmidhuber (1992)

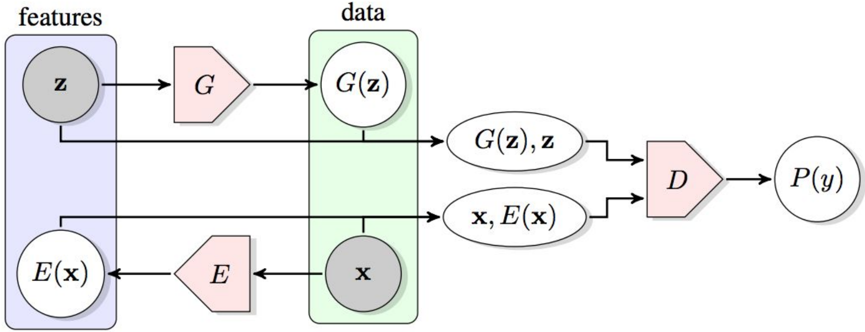
Prior distributions of generative models



Adversarial autoencoders:
Match aggregate approx posterior q(z)
[Makhzani et al. 2016]

Adversarial variational bayes:
Match approx posterior q(z|x)
[Mescheder et al. 2017]

Adversarial feature learning:
GAN loss in image space and latent space
[Dumoulin et al. 2017; Donahue et al. 2017]



References

Beutel et al. *Data decisions and theoretical implications when adversarially learning fair representations*. arXiv:1707.00075, 2017.

Denton and Birodkar. *Unsupervised Learning of Disentangled Representations from Video*. NIPS, 2017.

Donahue et al. *Adversarial Feature Learning*. ICLR, 2017.

Dumoulin et al. *Adversarially Learned Inference.* ICLR, 2017

Edwards & Storkey. *Censoring Representations with an Adversary*. ICLR, 2016.

Ganin and Lempitsky. *Unsupervised domain adaptation by backpropagation.* ICML, 2015.

Kim and Mnih. *Disentangling by Factorising*. ICML, 2018.

Madras et al. *Learning Adversarially Fair and Transferable Representations*. ICML, 2018.

Makhzani et al. *Adversarial Autoencoders*. ICLR Workshop, 2016.

Mescheder et al. *Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks*. ICML, 2017.

Schmidhuber. *Learning factorial codes by predictability minimization.* Neural Computation, 1992.

Tzeng et al. S*imultaneous deep transfer across domains and tasks.* ICCV, 2015.

Tzeng et al. *Adversarial discriminative domain adaptation.* CVPR, 2017.

Villegas, et al. D*ecomposing motion and content for natural video sequence prediction.* In ICLR, 2017.

Zhang et al. M*itigating Unwanted Biases with Adversarial Learning*. AIES, 2018.

Thanks!