Learning Object Representations by Mixing Scenes

Master Thesis Presentation

Lukas Zbinden

May 23rd, 2019

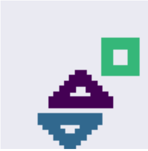
Supervisor: Prof. Dr. Paolo Favaro
Computer Vision Group, Institute of Computer Science
University of Bern

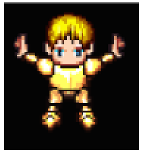
Agenda

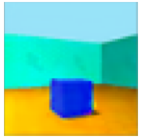
- Research Question
- Our approach: Learning Object Representations by Mixing Scenes (LORBMS)
- Prior Work
- Model and Architecture
- Experimental Results
- Conclusions and Future Work

Motivation

Datasets used by previous works:

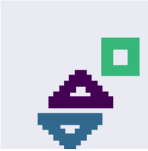


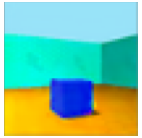




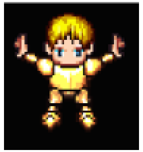
Motivation

Datasets used by previous works:


Shapes


Object in room


MNIST


Sprites


TFD


Multi-PIE


CelebA


dSprites


simple_superpos


Shapenet


3D chairs


CLEVR


3d faces

Motivation

Can we learn directly from natural image data?

Potential: unsupervised learning on Internet-scale data (i.e. billions of images)

Motivation

Can we learn directly from natural image data?

Potential: unsupervised learning on internet-scale data (i.e. billions of images)

MS COCO

http://cocodataset.org/#explore

Motivation

Our thesis: learn directly from natural image data

→ devise an unsupervised representation learning method
→ learn object representations by mixing everyday scenes

Motivation

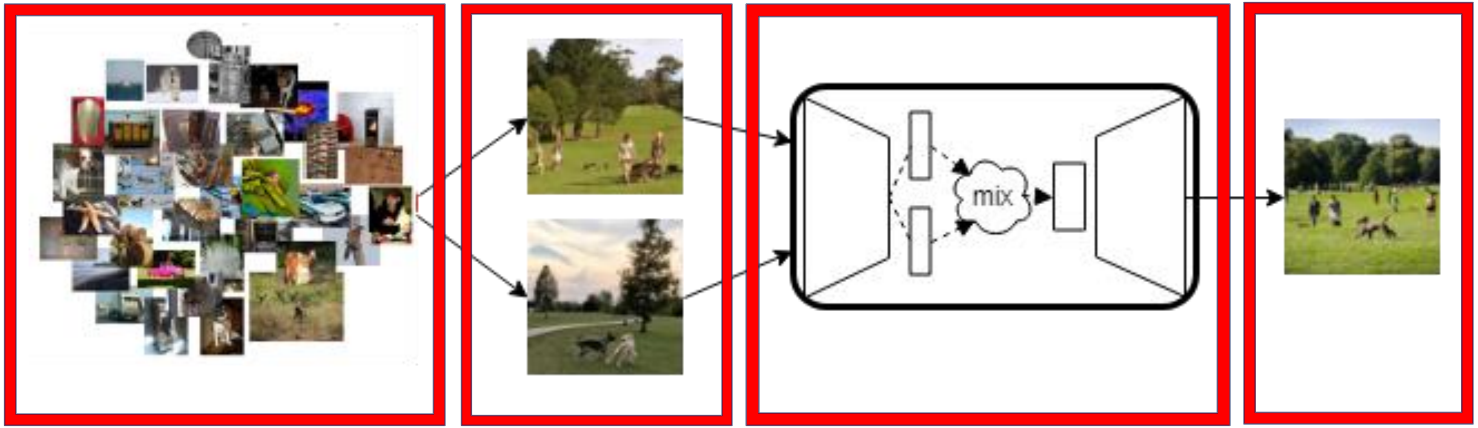
The proposed LORBMS system

1. natural dataset

2. pick similar images

3. LORBMS: mix images, generate new

4. new mixed scene



LORBMS Concept

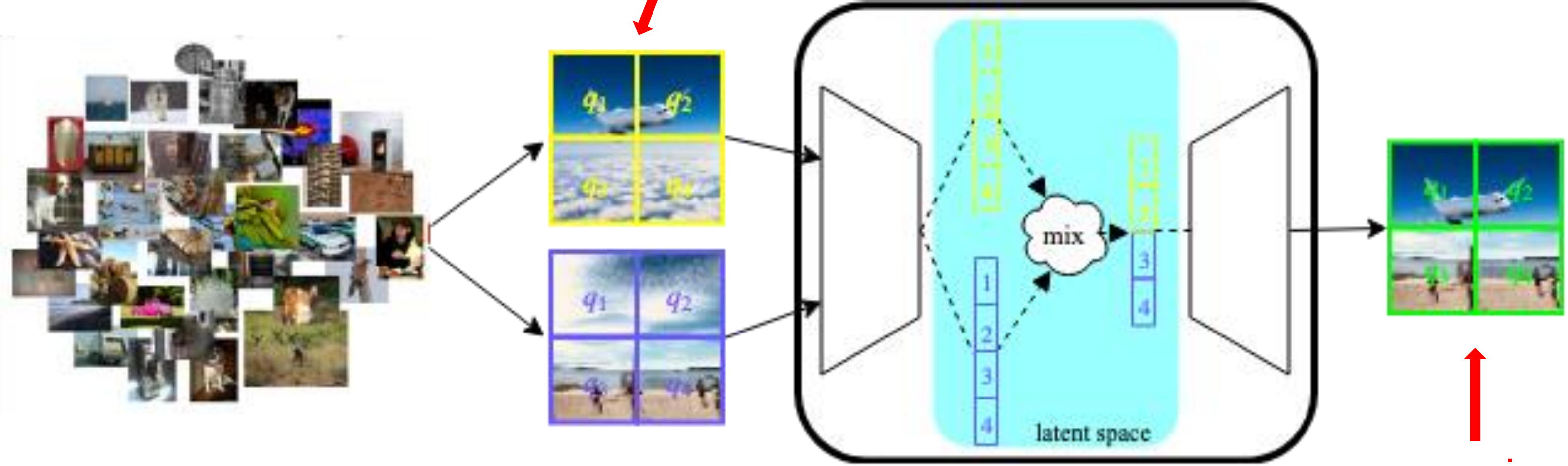
The 4 object assumption

LORBMS Concept

Concept

4 object assumption

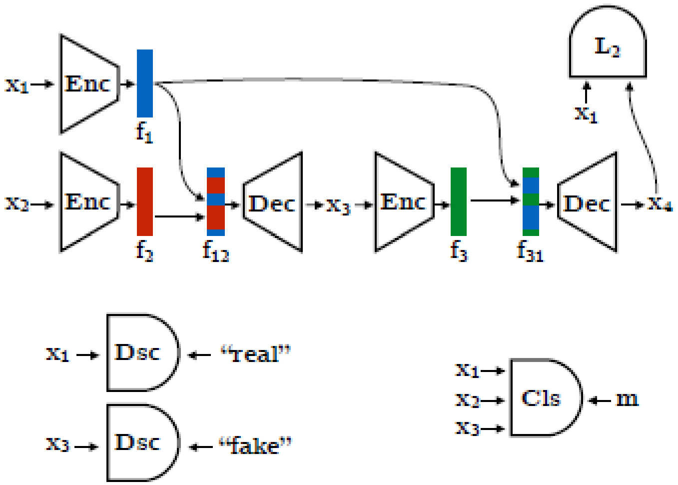
LORBMS: mix images, generate new

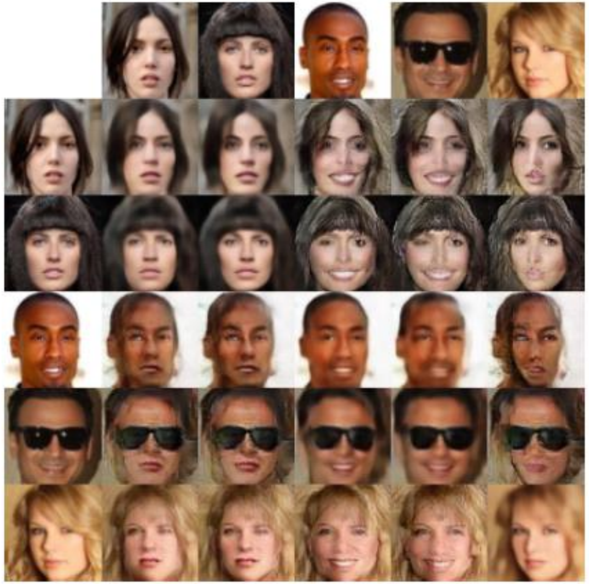


new mixed scene

Prior Work

Disentangling Factors of Variation by Mixing Them, Hu *et al.*



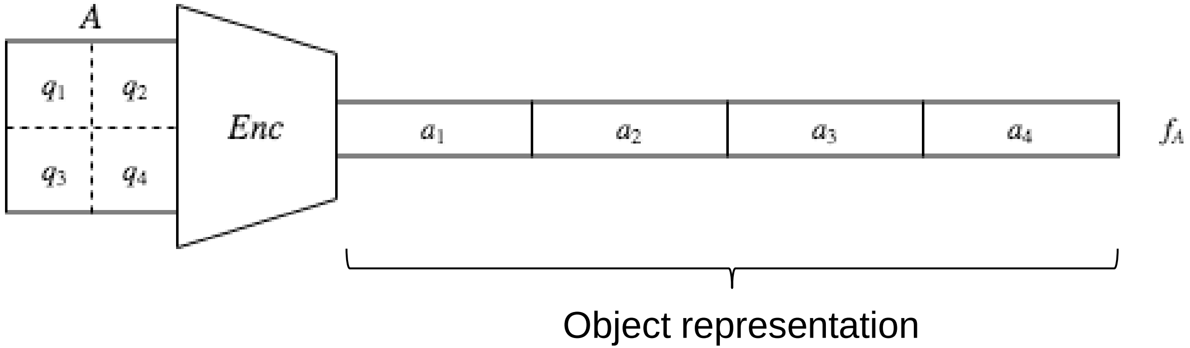


Pose/smile

Idea: leverage Hu's method and apply to natural data

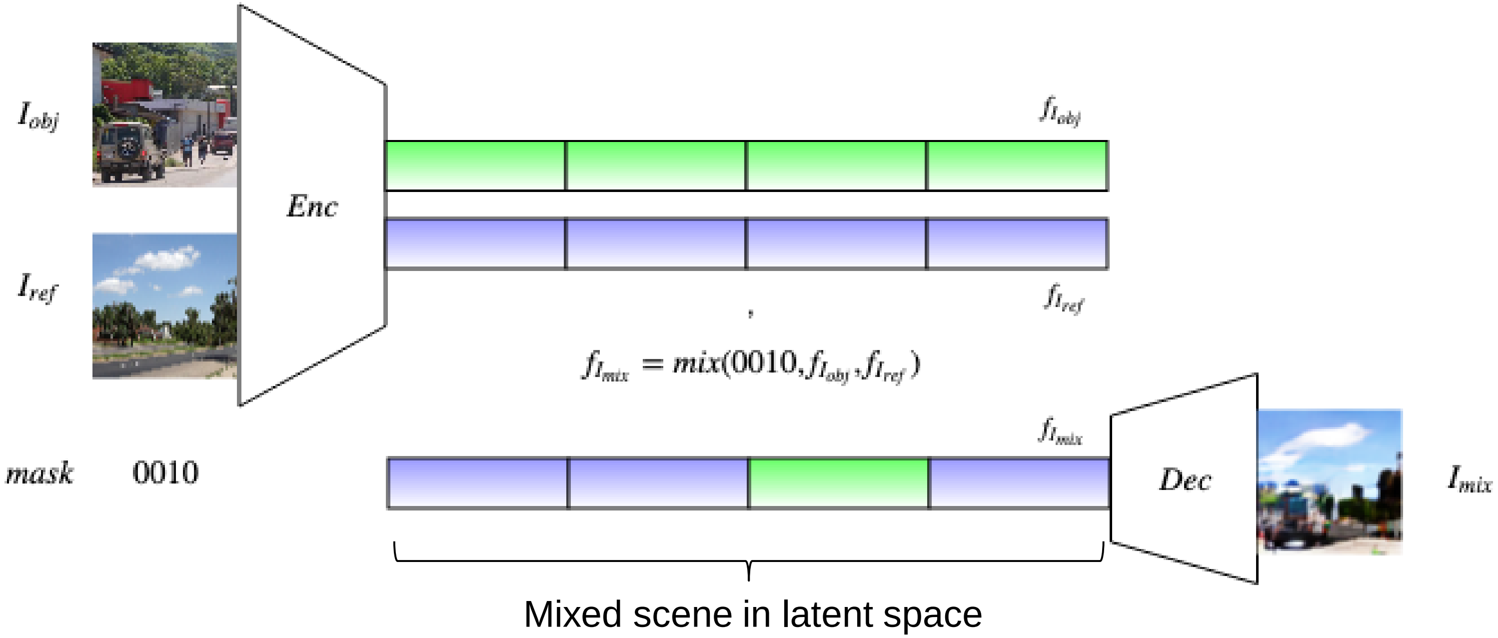
LORBMS Model Architecture

- Encoder

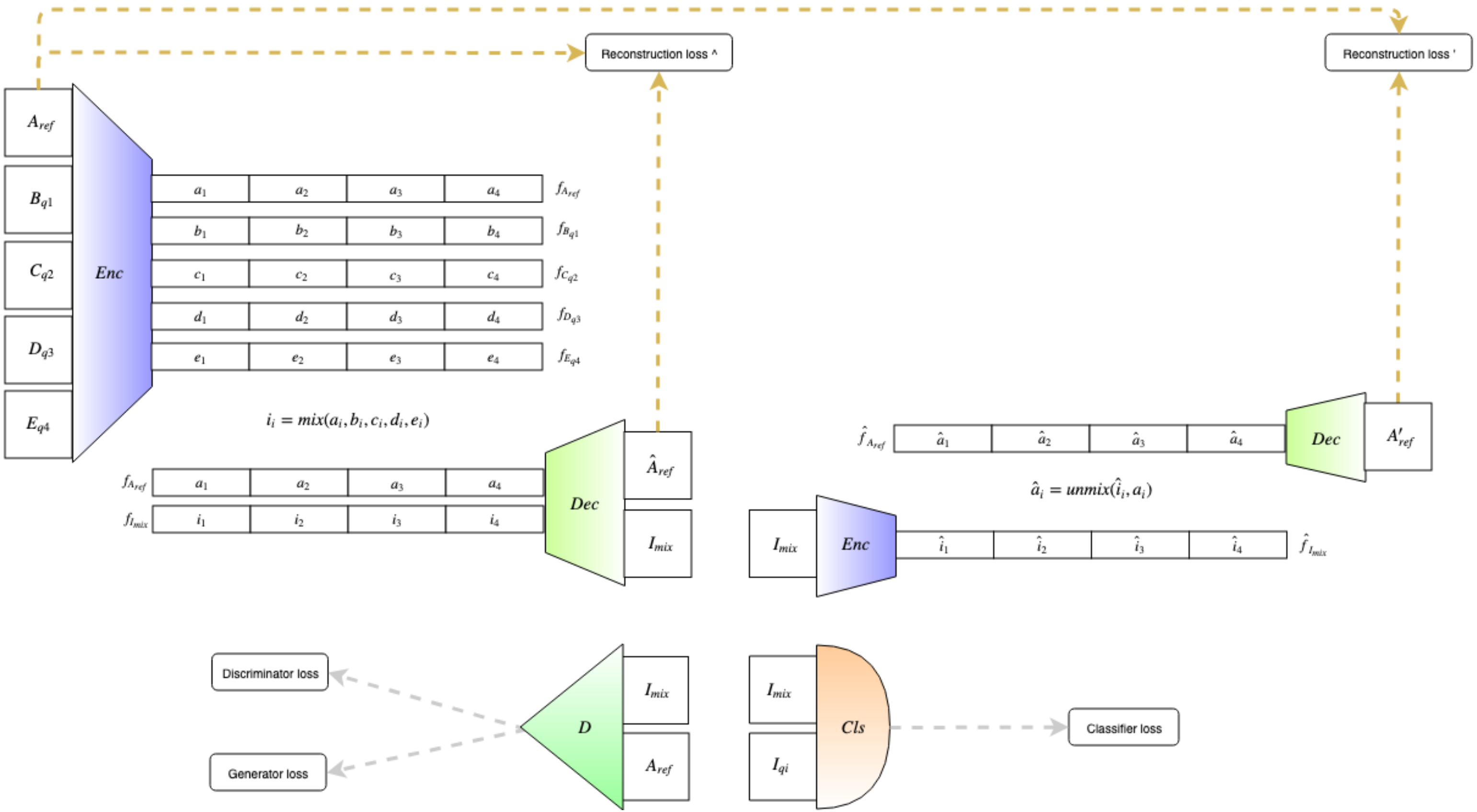


LORBMS Model Architecture

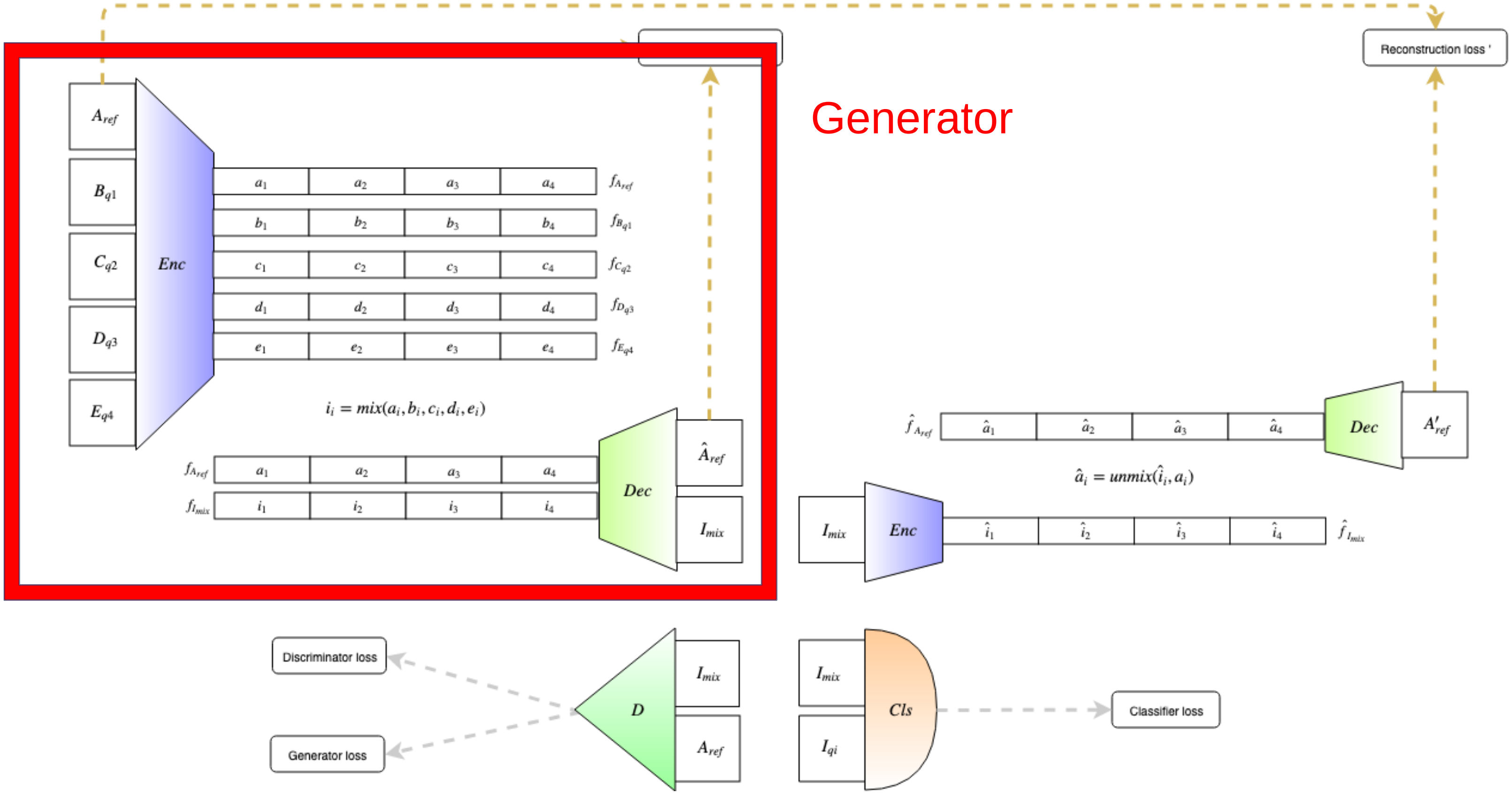
- Encoder + Decoder = Generator



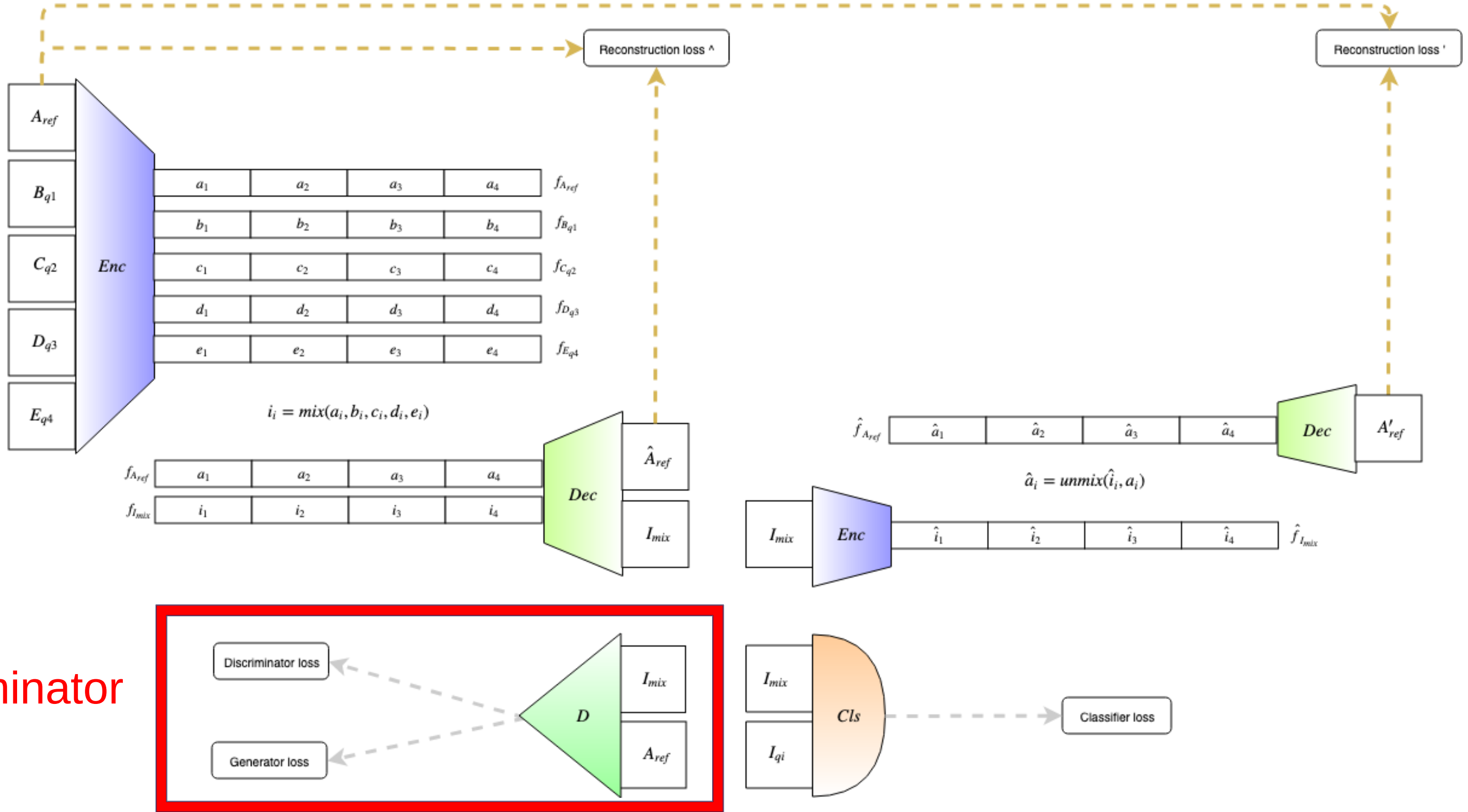
LORBMS Model Architecture



LORBMS Model Architecture

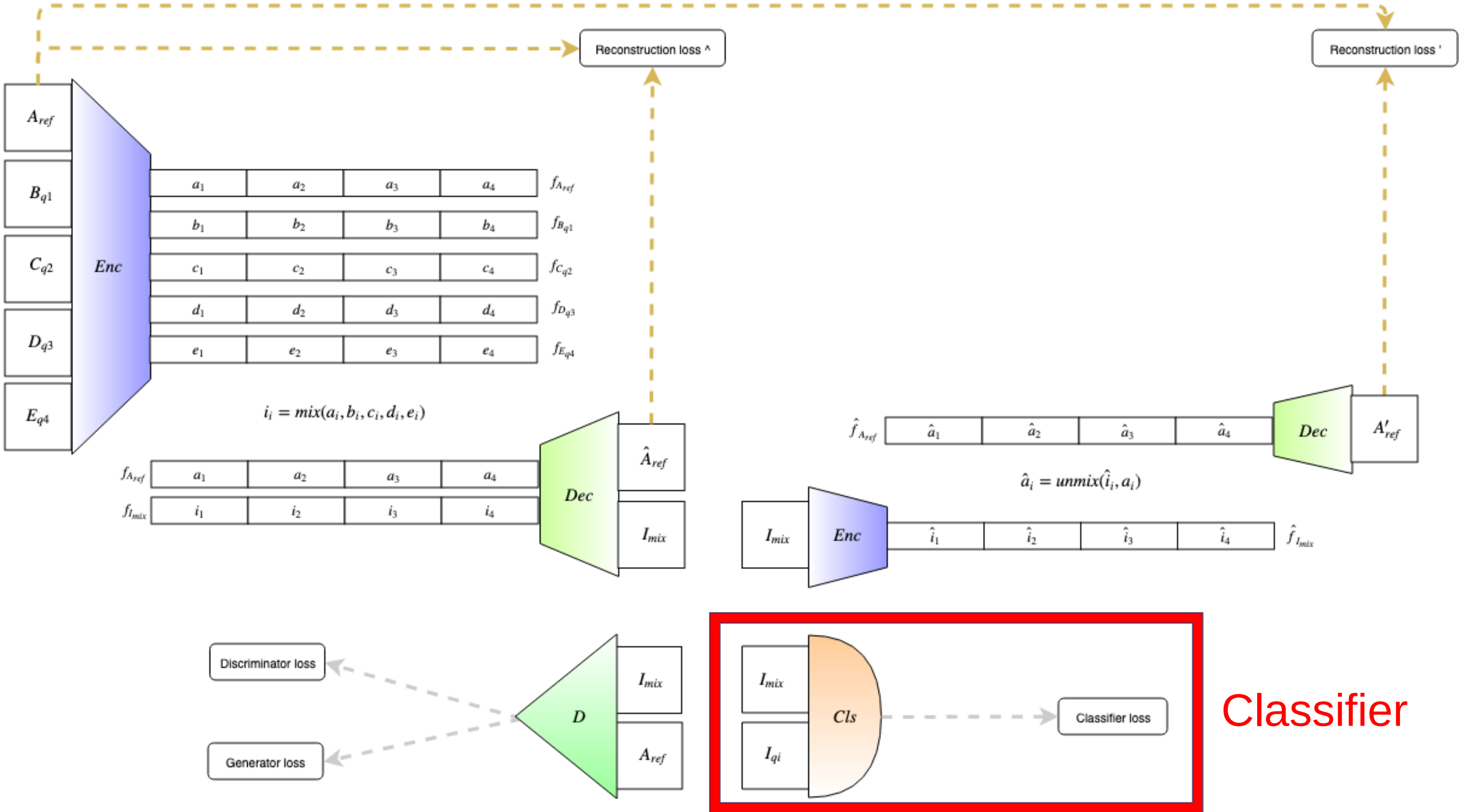


LORBMS Model Architecture



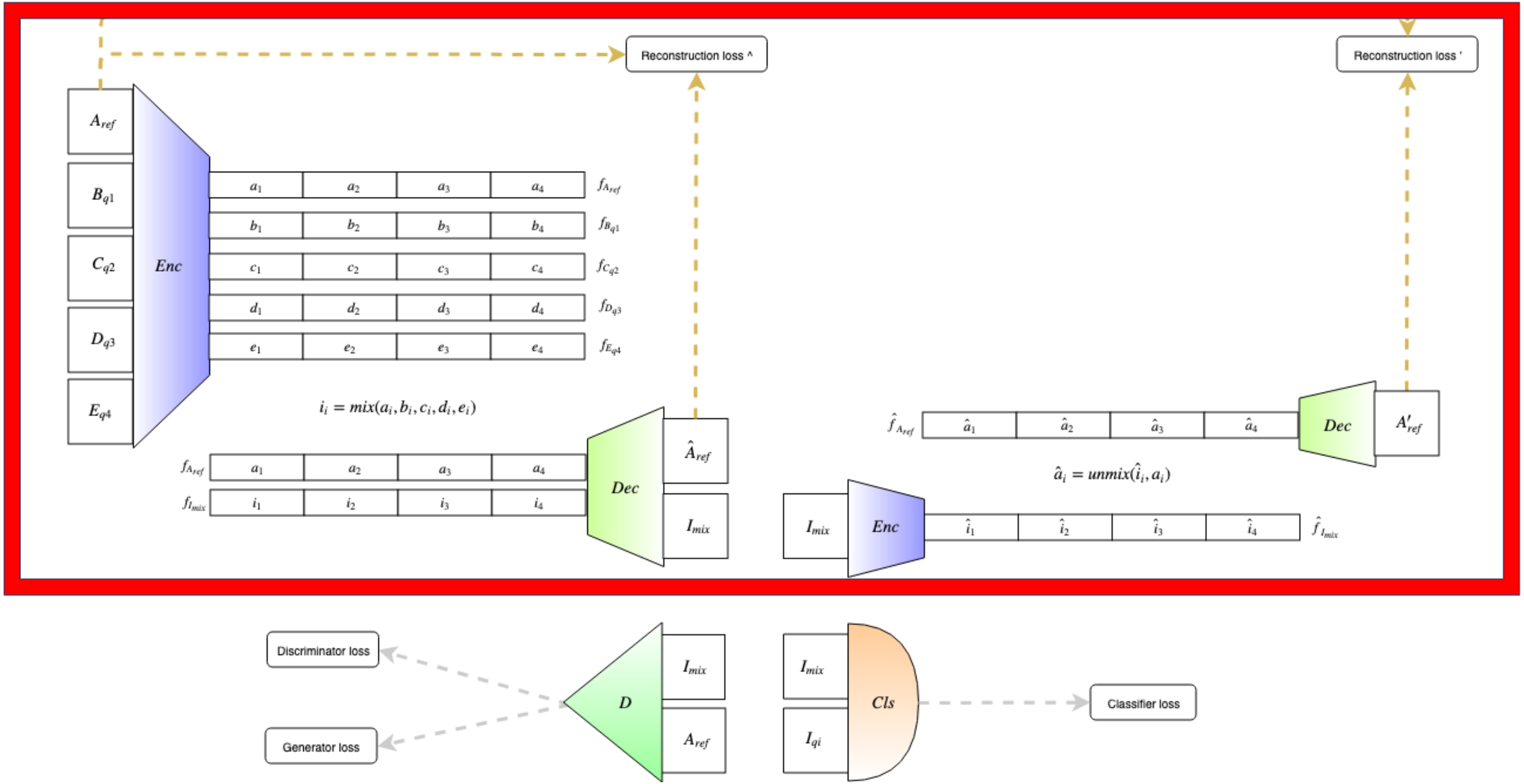
Discriminator

LORBMS Model Architecture



Classifier

LORBMS Model Architecture



LORBMS Training: GAN & Loss Functions

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{I_{ref} \sim p_{data}}\left[log\ D(I_{ref})\right] +$$

$$\mathbb{E}_{I_j \sim p_{data}}\left[log(1 - D(G(I_{ref}, I_{q1}, I_{q2}, I_{q3}, I_{q4})))\right]$$

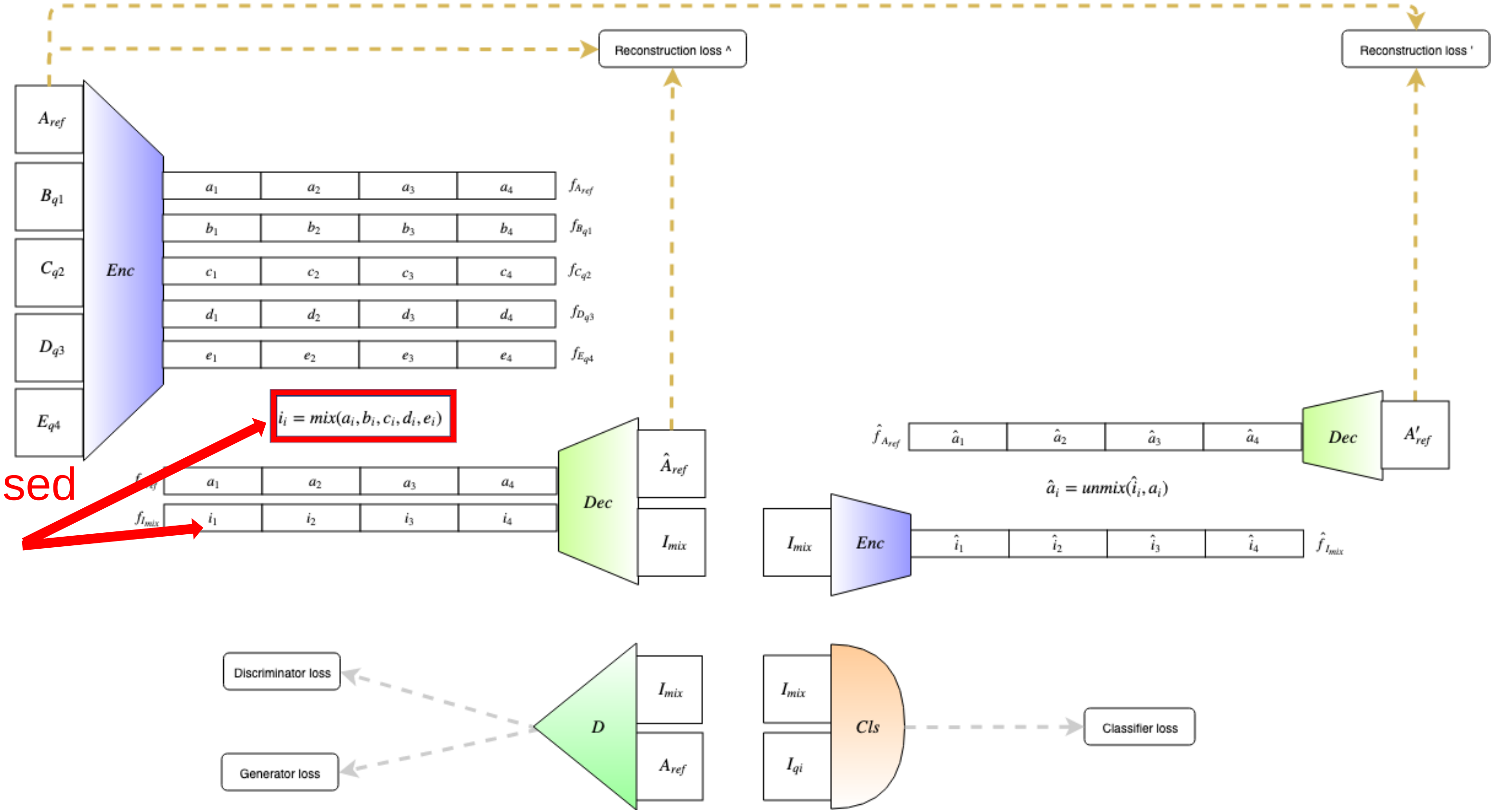
- Generator joint loss

$$\mathcal{L}_G = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}^{G} + \lambda_{Cls}\mathcal{L}_{Cls}$$

- Discriminator loss

$$\mathcal{L}_D = \mathcal{L}_{real}^{D} + \mathcal{L}_{fake}^{D}$$

Latent Space Mixing of Scenes



unsupervised mixing of scenes?

Latent Space Mixing of Scenes

- aim: assist the network in learning, mix a meaningful scene

- two step approach:

    1. Preprocessing: Visual similarity detection algorithm

    2. Training: Latent space scene mixing algorithm

Visual Similarity Detection Algorithm

Latent Space Sence Mixing Algorithm

- at training time:

  → mix the reference image with up to 3 quadrant replacement images

  → constraint #1: at least one quadrant remains from reference image

  → constraint #2: at least one quadrant is replaced

  → constraint #3: only „sufficiently similar“ replacements occur

Latent Space Sence Mixing Algorithm

- One quadrant

- Two quadrants

Latent Space Sence Mixing Algorithm

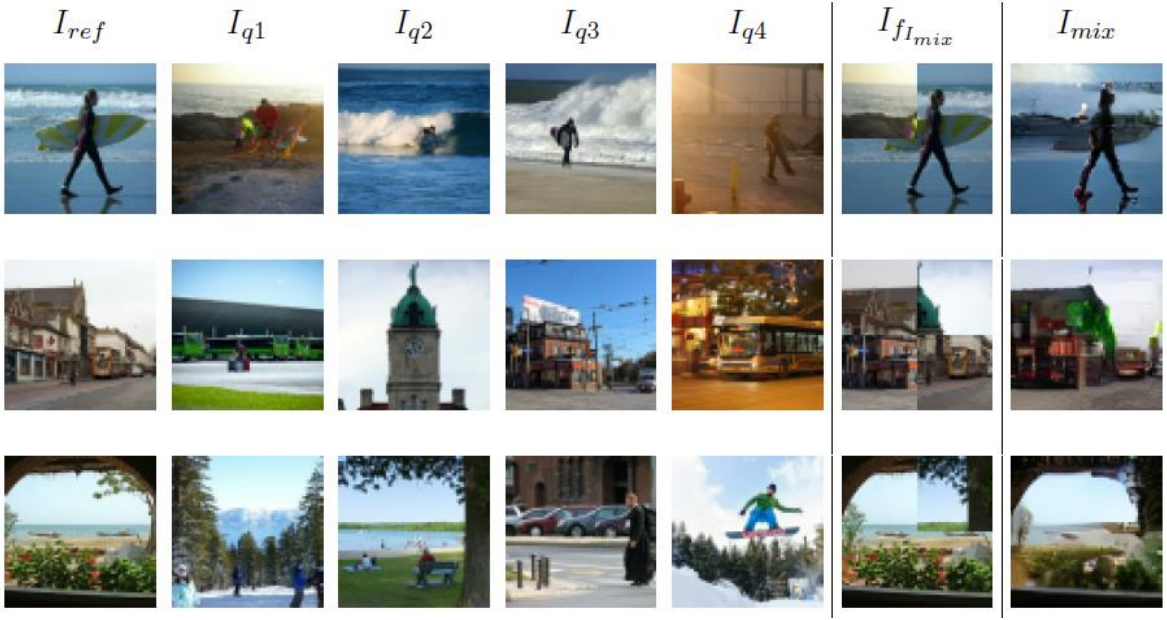
- Three quadrants

Experiments

- Qualitative and quantitative evaluations

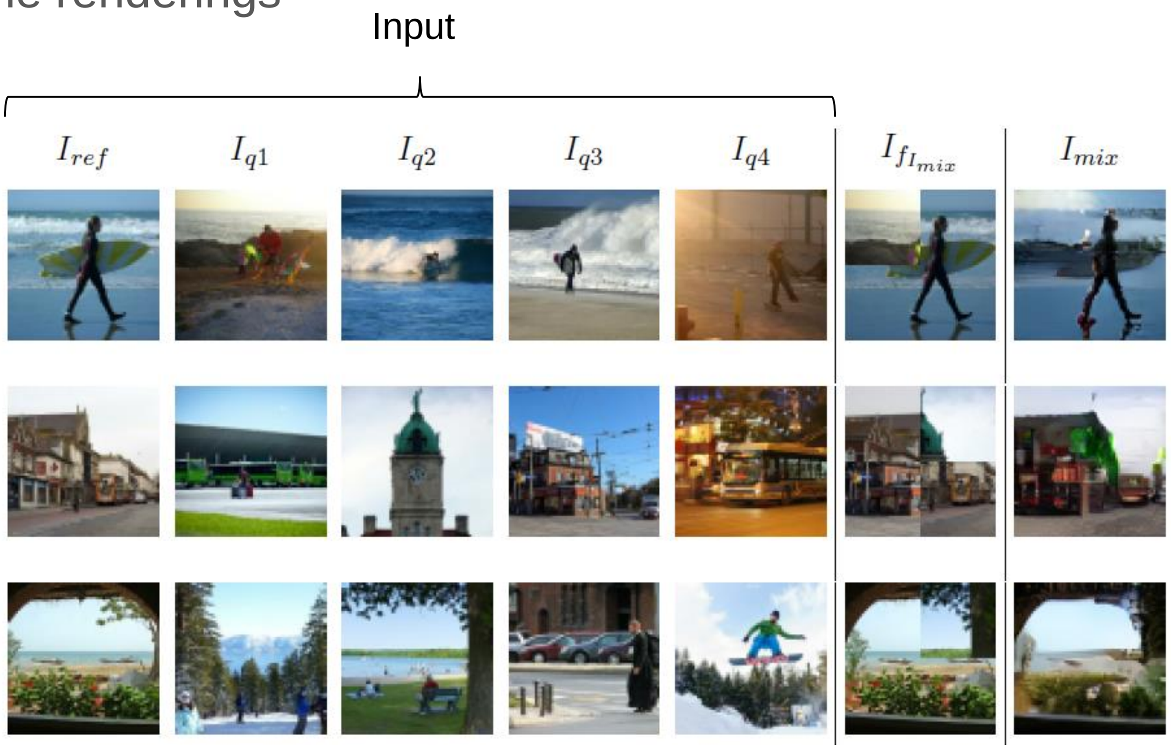
Experiments

- Mixed scene renderings



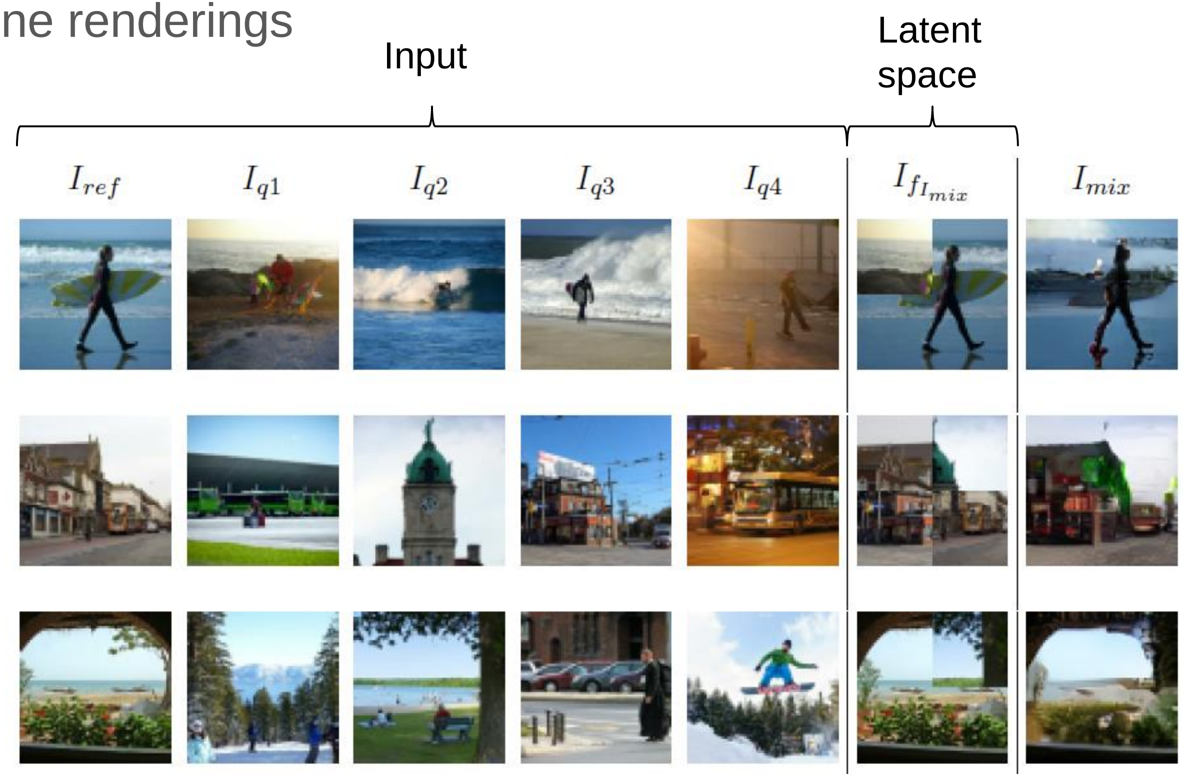
Experiments

- Mixed scene renderings



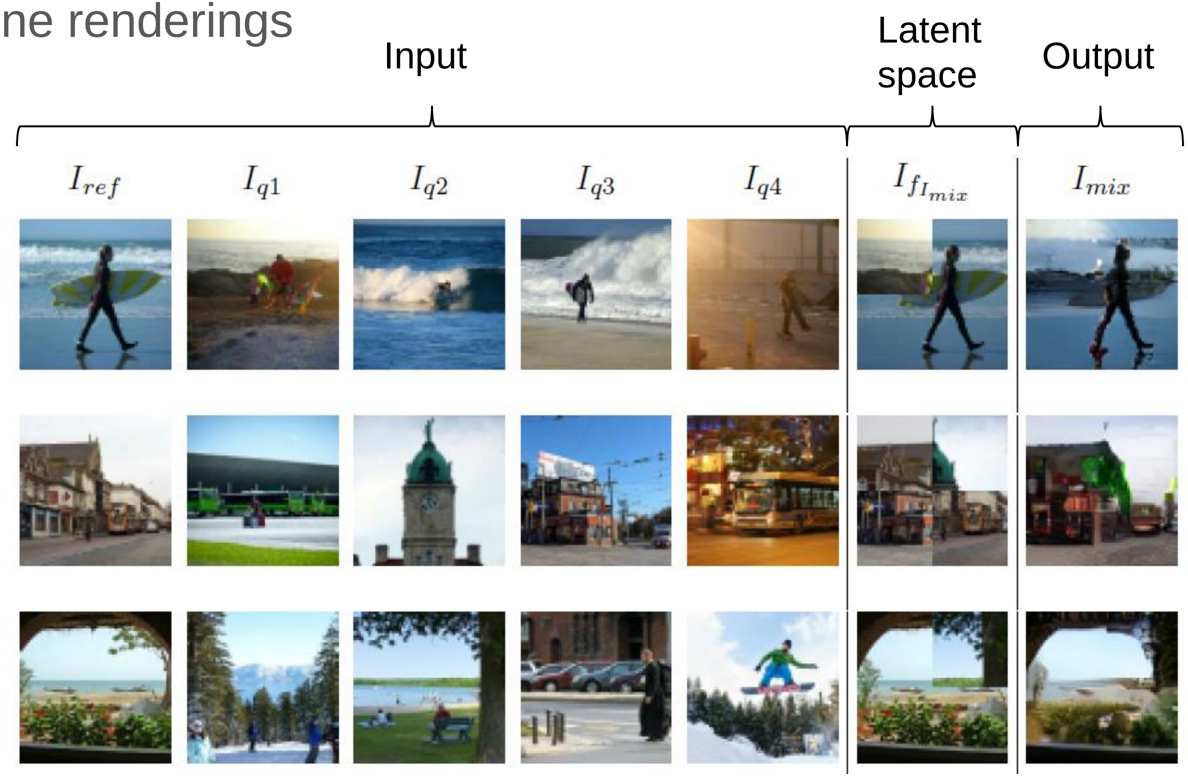
Experiments

- Mixed scene renderings



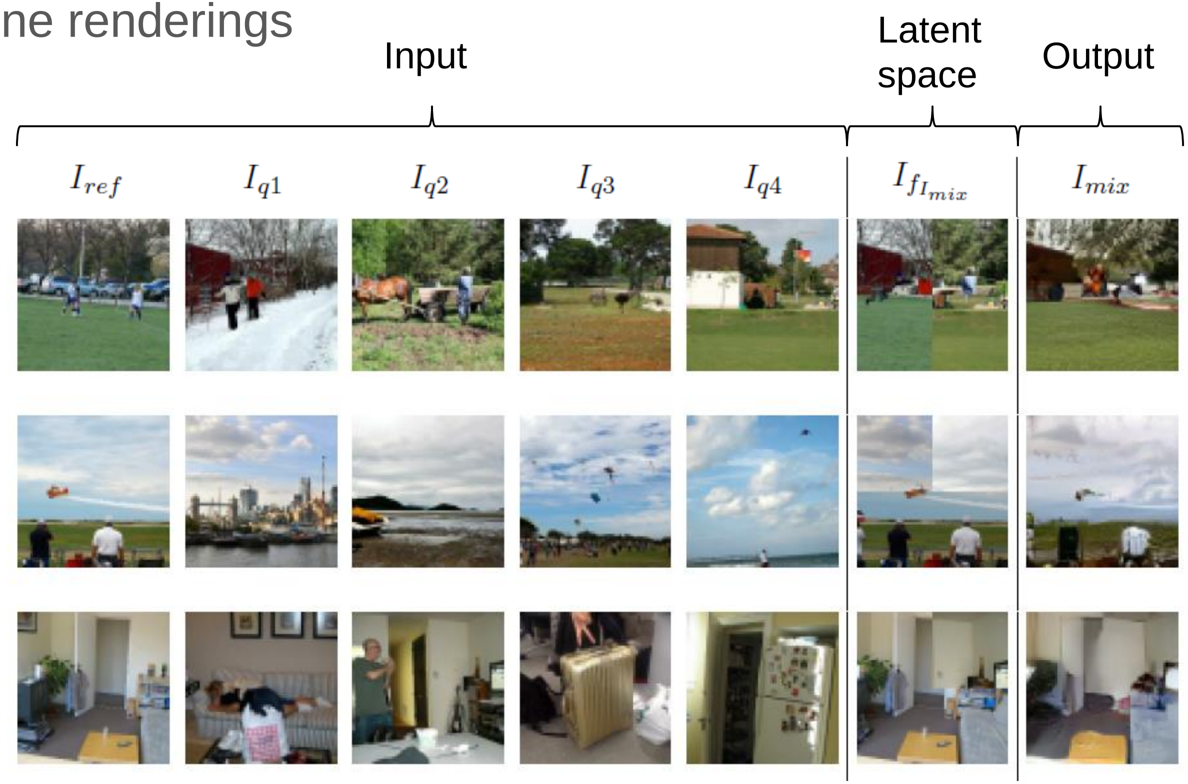
Experiments

- Mixed scene renderings



Experiments

- Mixed scene renderings



Experiments

- Mixed scene renderings - failings

Experiments

caption by PowerPoint

- Object transfer

 +  + 0010 =  *«A truck on a city street»*

Experiments

- Object transfer

 +  + 0010 =  *«A truck on a city street»*

 +  + 1100 =  *«A group of people standing around a plane»*

Experiments

- Object transfer

 +  + 0010 =  *«A truck on a city street»*

 +  + 1100 =  *«A group of people standing around a plane»*

 +  + 0010 =  *«A truck driving down a dirt road»*

 +  + 0001 =  *«A picture containing sky, indoor»*

Experiments

- Object transfer - failings

 +  + 1010 =  *«A blurry image of a kitchen»*

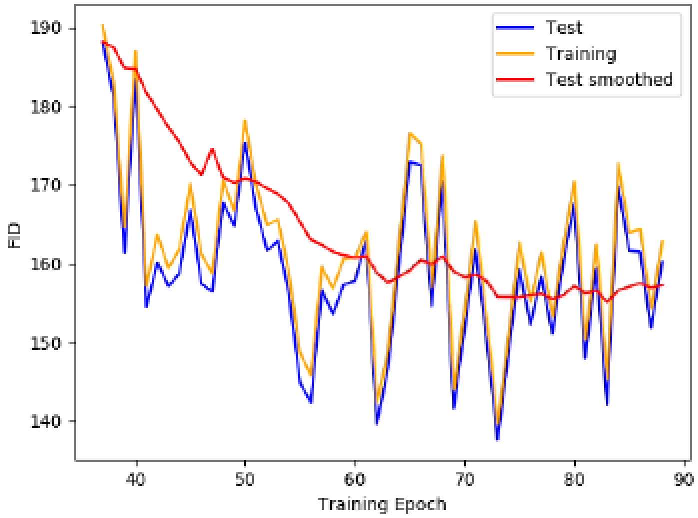
 +  + 0010 =  *«A group of people on a beach»*

 +  + 0101 =  *«A blurry image of a person»*

Experiments

- FID of generated images

	Mean	SD	Best
FID	158.5	±10.9	137.7
IS	6.9	±0.5	7.6



Experiments

- Transfer learning on STL-10

Experiments

- Transfer learning on STL-10

Random

Model	Accuracy	SD	Cls
Random	10.0%	-	-
Random encoder	43.4%	±0.4	17,290
Random encoder (finetuned)	56.5%	±0.6	17,290
STL-10 encoder	**78.7%**	±0.1	17,290
PASCAL encoder	47.1%	±0.2	17,290
STL-10 AlexNet	60.9%	±0.1	40,970
ImageNet AlexNet	62.4%	±0.3	40,970
Jenni & Favaro [34] (frozen)	**76.9%**	±0.1	40,970
Swersky *et al.* [69]	70.1%	±0.6	-
Ours (discriminator)	**62.8%**	±0.3	120,970
Ours (discriminator finetuned)	62.6%	±0.2	120,970
Ours (encoder)	36.5%	±0.2	17,290
Ours (encoder finetuned)	54.2%	±0.2	17,290
Ours (knowledge transfer [57], discriminator)	42.9%	±0.3	40,970
Ours (knowledge transfer [57], encoder)	38.5%	±0.1	40,970

Experiments

- Transfer learning on STL-10

	Model	Accuracy	SD	Cls
Random	Random	10.0%	-	-
	Random encoder	43.4%	±0.4	17,290
	Random encoder (finetuned)	56.5%	±0.6	17,290
Supervised	STL-10 encoder	**78.7%**	±0.1	17,290
	PASCAL encoder	47.1%	±0.2	17,290
	STL-10 AlexNet	60.9%	±0.1	40,970
	ImageNet AlexNet	62.4%	±0.3	40,970
	Jenni & Favaro [34] (frozen)	**76.9%**	±0.1	40,970
	Swersky *et al.* [69]	70.1%	±0.6	-
	Ours (discriminator)	**62.8%**	±0.3	120,970
	Ours (discriminator finetuned)	62.6%	±0.2	120,970
	Ours (encoder)	36.5%	±0.2	17,290
	Ours (encoder finetuned)	54.2%	±0.2	17,290
	Ours (knowledge transfer [57], discriminator)	42.9%	±0.3	40,970
	Ours (knowledge transfer [57], encoder)	38.5%	±0.1	40,970

Experiments

- Transfer learning on STL-10

Random

Supervised

SOTA

Model	Accuracy	SD	Cls
Random	10.0%	-	-
Random encoder	43.4%	±0.4	17,290
Random encoder (finetuned)	56.5%	±0.6	17,290
STL-10 encoder	**78.7%**	±0.1	17,290
PASCAL encoder	47.1%	±0.2	17,290
STL-10 AlexNet	60.9%	±0.1	40,970
ImageNet AlexNet	62.4%	±0.3	40,970
Jenni & Favaro [34] (frozen)	**76.9%**	±0.1	40,970
Swersky *et al.* [69]	70.1%	±0.6	-
Ours (discriminator)	**62.8%**	±0.3	120,970
Ours (discriminator finetuned)	62.6%	±0.2	120,970
Ours (encoder)	36.5%	±0.2	17,290
Ours (encoder finetuned)	54.2%	±0.2	17,290
Ours (knowledge transfer [57], discriminator)	42.9%	±0.3	40,970
Ours (knowledge transfer [57], encoder)	38.5%	±0.1	40,970

Experiments

- Transfer learning on STL-10

Group	Model	Accuracy	SD	Cls
Random	Random	10.0%	-	-
	Random encoder	43.4%	±0.4	17,290
	Random encoder (finetuned)	56.5%	±0.6	17,290
Supervised	STL-10 encoder	**78.7%**	±0.1	17,290
	PASCAL encoder	47.1%	±0.2	17,290
	STL-10 AlexNet	60.9%	±0.1	40,970
	ImageNet AlexNet	62.4%	±0.3	40,970
SOTA	Jenni & Favaro [34] (frozen)	**76.9%**	±0.1	40,970
	Swersky *et al.* [69]	70.1%	±0.6	-
Ours	Ours (discriminator)	**62.8%**	±0.3	120,970
	Ours (discriminator finetuned)	62.6%	±0.2	120,970
	Ours (encoder)	36.5%	±0.2	17,290
	Ours (encoder finetuned)	54.2%	±0.2	17,290
	Ours (knowledge transfer [57], discriminator)	42.9%	±0.3	40,970
	Ours (knowledge transfer [57], encoder)	38.5%	±0.1	40,970

Experiments

- Transfer learning on STL-10

	Model	Accuracy	SD	Cls
Random	Random	10.0%	-	-
	Random encoder	43.4%	±0.4	17,290
	Random encoder (finetuned)	56.5%	±0.6	17,290
Supervised	STL-10 encoder	**78.7%**	±0.1	17,290
	PASCAL encoder	47.1%	±0.2	17,290
	STL-10 AlexNet	60.9%	±0.1	40,970
	ImageNet AlexNet	62.4%	±0.3	40,970
SOTA	Jenni & Favaro [34] (frozen)	**76.9%**	±0.1	40,970
	Swersky *et al.* [69]	70.1%	±0.6	-
Ours	Ours (discriminator)	**62.8%**	±0.3	120,970
	Ours (discriminator finetuned)	62.6%	±0.2	120,970
	Ours (encoder)	36.5%	±0.2	17,290
	Ours (encoder finetuned)	54.2%	±0.2	17,290
Ours*	Ours (knowledge transfer [57], discriminator)	42.9%	±0.3	40,970
	Ours (knowledge transfer [57], encoder)	38.5%	±0.1	40,970

Experiments

- Transfer learning on STL-10

	Model	Accuracy	SD	Cls
Random	Random	10.0%	-	-
	Random encoder	43.4%	±0.4	17,290
	Random encoder (finetuned)	56.5%	±0.6	17,290
Supervised	STL-10 encoder	**78.7%**	±0.1	17,290
	PASCAL encoder	47.1%	±0.2	17,290
	STL-10 AlexNet	60.9%	±0.1	40,970
	ImageNet AlexNet	62.4%	±0.3	40,970
SOTA	Jenni & Favaro [34] (frozen)	**76.9%**	±0.1	40,970
	Swersky *et al.* [69]	70.1%	±0.6	-
Ours	Ours (discriminator)	**62.8%**	±0.3	120,970
	Ours (discriminator finetuned)	62.6%	±0.2	120,970
	Ours (encoder)	36.5%	±0.2	17,290
	Ours (encoder finetuned)	54.2%	±0.2	17,290
Ours*	Ours (knowledge transfer [57], discriminator)	42.9%	±0.3	40,970
	Ours (knowledge transfer [57], encoder)	38.5%	±0.1	40,970

AlexNet

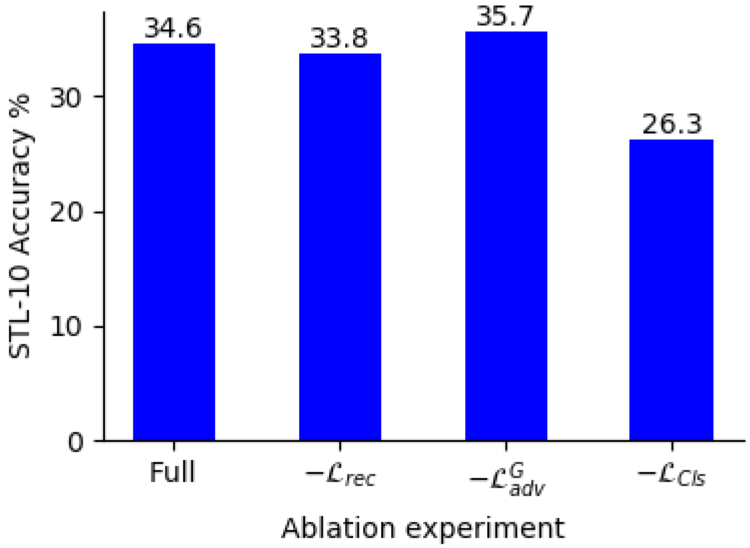
Experiments

- Ablation Analysis

$$\mathcal{L}_G = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}^{G} + \lambda_{Cls}\mathcal{L}_{Cls}$$

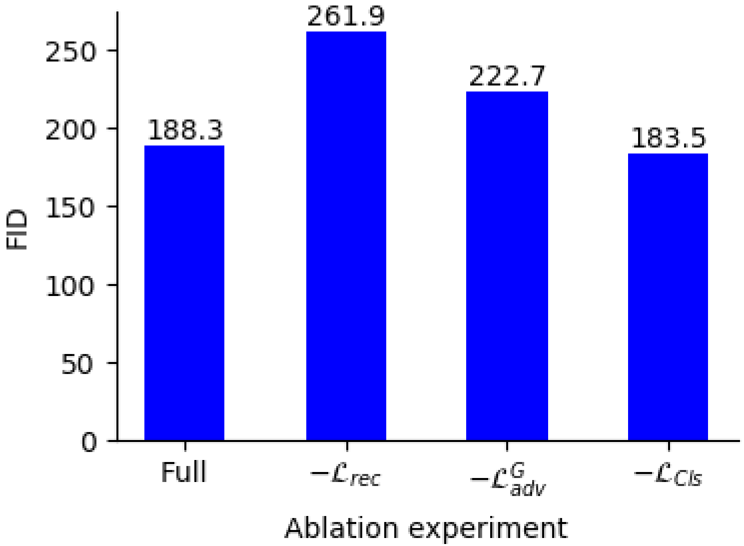
Experiments

- Ablation Analysis on STL-10, after 10 epochs



Experiments

- Ablation Analysis with FID on COCO test set, after 10 epochs



Conclusions

- novel method for unsupervised representation learning
- learns directly from Internet-scale natural image data by mixing scenes
- experiments demonstrate capability of rendering realistic scenes
    - degree of realism offers room for improvement
- learnt representations for TL not at SOTA level

Future Work

- more experiments could yield substantial improvements considering:

  → standard architectures, increased model capacity, more data augmentations, larger hyperparameter search

- introduce explicit notion of object location

- train model iteratively on datasets of increasing complexity

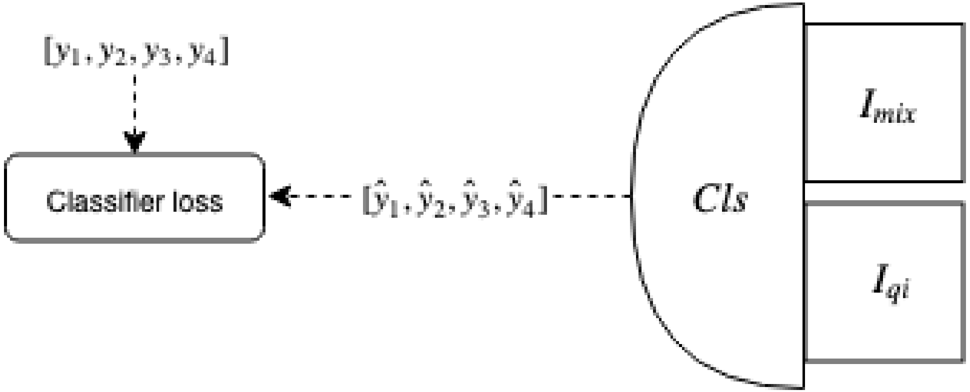
  → finetuning across multiple datasets

- later: disentangle not only objects but its attributes as well

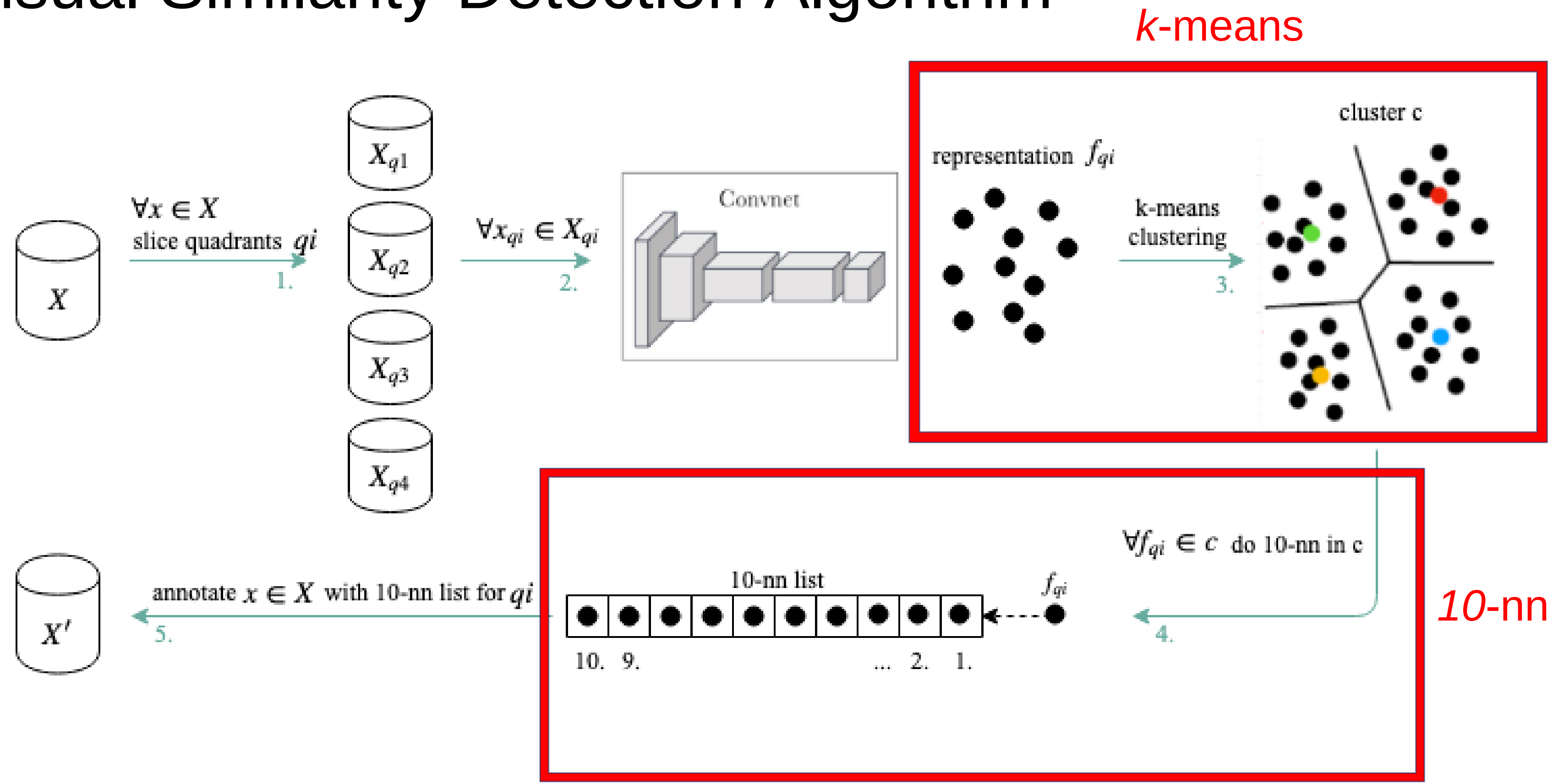
Thank you for your attention.

LORBMS Model Architecture

- Classifier



Visual Similarity Detection Algorithm



LORBMS Model Architecture

- Vanilla GAN

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} \left[log\, D(x) \right] + \mathbb{E}_{z \sim p_z(z)} \left[log(1 - D(G(z))) \right]$$

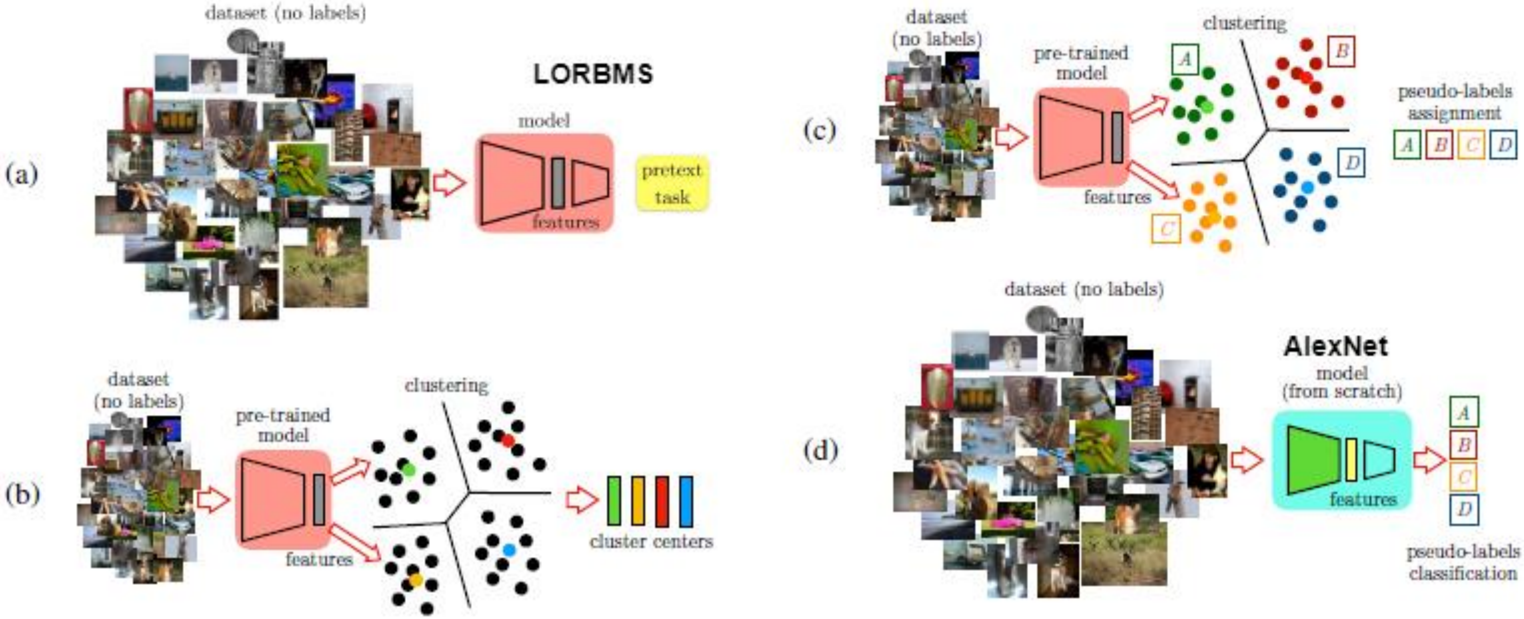
- LORBMS GAN

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{I_{ref} \sim p_{data}} \left[log\, D(I_{ref}) \right] +$$

$$\mathbb{E}_{I_j \sim p_{data}} \left[log(1 - D(G(I_{ref}, I_{q1}, I_{q2}, I_{q3}, I_{q4}))) \right]$$

Experiments

- Transfer learning on STL-10: Knowledge transfer from LORBMS to AlexNet



adapted from Noroozi *et.al.*, *Boosting Self-Supervised Learning via Knowledge Transfer*

Loss functions

$$\mathcal{L}_G = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}^{G} + \lambda_{Cls}\mathcal{L}_{Cls}$$

$$\mathcal{L}_{rec} = \mathbb{E}_{I_{ref}\sim p_{data}(I_{ref})}\left[\| I_{ref} - \hat{I}_{ref}\|_p + \| I_{ref} - I'_{ref}\|_p\right]$$

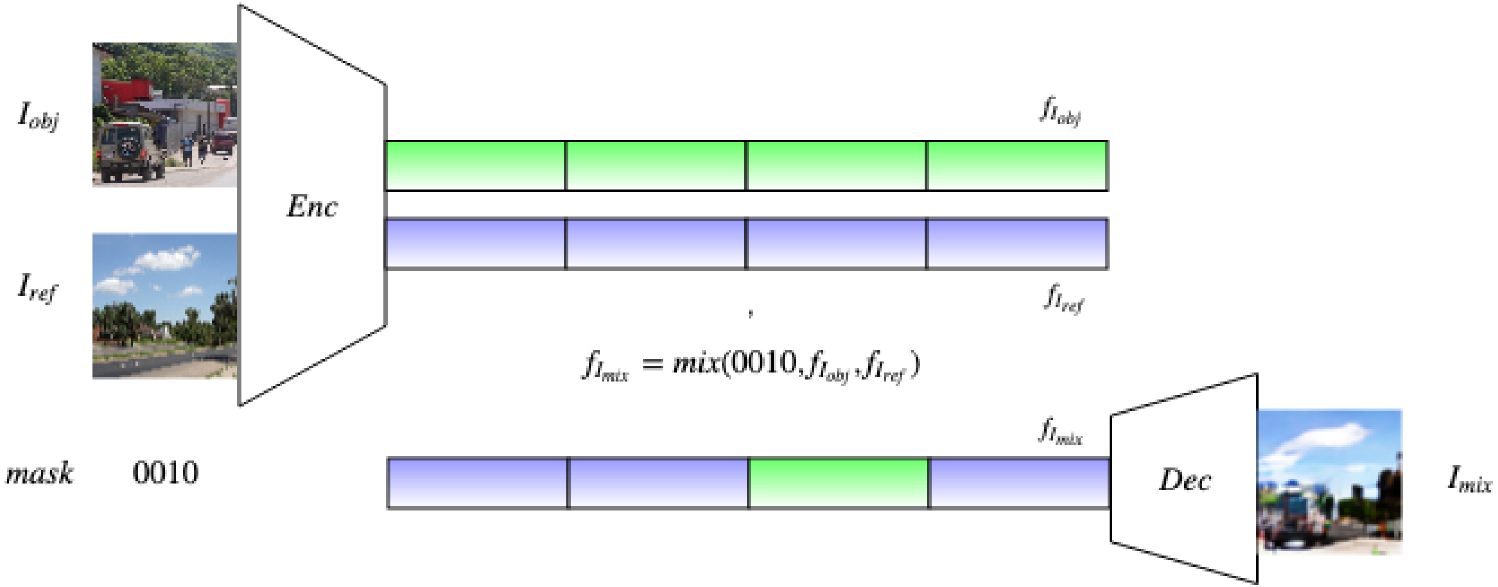
$$\mathcal{L}_{adv}^{G} = \mathbb{E}_{I_j\sim p_{data}(I_j)}\left[\mathcal{L}_{bce}(D(G(I_{ref}, I_{q1}, I_{q2}, I_{q3}, I_{q4})), 1)\right]$$

$$\mathcal{L}_{Cls} = \mathbb{E}_{I_j\sim p_{data}(I_j)}\left[\sum_j \lambda_j \mathcal{L}_{bce}(\hat{y}^j, y^j)\right]$$

$$\mathcal{L}_{bce}(p, y) = -(ylog(p) + (1-y)log(1-p))$$

Experiments

- Object transfer



Visual Similarity Detection Algorithm

- the search for „quadrant replacement“ candidates

Challenges

- unsupervised learning
- learning object representations given natural images
- disentangle factors of variation
- unaligned natural dataset
- generalization (inference)