

Mix and match networks: encoder-decoder alignment for zero-pair image-to-image translation

Luis Herranz

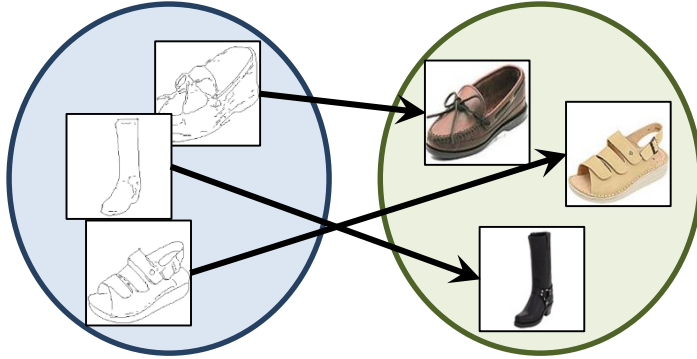
Computer Vision Center, UAB

December 2018

#DLBCN 2018

Progress in image-to-image (I2I) translation

Paired I2I translation



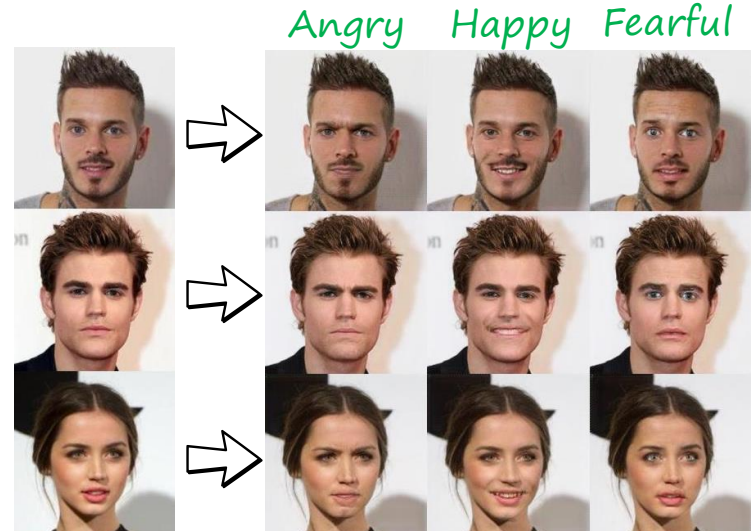
Unpaired I2I translation



Diversity in I2I translation



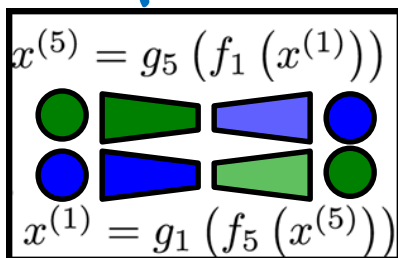
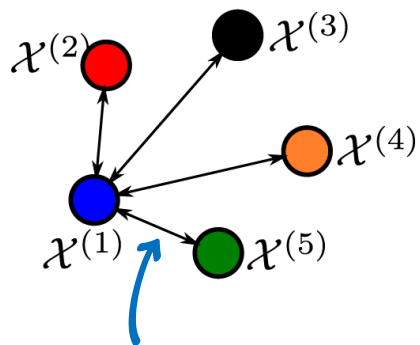
Multi-domain I2I translation



Unseen multi-domain I2I translations

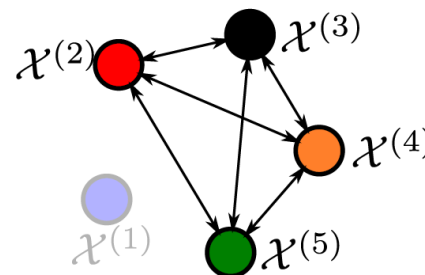
Only these translations
are trained (seen)

Train

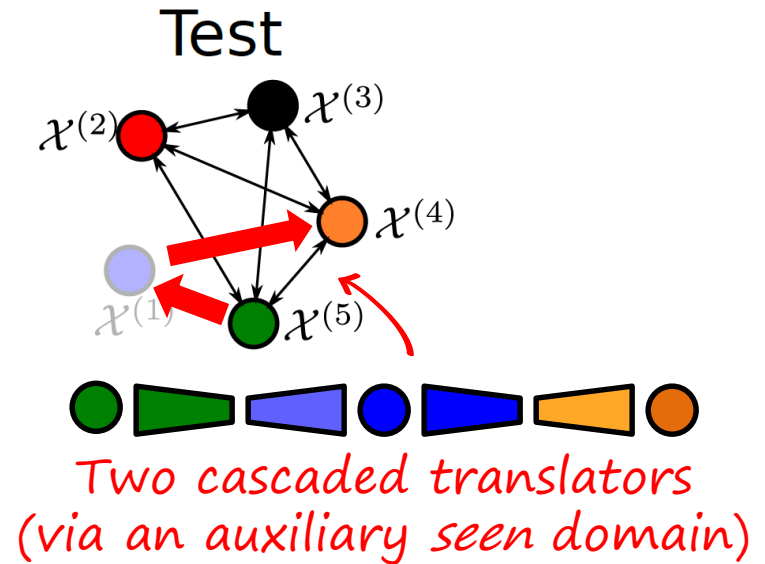
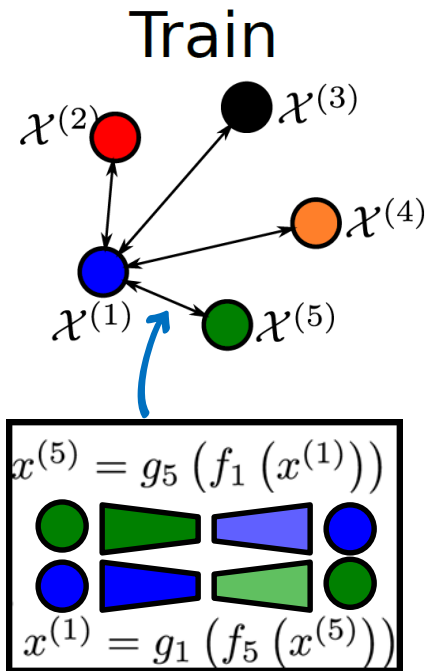


Evaluate on these unseen
translations

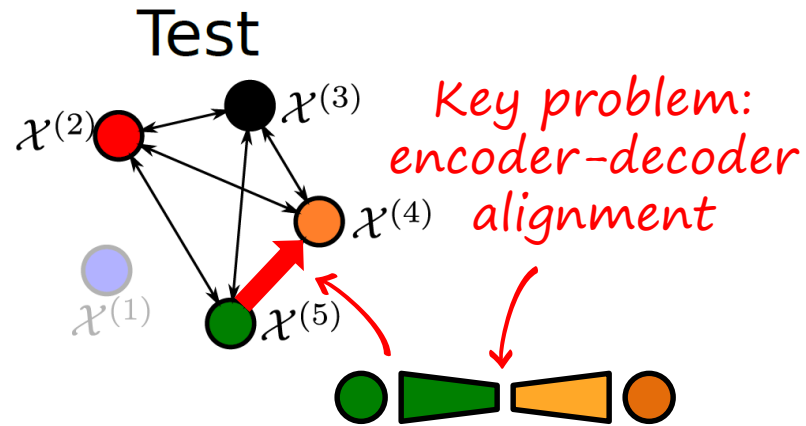
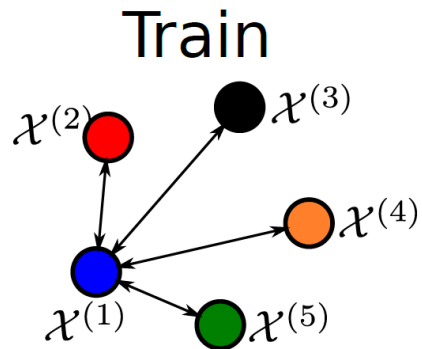
Test



Cascading I2I translators



Mix and match networks



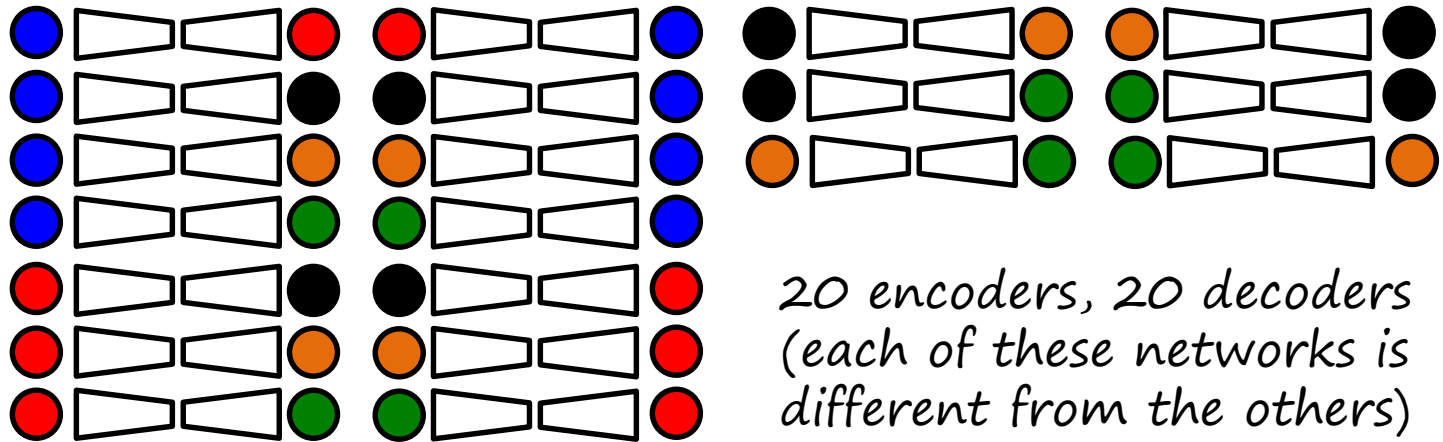
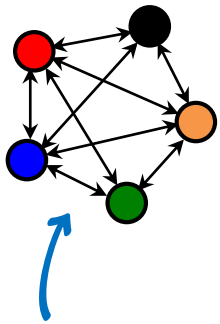
*Mix&match encoder-decoders
(they haven't seen each other
during training)*

Application: many-to-many translations

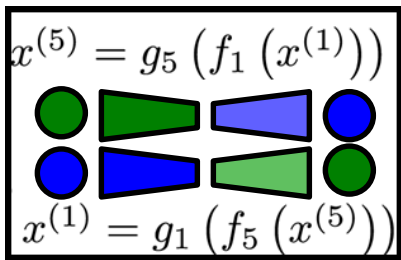
We could train all possible translators

Problems:

- No sharing
- Poor scalability: number of networks $O(N^2)$



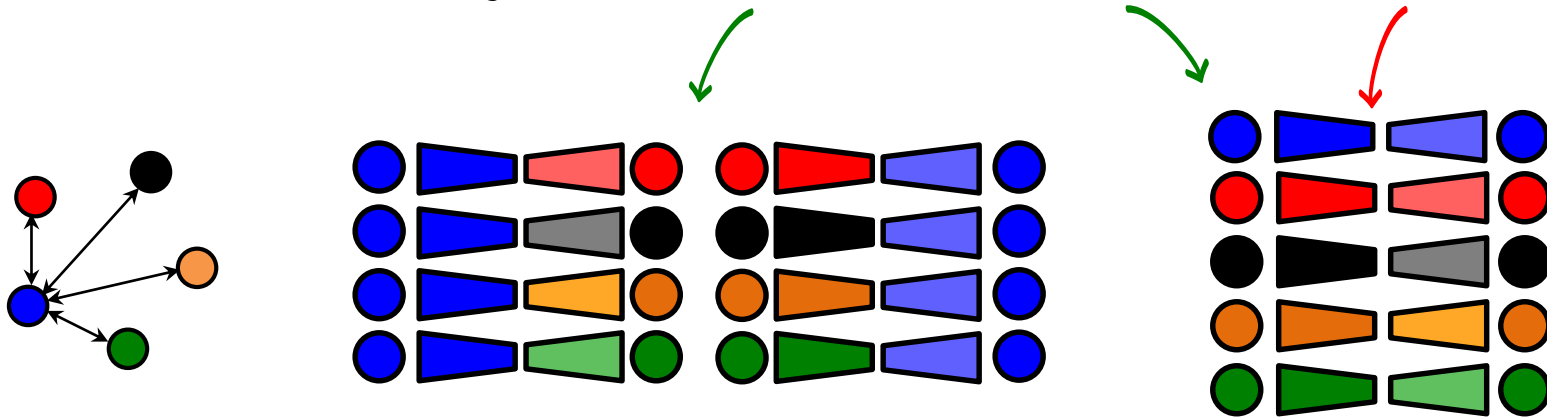
20 encoders, 20 decoders
(each of these networks is different from the others)



Mix and match networks

Unseen encoder-decoder alignment

- Latent representation should be **domain-independent**
- Achieved using **shared encoder/decoders** and **autoencoders**

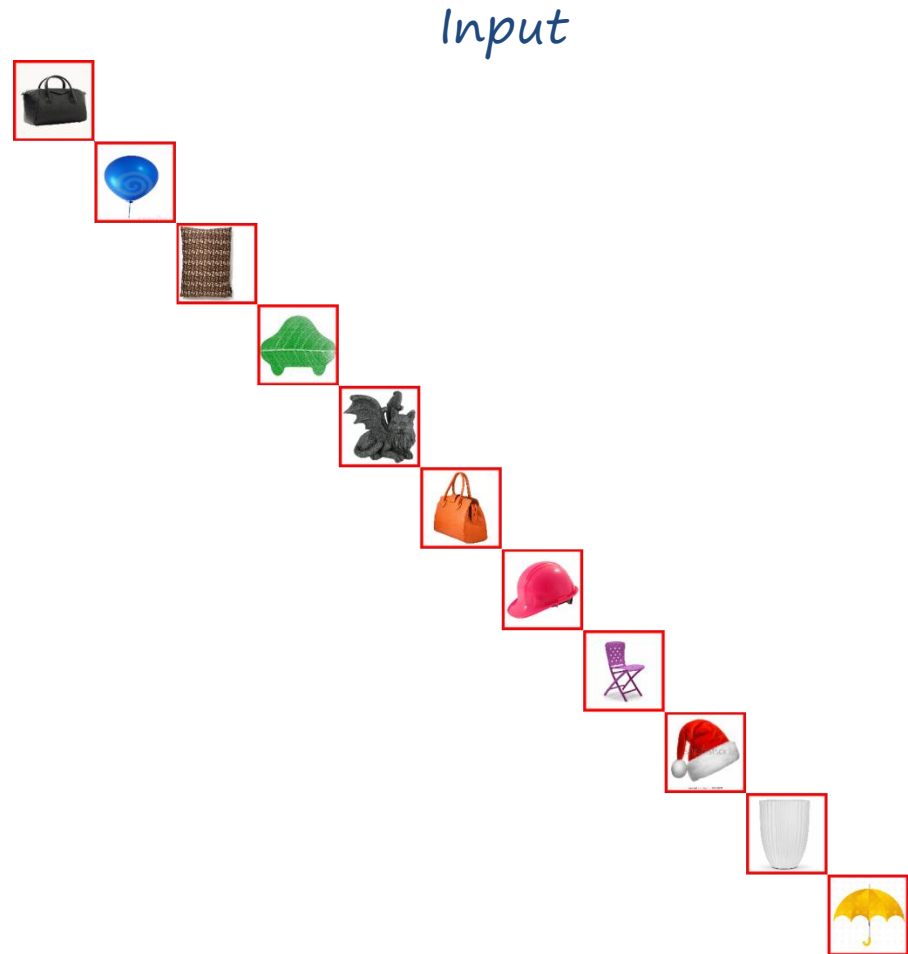


5 encoders, 5 decoders

- Scalable: number of networks $O(N)$

Example: scalable recolorization

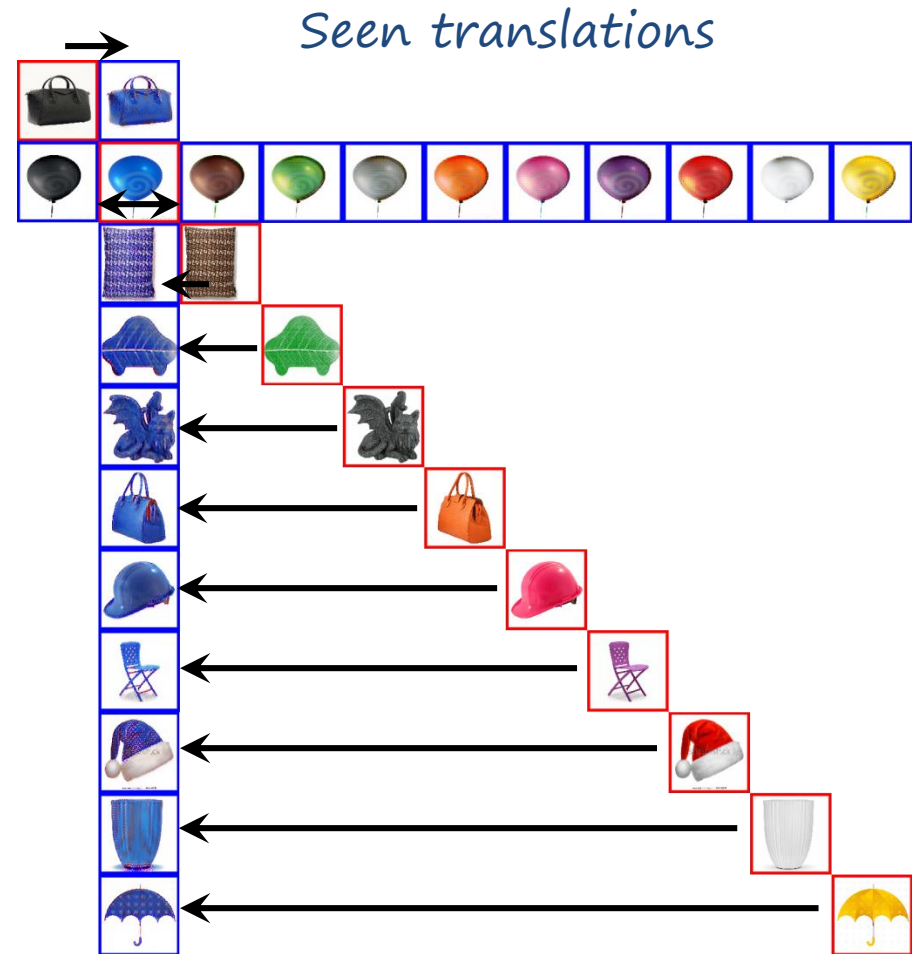
Unpaired translation
11 colors (i.e. 11 domains)



Example: scalable recolorization

Unpaired translation
11 colors (i.e. 11 domains)

Requires training 11
encoders and 11 decoders



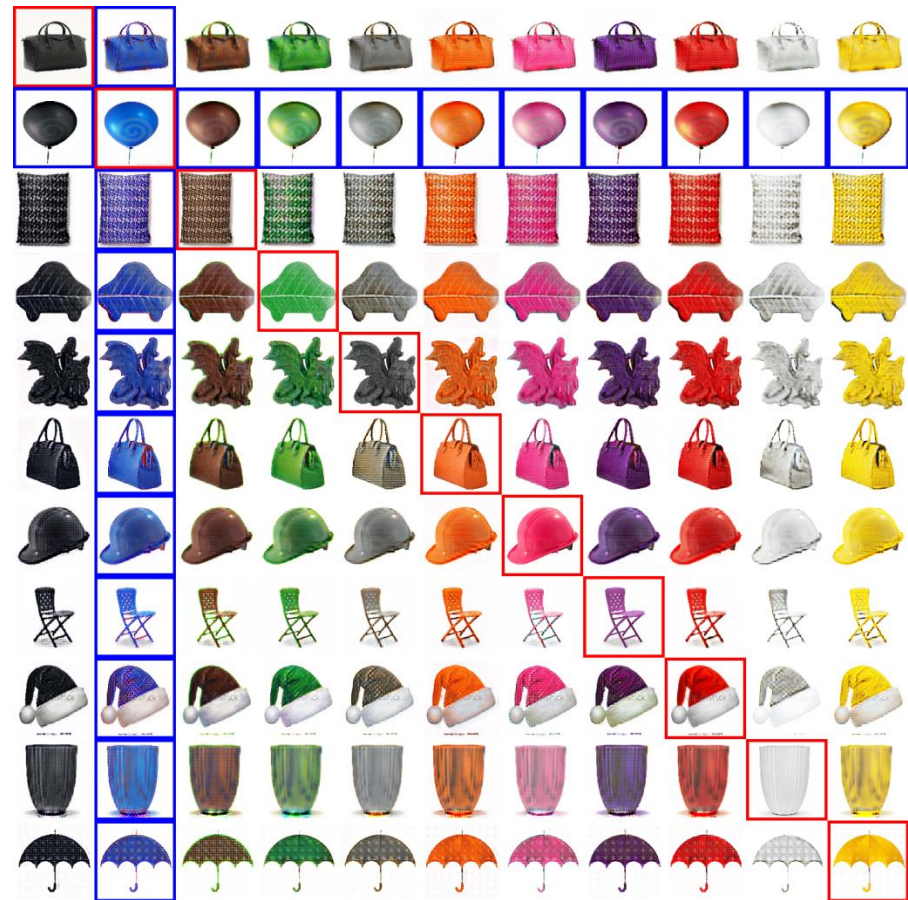
Example: scalable recolorization

Unpaired translation
11 colors (i.e. 11 domains)

Requires training 11
encoders and 11 decoders

*CycleGANs for all
combinations would require
55 encoders and 55 decoders*

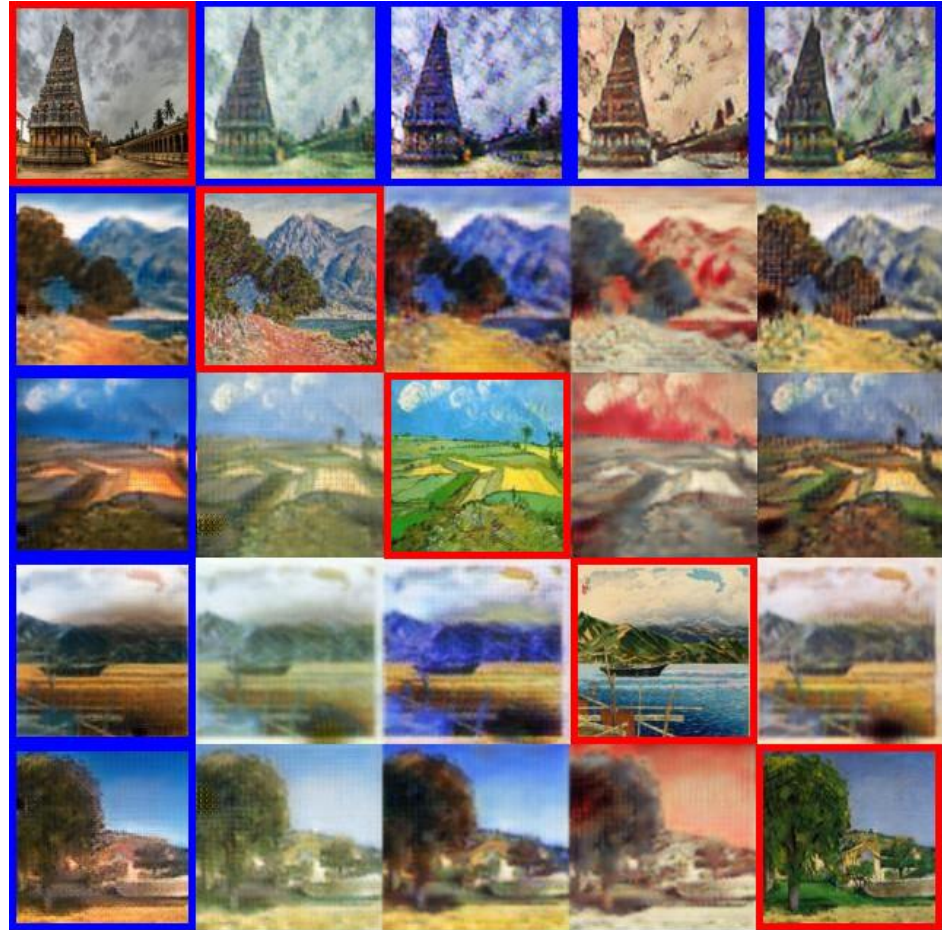
Seen+unseen translations



Example: scalable style transfer

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

*(5 encoders and 5
decoders)*

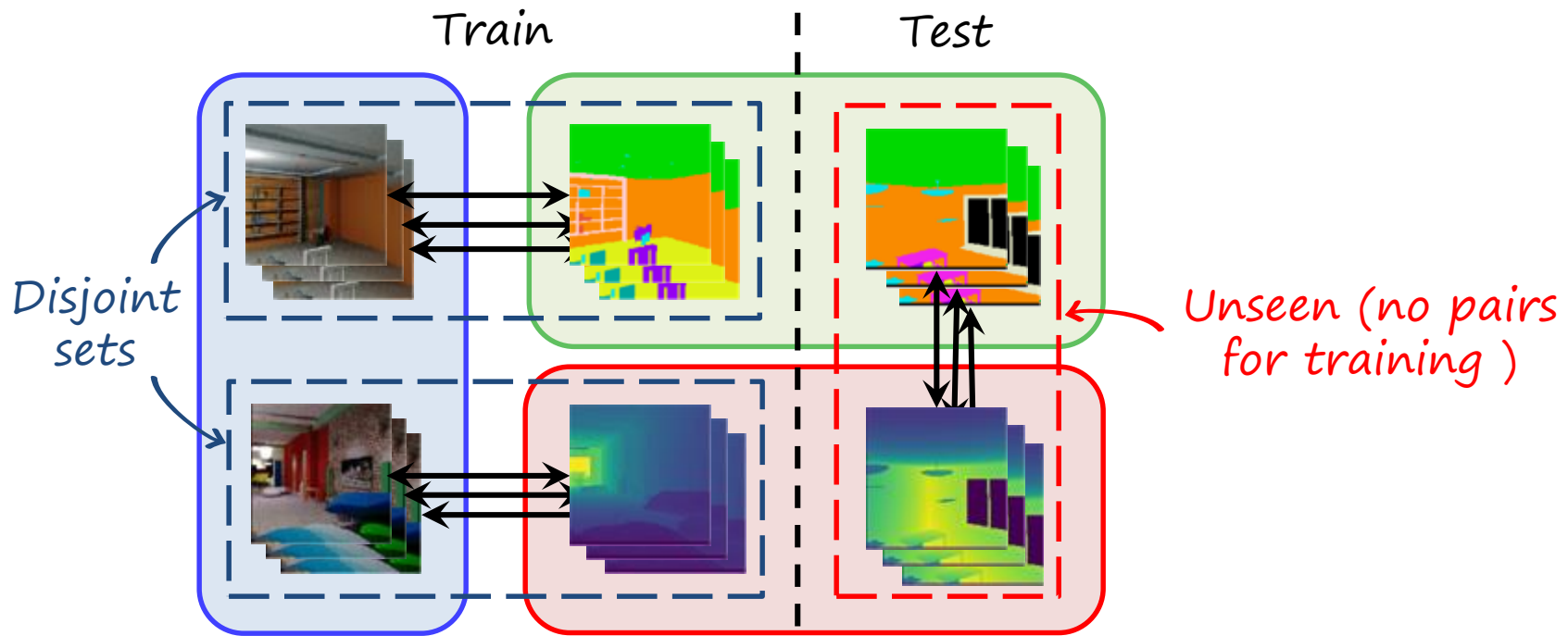


Application: cross-modal zero-pair translation

Cross-modal translation setting (*RGB*, *segmentation* and *depth*)

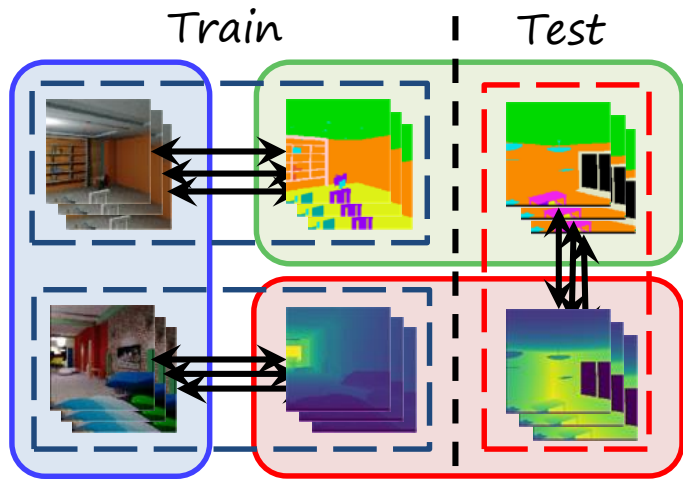
Paired data available for (*RGB*, *segm.*) and (*RGB*, *depth*)

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



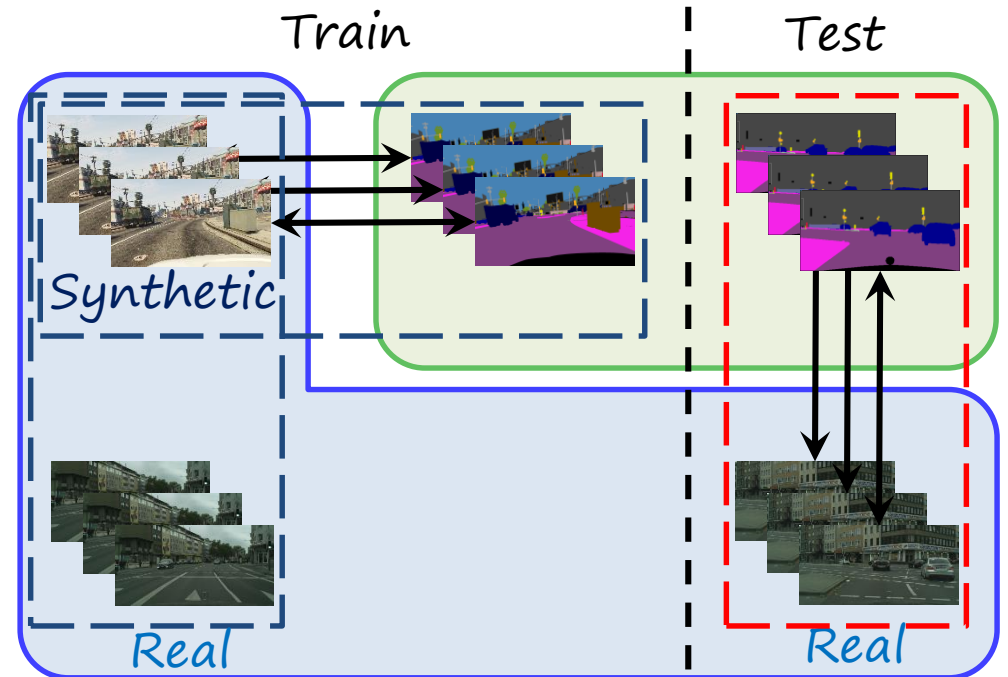
Relation to unsupervised domain adaptation

- Three modalities (*RGB*, *segm.*, *depth*)
- Pairs are available (supervised)



Cross-modal zero-pair translation

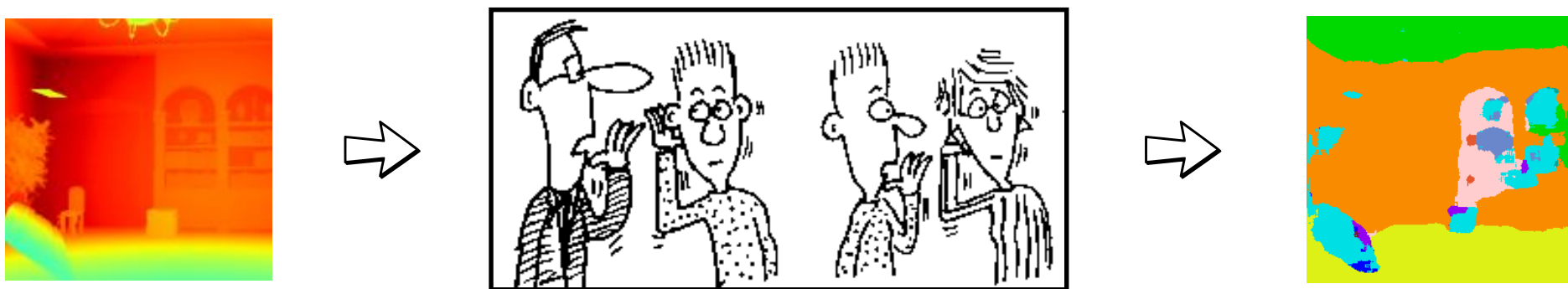
- Two modalities (*RGB*, *segm.*)
- Two RGB domains (*synth*, *real*)
- No pairs between domains



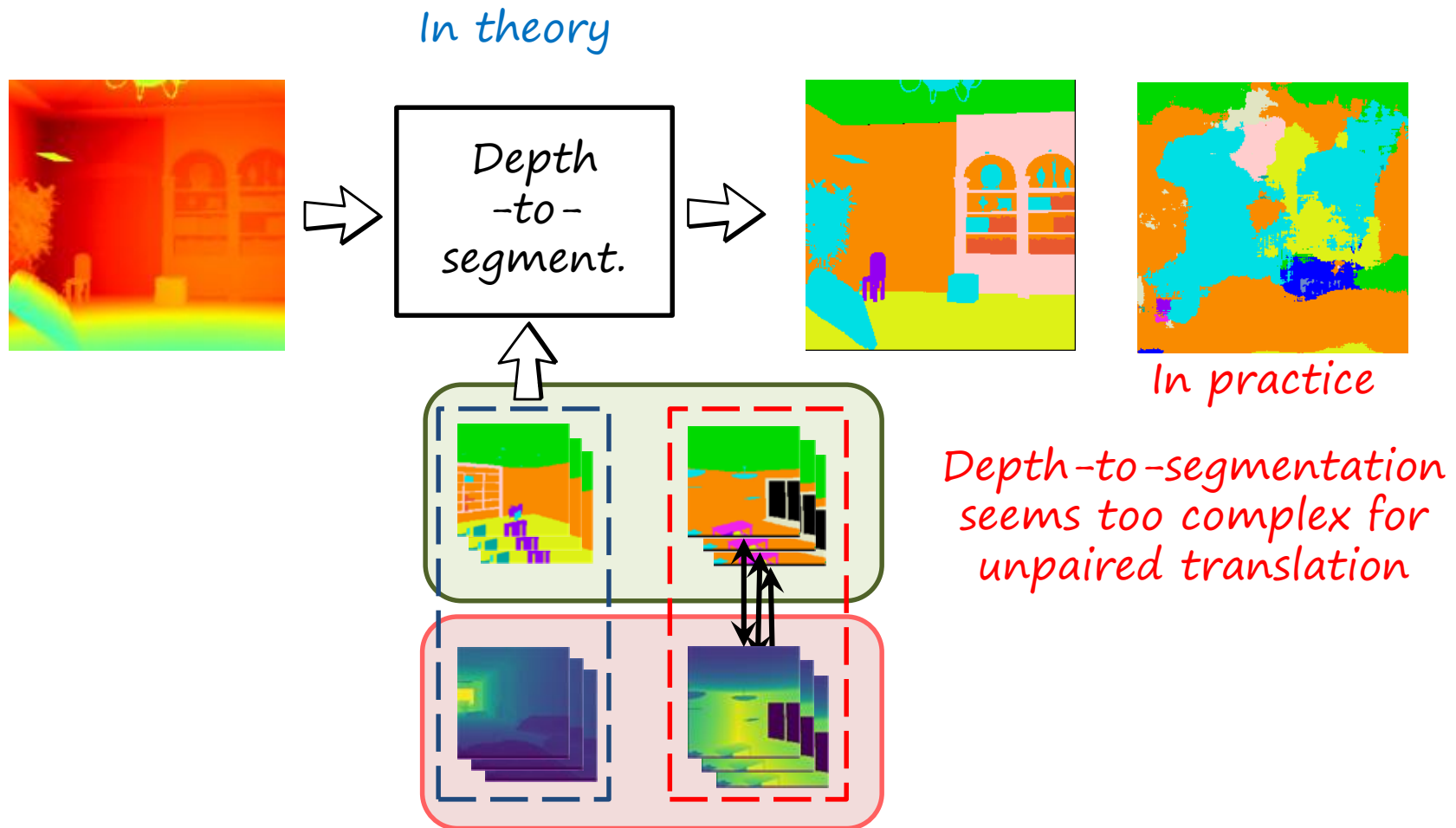
Unsupervised domain adaptation for segmentation

Zero-pair translation with two cascaded paired translations

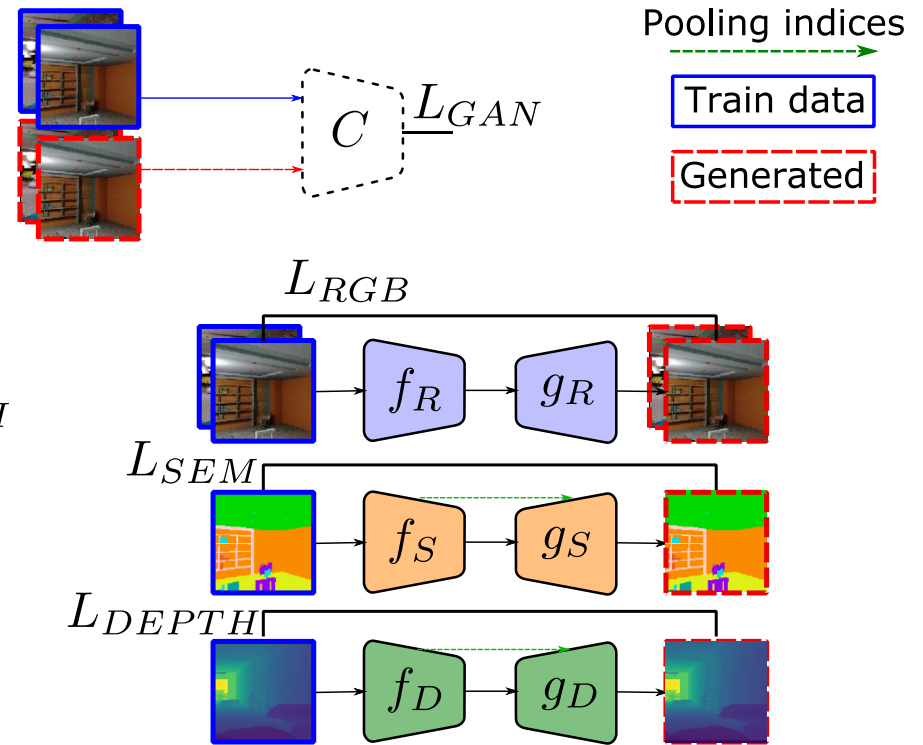
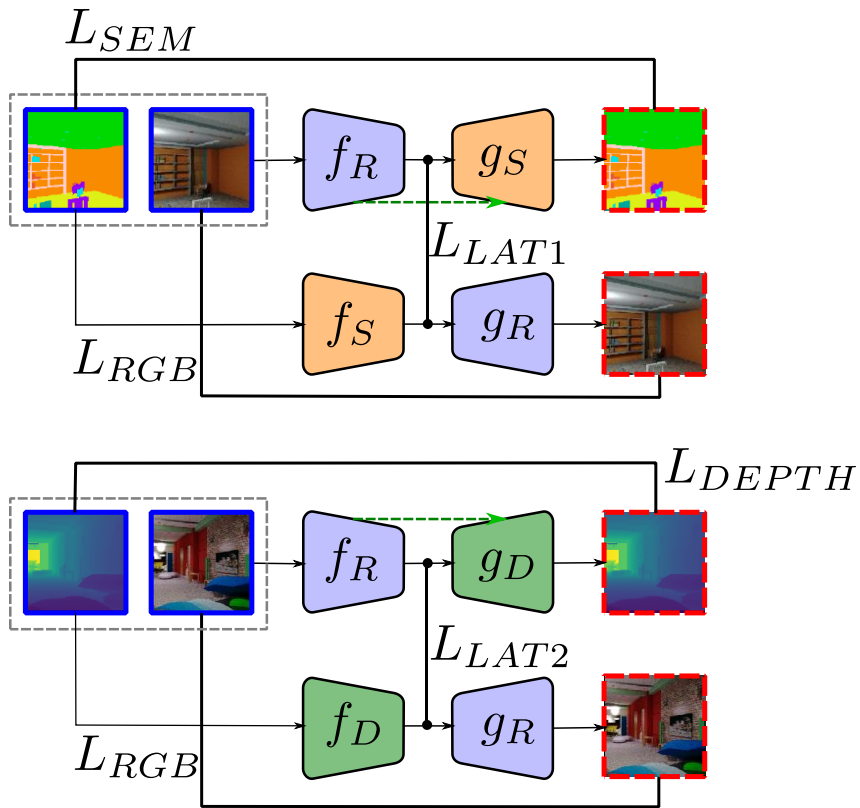
In theory



Zero-pair translation with unpaired translation

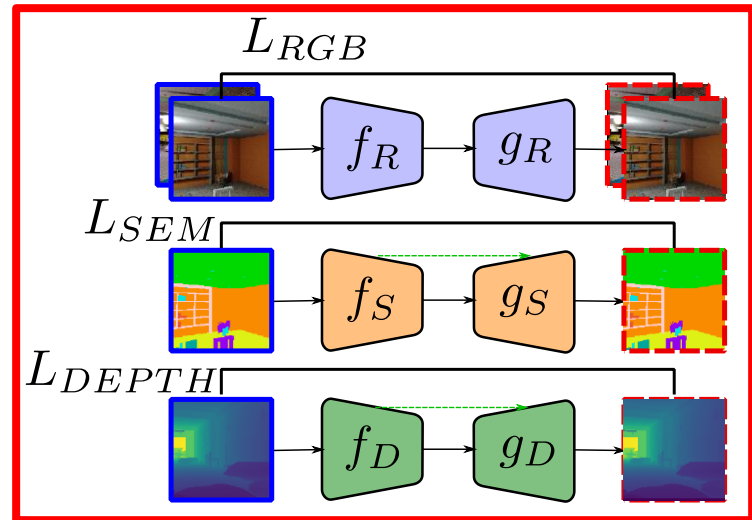
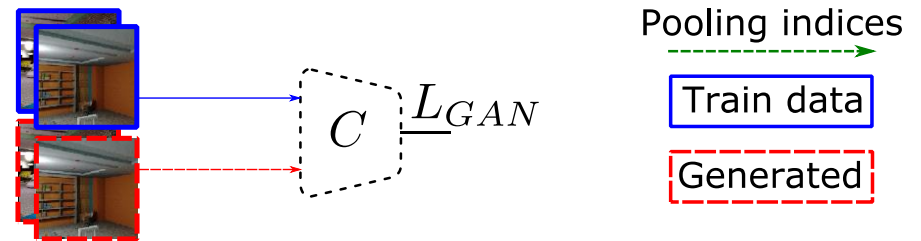
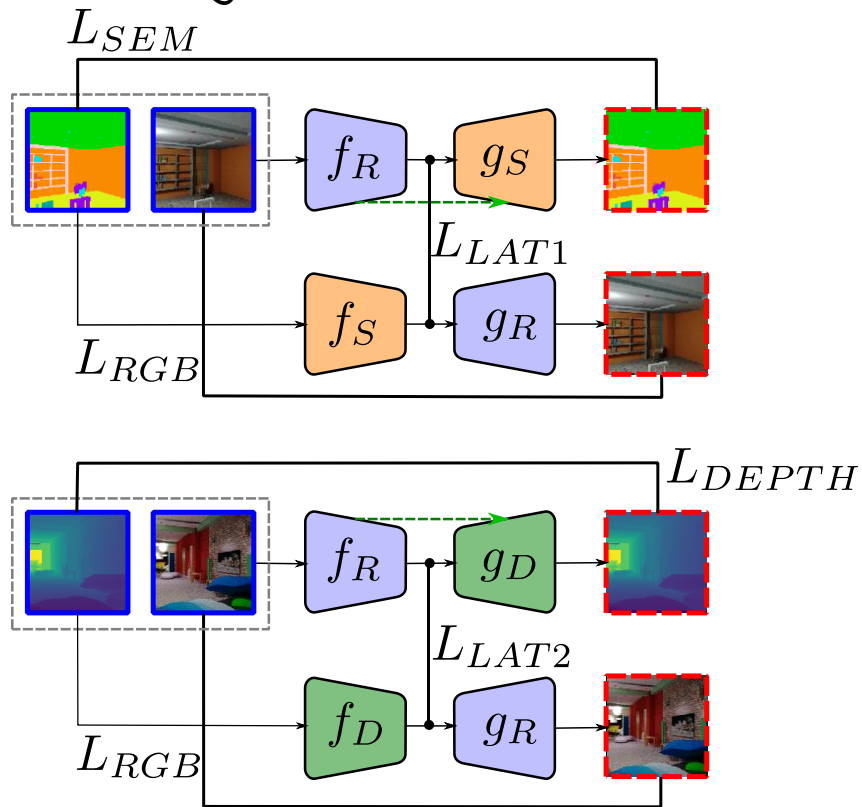


Zero-pair translation with mix and match networks



Zero-pair translation with mix and match networks

Training for encoder-decoder alignment: *Autoencoders*

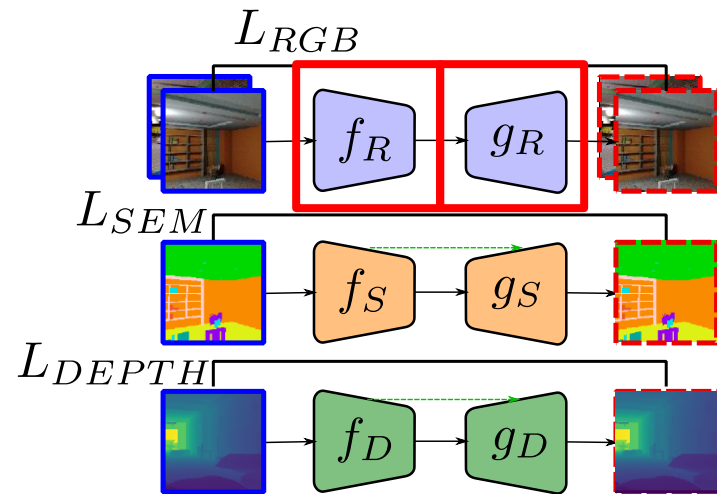
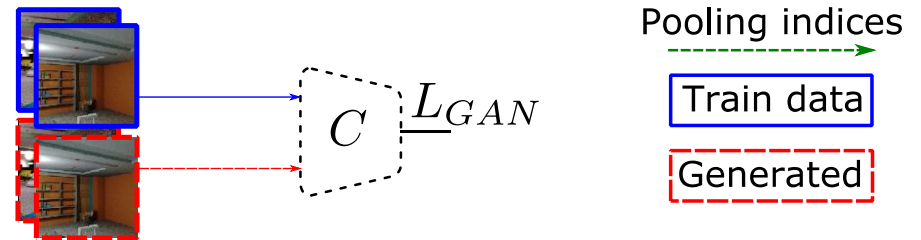
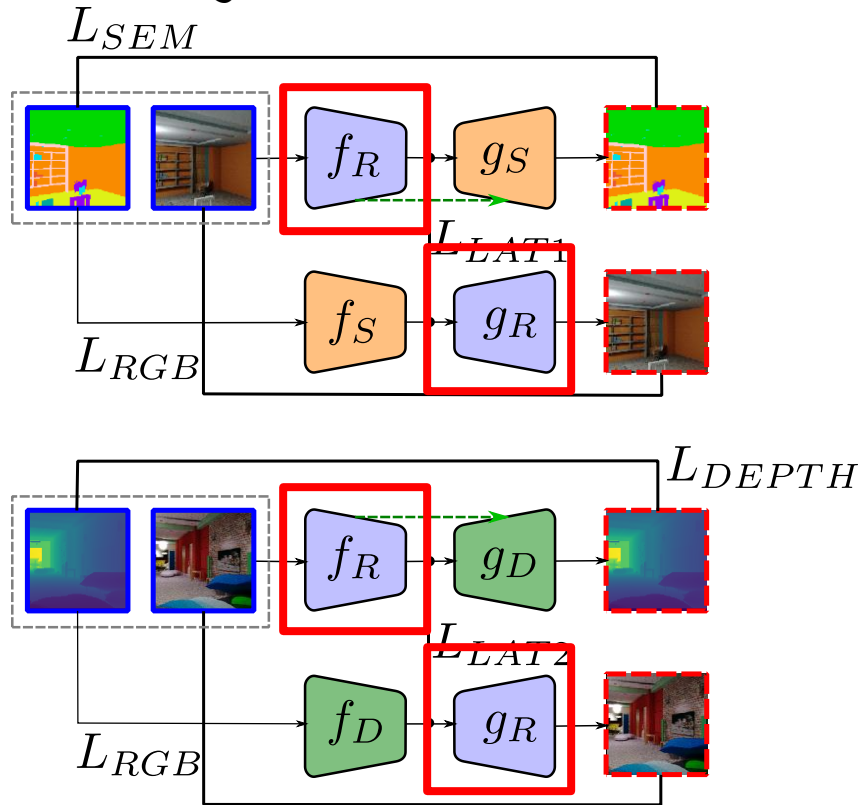


Zero-pair translation with mix and match networks

Training for encoder-decoder alignment:

Autoencoders

Shared encoder/decoders

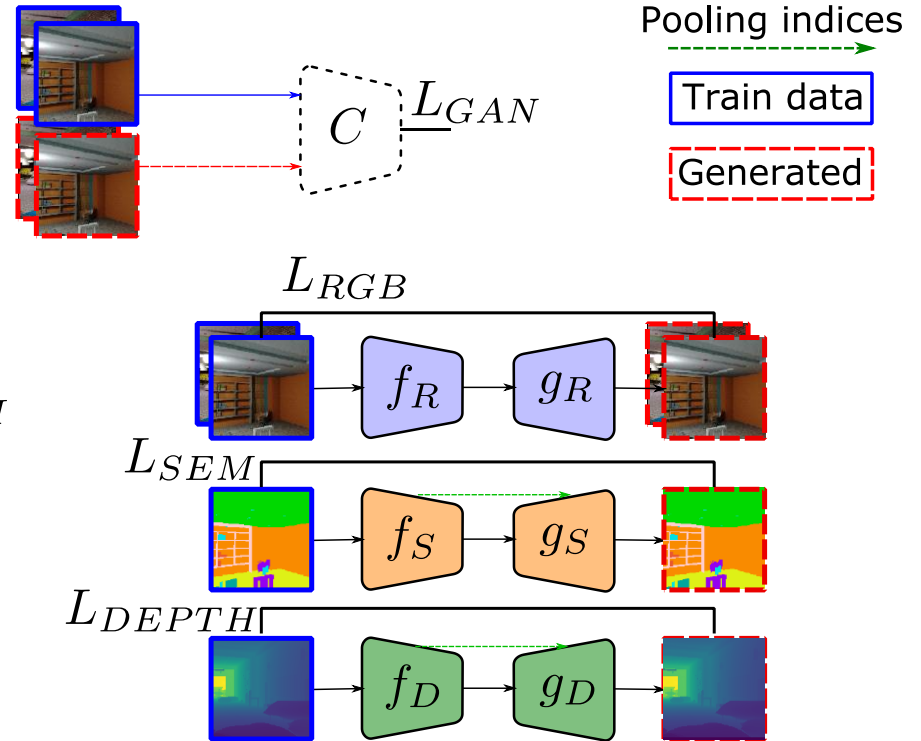
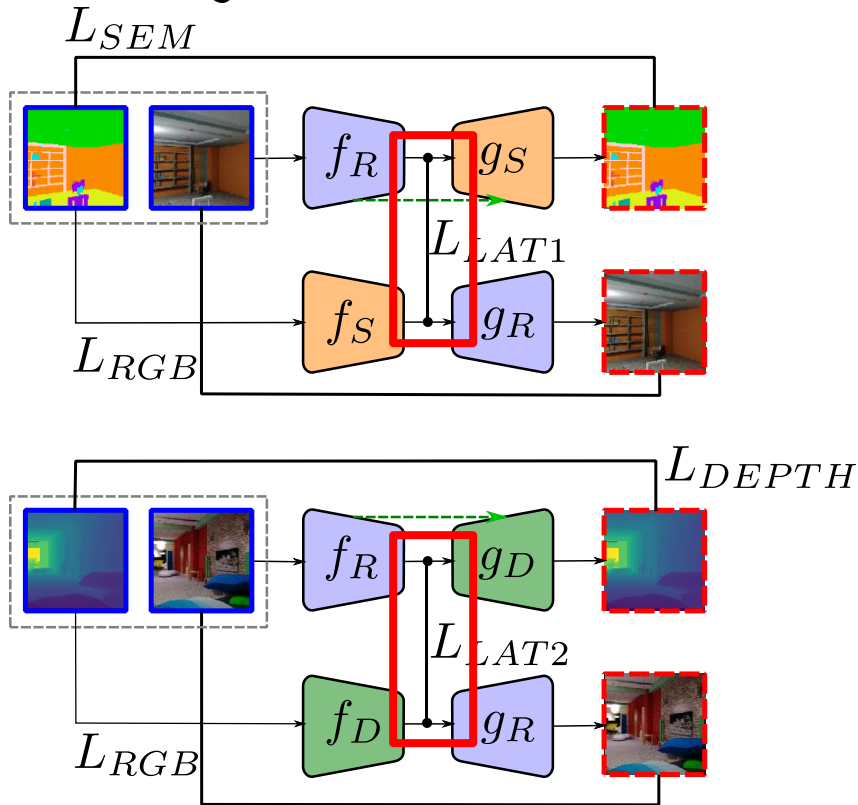


Zero-pair translation with mix and match networks

Training for encoder-decoder alignment:

Autoencoders
Shared encoder/decoders

Latent losses

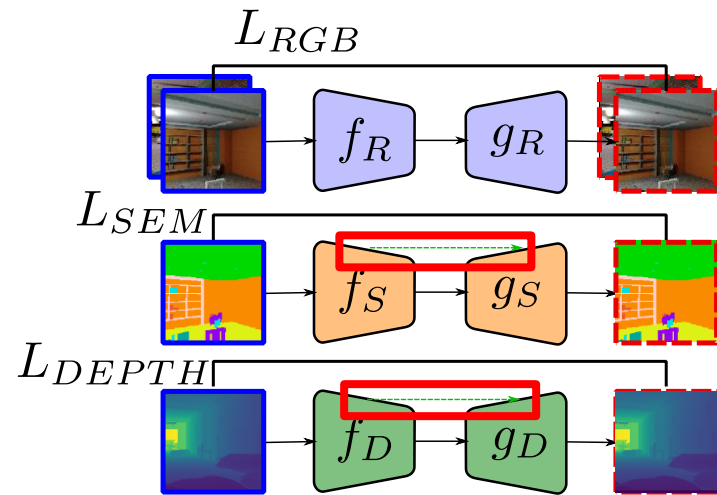
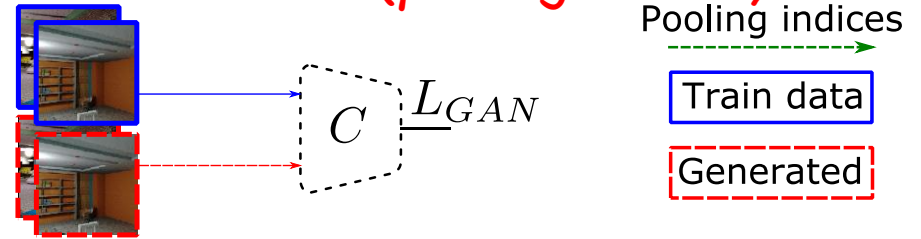
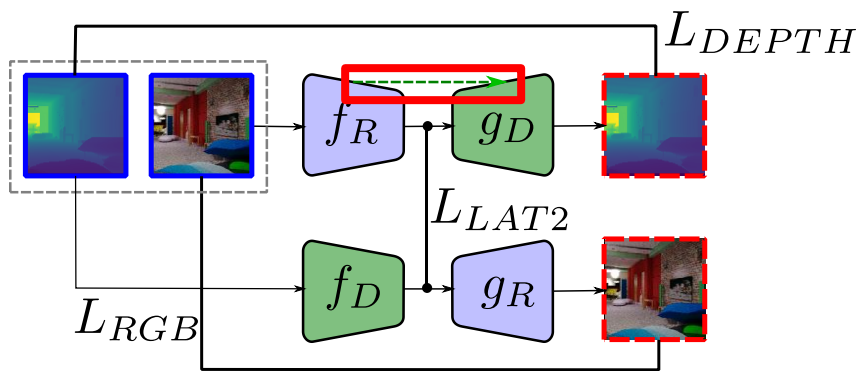
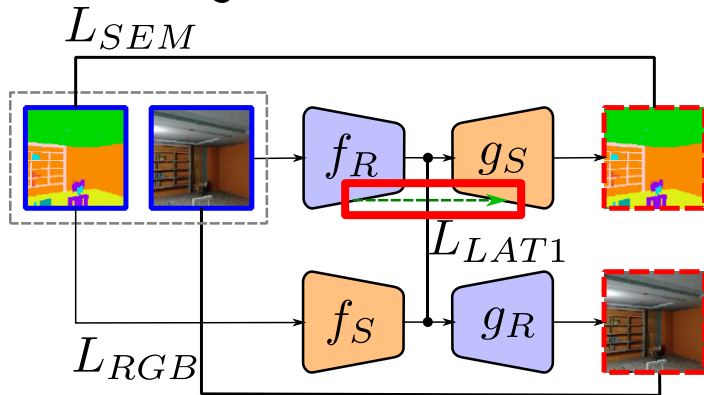


Zero-pair translation with mix and match networks

Training for encoder-decoder alignment:

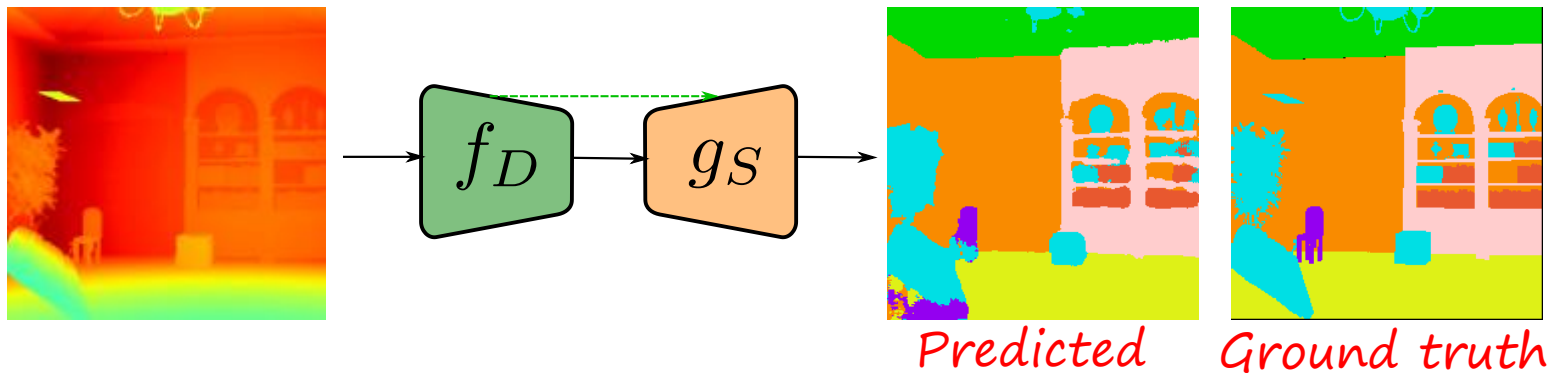
Autoencoders
Shared encoder/decoders

Latent losses
Robust side information (pooling indices)



Zero-pair translation with mix and match networks

Test on zero-pair translation depth-to-segmentation



Comparison: depth-to-segmentation

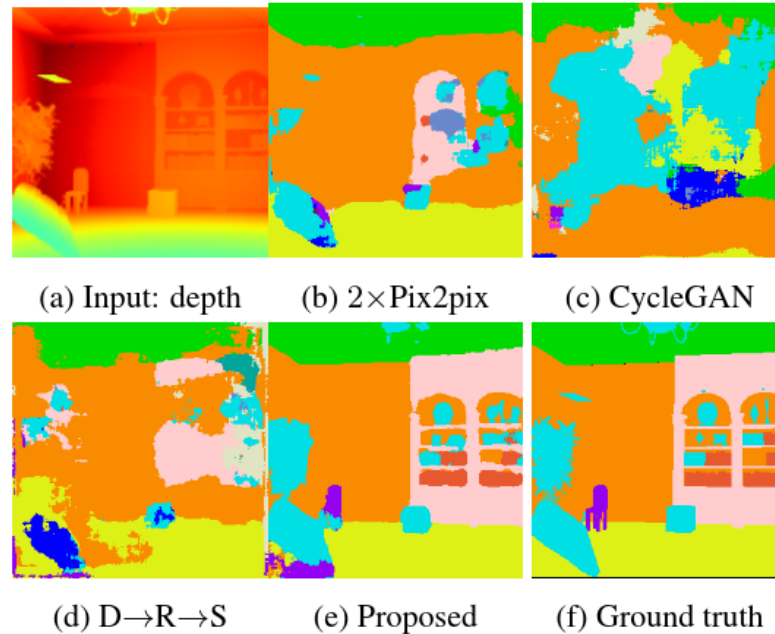
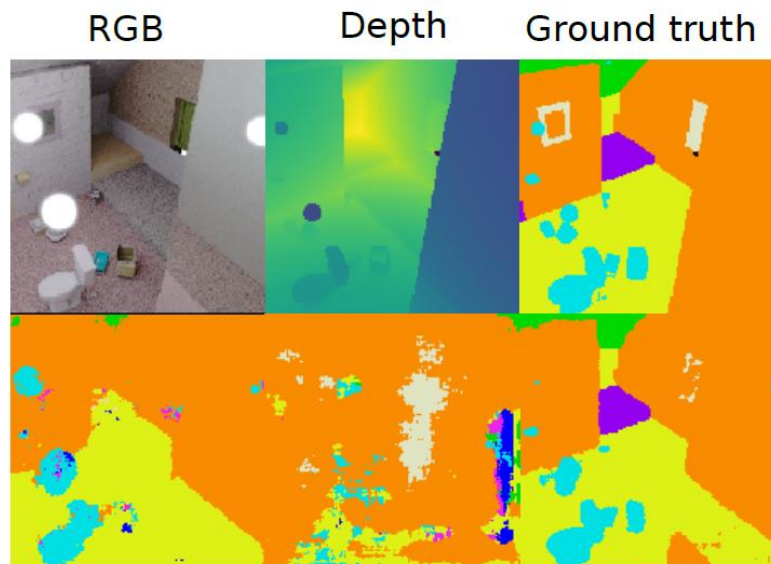


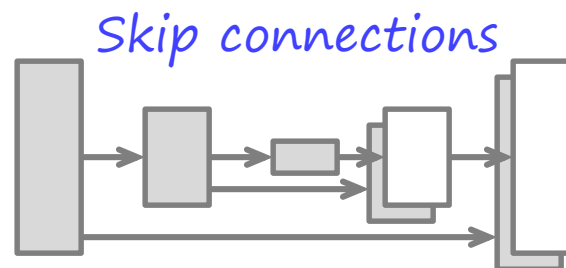
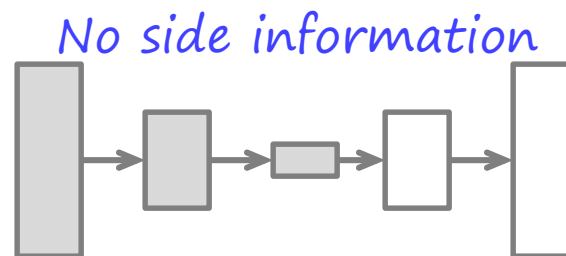
Figure 1: Zero-pair depth \rightarrow segmentation, trained on (depth,RGB) and (RGB,segmentation).

Side information in mix and match networks

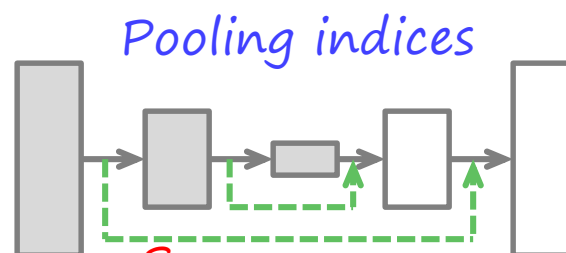


No side information Skip connections Pooling indices

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%



Decoder conditioned on seen encoder(s)



Seems more invariant and robust

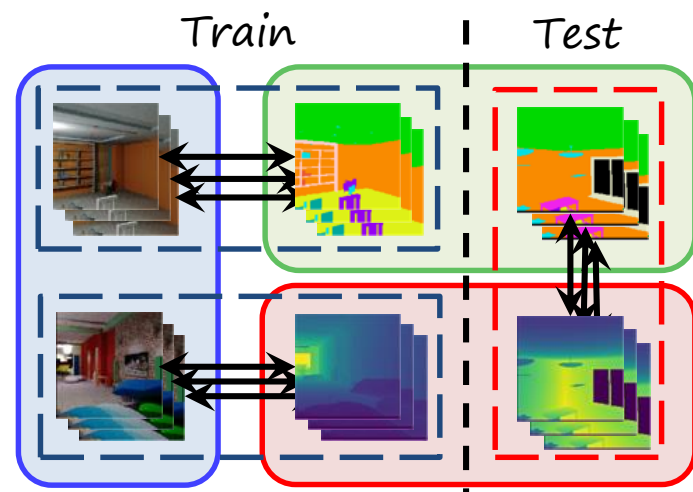
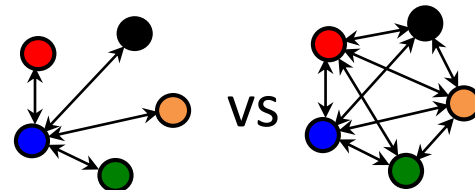
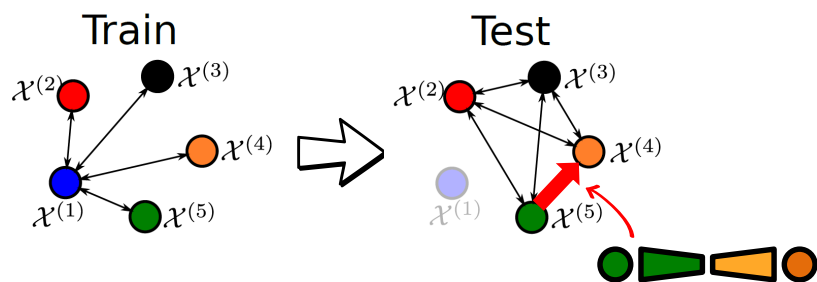
Quantitative evaluation

Method	Conn.	L_{SEM}	Bed	Book	Ceiling	Chair	Floor	Furniture	Object	Picture	Sofa	Table	TV	Wall	Window	mIoU	Global
Baselines																	
CycleGAN [34]	SC	CE	2.79	0.00	16.9	6.81	4.48	0.92	7.43	0.57	9.48	0.92	0.31	17.4	15.1	6.34	14.2
2×pix2pix [10]	SC	CE	34.6	1.88	70.9	20.9	63.6	17.6	14.1	0.03	38.4	10.0	4.33	67.7	20.5	25.4	57.6
M&MNet $D \rightarrow R \rightarrow S$	PI	CE	0.02	0.00	8.76	0.10	2.91	2.06	1.65	0.19	0.02	0.28	0.02	58.2	3.3	5.96	32.3
M&MNet $D \rightarrow R \rightarrow S$	SC	CE	25.4	0.26	82.7	0.44	56.6	6.30	23.6	5.42	0.54	21.9	10.0	68.6	19.6	24.7	59.7
Zero-pair																	
M&MNet $D \rightarrow S$	PI	CE	50.8	18.9	89.8	31.6	88.7	48.3	44.9	62.1	17.8	49.9	51.9	86.2	79.2	55.4	80.4
Multi-modal																	
M&MNet $(R, D) \rightarrow S$	PI	CE	49.9	25.5	88.2	31.8	86.8	56.0	45.4	70.5	17.4	46.2	57.3	87.9	79.8	57.1	81.2

Table 3: Zero-pair depth-to-semantic segmentation. **SC**: skip connections, **PI**: pooling indexes, **CE**: cross-entropy

Summary

- Infer unseen translations by mixing and matching nets
 - Encoder-decoder alignment
- Applications
 - Scalability in pairwise translations
 - Cross-modal zero-pair translation
- Tricks for alignment
 - Shared networks, autoencoders
 - Latent space loss
 - Robust side information (pooling indices)



THANK YOU!

www.cvc.uab.es/lamp
lherranz@cvc.uab.es

More details at <http://www.lherranz.org/2018/08/31/mixmatchnets>



Yaxing Wang



Joost van de Weijer