Image-to-image translation

Luis Herranz, Joost van de Weijer
Learning and machine perception (LAMP) group
Computer Vision Center (Barcelona)

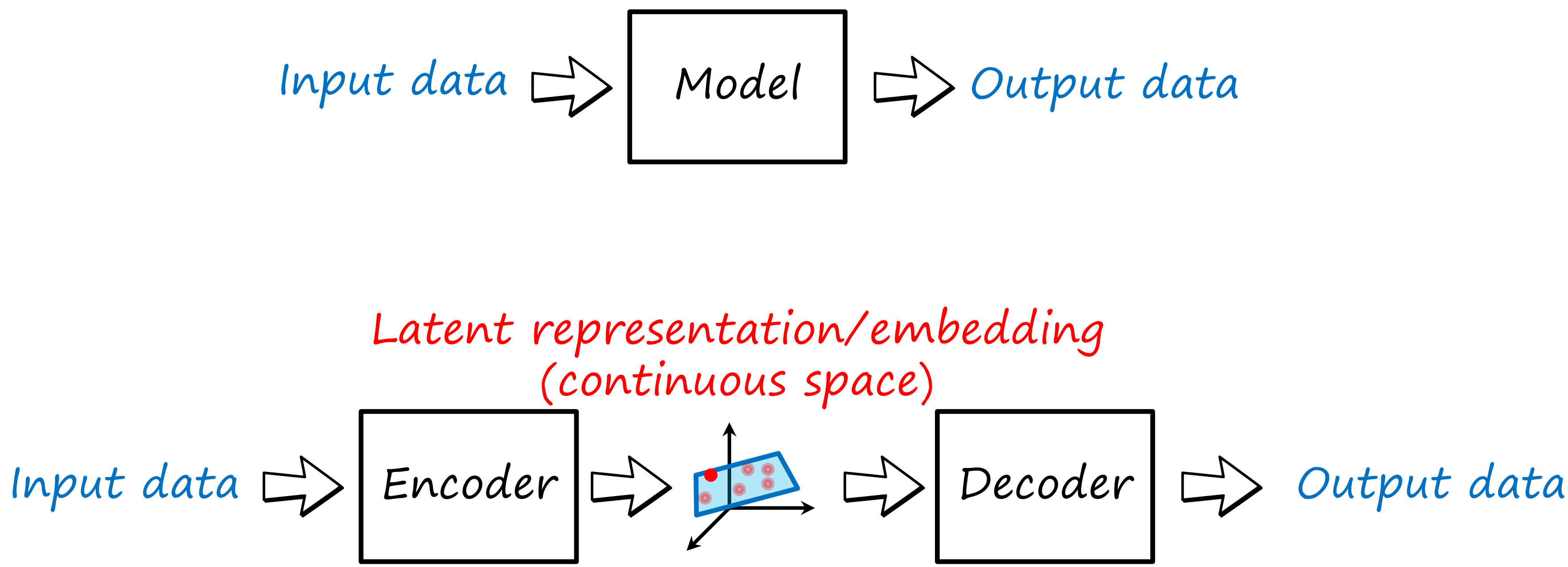
Outline

- M2CR project framework
- Paired image-to-image translation (pix2pix)
- Unpaired image-to-image translation (cycleGAN)
- Unseen translations (mix&match networks)

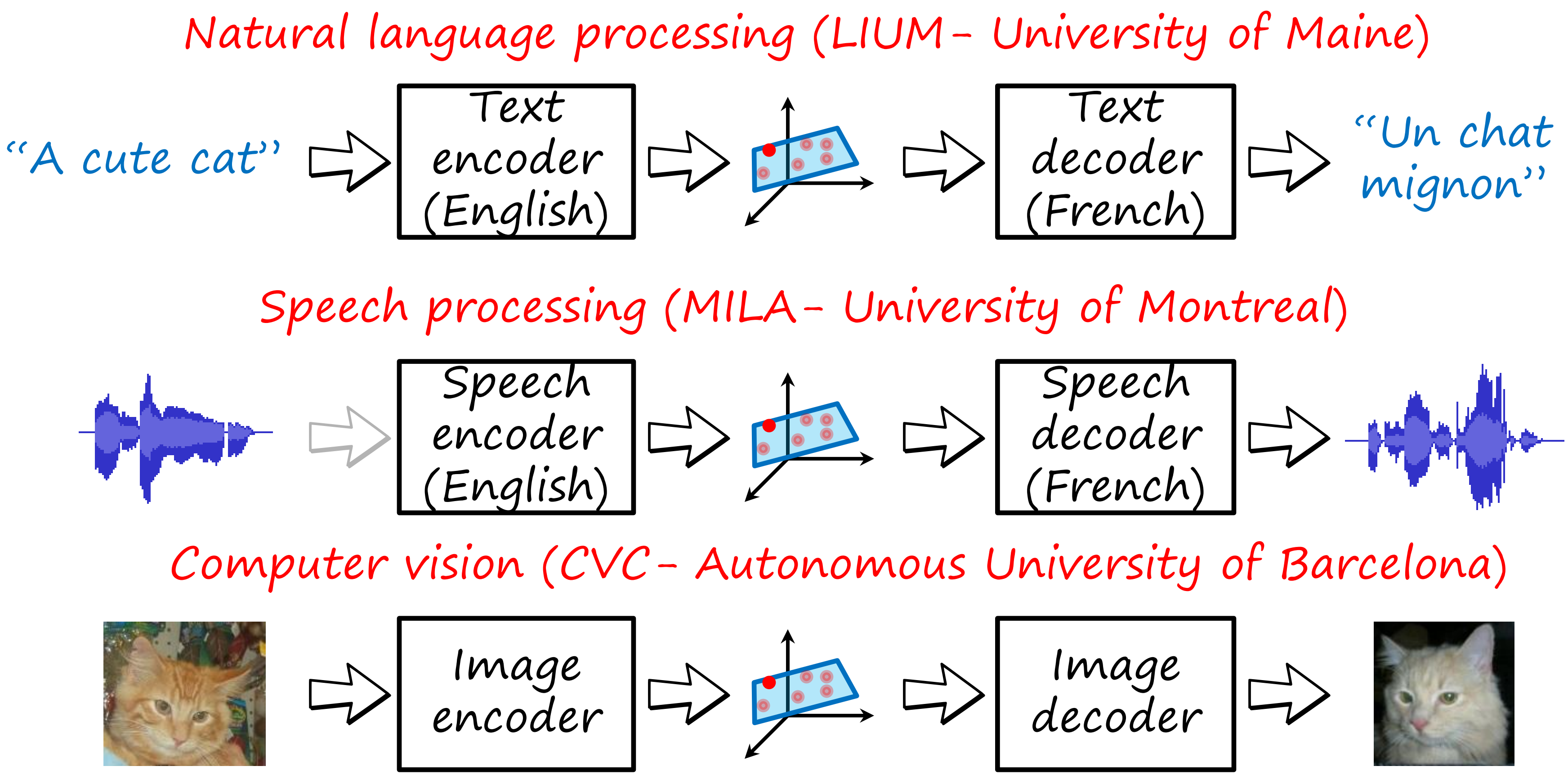
Outline

- M2CR project framework
- Paired image-to-image translation (pix2pix)
- Unpaired image-to-image translation (cycleGAN)
- Unseen translations (mix&match networks)

Encoder-decoder framework

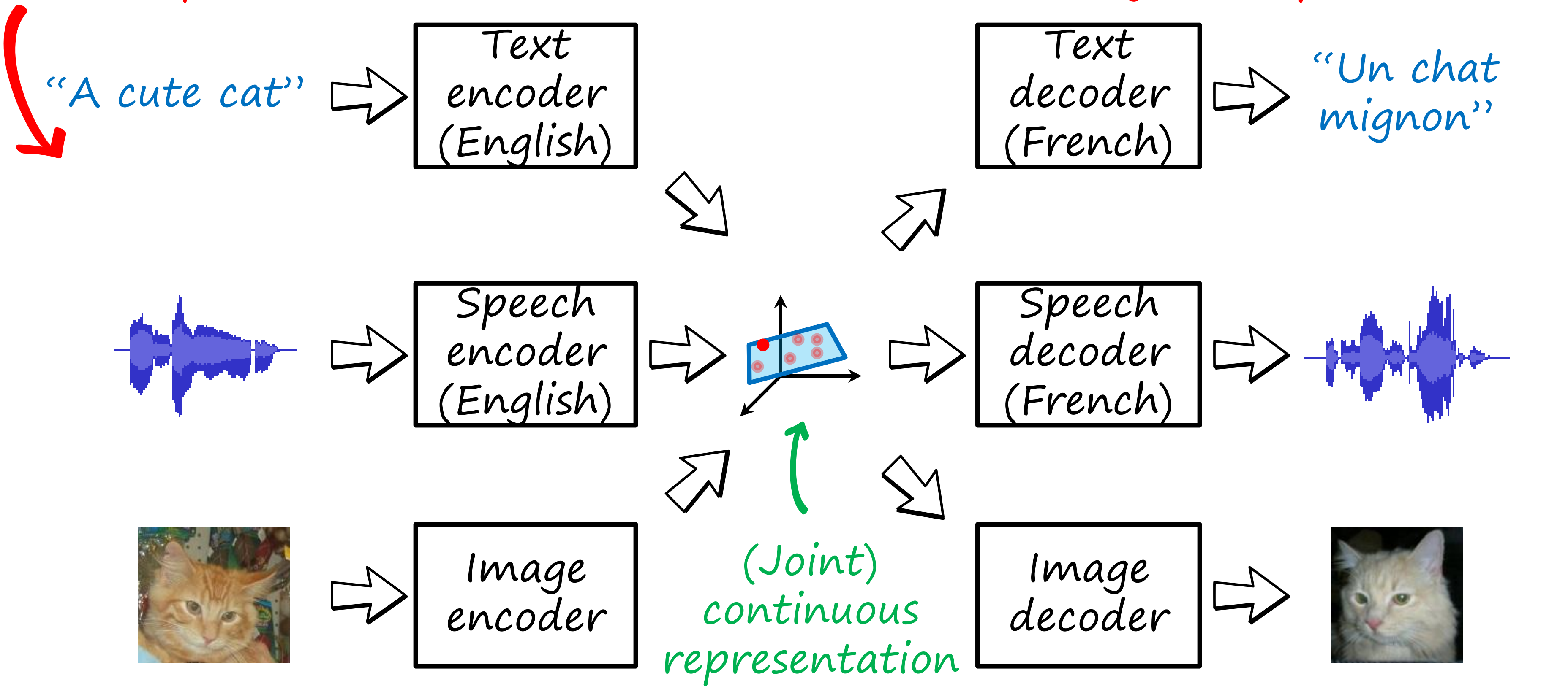


M2CR partners

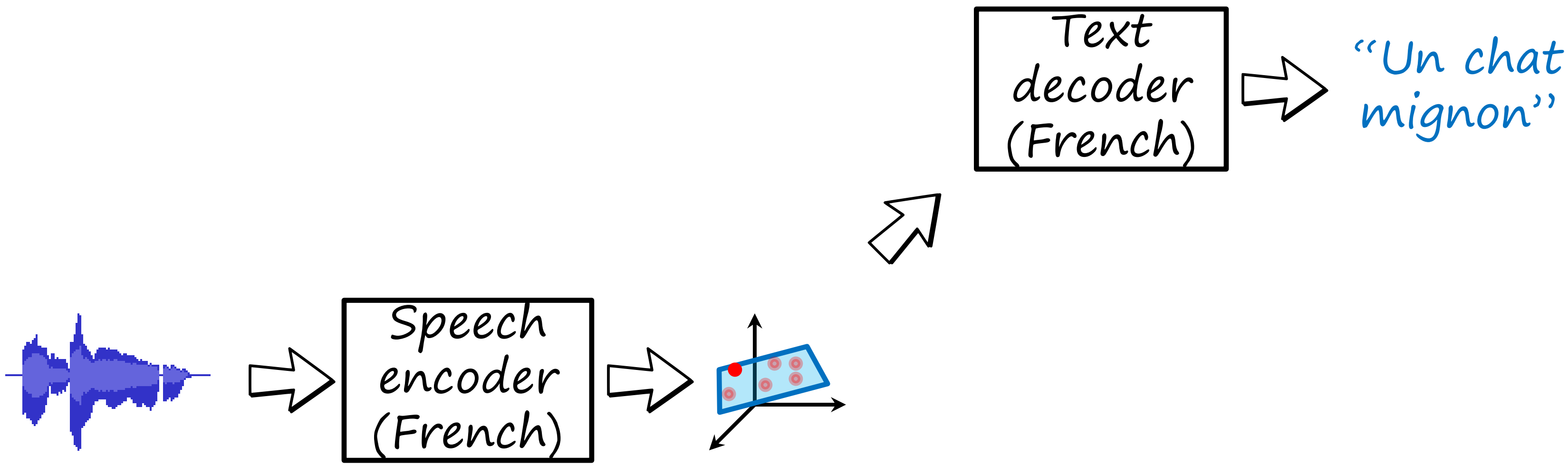


M2CR: multilingual multimodal continuous representations

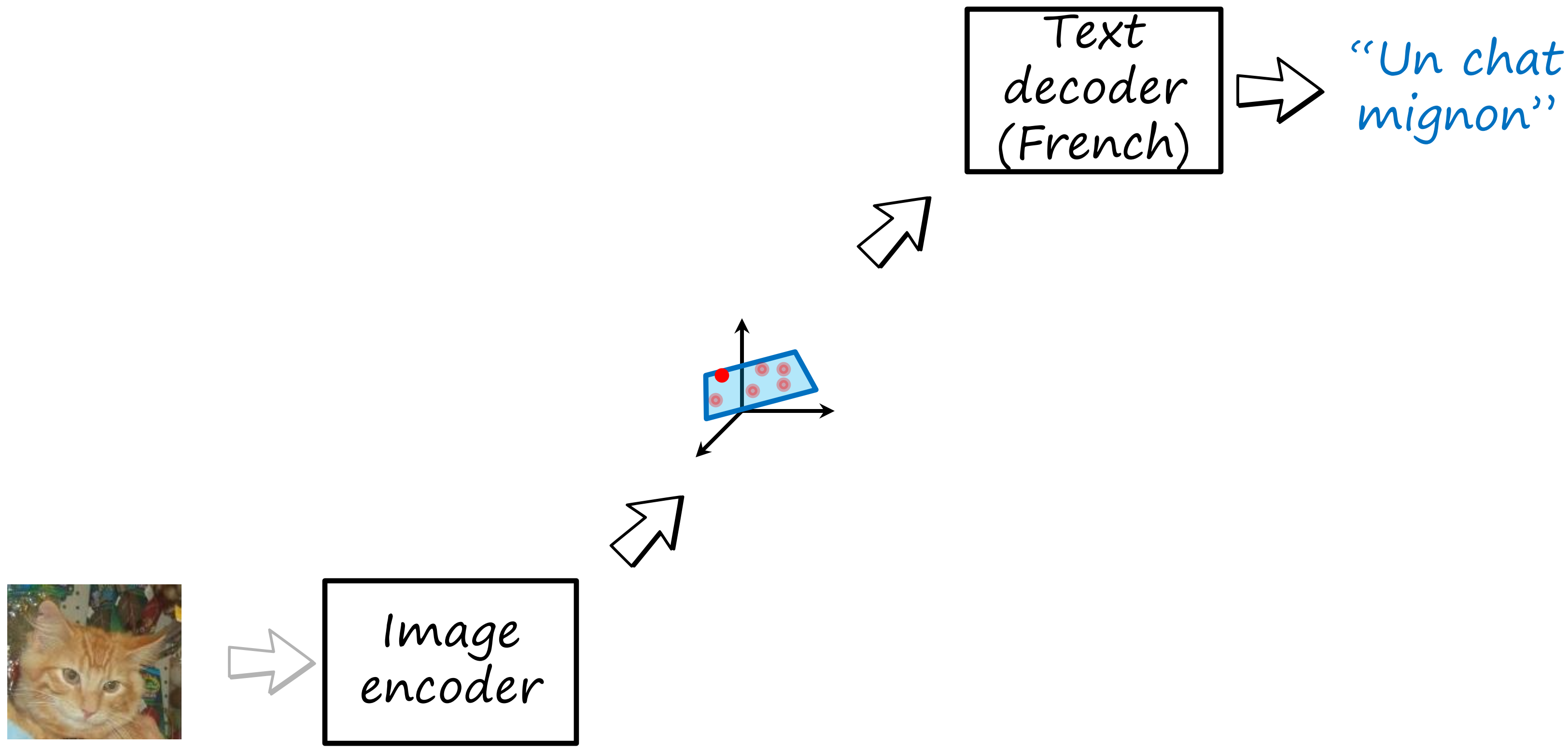
Humans perceive, understand and communicate through multiple **modalities**



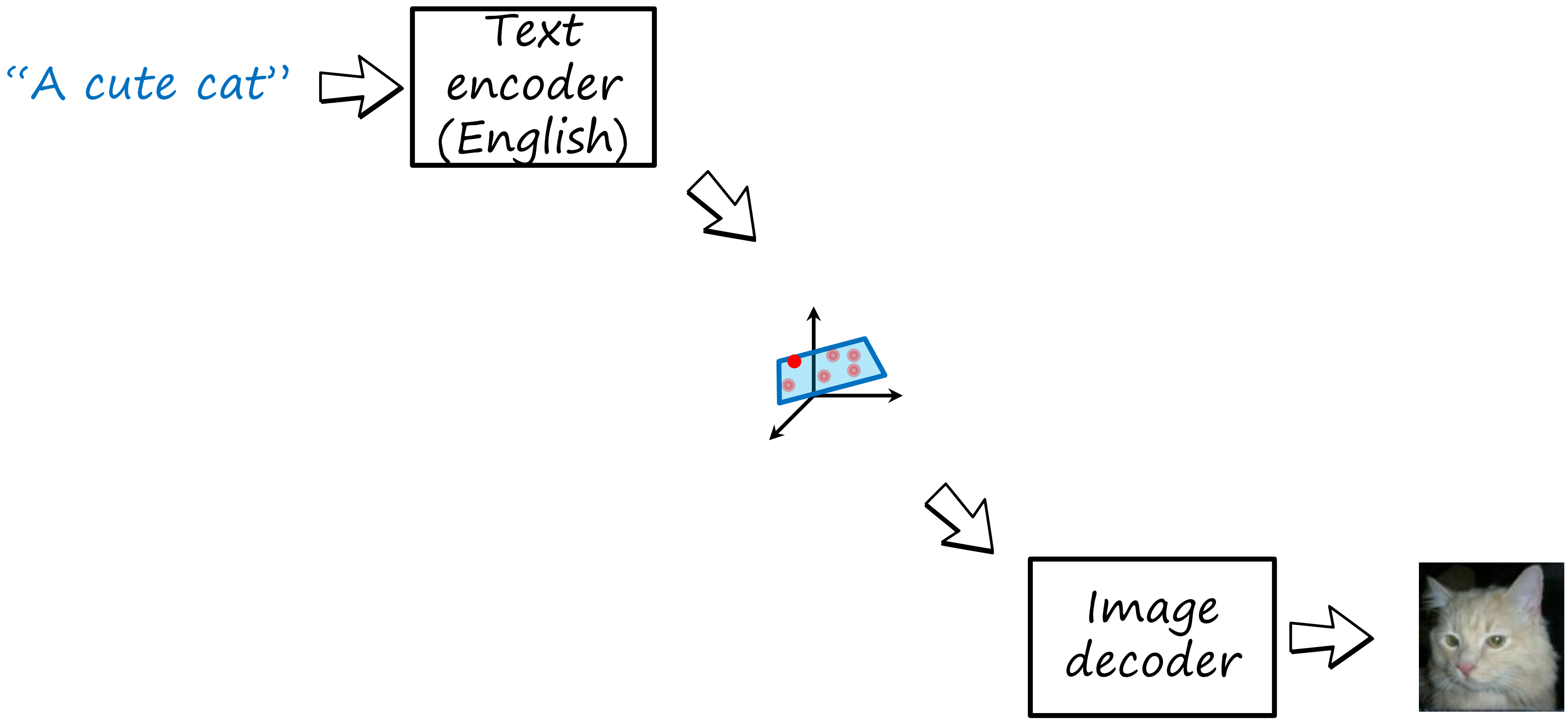
Cross-modal translation
Example: speech recognition



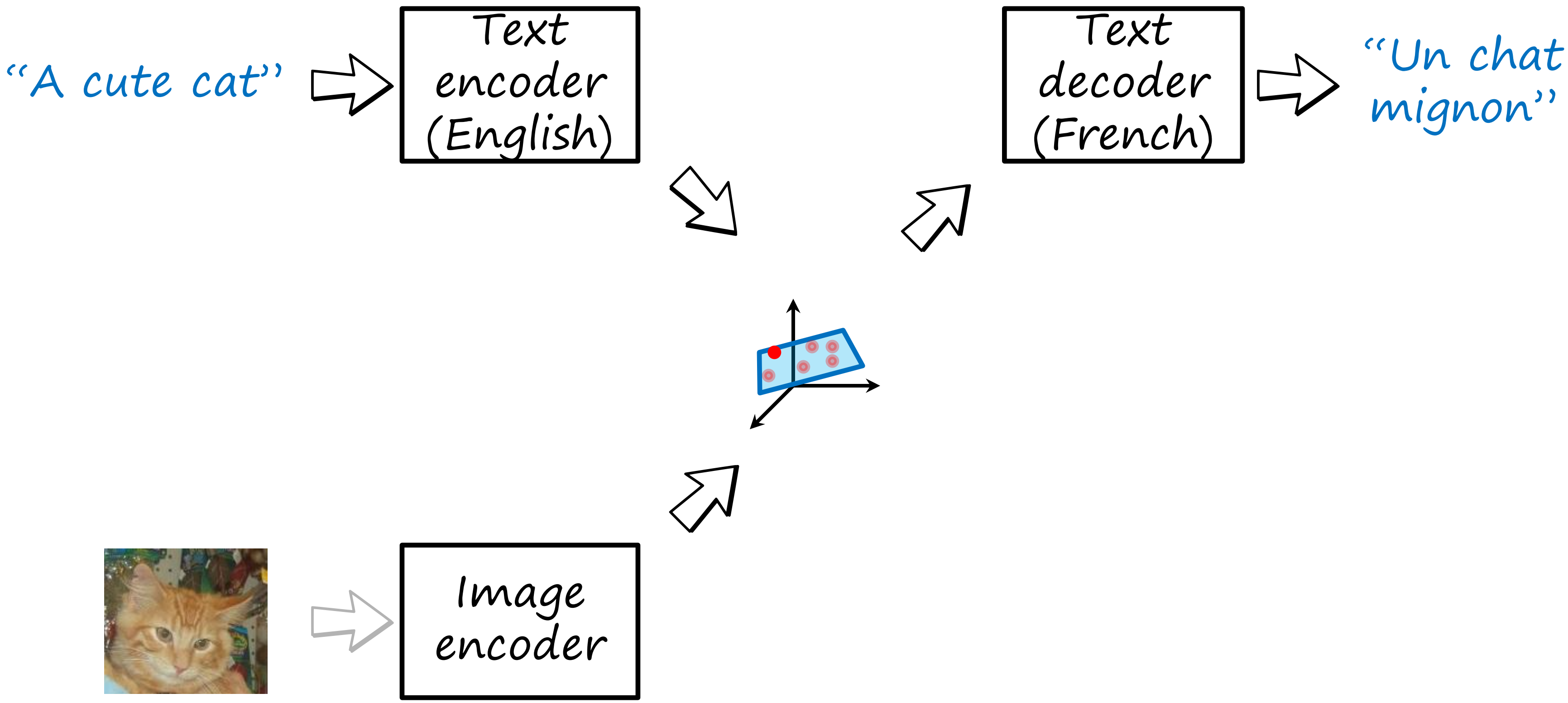
Cross-modal translation
Example: image captioning



M2CR: multilingual multimodal continuous representations



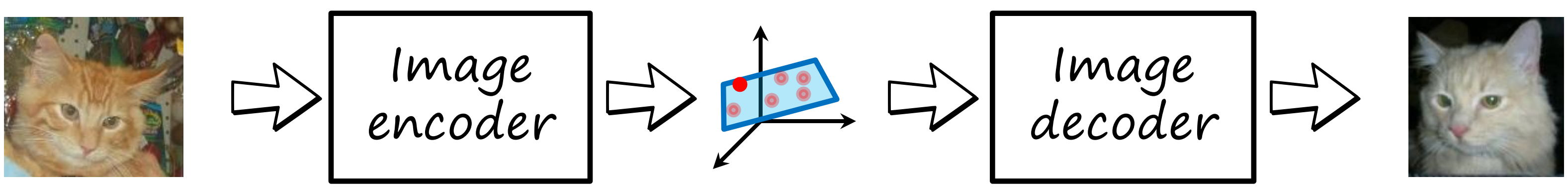
Multimodal translation
Example: text+image to text



Challenges

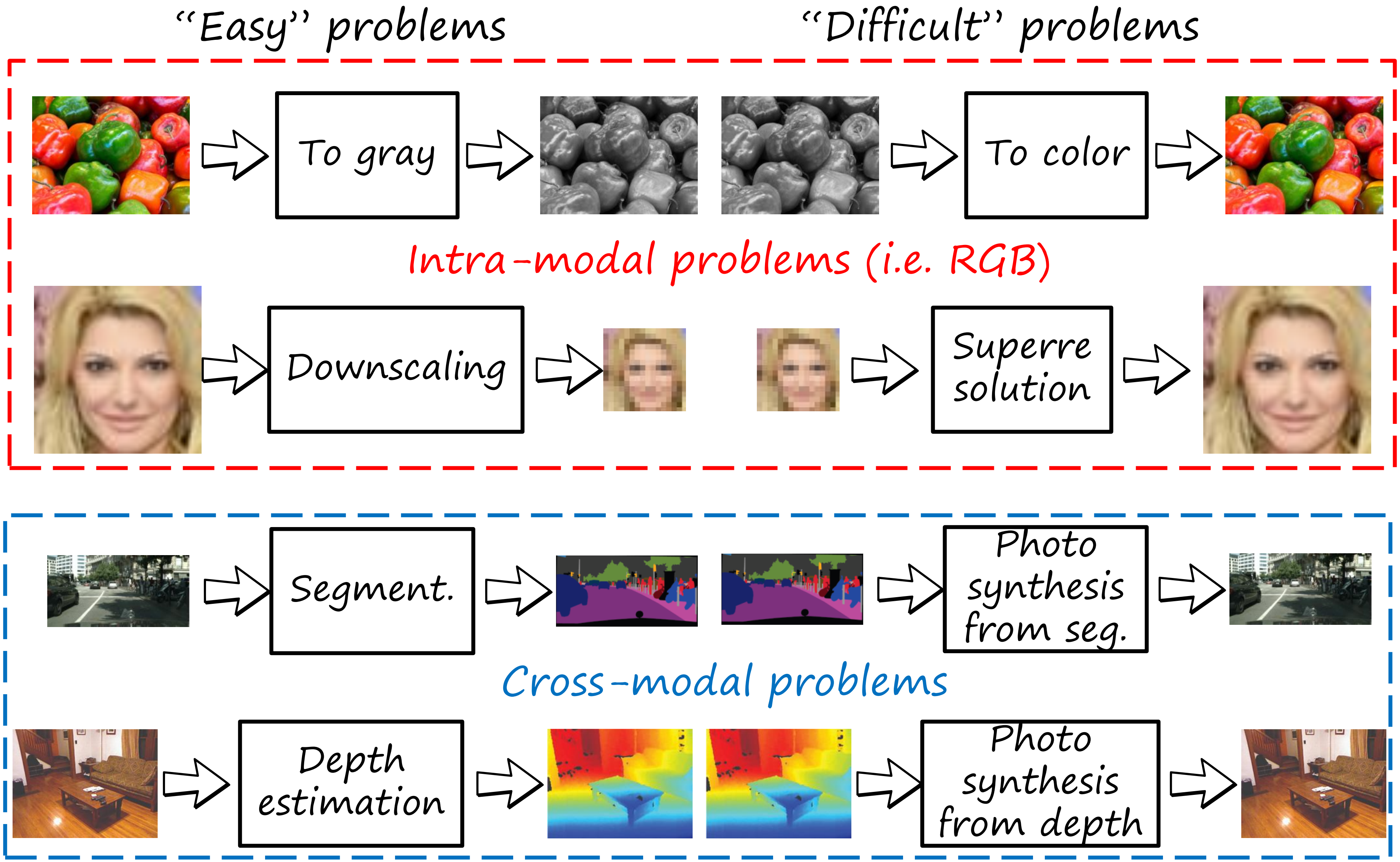
- Heterogeneous modalities
  – Images: fixed-size 2D data in a continuous space
  – Speech: variable-length 1D in a continuous space
  – Language: variable-length discrete (one-hot) data
- Heterogeneous encoders-decoders
  – Text, speech: recurrent neural networks (RNNs)
  – Images: convolutional neural networks (CNNs)
- How to combine modalities properly
  – Usually depends on the particular task

This talk: image-to-image translation



Outline

- M2CR project framework
- Paired image-to-image translation (pix2pix)
- Unpaired image-to-image translation (cycleGAN)
- Unseen translations (mix&match networks)


"Easy" problems
"Difficult" problems
To gray
To color
Intra-modal problems (i.e. RGB)
Downscaling
Superresolution
Segment.
Photo synthesis from seg.
Cross-modal problems
Depth estimation
Photo synthesis from depth

How to solve the inverse problem?

- Just signal processing: not possible

- Machine learning: do you have enough **data**?
  - Learn **priors** (e.g. faces)
  - Discover the data **manifold**
  - More realistic approximation

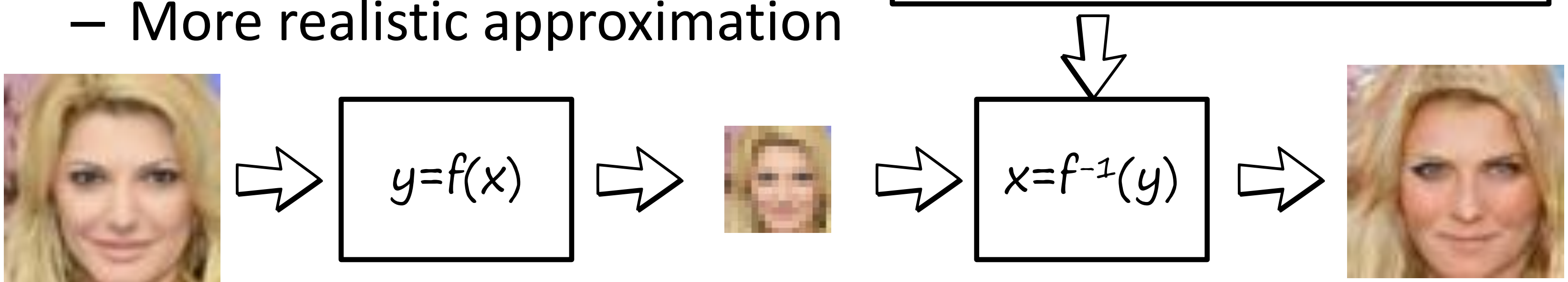


Image-to-image translation

- General purpose
- Learns from image pairs (input, output)

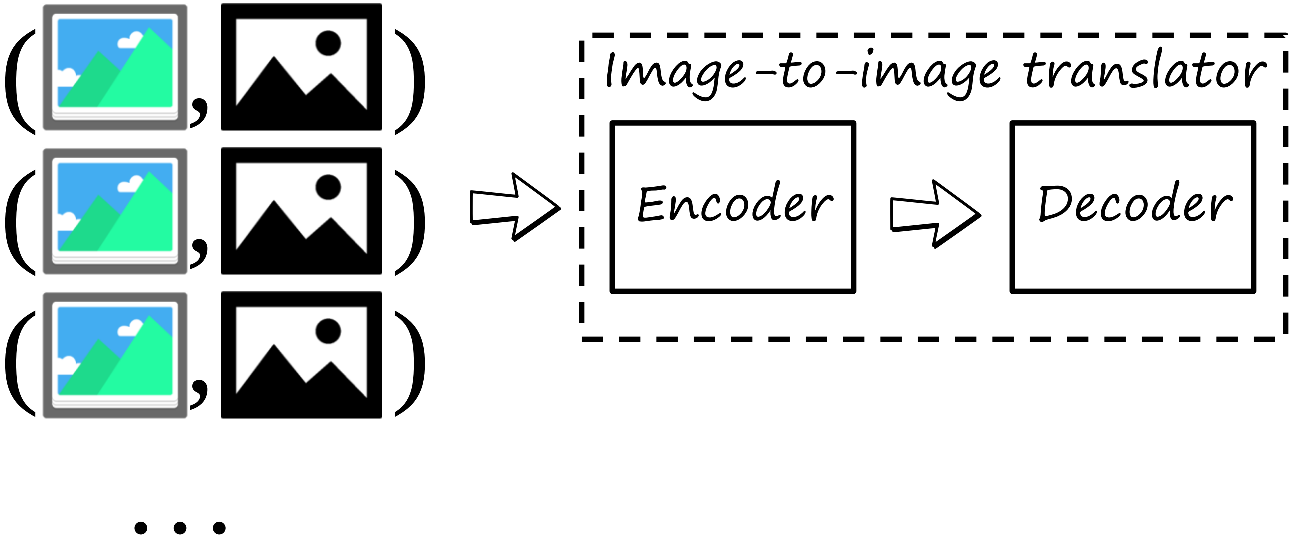
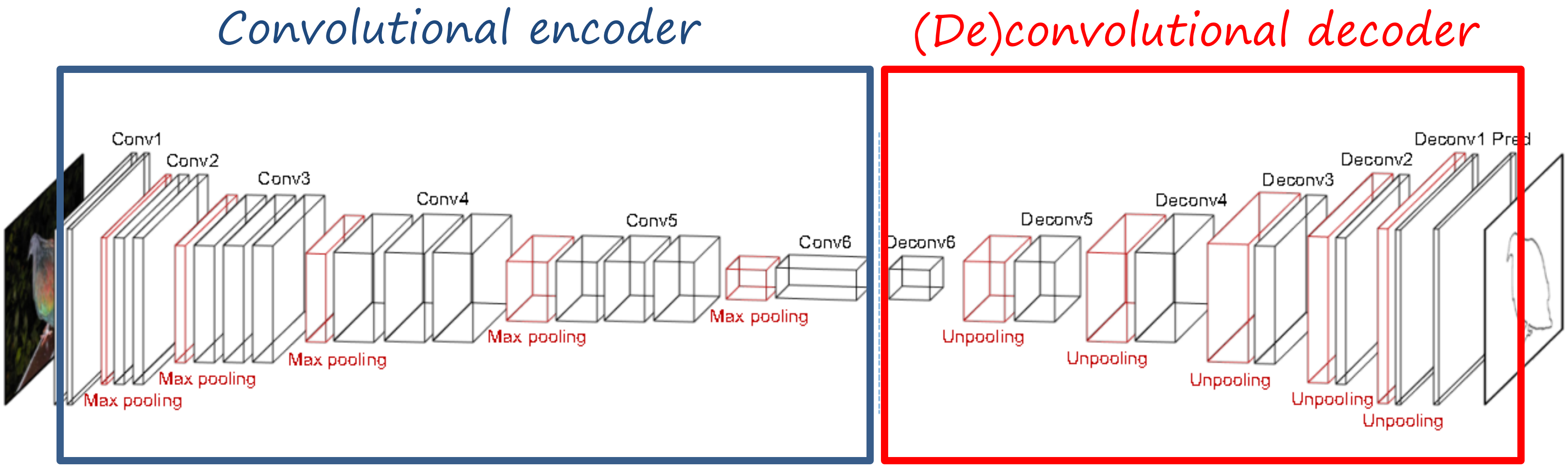
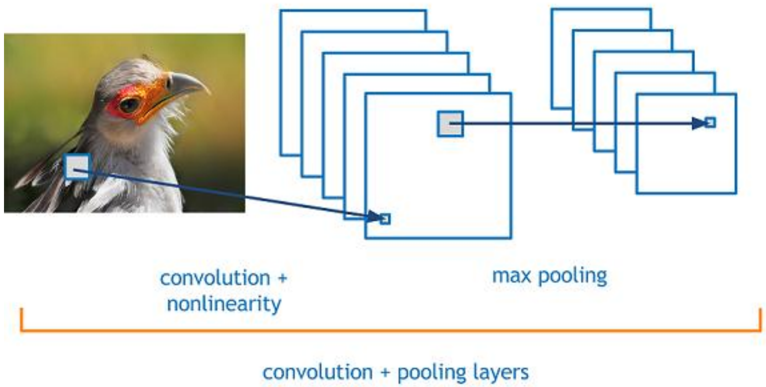
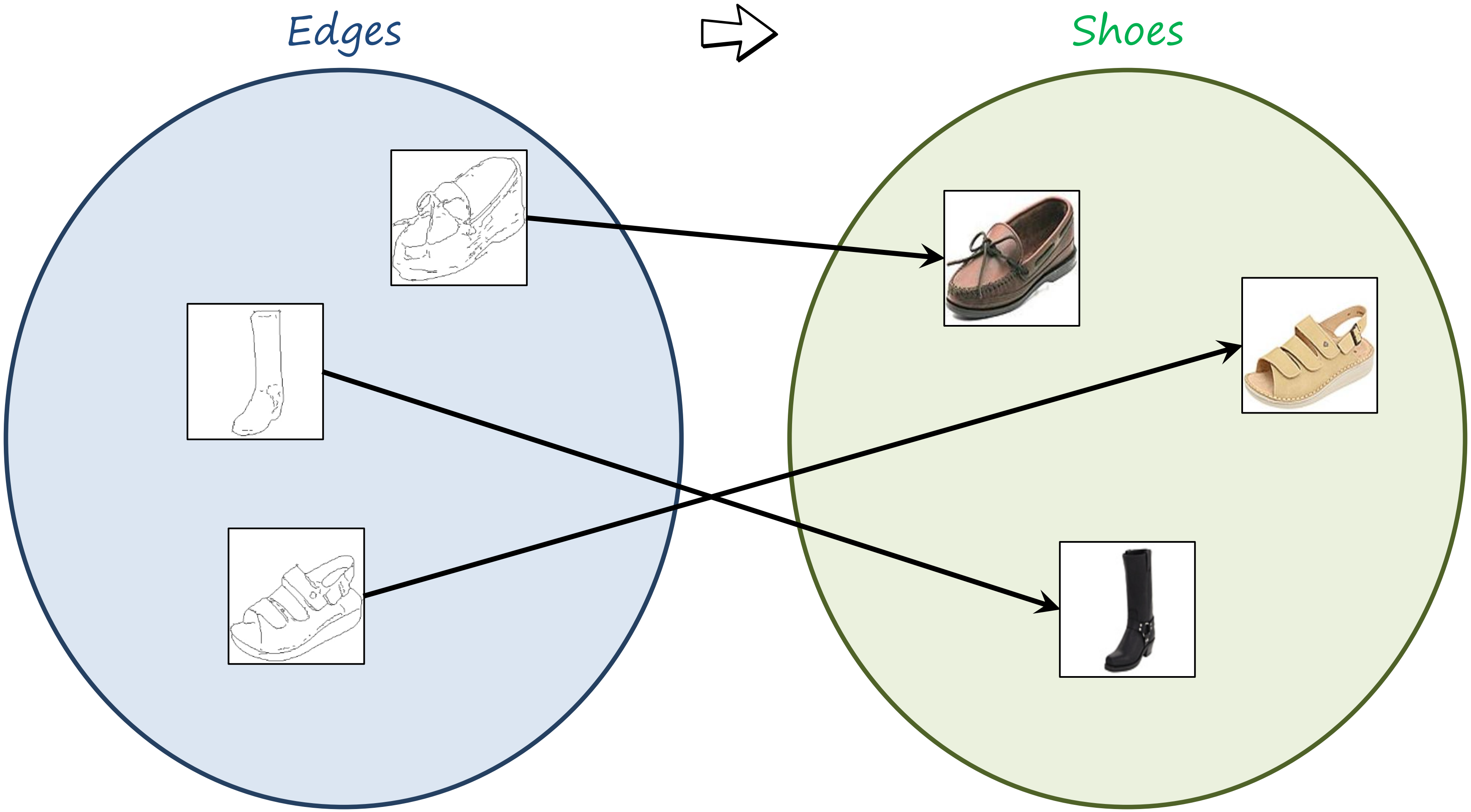


Image encoder-decoder



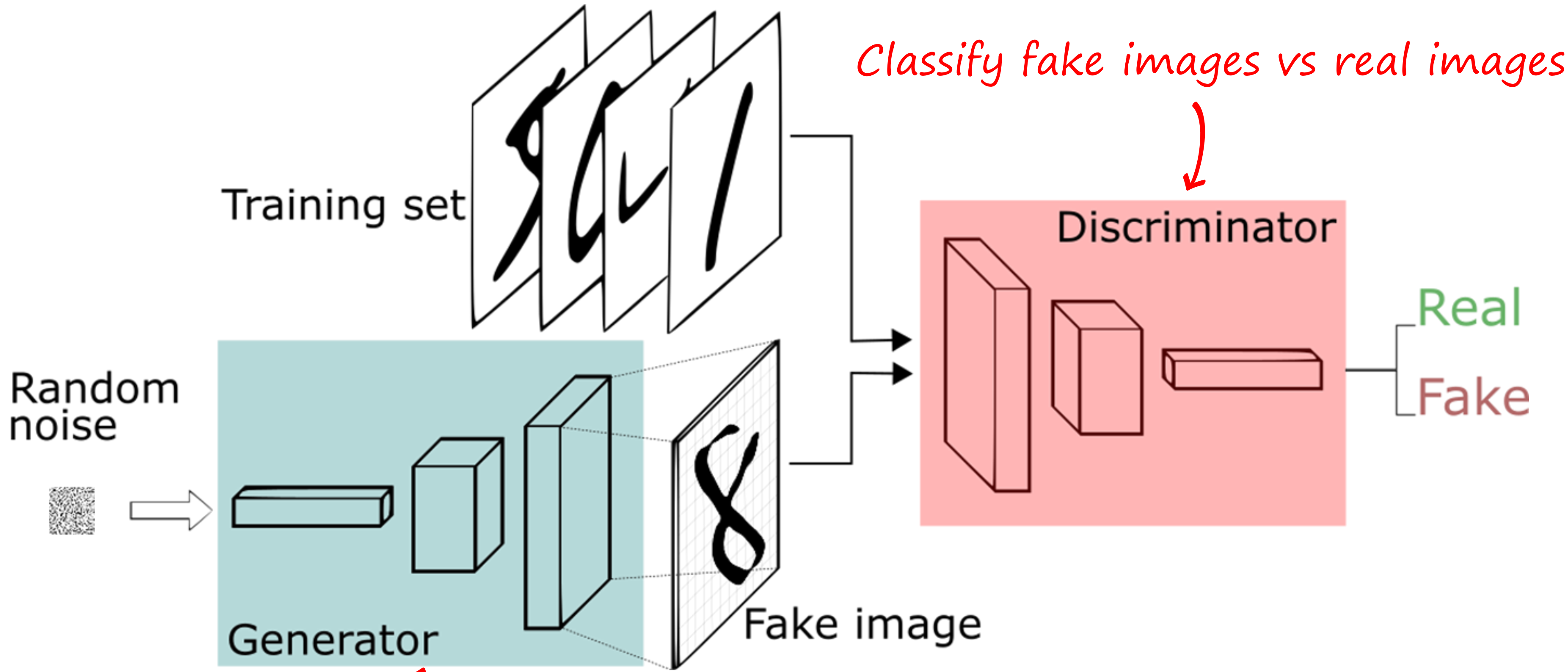


Paired image-to-image translation



Generative Adversarial Networks

Classify fake images vs real images



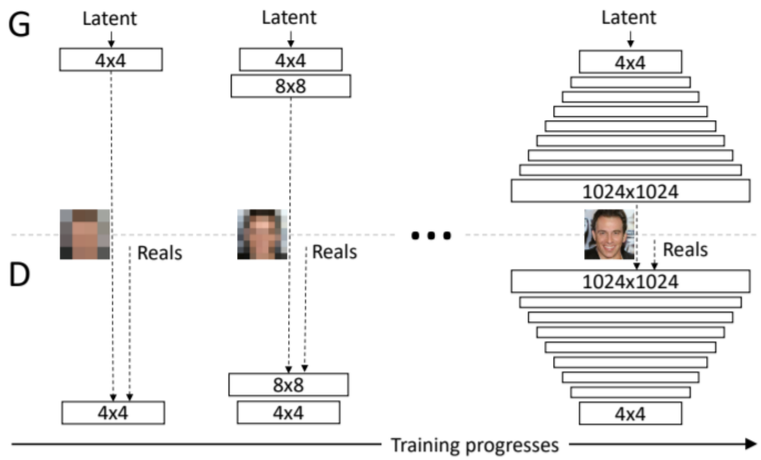
Generate fake samples to fool the discriminator

Goodfellow et al., "Generative Adversarial Networks", NIPS 2014
Figure from https://deeplearning4j.org/generative-adversarial-network

Generative Adversarial Networks

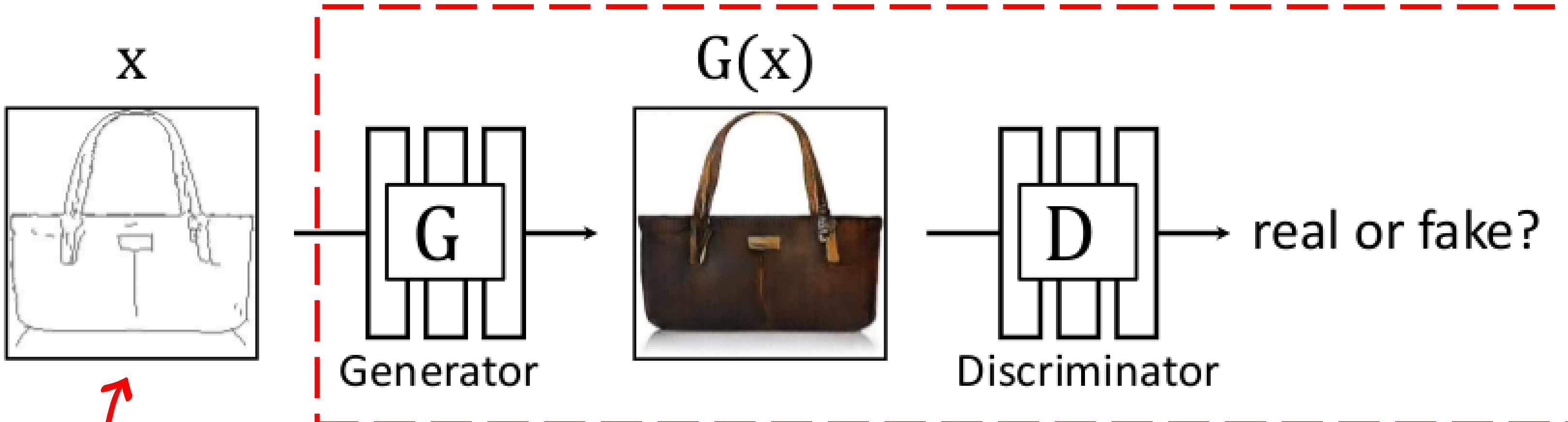
Wasserstein GAN (WGAN-GP)

Progressive growing of GANs



and many more…

pix2pix: image-to-image translation with conditional GANs



Input condition (instead of random noise)

This is still a GAN

G: generate fake samples that can fool D

D: classify fake samples vs. real images

Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017
Slide adapted from Zhu and Isola

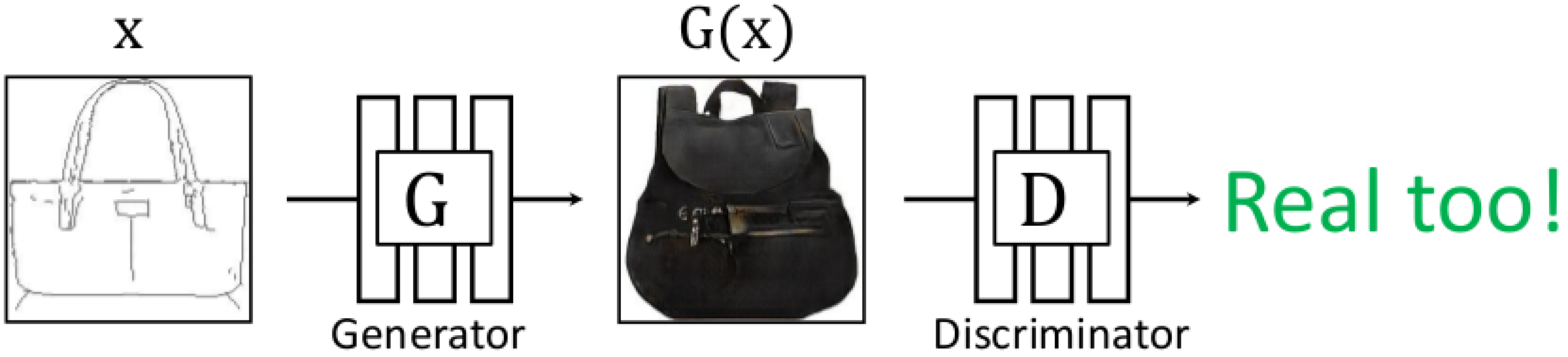
pix2pix: image-to-image translation with conditional GANs

$$\min_G \max_D \mathbb{E}_{x,y}[\log D(G(x)) + \log(1 - D(y))]$$

Optimization problem in a conventional GAN

[Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017]
Slide adapted from Zhu and Isola

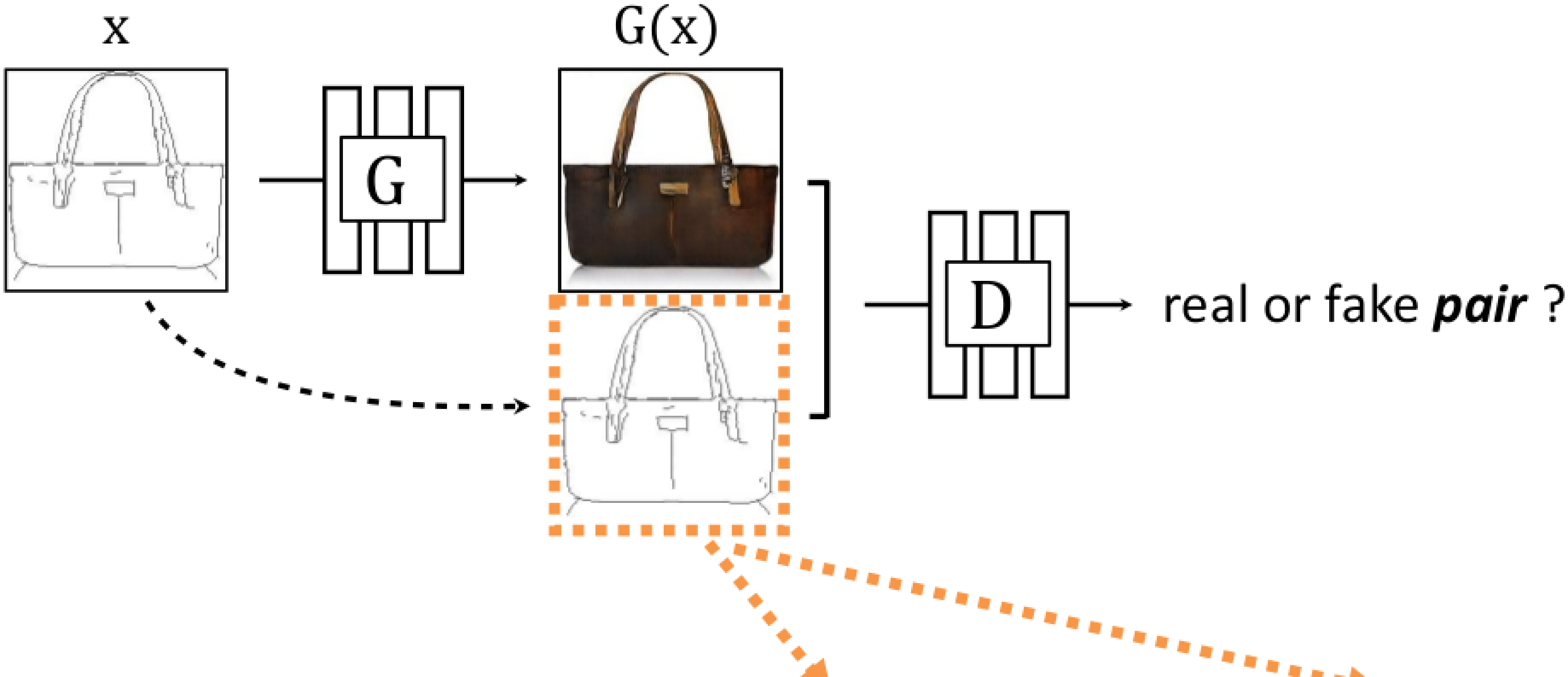
pix2pix: image-to-image translation with conditional GANs



$$\min_{G} \max_{D} \mathbb{E}_{x,y}[\log D(G(x)) + log(1 - D(y))]$$

[Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017]
Slide adapted from Zhu and Isola

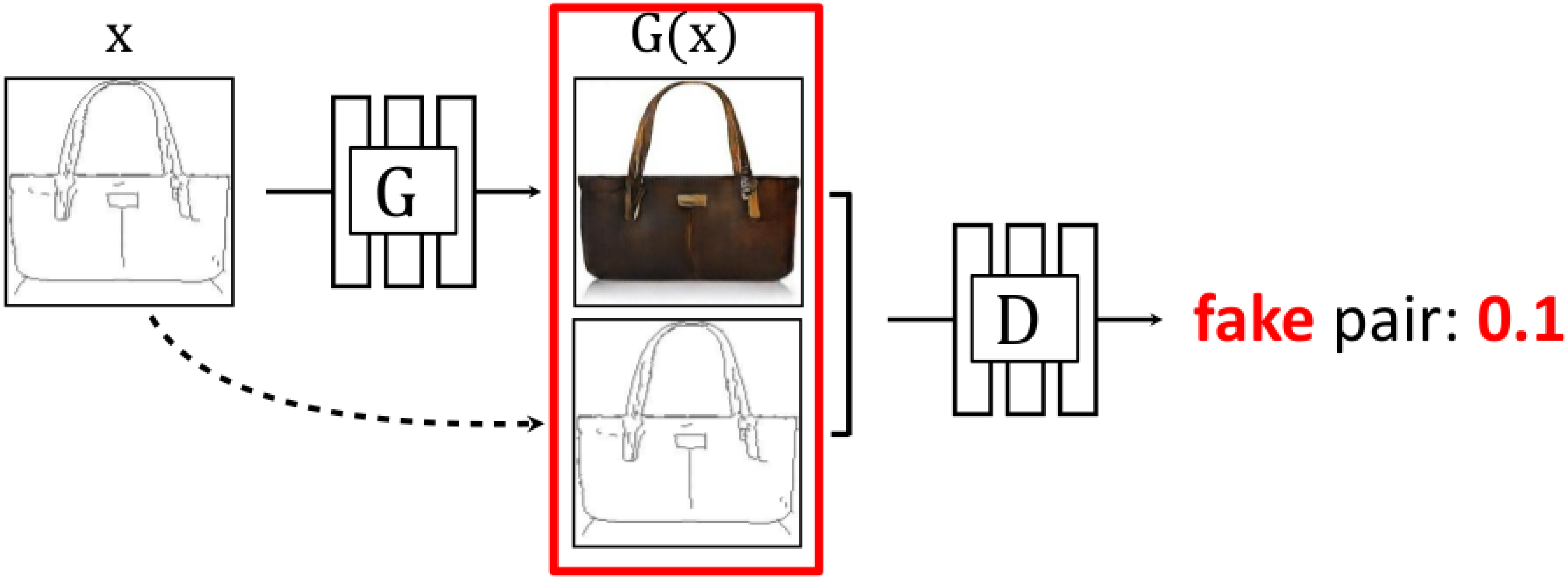
pix2pix: image-to-image translation with conditional GANs



$$\min_G \max_D \mathbb{E}_{x,y}[\log D(\boxed{x}, G(x)) + \log(1 - D(\boxed{x}, y))]$$

Isola et al., “Image-to-Image Translation with Conditional Adversarial Networks”, CVPR 2017
Slide adapted from Zhu and Isola

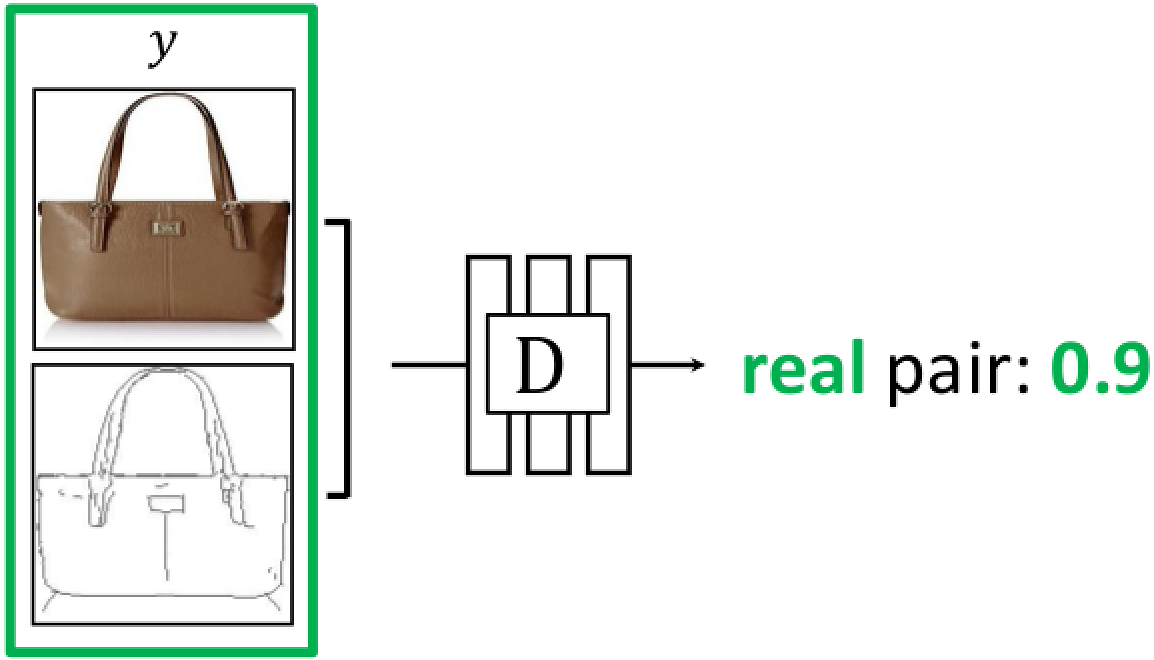
pix2pix: image-to-image translation with conditional GANs



$$\min_G \max_D \mathbb{E}_{x,y}[\boxed{\log D(x, G(x))} + \log(1 - D(x, y))]$$

[Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017](#)
Slide adapted from Zhu and Isola

pix2pix: image-to-image translation with conditional GANs



$$\min_G \max_D \mathbb{E}_{x,y}[\log D(x, G(x)) + \boxed{\log(1 - D(x, y))}]$$

[Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017](#)
Slide adapted from Zhu and Isola

pix2pix: image-to-image translation with conditional GANs

- Training details
  - Conditional GAN + L1

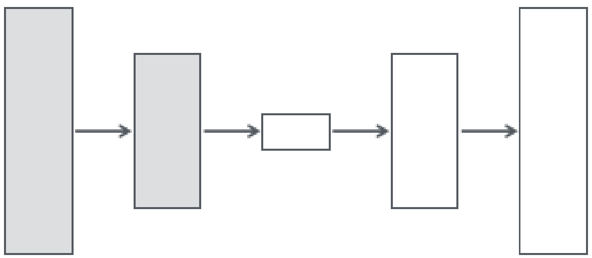
$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G).$$

  - Stable training + fast convergence.

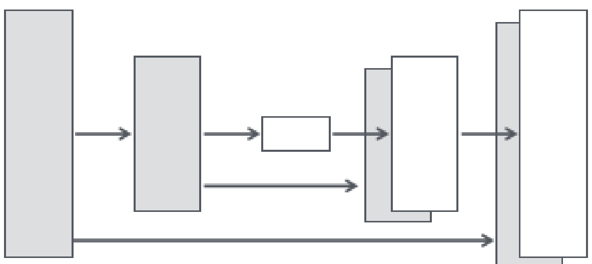
[Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017](#)
Slide adapted from Zhu and Isola

pix2pix: image-to-image translation with conditional GANs

Generator



Encoder-decoder

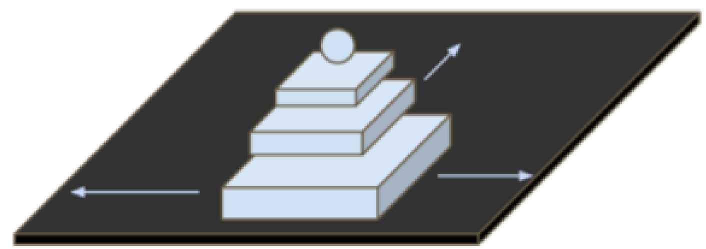


U-Net

[Ronneberger et al.]

Discriminator

PatchGAN (FCN)



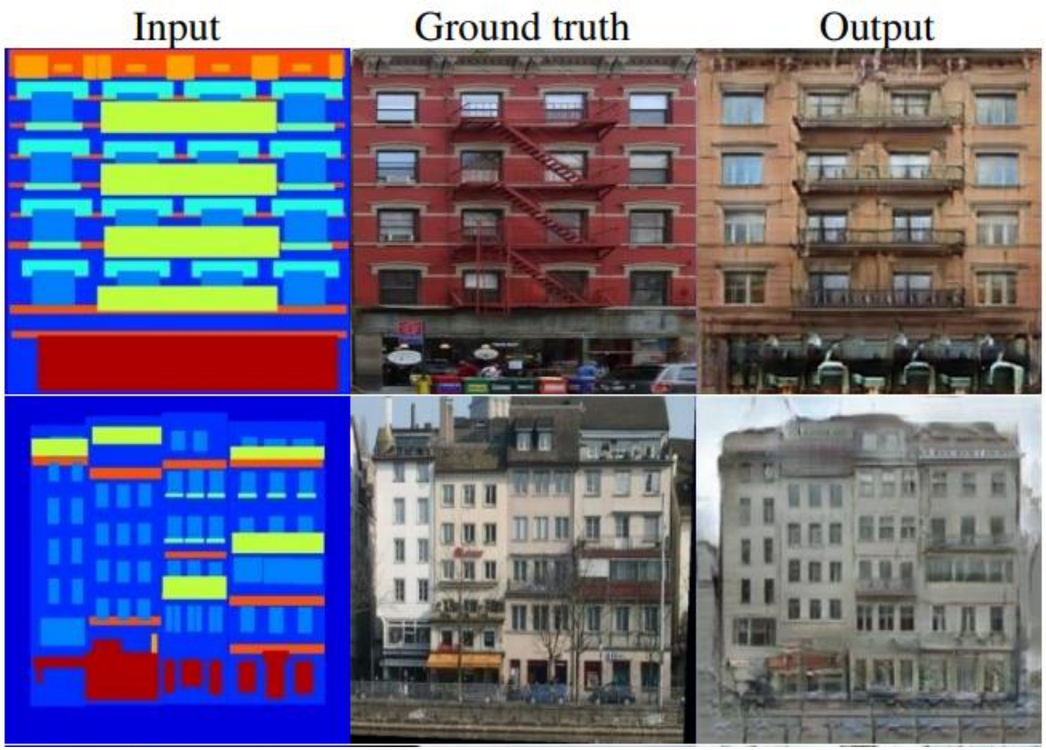
Skip connections (decoder conditioned on encoder activations)

[Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017]()

Slide adapted from Zhu and Isola

pix2pix: image-to-image translation with conditional GANs

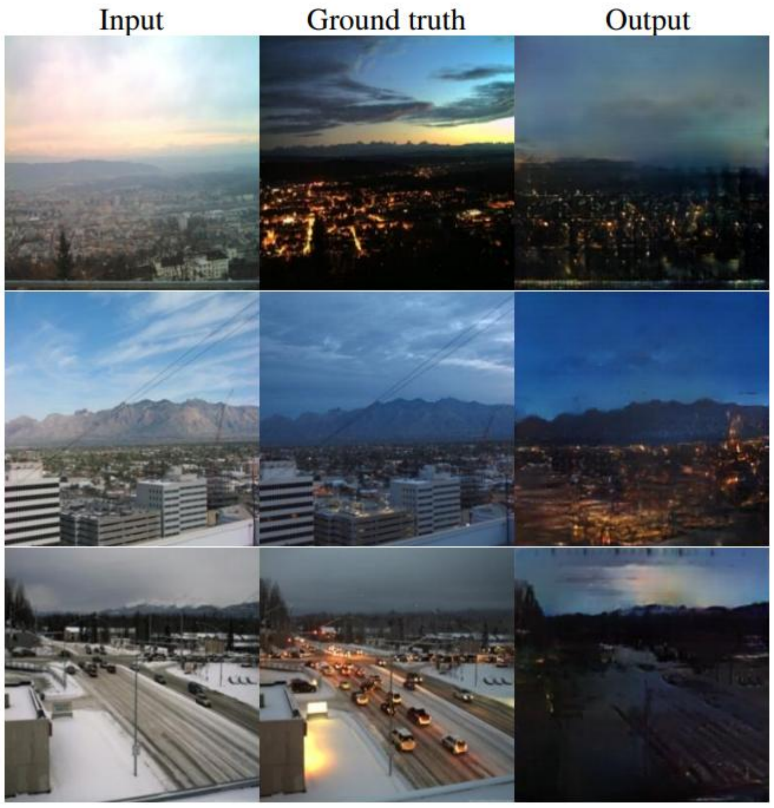
- Examples



Isola et al, "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017
Figures from https://affinelayer.com/pix2pix/

pix2pix: image-to-image translation with conditional GANs

- Examples



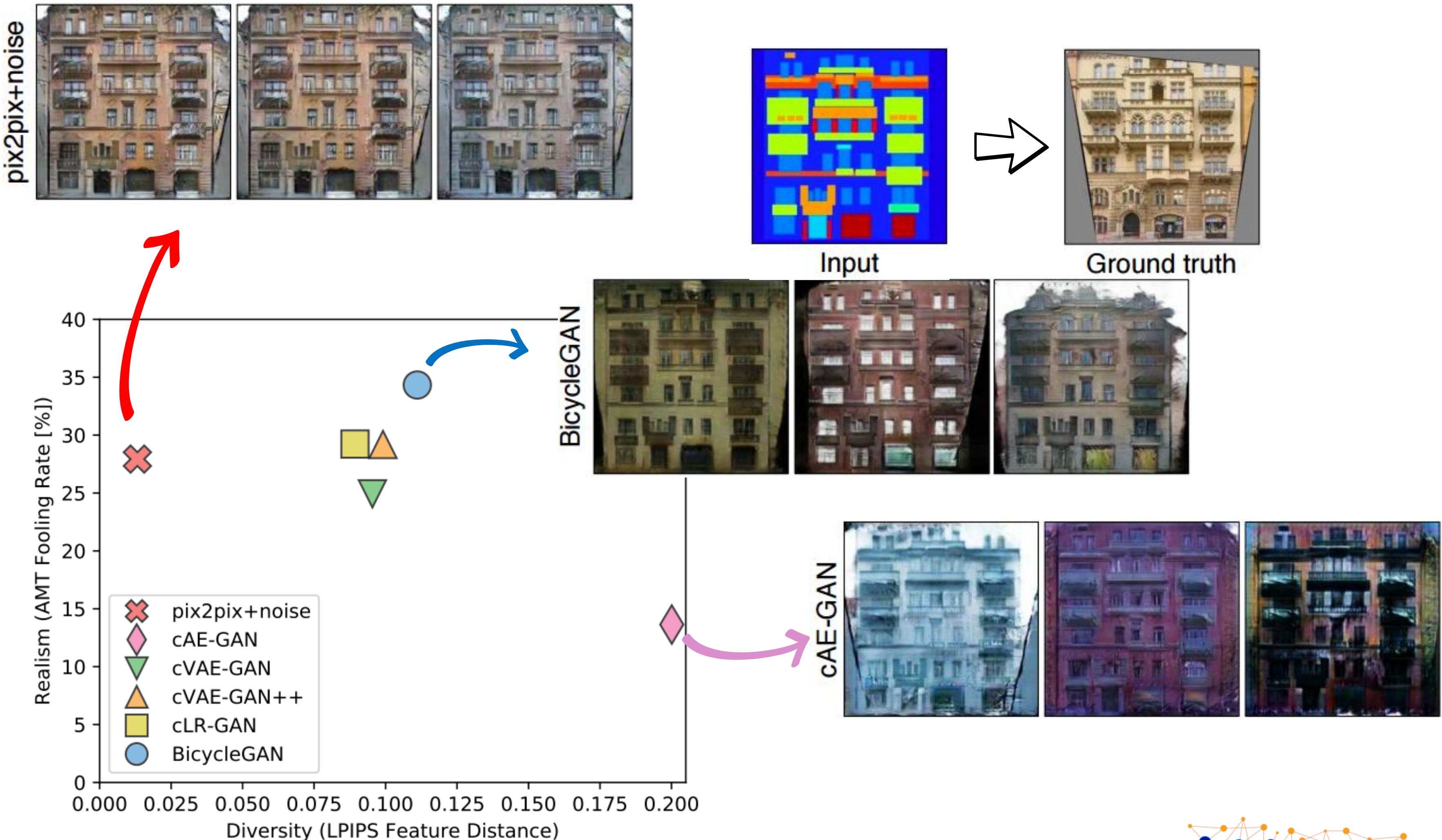
Isola et al, "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017
Figures from https://affinelayer.com/pix2pix/

pix2pix: image-to-image translation with conditional GANs

- Encoder-decoder (w/o skip) vs UNet (w/ skip)
- Loss: L1 vs L1+cGAN

[Chen et al, "VGAN-Based Image Representation Learning for Privacy-Preserving Facial Expression Recognition", arxiv 2018](#)

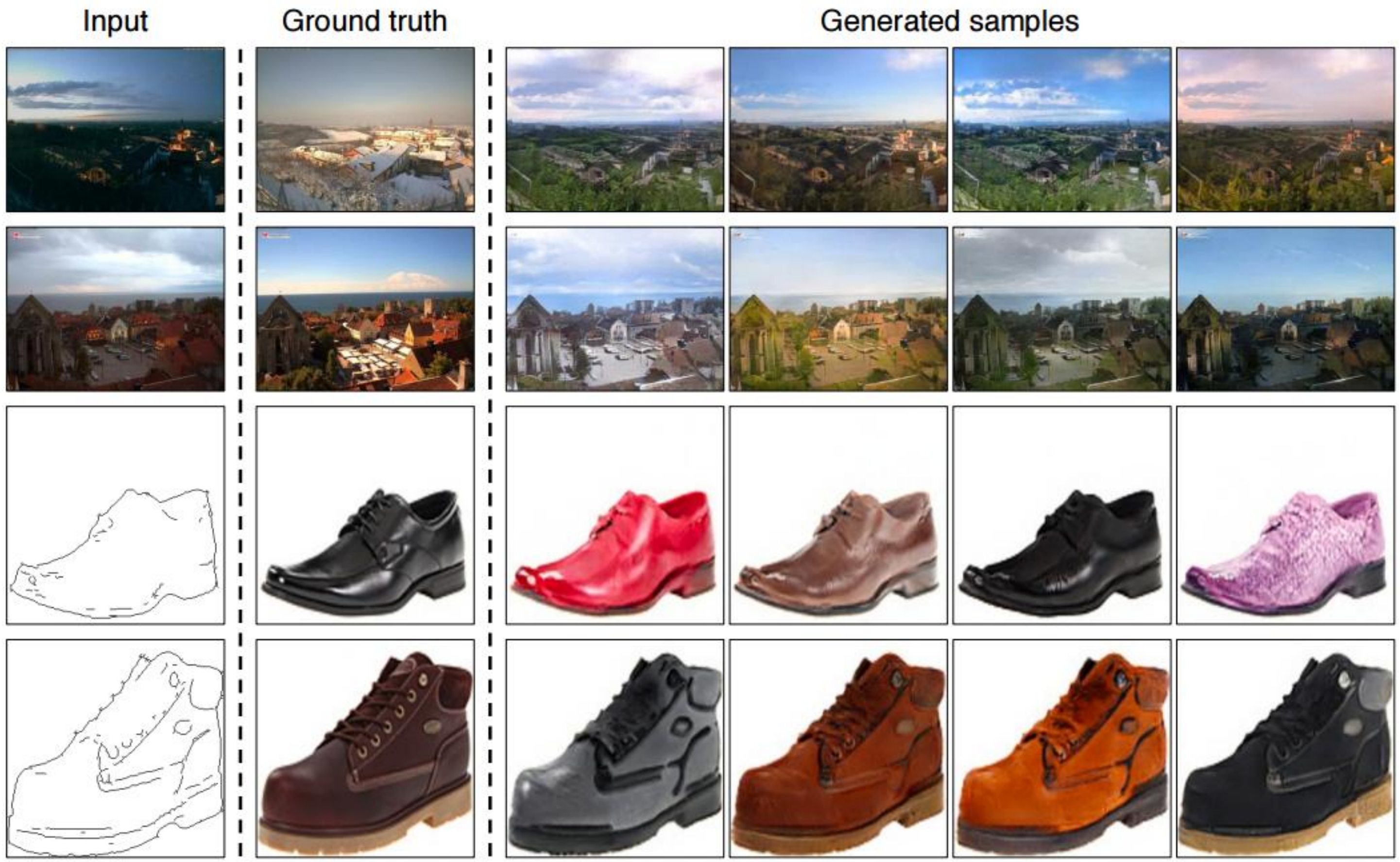
Diversity in image-to-image translation



Zhu et al. "Toward Multimodal Image-to-Image Translation", NIPS 2017

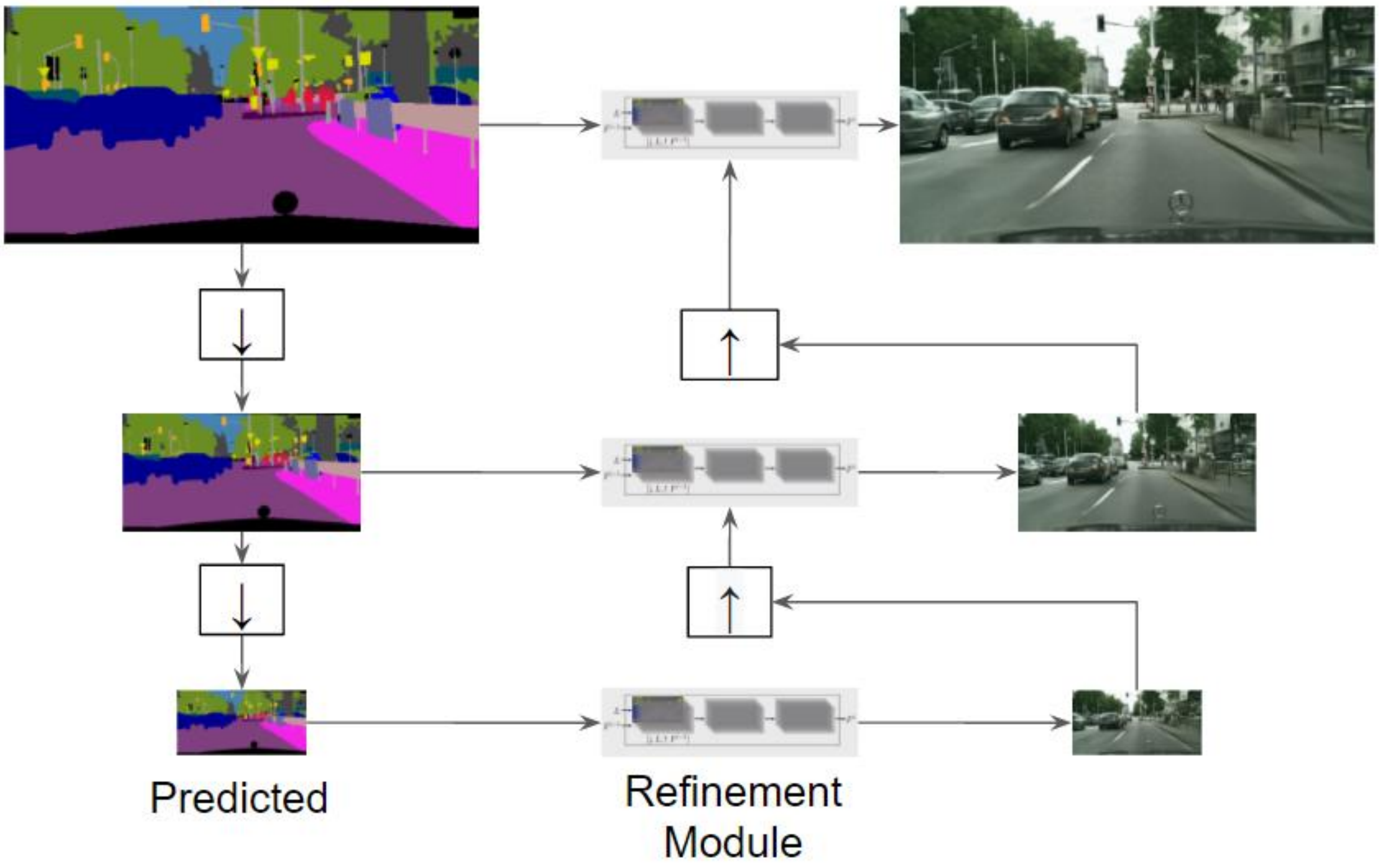
Diversity in image-to-image translation

- More results. Bicycle GAN



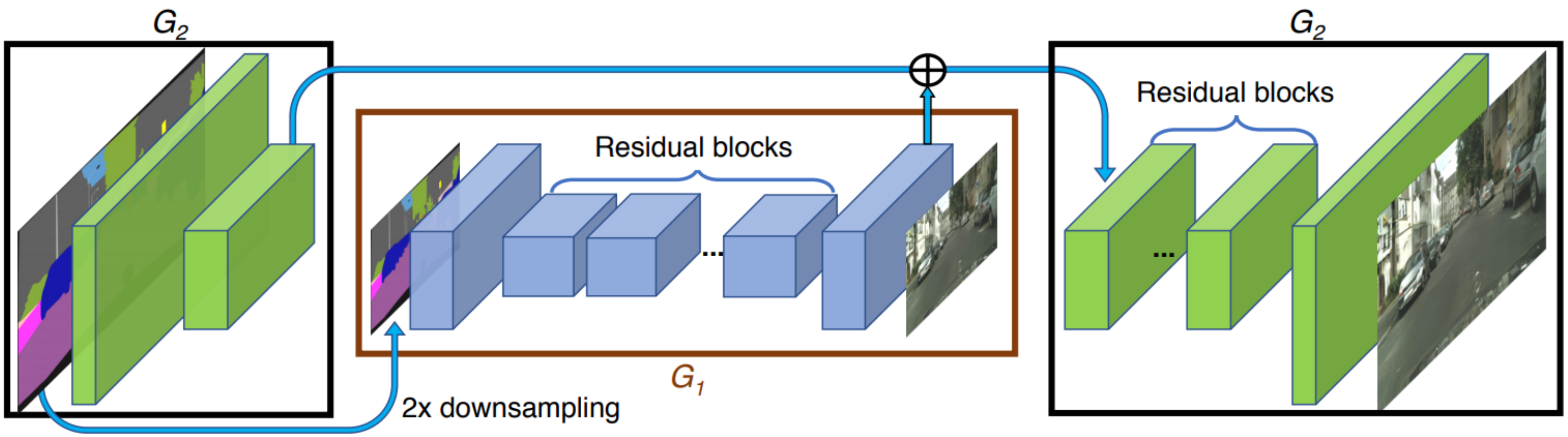
Zhu et al. “Toward Multimodal Image-to-Image Translation”, NIPS 2017

Cascade refinement networks



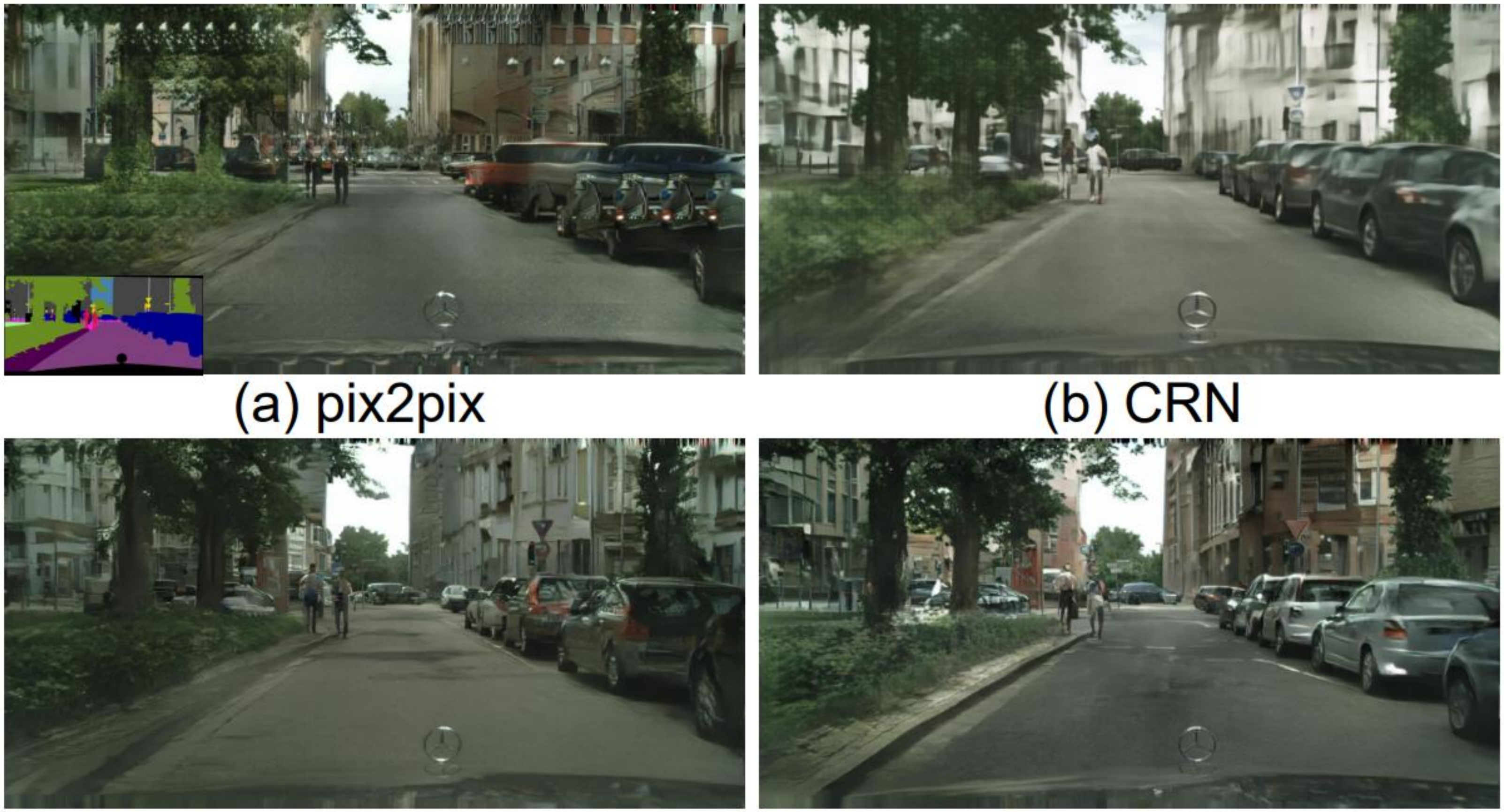
[Chen and Koltun, "Photographic Image Synthesis with Cascaded Refinement Networks", ICCV 2017]

pix2pixHD



[Wang et al, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs", arxiv 2017](#)

Comparison image-to-image translation

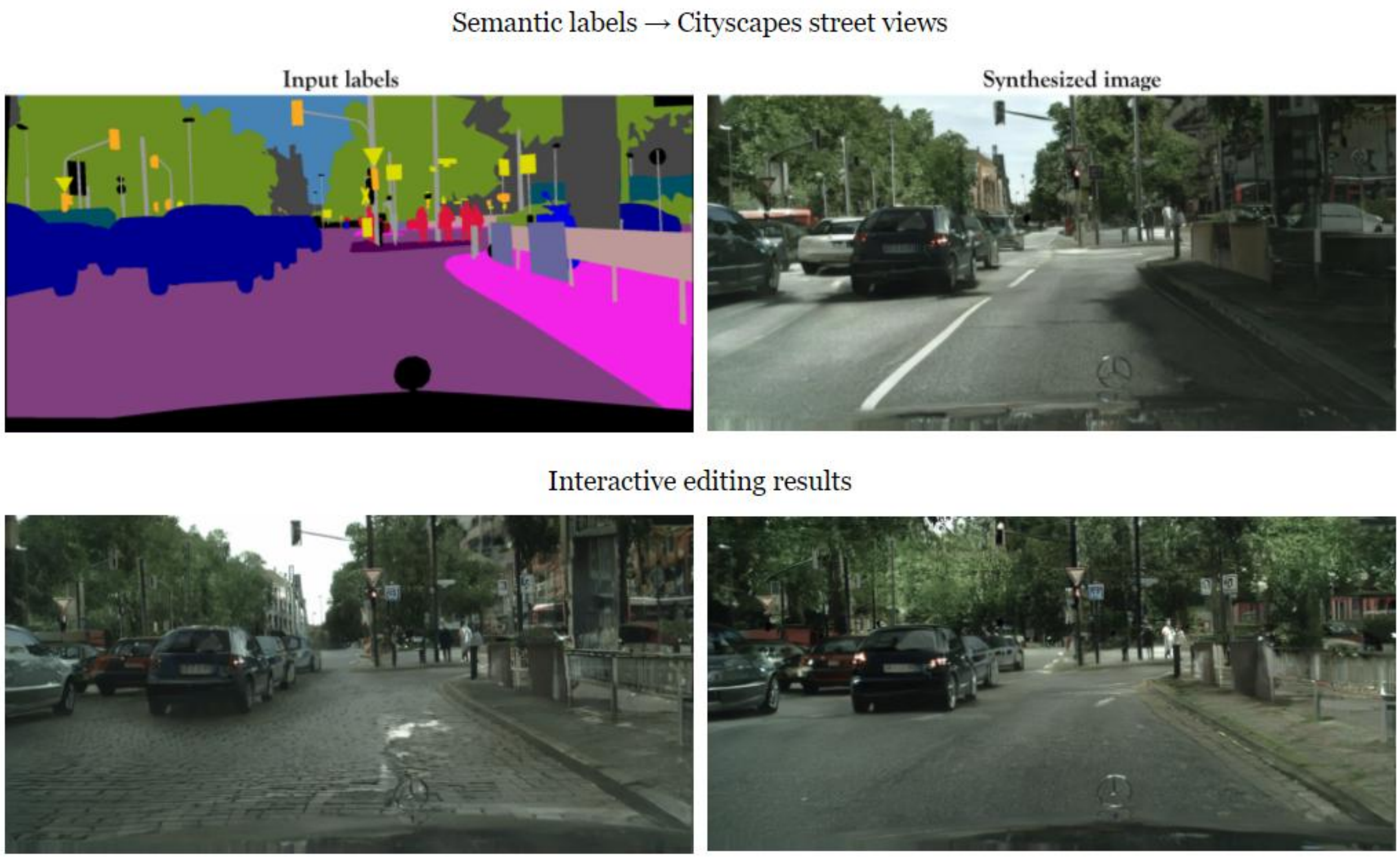


(a) pix2pix (b) CRN

(c) Ours (w/o VGG loss) (d) Ours (w/ VGG loss)

Wang et al, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs", arxiv 2017

pix2pixHD: interactive image-to-image translation

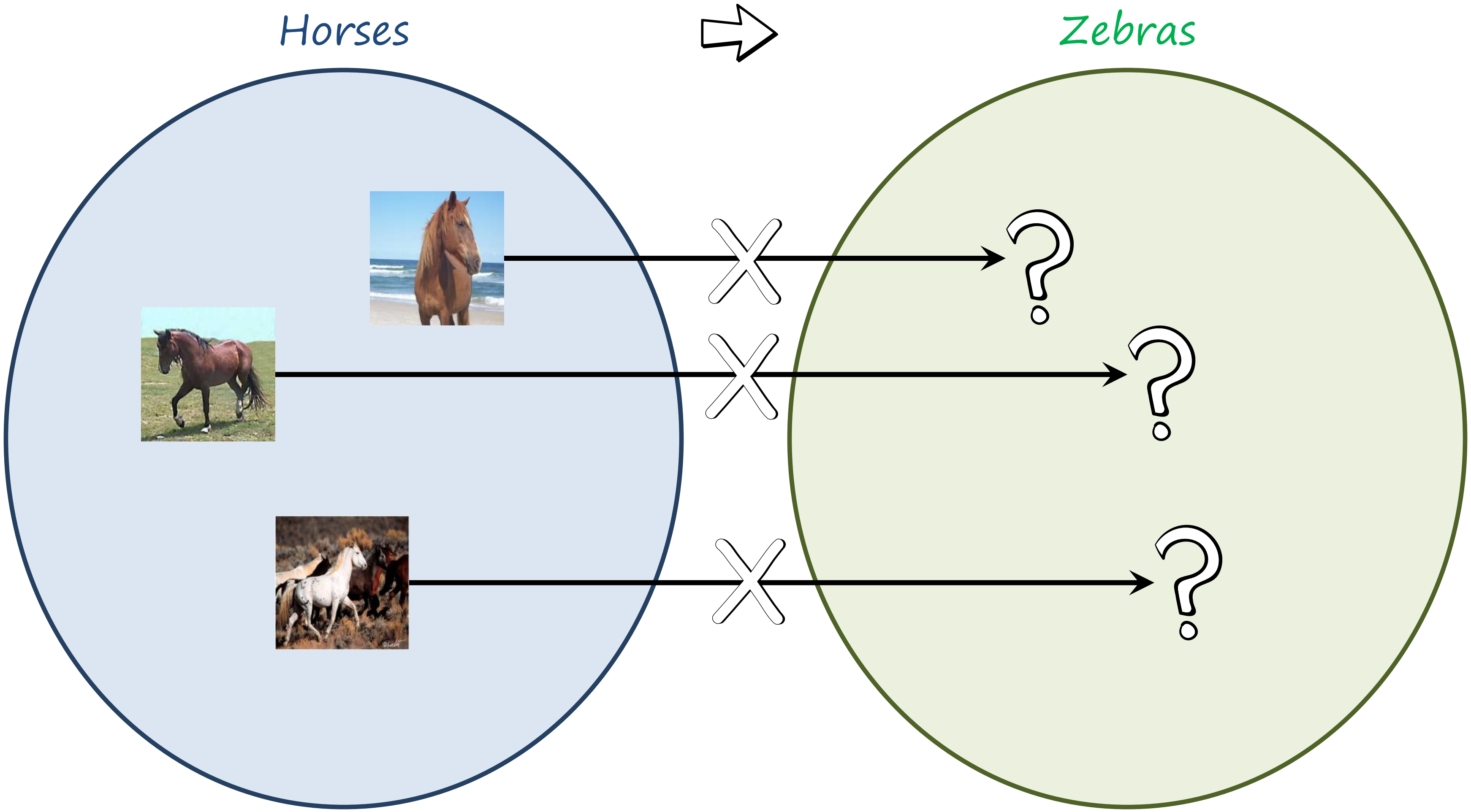


[Wang et al, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs", arxiv 2017]

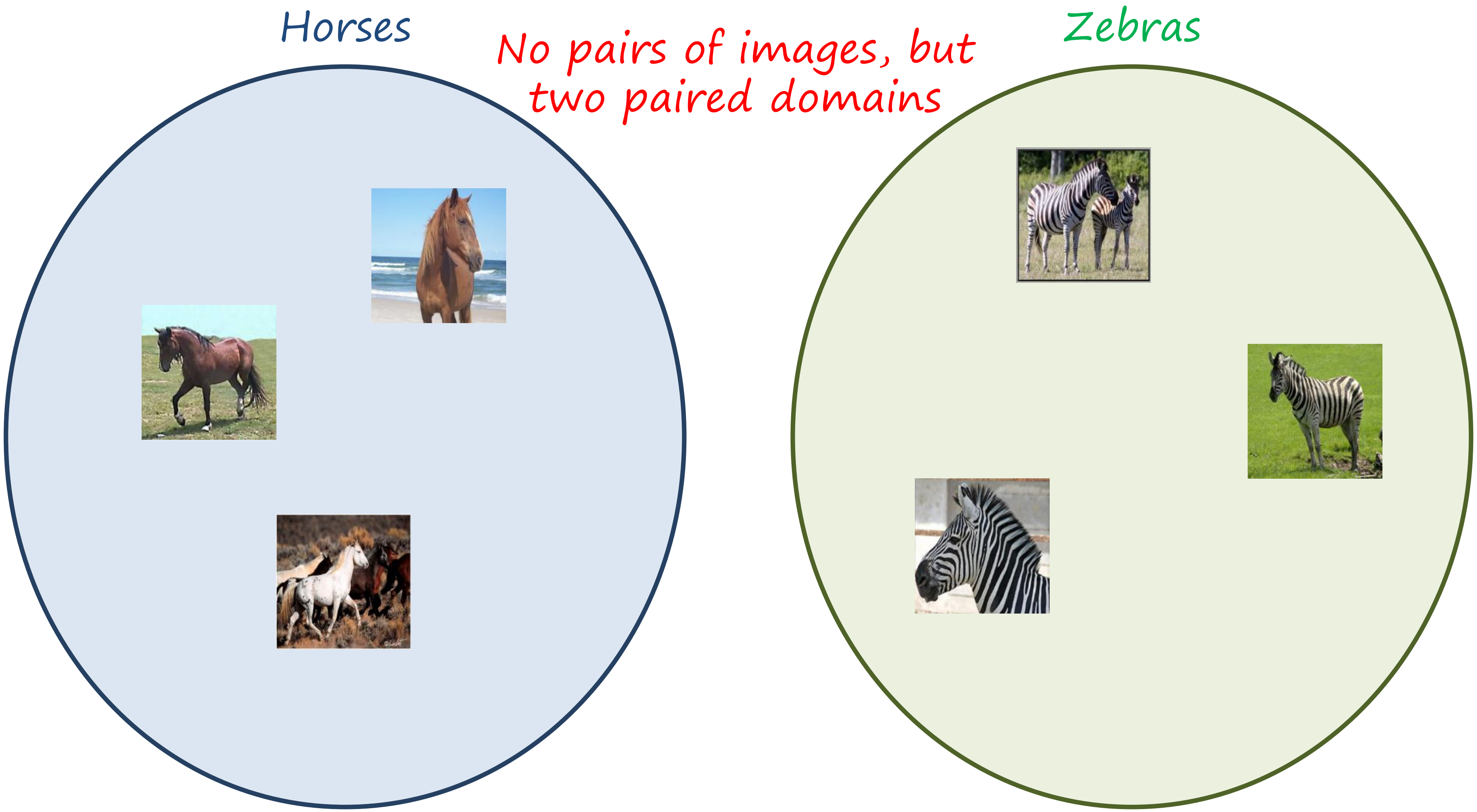
Outline

- M2CR project framework
- Paired image-to-image translation (pix2pix)
- Unpaired image-to-image translation (cycleGAN)
- Unseen translations (mix&match networks)

Unpaired image-to-image translation

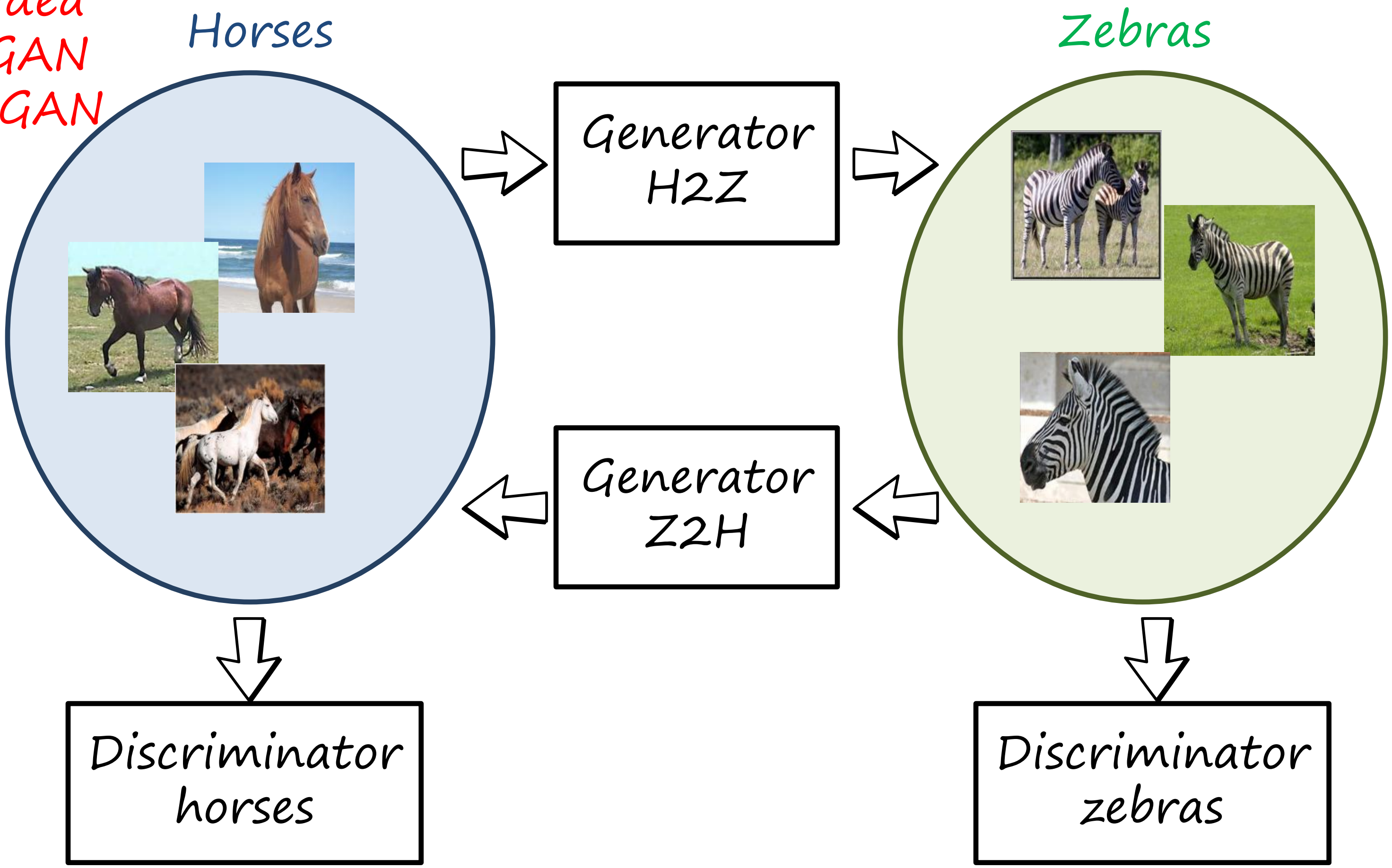


Unpaired image-to-image translation



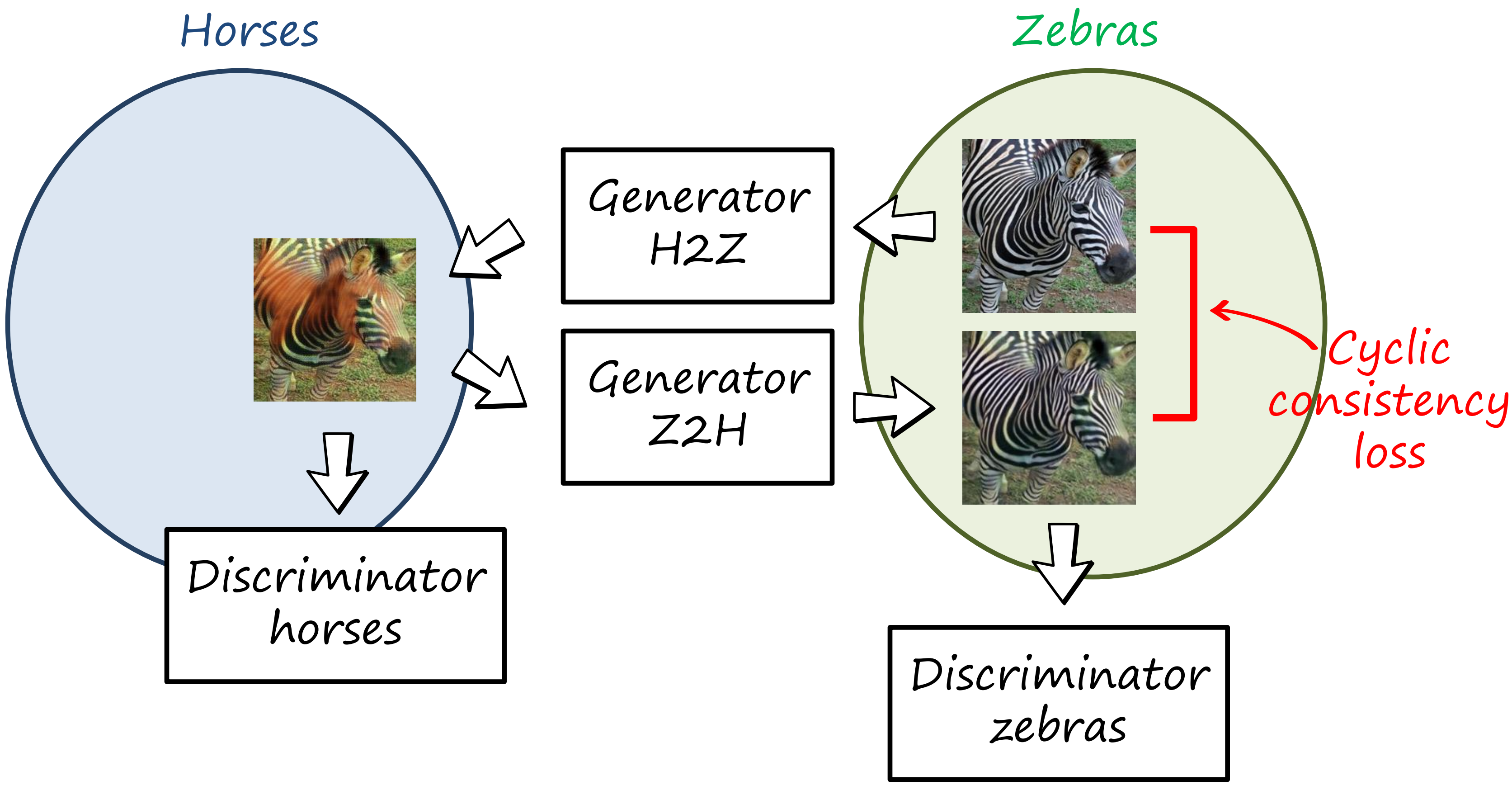
CycleGAN: unpaired image-to-image translation using cycle consistency

Similar idea in DiscoGAN and DualGAN



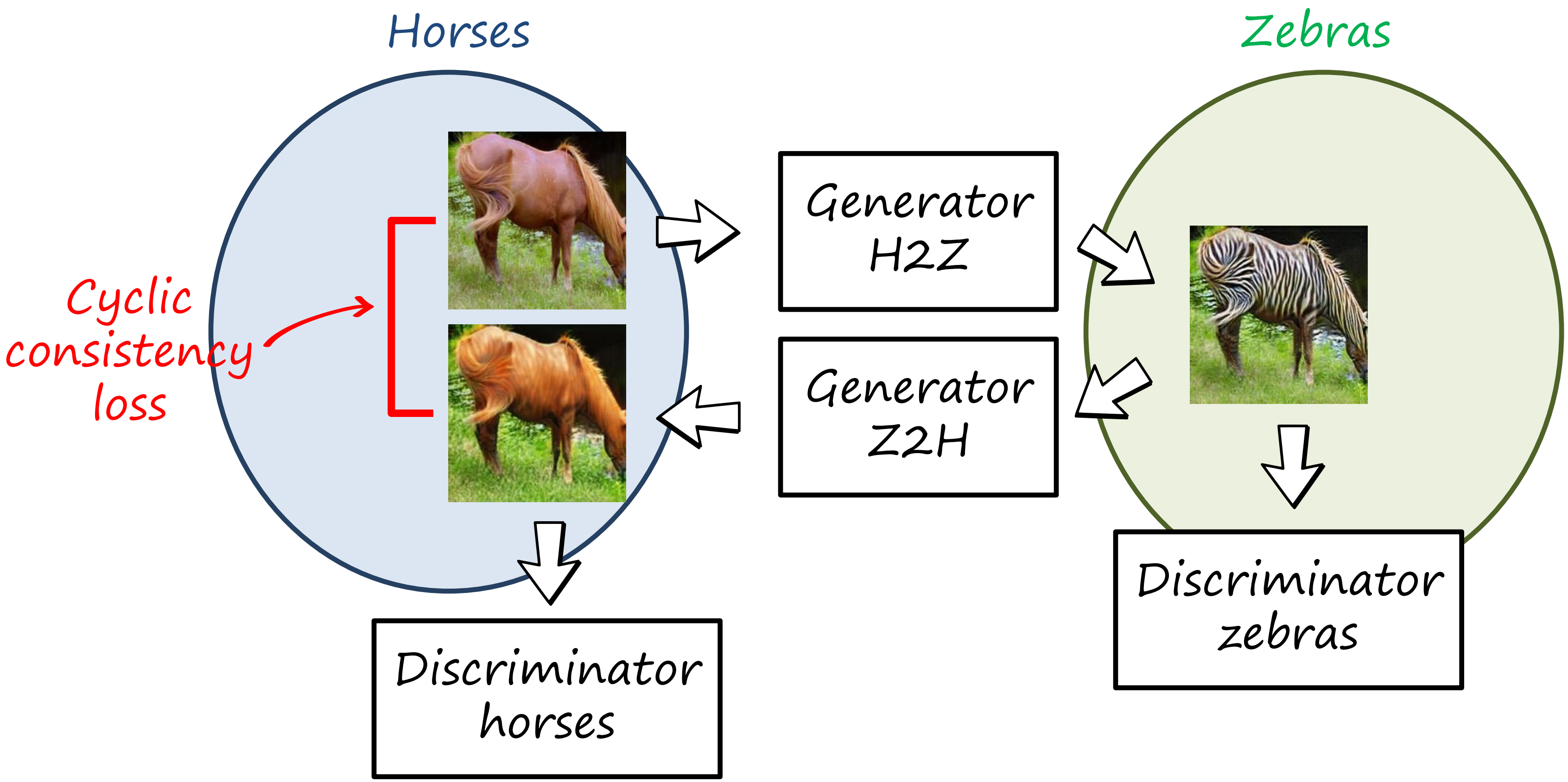
[Zhu et al, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", ICCV 2017]

CycleGAN: unpaired image-to-image translation using cycle consistency



[Zhu et al, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", ICCV 2017]

CycleGAN: unpaired image-to-image translation using cycle consistency

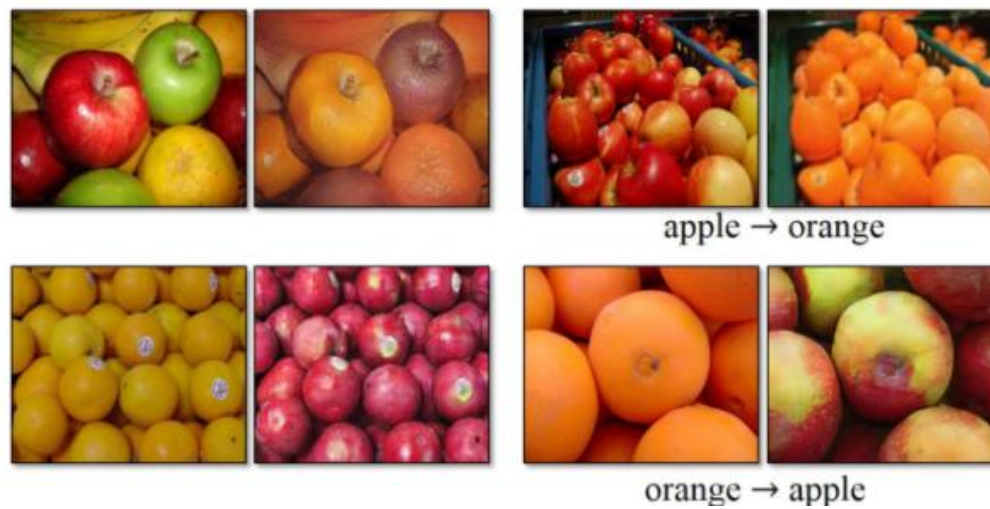


Zhu et al, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", ICCV 2017

CycleGAN: unpaired image-to-image translation using cycle consistency

- Results

Monet Paintings to Photos

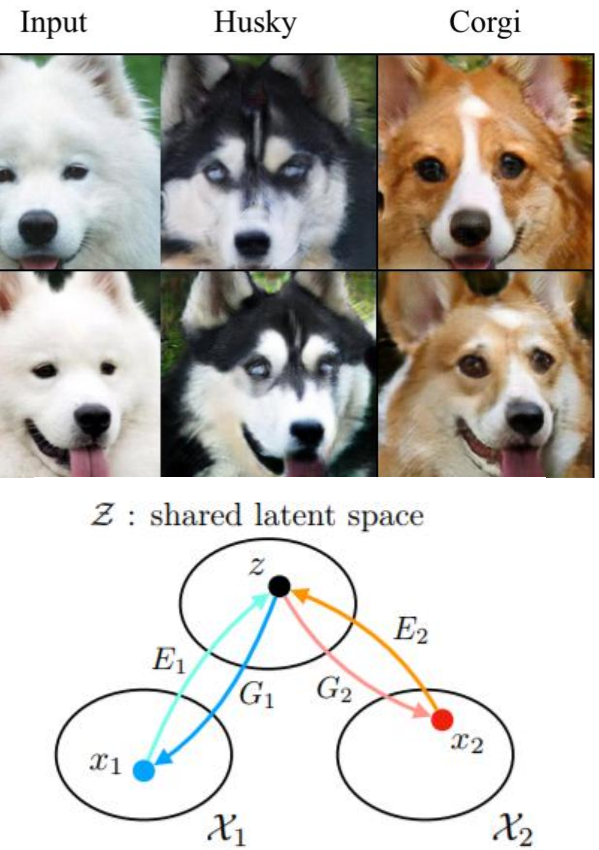


[Zhu et al, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", ICCV 2017]()

More unpaired image translation

Domain Transfer Network (DTN)

UNIT



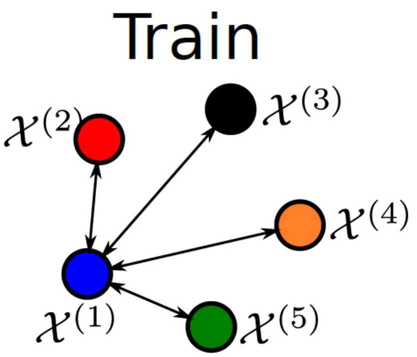
and many more…

Outline

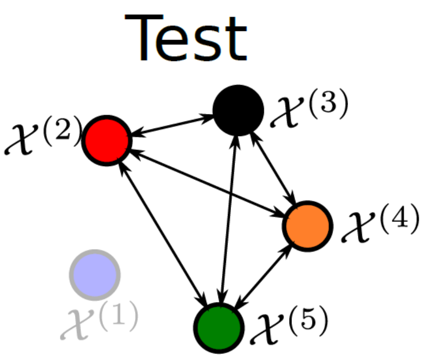
- M2CR project framework
- Paired image-to-image translation (pix2pix)
- Unpaired image-to-image translation (cycleGAN)
- Unseen translations (mix&match networks)

Unseen translations

Only these transformations are trained



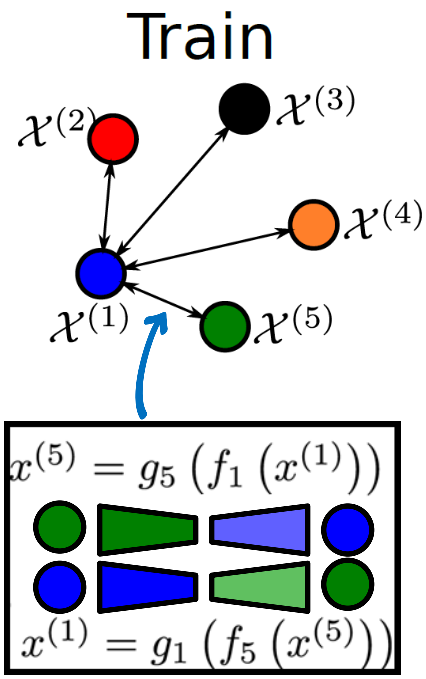
Evaluate on these ***unseen*** *translations*

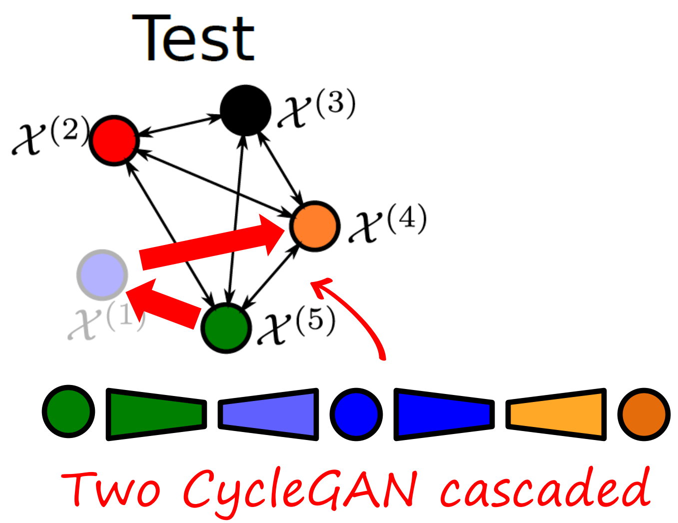


[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]()

Cascading image-to-image translators

Image-to-image (e.g. CycleGAN)



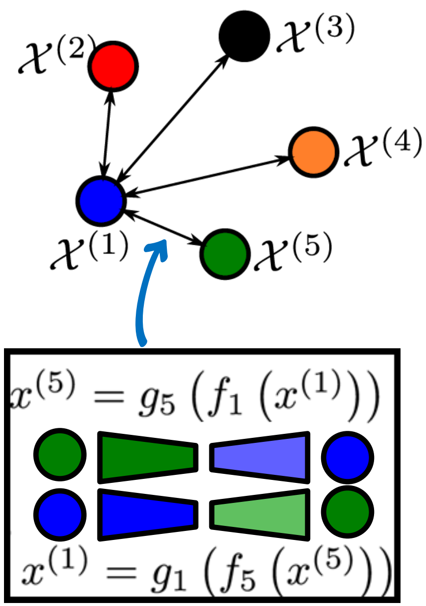


[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]()

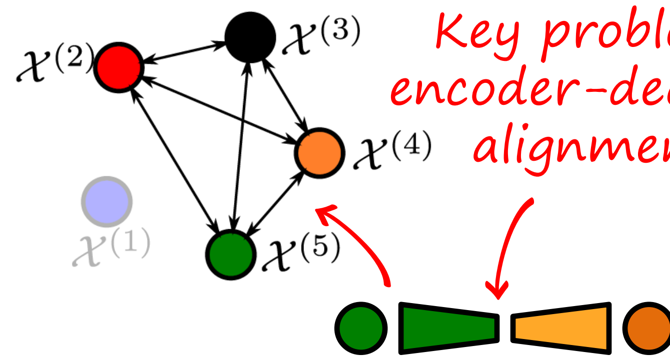
Mix and match networks

Image-to-image (e.g. CycleGAN)

Train



Test



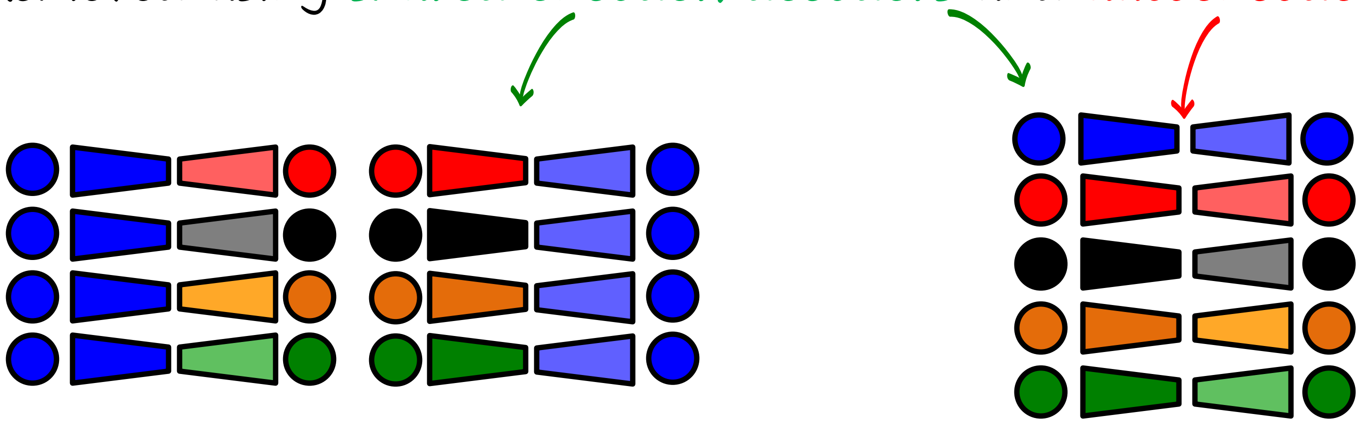
Key problem:
encoder-decoder
alignment

Mix&match encoder-decoders
(they haven't seen each other
during training)

Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018

Mix and match networks

Unseen encoder-decoder alignment
- Scalable: number of networks O(N)
- Latent representation should be domain-independent
- Achieved using shared encoder/decoders and autoencoders

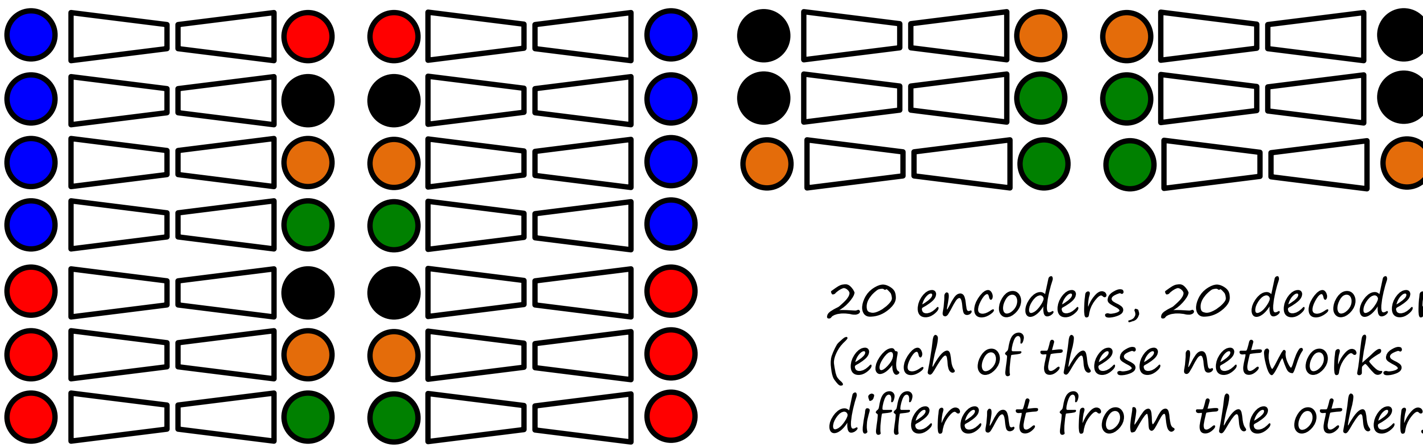


5 encoders, 5 decoders

Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018

Training all possible translators

Since it is unpaired, we could train all possible translators. Problems:
- No sharing
- Poor scalability: number of networks $O(N^2)$

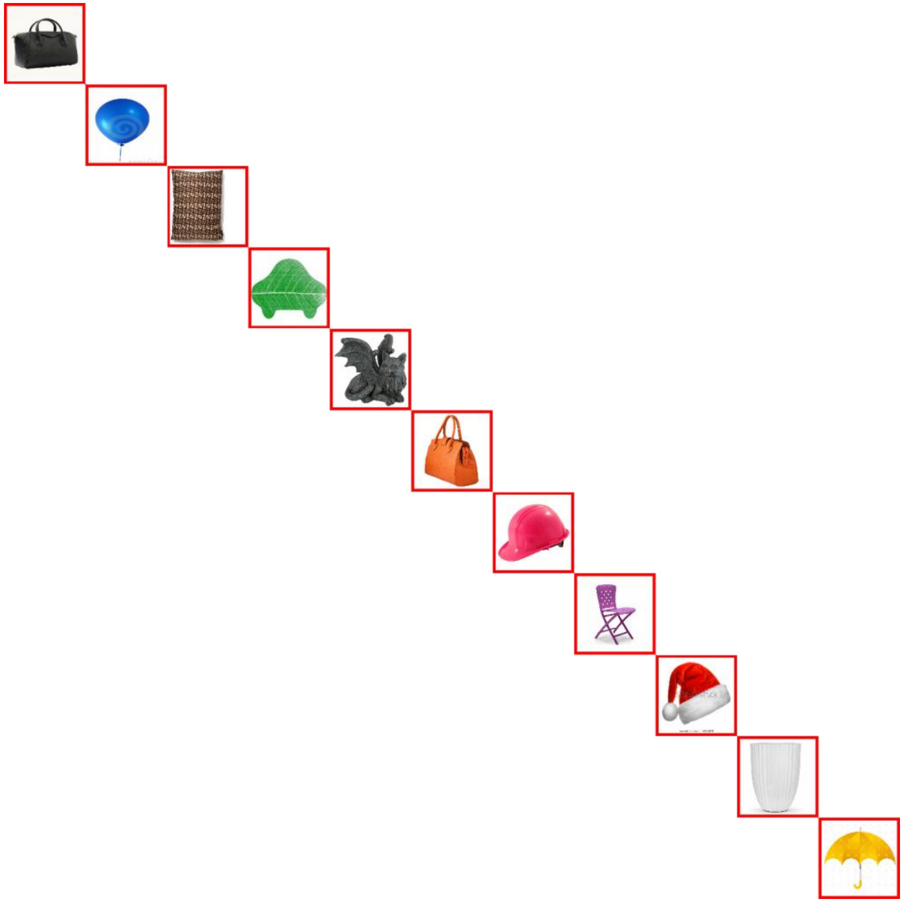


20 encoders, 20 decoders
(each of these networks is
different from the others)

Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018

Example: scalable recolorization

Unpaired translation
Eleven colors (i.e. domains)

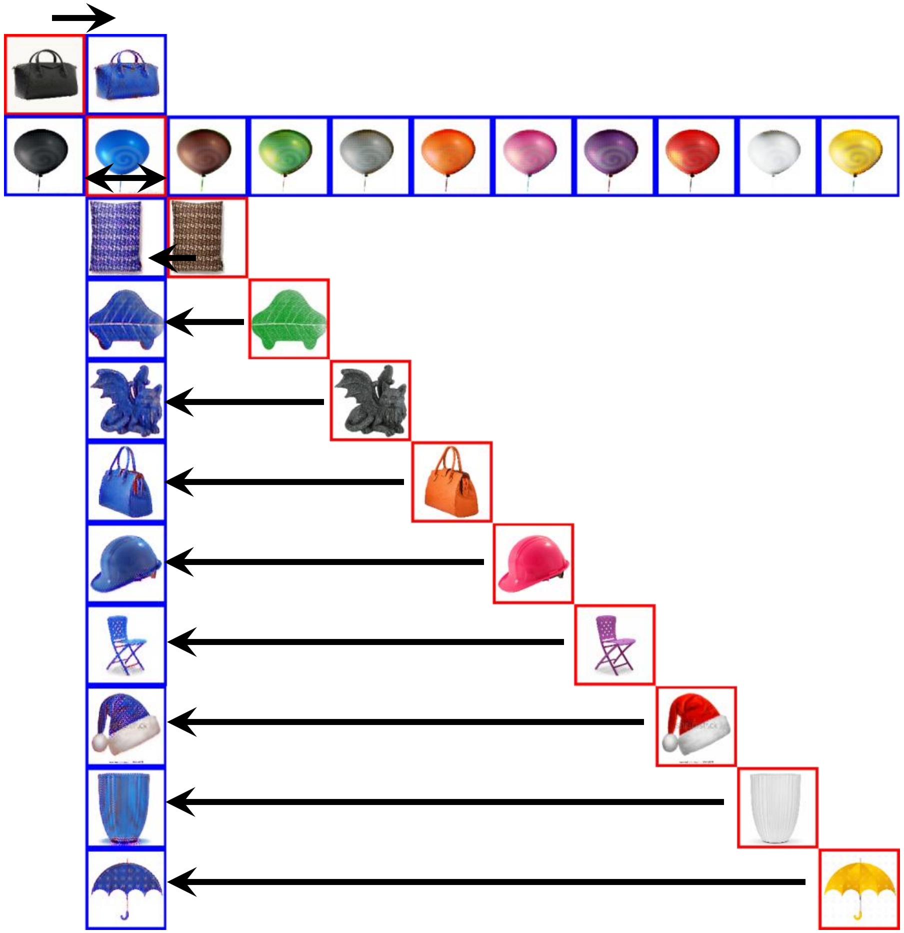


[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]

Example: scalable recolorization

Unpaired translation
Eleven colors (i.e. domains)

Requires training 10 encoders and 10 decoders



[Wang, van de Weijer, Herranz, “Encoder-decoder alignment for zero-pair image-to-image translation”, CVPR 2018]()

Example: scalable recolorization

Unpaired translation
Eleven colors (i.e. domains)

Requires training 10 encoders and 10 decoders

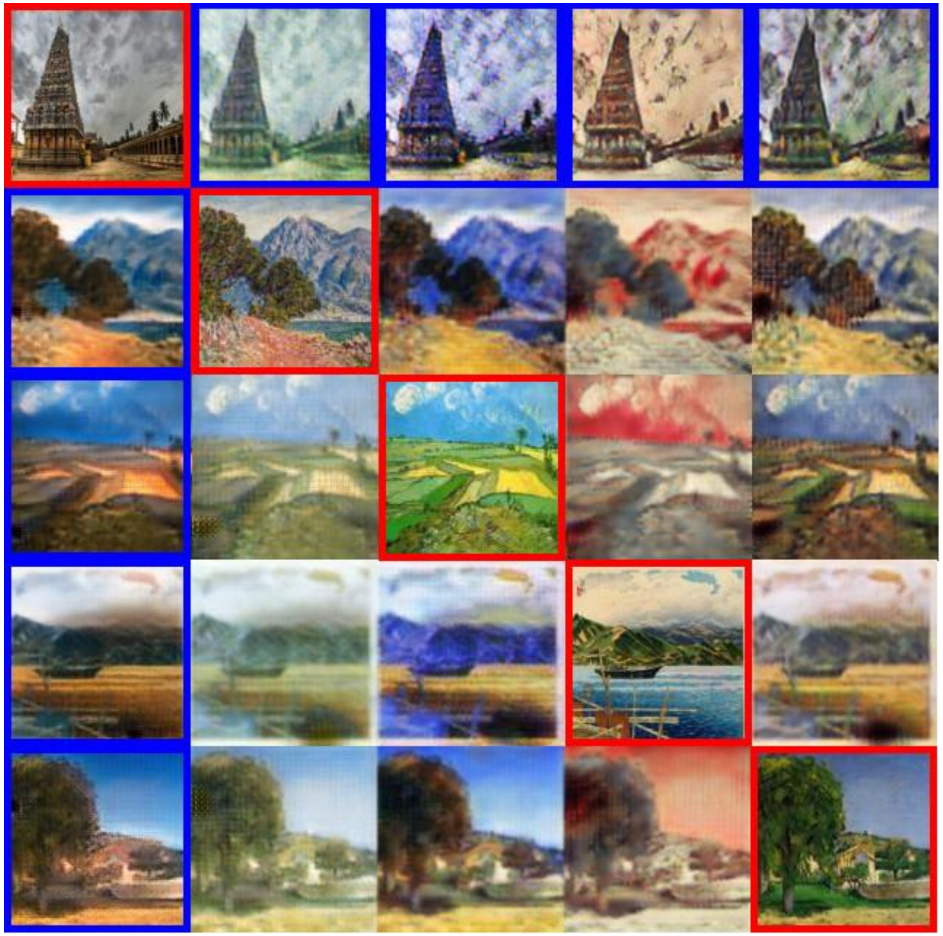
CycleGANs for all combinations would require 55 encoders and 55 decoders

[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]

Example: scalable style transfer

Unpaired translation
Five domains
(photo, Monet, van Gogh, Ukiyo-e, Cezanne)

(4 encoders and 4 decoders)



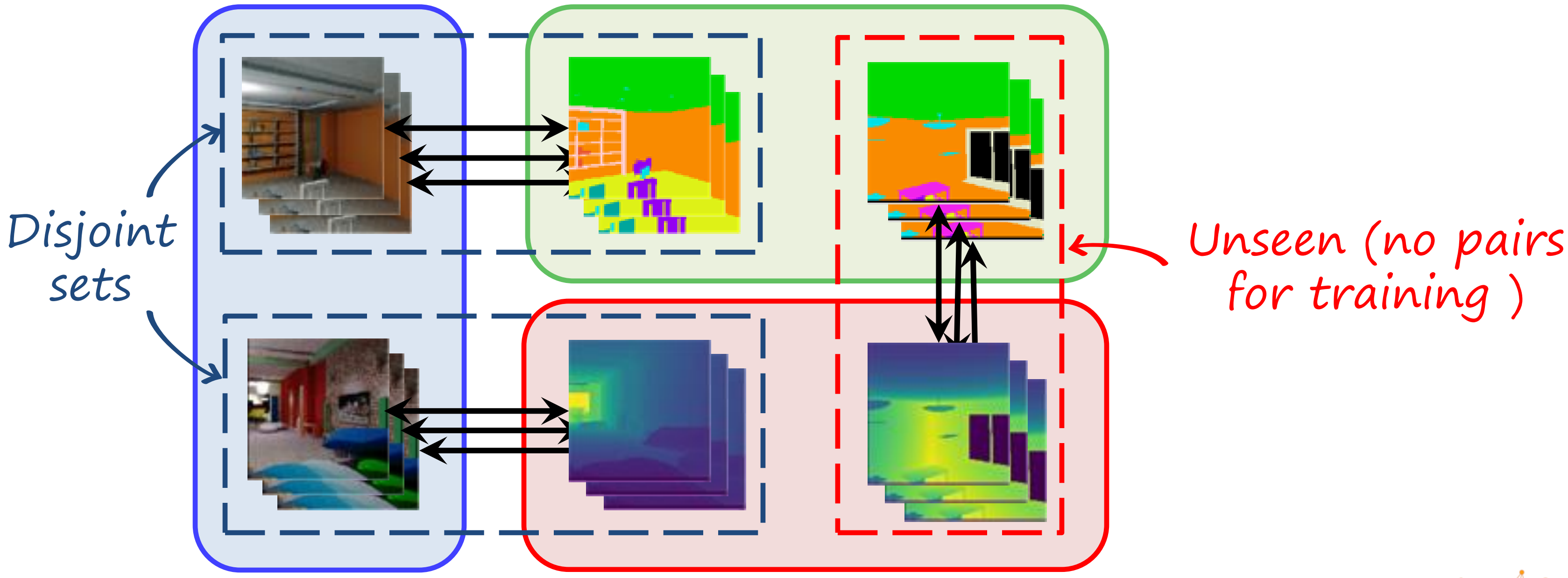
[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]()

Zero-pair translation

Cross-modal translation setting

Paired data available for (RGB, depth) and (RGB, segm.)

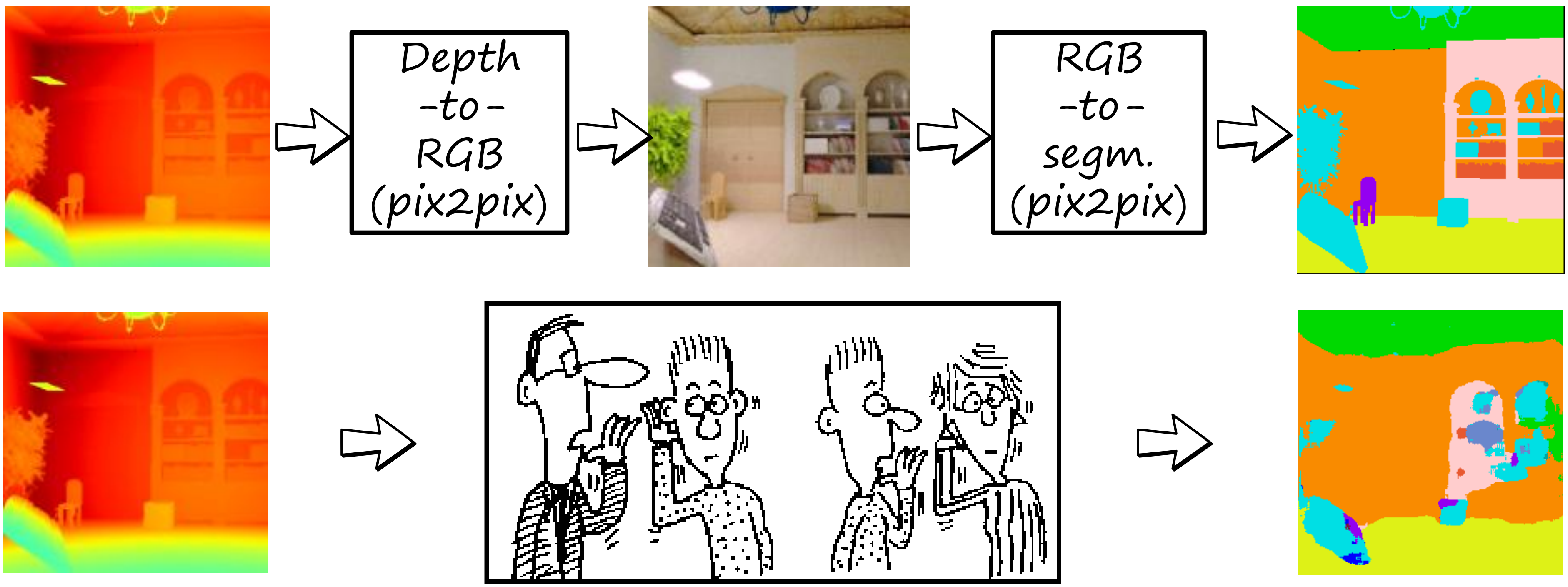
Evaluate on the **unseen zero-pair translations** (depth, segm.)



Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018

Zero-pair translation with two cascaded pix2pix (paired translations)

In theory

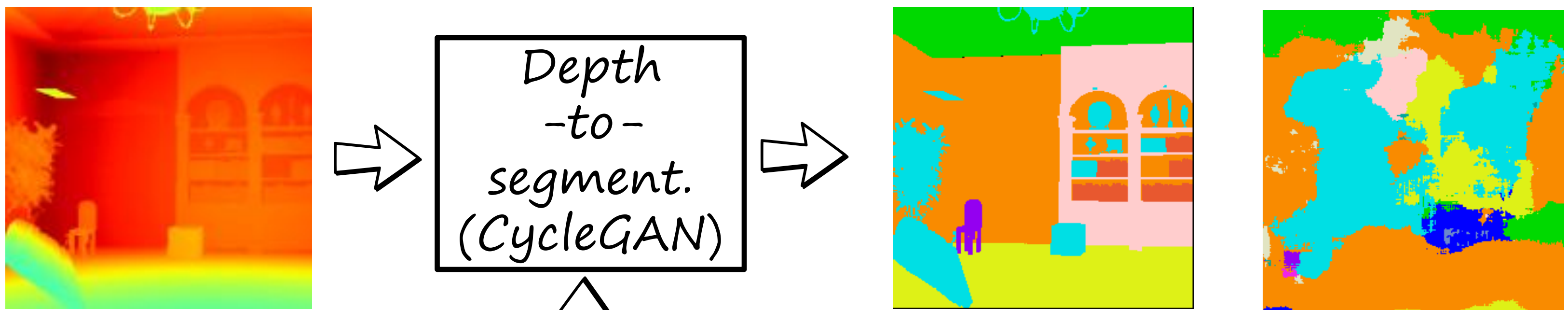


In practice

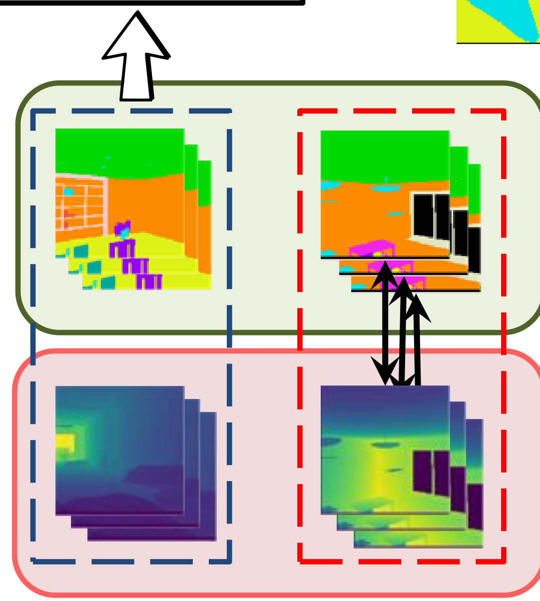
[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]

Zero-pair translation with CycleGAN (unpaired translation)

In theory



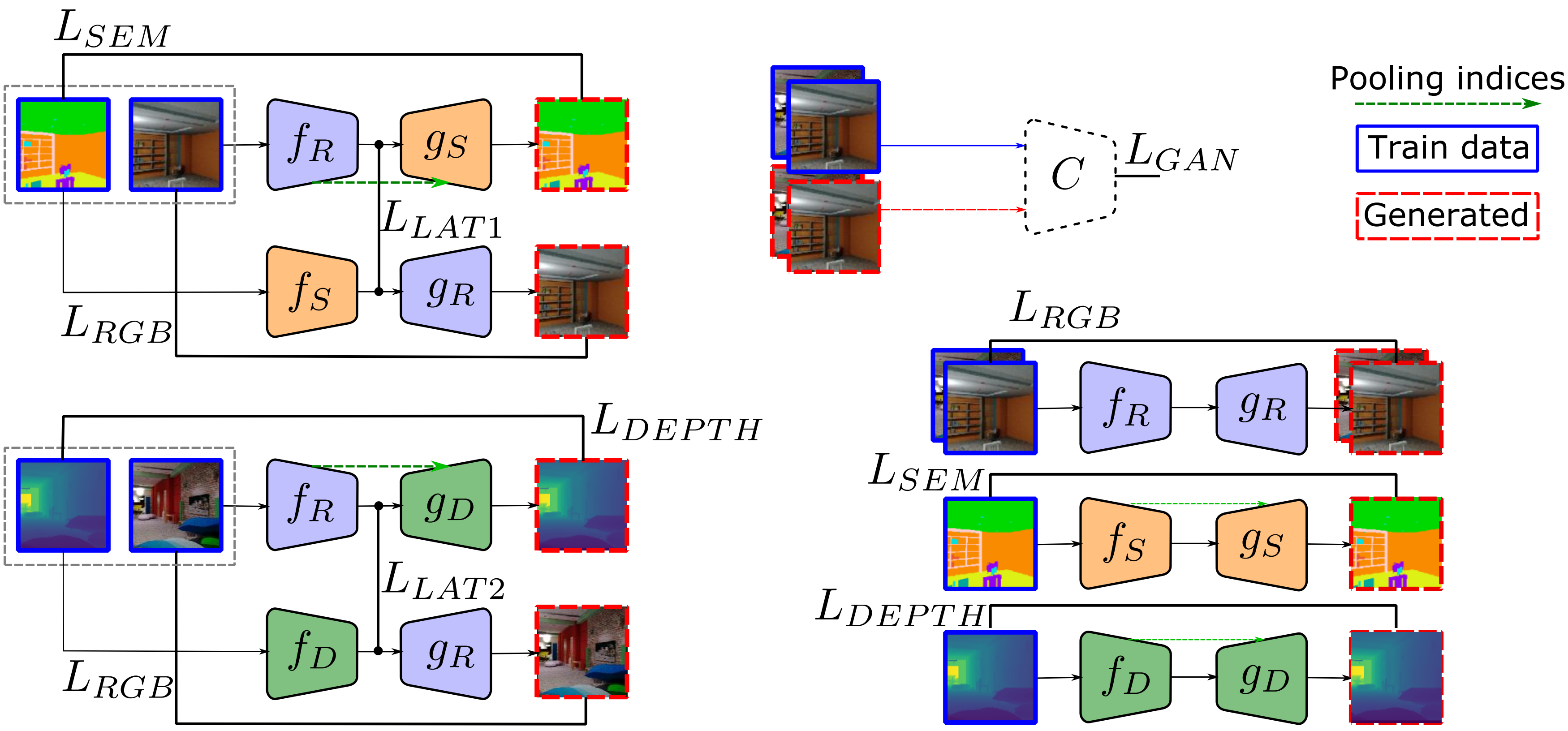
In practice



Depth-to-segmentation is too complex for CycleGAN

[Wang, van de Weijer, Herranz, “Encoder-decoder alignment for zero-pair image-to-image translation”, CVPR 2018]

Zero-pair translation with mix and match networks

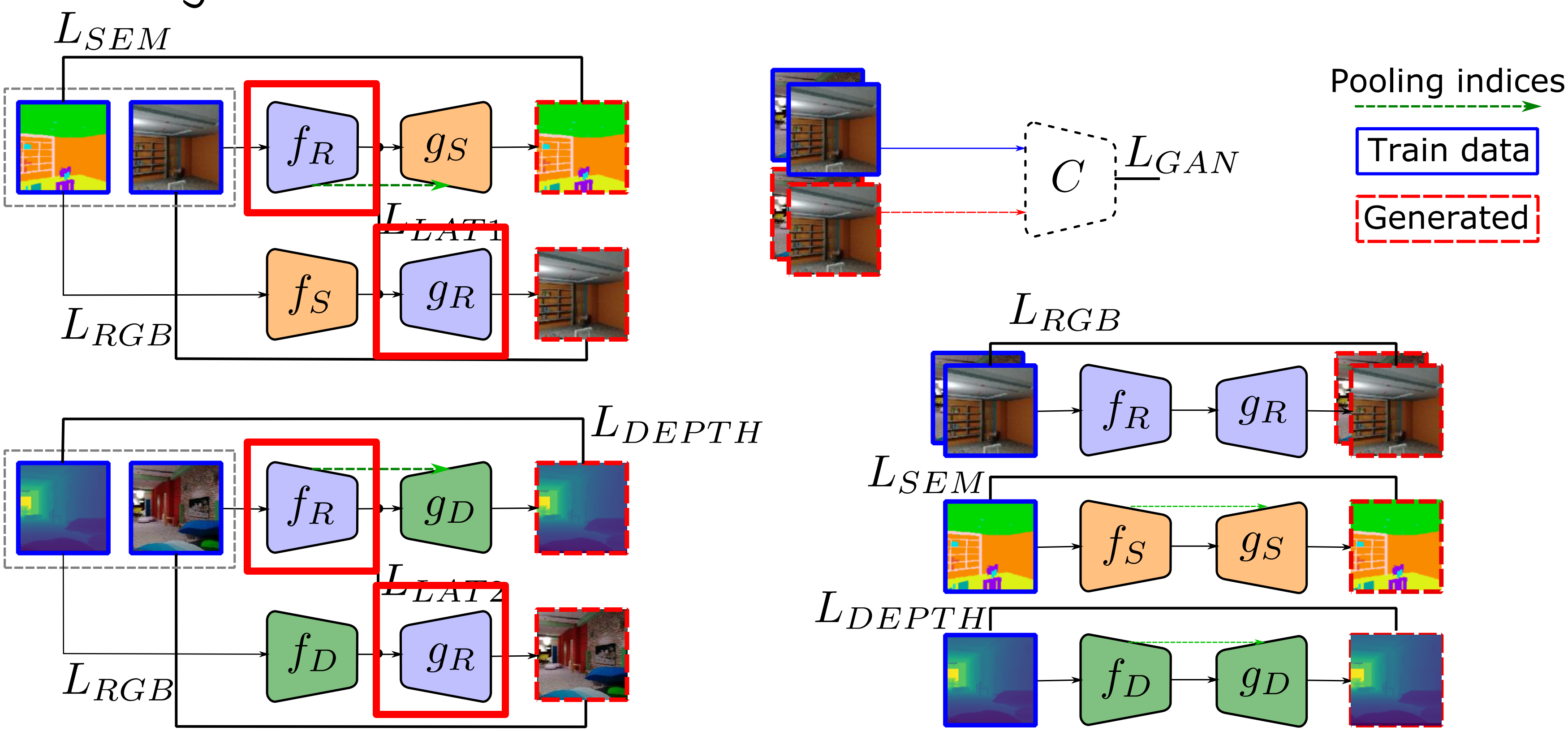


[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]()

Zero-pair translation with mix and match networks

Training for encoder-decoder alignment:

Shared encoder/decoders



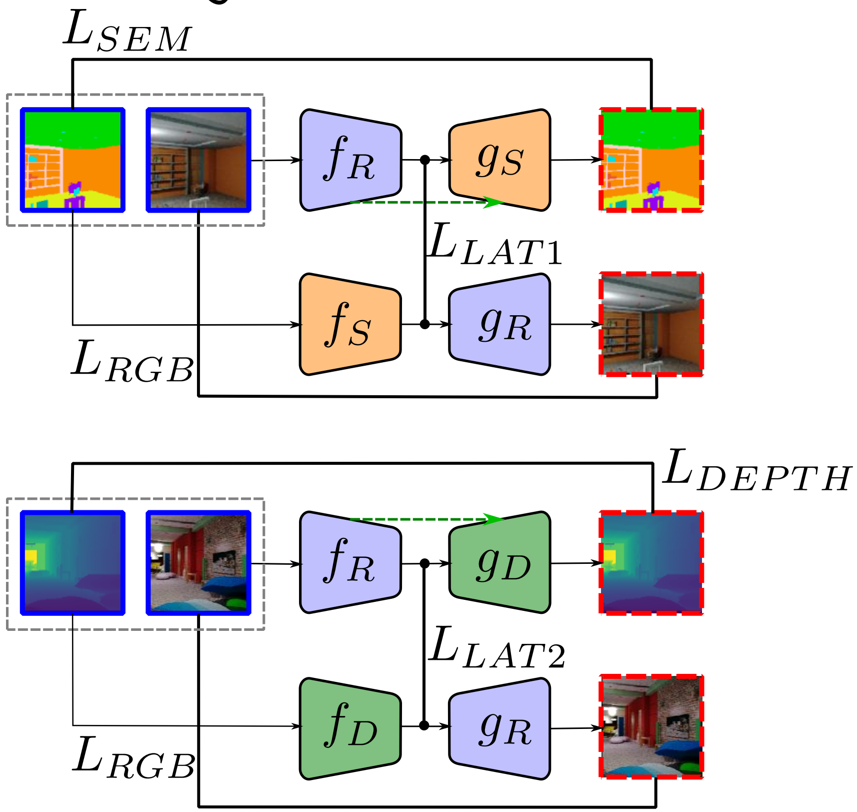
[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]()

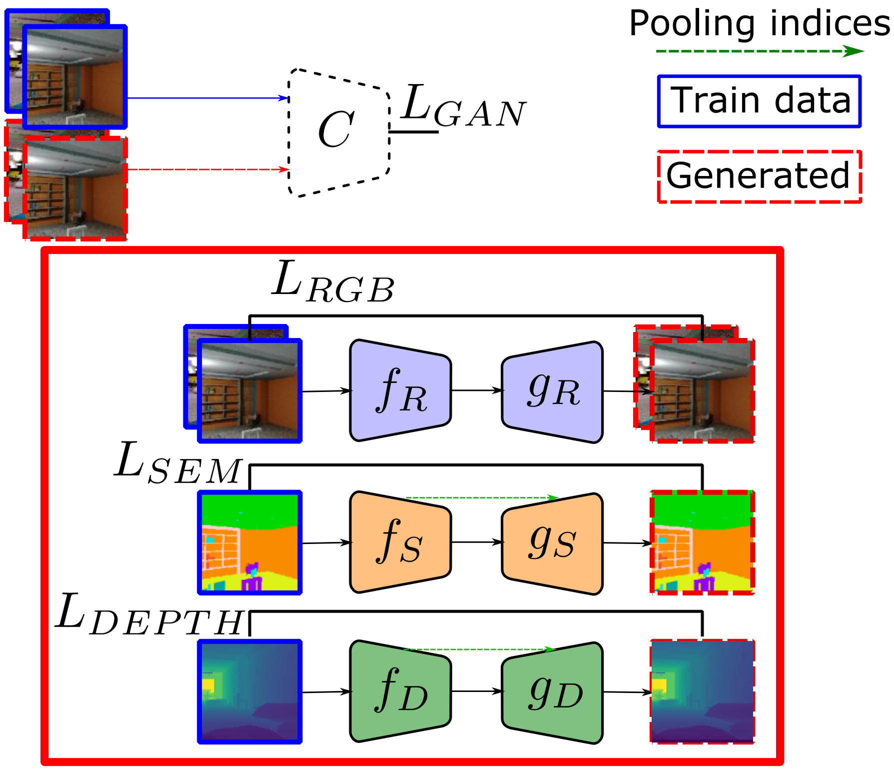
Zero-pair translation with mix and match networks

Training for encoder-decoder alignment:

Shared encoder/decoders

Autoencoders





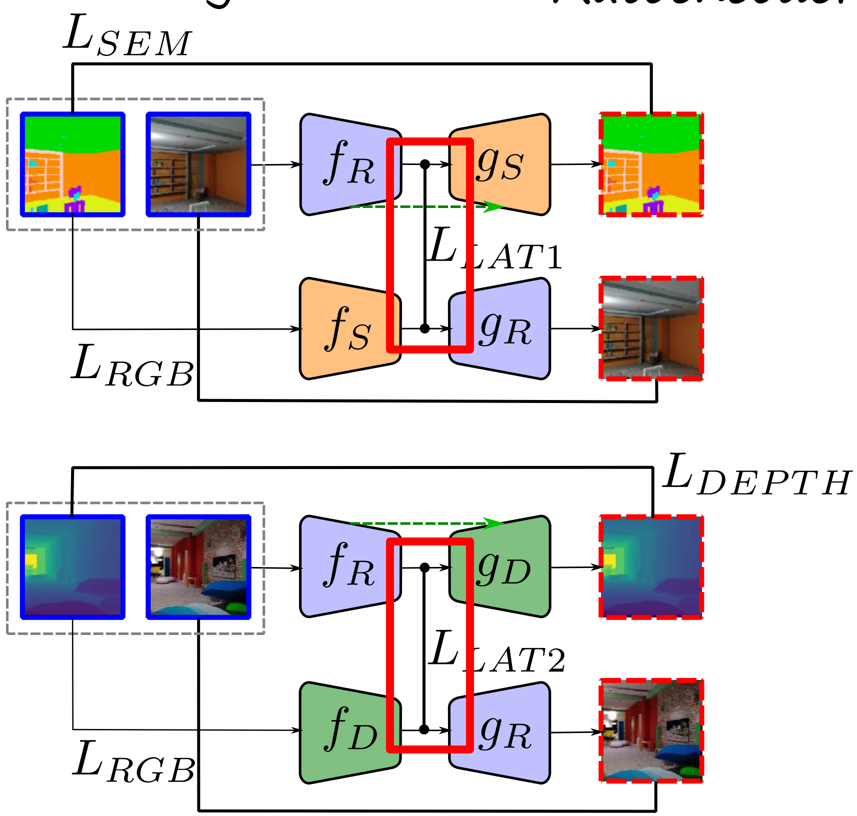
[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]

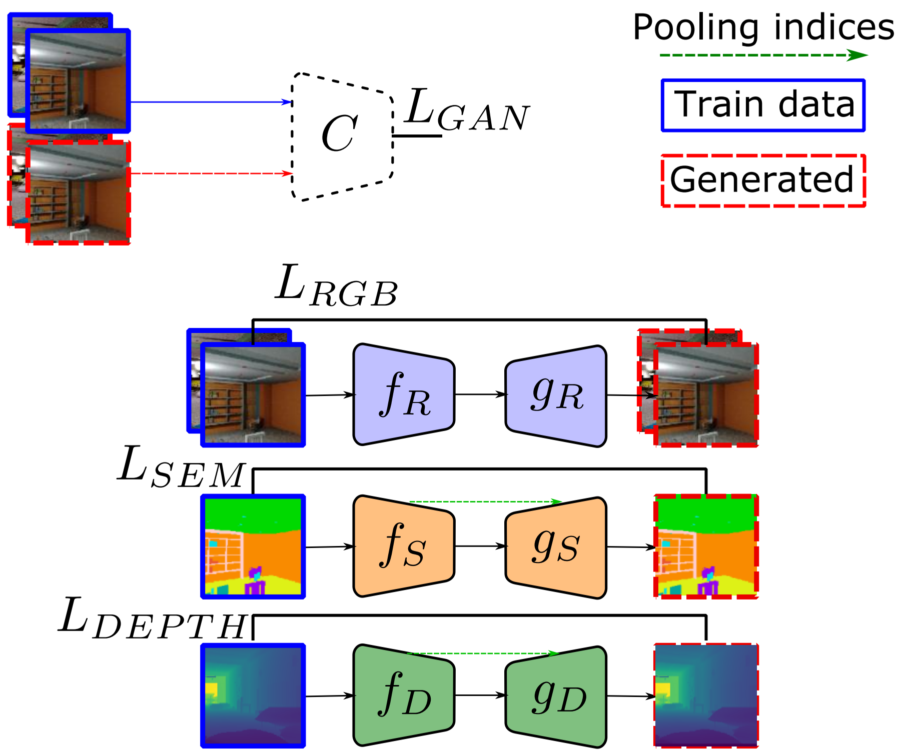
Zero-pair translation with mix and match networks

Training for encoder-decoder alignment:

Shared encoder/decoders
Autoencoders

Latent losses





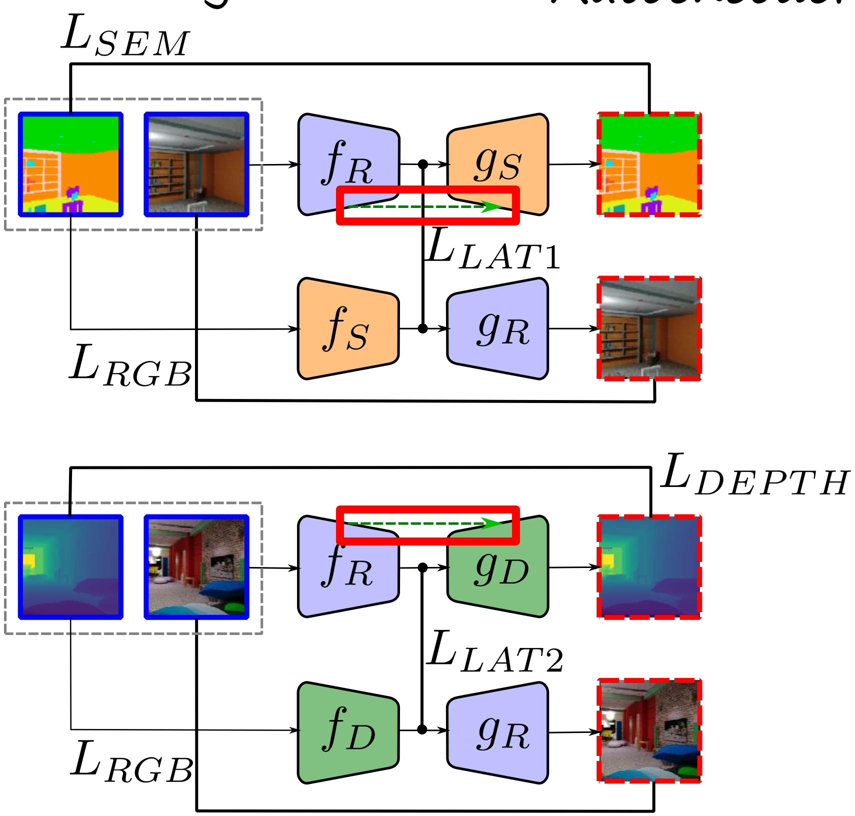
Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018

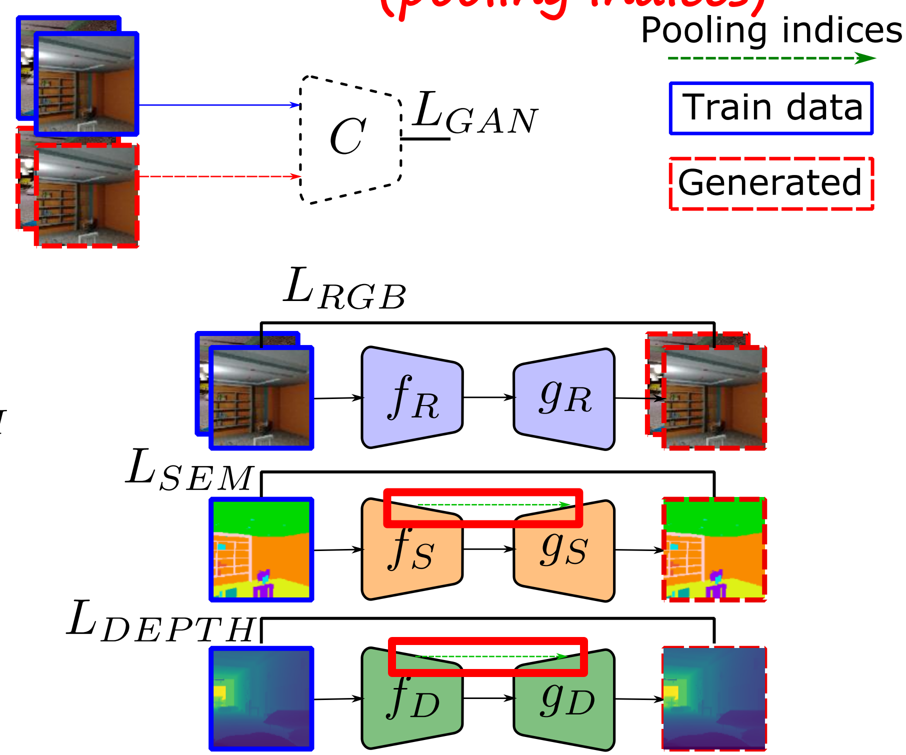
Zero-pair translation with mix and match networks

Training for encoder-decoder alignment:

Shared encoder/decoders
Autoencoders

Latent losses
Robust side information (pooling indices)

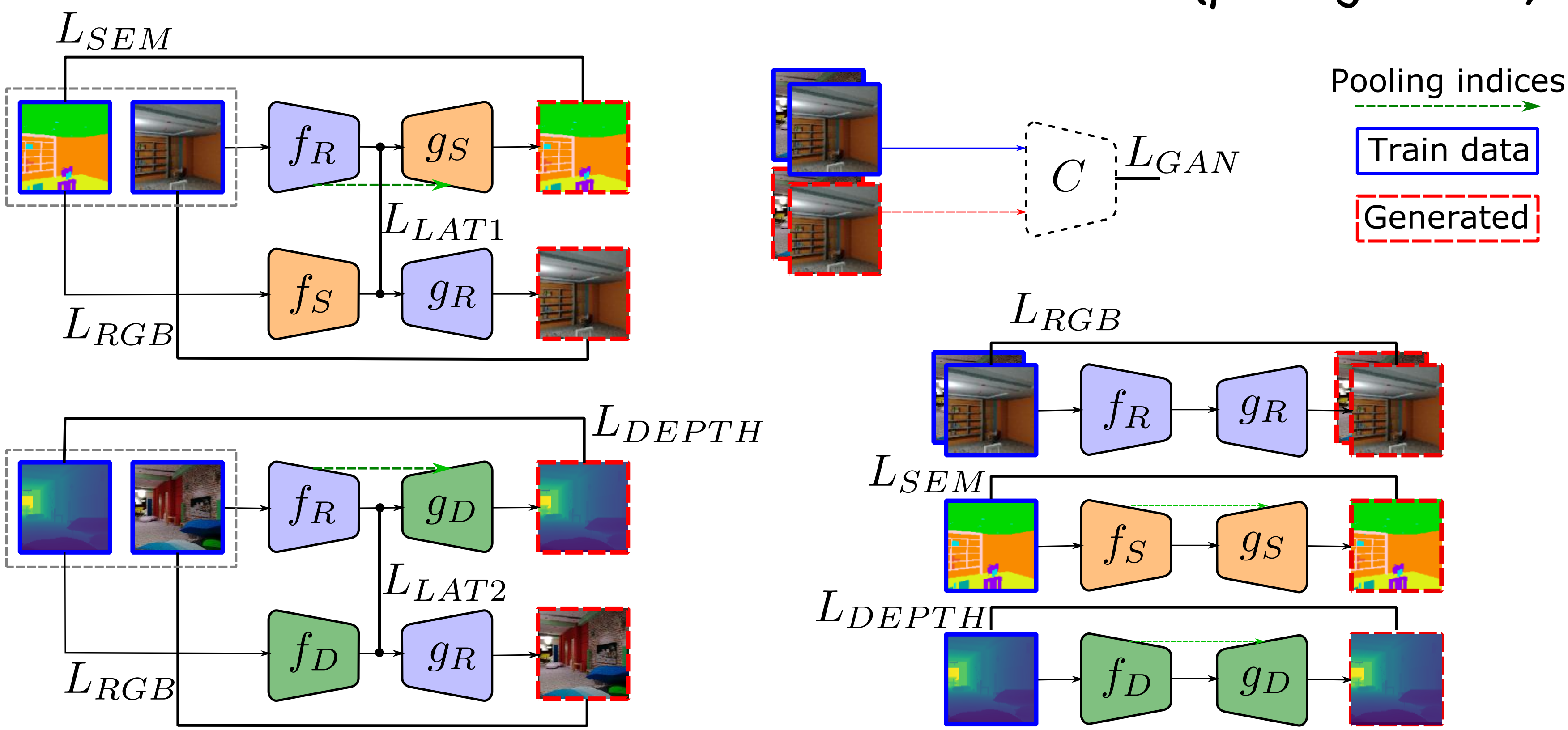




[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]

Zero-pair translation with mix and match networks

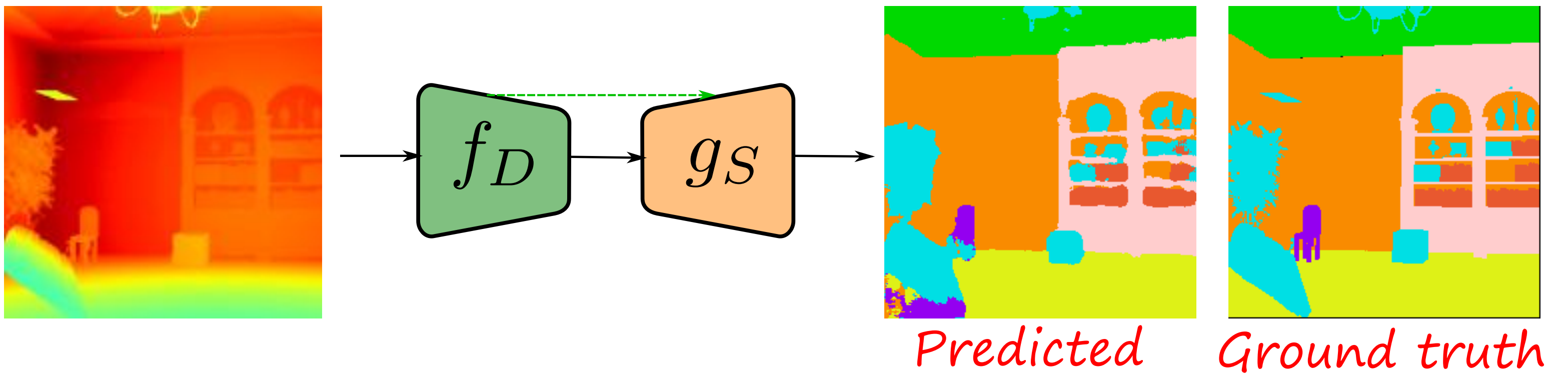
Training for encoder-decoder alignment: ***shared encoder/decoders*, *autoencoders*,** *latent losses* and ***robust side information (pooling indices)***



[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018](#)

Zero-pair translation with mix and match networks

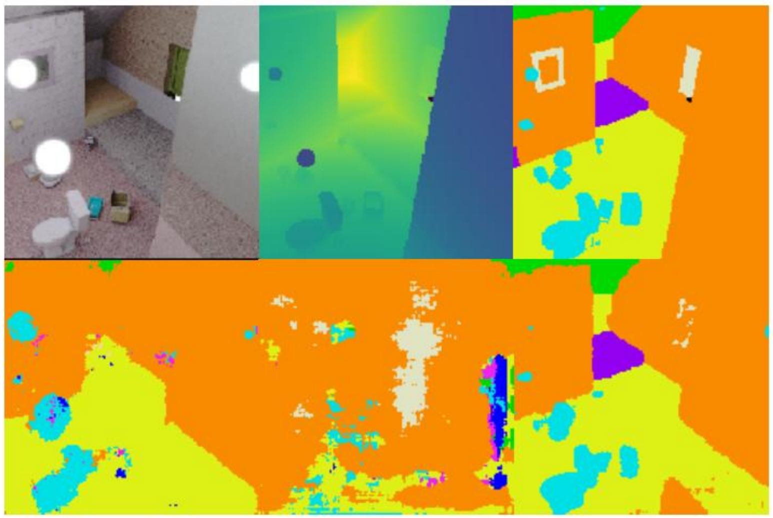
Test on zero-pair translation depth-to-segmentation



Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018

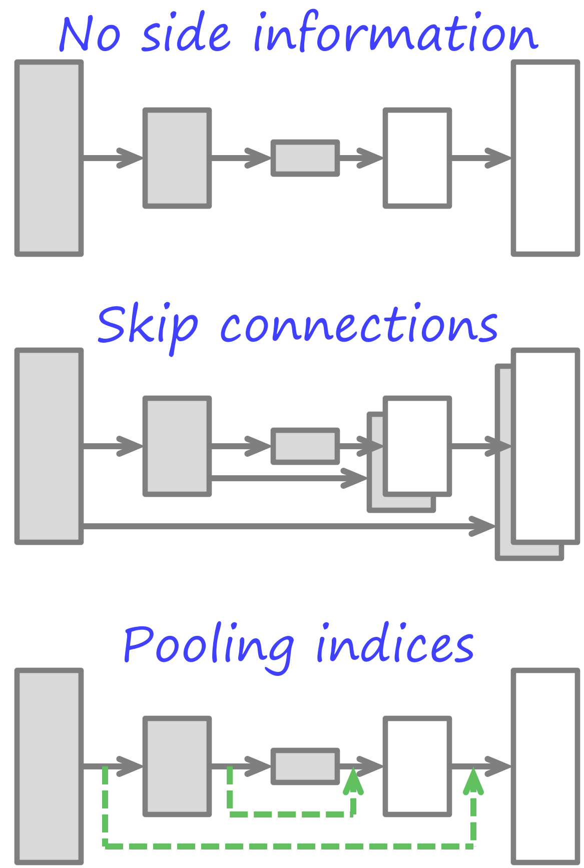
Side information in mix and match networks

RGB | Depth | Ground truth



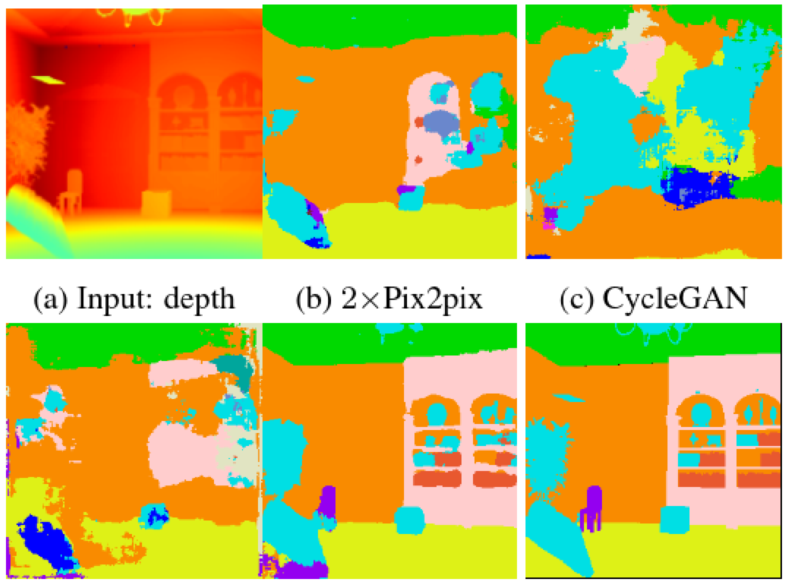
No side information | Skip connections | Pooling indices

Side information	Pretrained	mIoU	Global
-	N	32.2%	63.5%
Skip connections	N	14.1%	52.6%
Pooling indices	N	45.6%	73.4%
Pooling indices	Y	49.5%	80.0%



Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018

Comparison: depth-to-segmentation



(a) Input: depth (b) 2×Pix2pix (c) CycleGAN

(d) D→R→S (e) Proposed (f) Ground truth

Figure 1: Zero-pair depth→segmentation, trained on (depth,RGB) and (RGB,segmentation).

[Wang, van de Weijer, Herranz, "Encoder-decoder alignment for zero-pair image-to-image translation", CVPR 2018]()

Thanks!

Computer Vision Center
Edifici O, Campus UAB, Barcelona
http://www.cvc.uab.es

Learning and Machine Perception (LAMP) team
http://www.cvc.uab.es/lamp