I|S|T AUSTRIA
Institute of Science and Technology

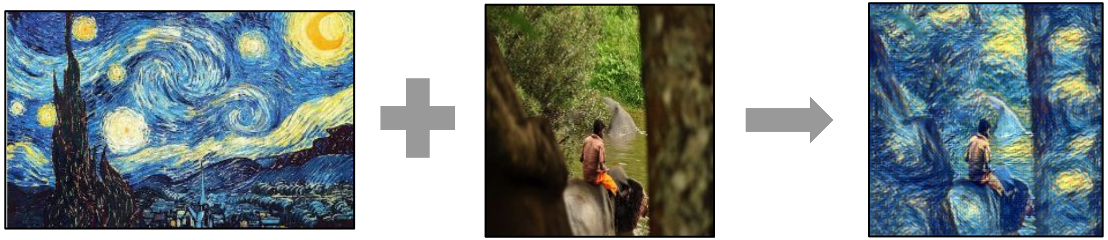
XGAN: Unsupervised Image-to-Image Translation for Many-to-Many mappings

Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, Kevin Murphy

Domain Adaptation for Visual Understanding Workshop
Stockholm, July 13rd, 2018

Introduction

Style Transfer = image-to-image transfer



Two objectives
Style representation ~ Texture
Content representation ~ Structure

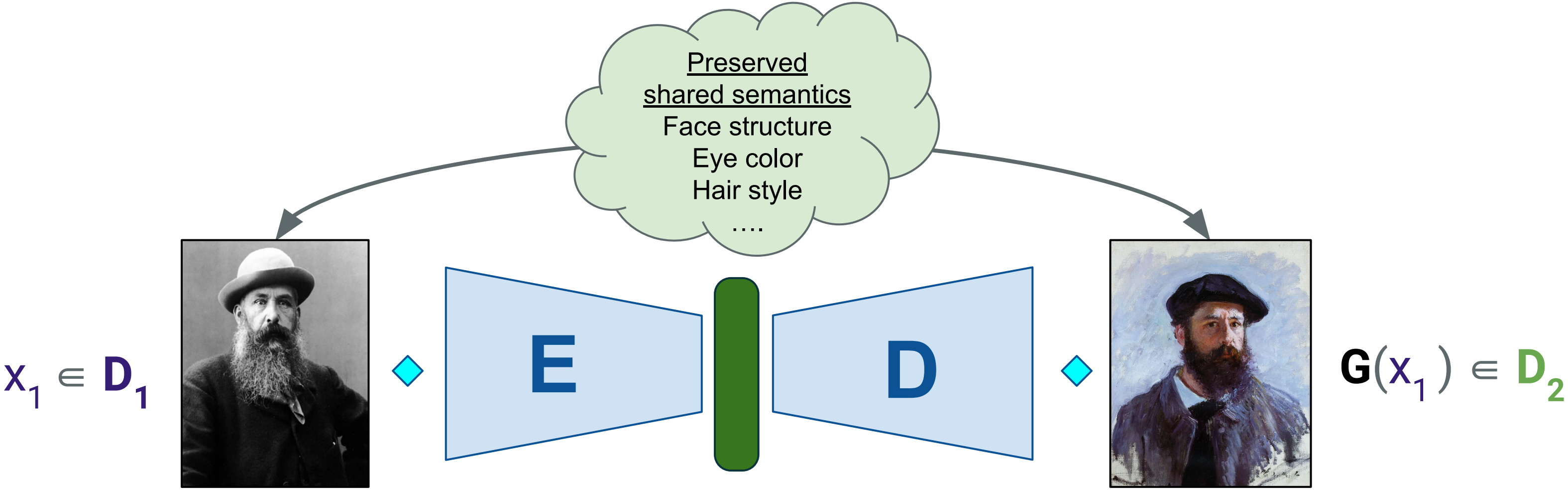
Semantic Style Transfer = corpus-level + feature-level style

Target Domain
(style)

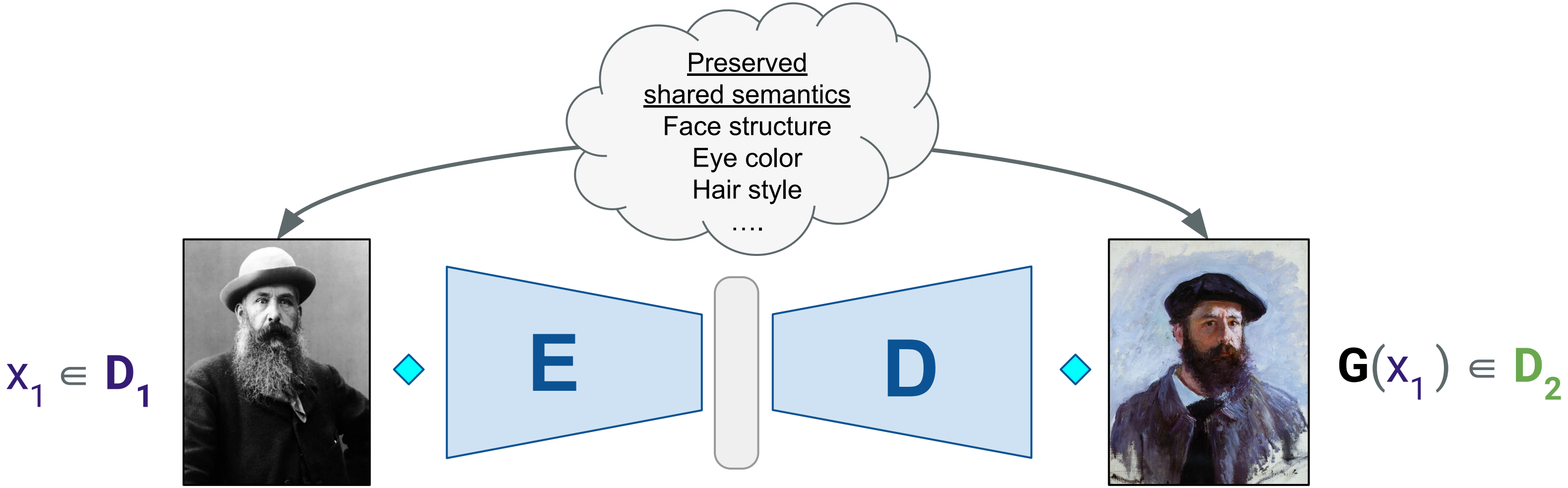
Source Domain
(content)

High-level goal
Transfer the style from one domain to another conditioned on the input content

Problem formulation



Problem formulation



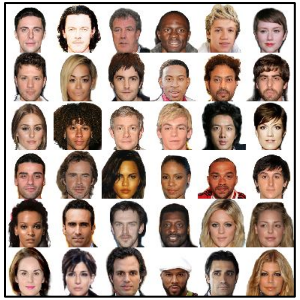
Main difficulties

- No quantitative evaluation of the generated samples (Inception Score…)
- Lack of supervision (paired samples ? semantic labels ?)

Datasets and Applications

Toy Dataset (SVHN → MNIST)

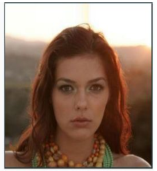
Main Dataset (Face → Cartoon)



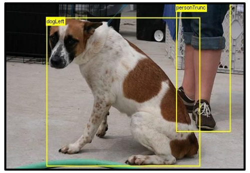
VGGFaces

CartoonSet
public release at:
google.github.io/cartoonset/

Other Examples...



Face → Drawn Portraits

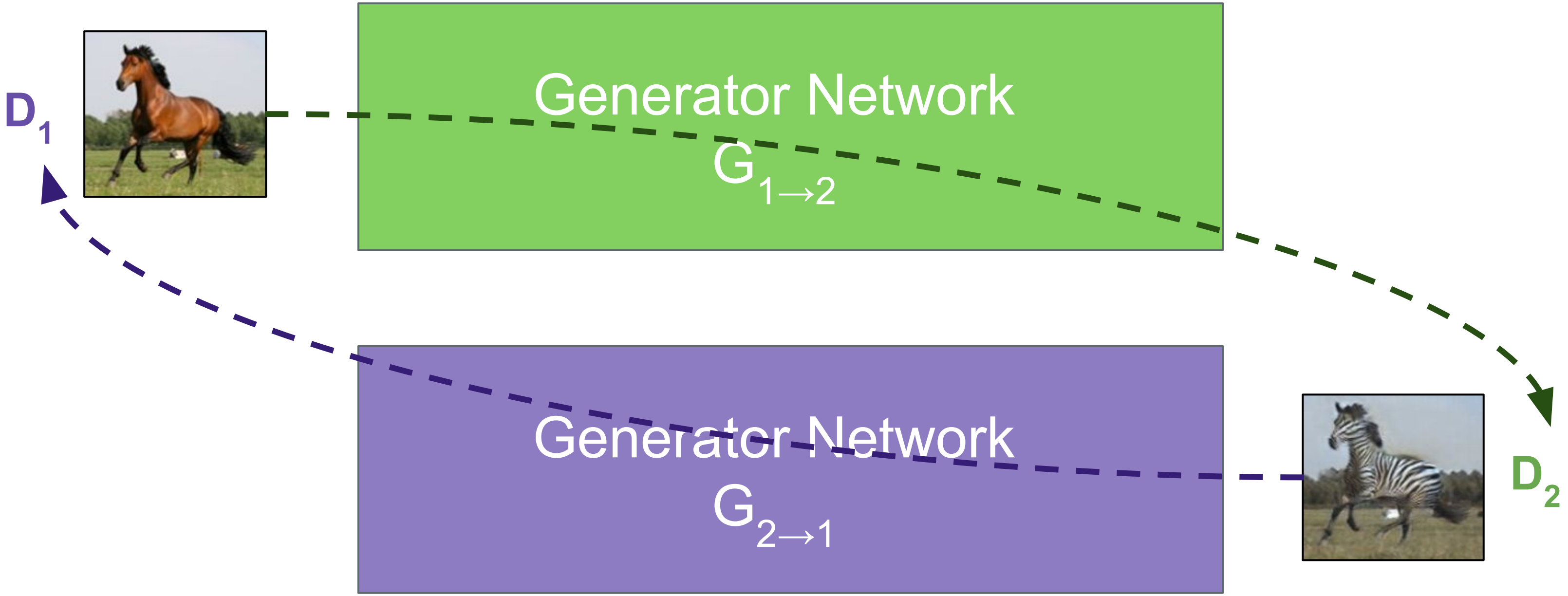


Dog (PASCAL) → Paintings (VGG)

Related Work

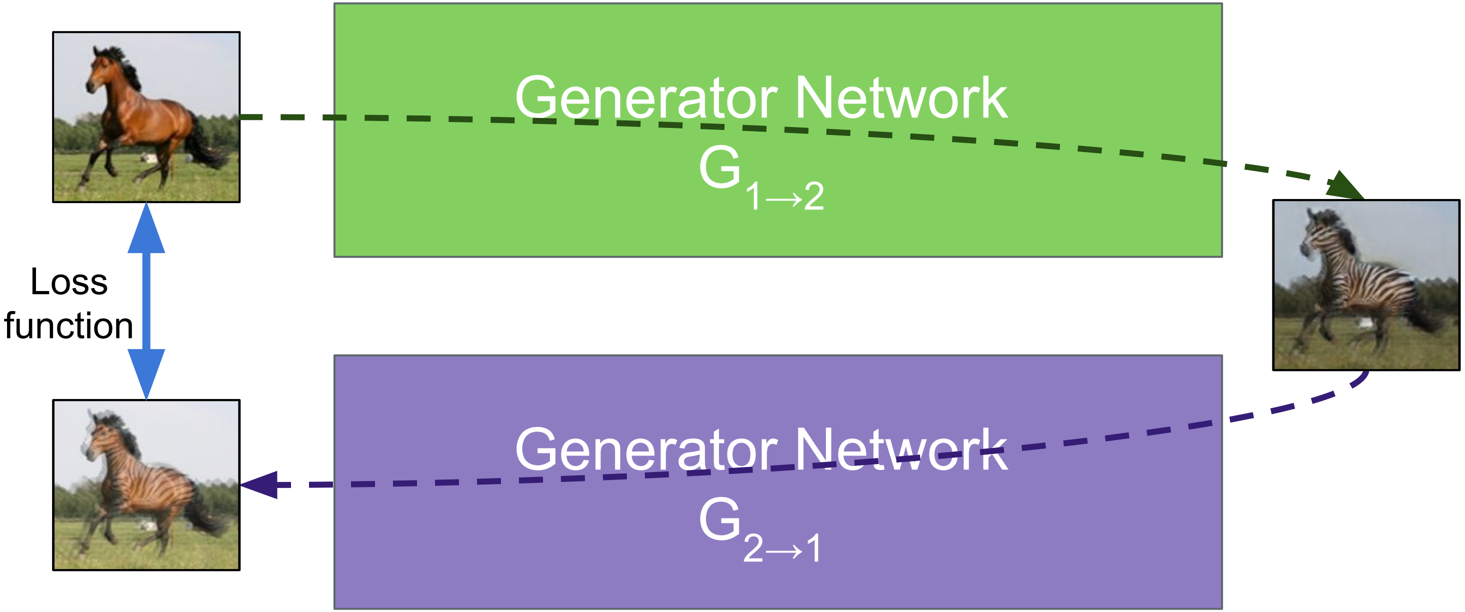
CycleGANs: Cyclic Consistency (+ DualGAN, DiscoGAN)

"Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", Zhu et al., ICCV'17



CycleGANs: Cyclic Consistency (+ DualGAN, DiscoGAN)

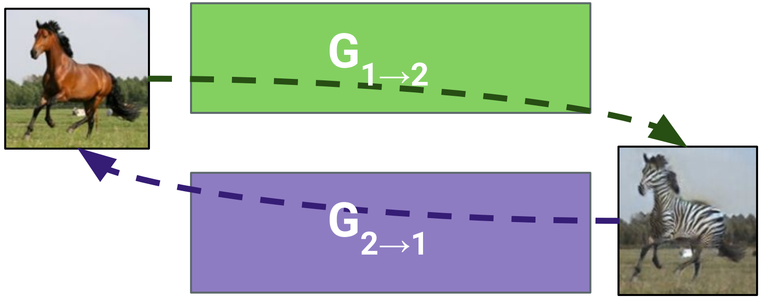
“Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”, Zhu et al., ICCV’17



CycleGANs: Cyclic Consistency (+ DualGAN, DiscoGAN)

"Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", Zhu et al., ICCV'17

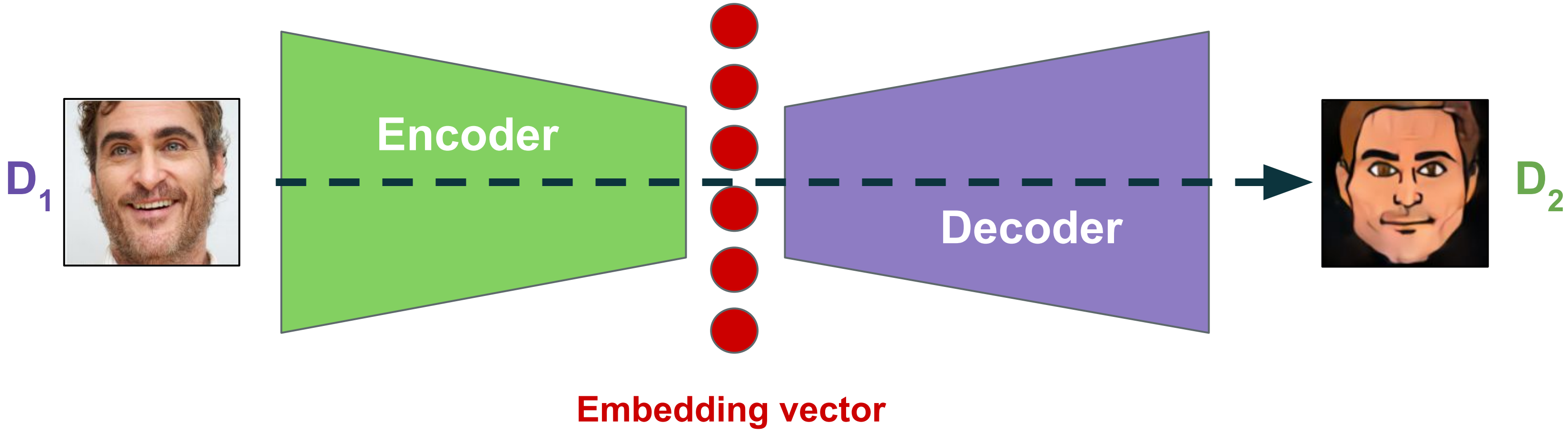
- Learn both mappings simultaneously
- **Cycle-consistency** loss: $G_{2\rightarrow 1} \circ G_{1\rightarrow 2} = id$



- [✓] **Self-supervised** method
- [✕] Two distinct generators, no sharing
- [✕] In practice, **pixel-level** structure hard to modify

Domain Transfer Network: Semantic Consistency

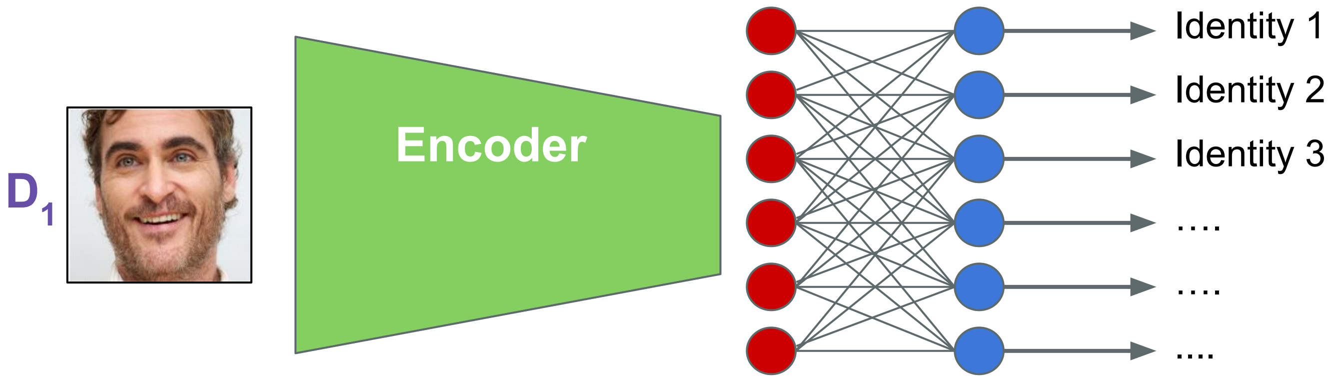
"Unsupervised Cross-Domain Image Generation", Taigman et al., ICLR'17



Domain Transfer Network: Semantic Consistency

"Unsupervised Cross-Domain Image Generation", Taigman et al., ICLR'17

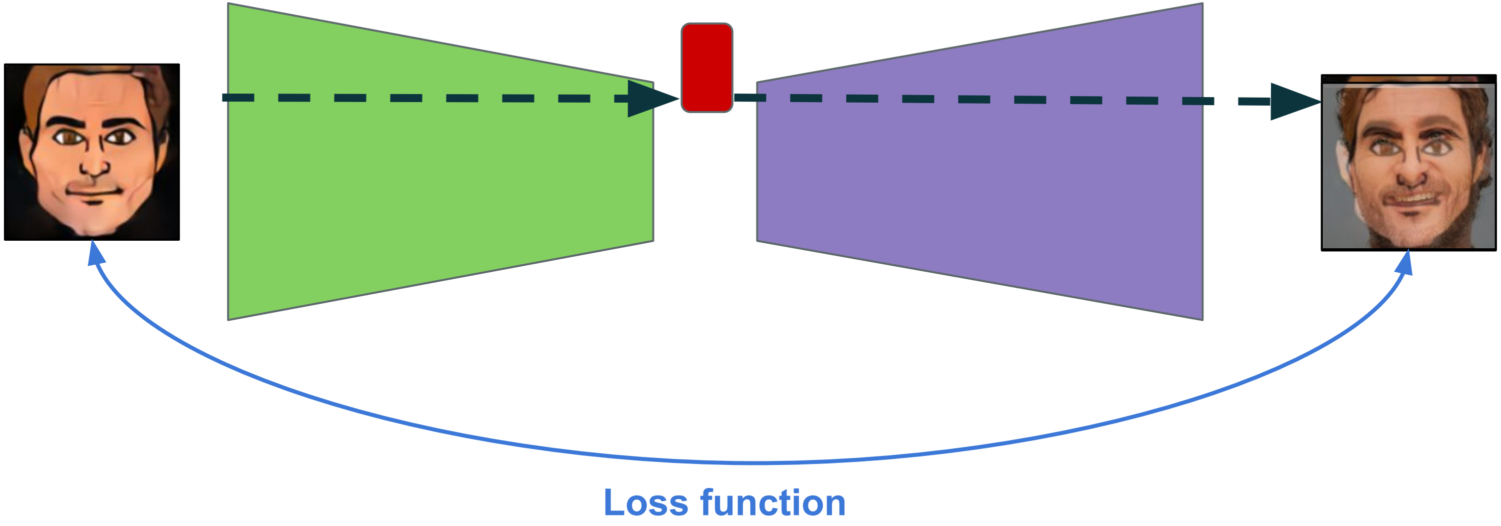
Fixed encoder, pre-trained on Face recognition



Domain Transfer Network: Semantic Consistency

"Unsupervised Cross-Domain Image Generation", Taigman et al., ICLR'17

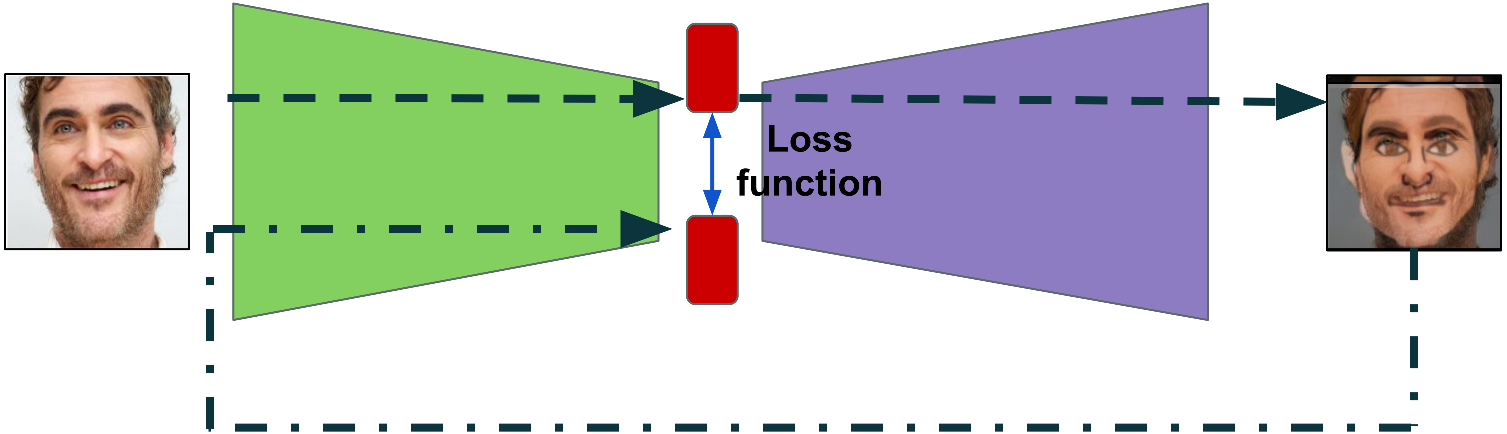
First loss: Reconstruction loss for inputs from the target domain



Domain Transfer Network: Semantic Consistency

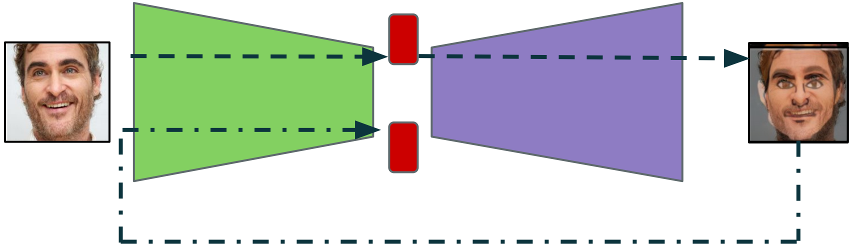
“Unsupervised Cross-Domain Image Generation”, Taigman et al., ICLR’17

Second loss: semantic consistency loss at the feature-level

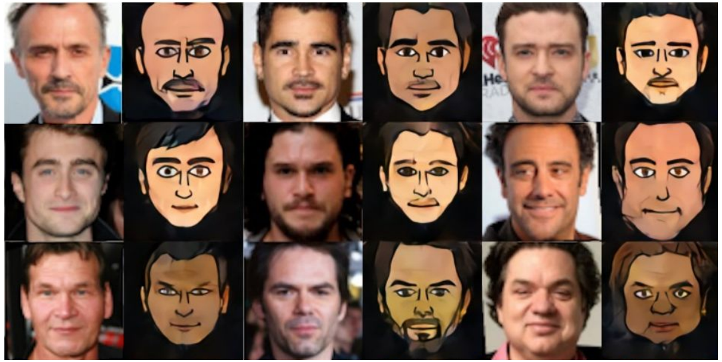


Domain Transfer Network: Semantic Consistency

"Unsupervised Cross-Domain Image Generation", Taigman et al., ICLR'17



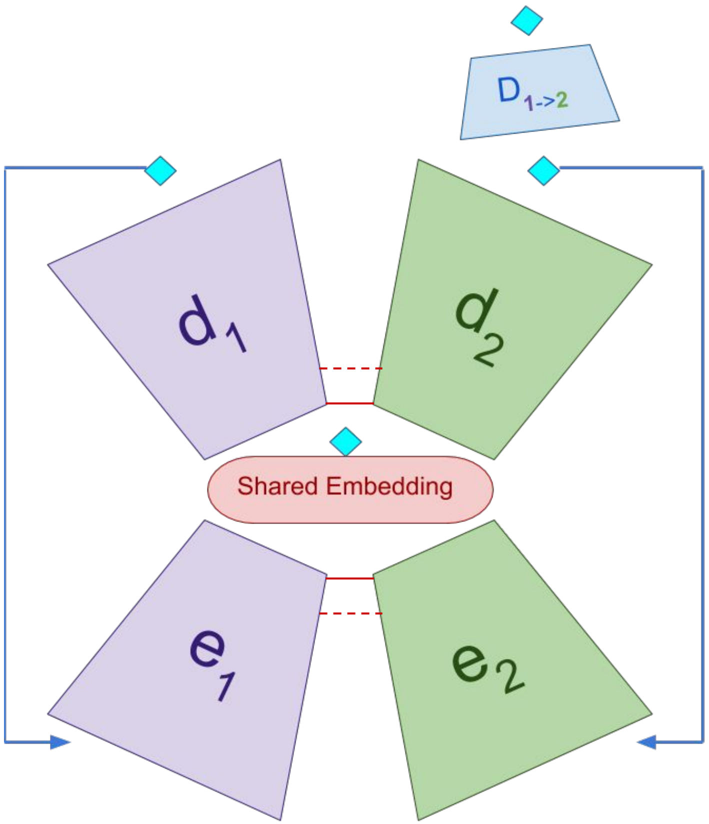
- Fixed pre-trained encoder
- **Feature-level** consistency



- [✓] **Feature-level** transformation
- [✓] **Semantic consistency** loss
- [✗] Fixed encoder for **both** domains

Proposed Model

Proposed Model - «XGAN» (“Cross-GAN”)



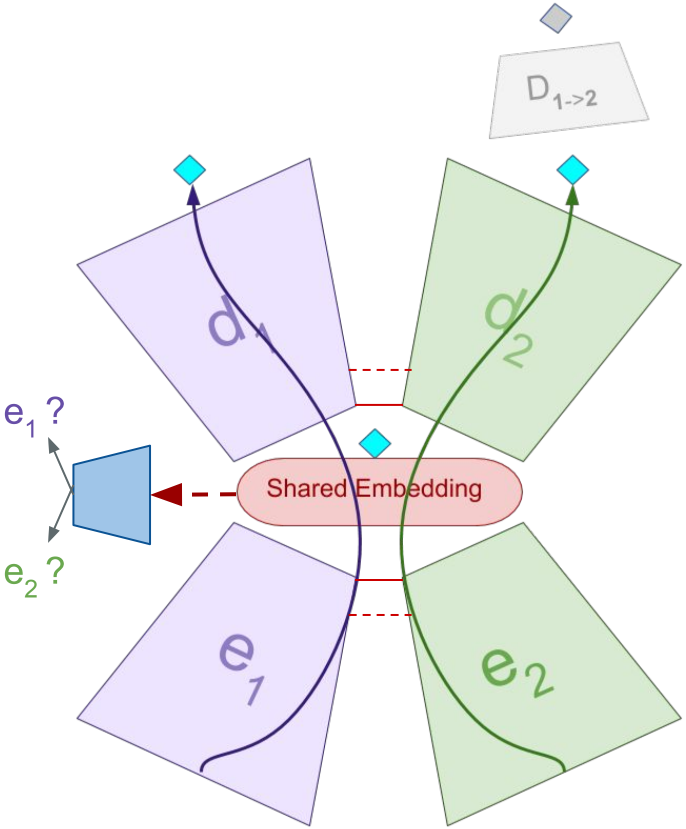
Intuition

- Learn a **joint embedding** on both domains
- Cross-domain encoder/decoder pair

Supervision

- **Self-supervision**: the transformation should be invariant under the **embedding**

Proposed Model - «XGAN»



Domain-adversarial auto-encoder

- Reconstruction losses

Embeddings encode **enough information** to reconstruct the inputs perfectly

- Domain-adversarial loss

Embeddings should lie in a **common subspace**

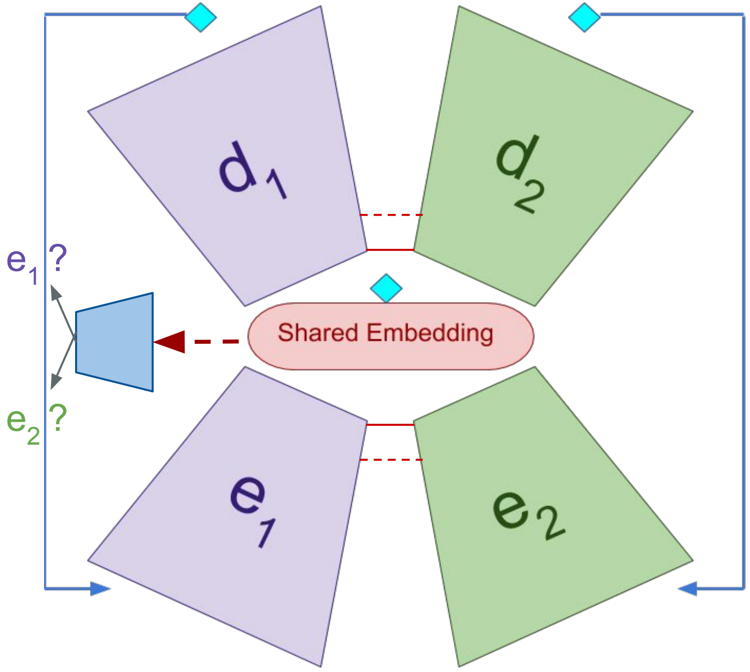
Proposed Model - «XGAN»

Domain adversarial Neural Network

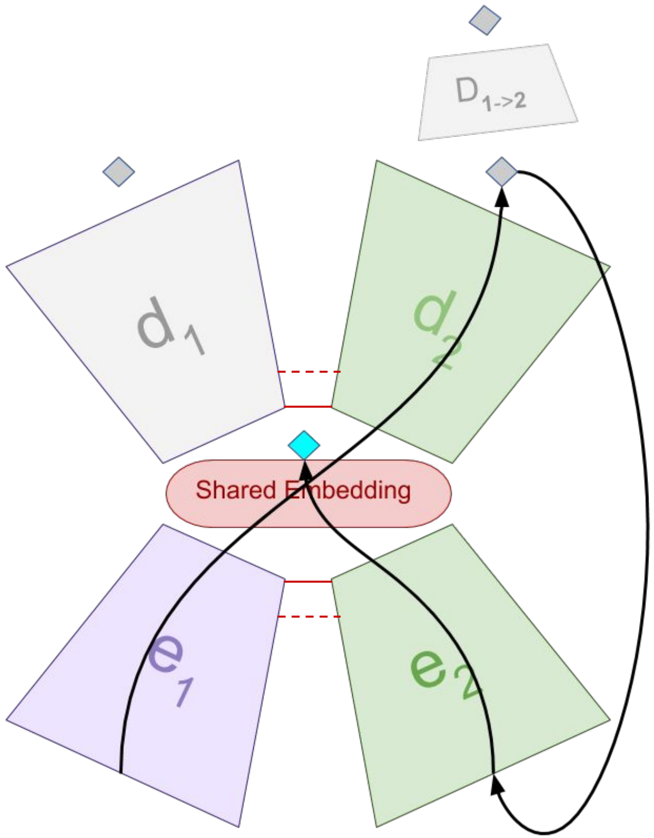
"Domain Adversarial Training of Neural Networks", Y.Ganing et al., JMLR'16

- Classifier c_{DANN} distinguishes between embeddings from D_1 or D_2

- Adversarial training via gradient reversal layer (very stable in practice)



Proposed Model - «XGAN»



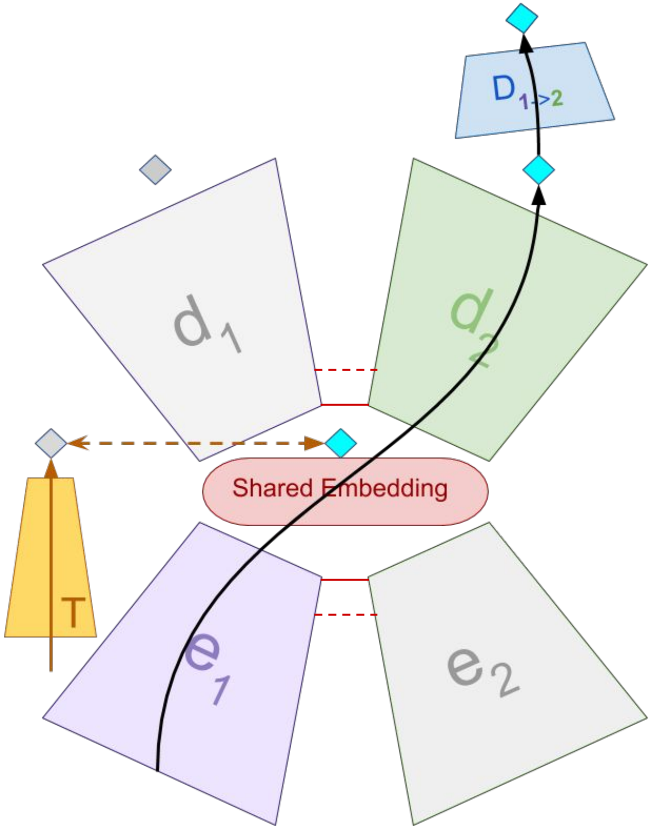
Semantic consistency

- Semantic consistency loss $\mathbf{D_1} \rightarrow \mathbf{D_2}$

The learned embedding is preserved through the domain transformation: **Feature-level self-supervision**

- And its mirrored counterpart $\mathbf{D_2} \rightarrow \mathbf{D_1}$

Proposed Model - «XGAN»



Optional refinements

- GAN loss (add discriminator $\mathbf{D}_{1 \rightarrow 2}$)

Produce realistic source → target samples

- Teacher network (e.g., FaceNet)

Incorporate prior semantic knowledge from the source domain

Qualitative experiments

Comparison with baselines

	CycleGAN	DTN	XGAN
Mappings	both	$D_{1 \rightarrow 2}$	both
Shared representation	No	Fixed	Yes
Supervision	None	Fixed embedding	Optional teacher network
Transformation	Pixel-level	Feature-level	Feature-level

Baseline 1 - CycleGAN

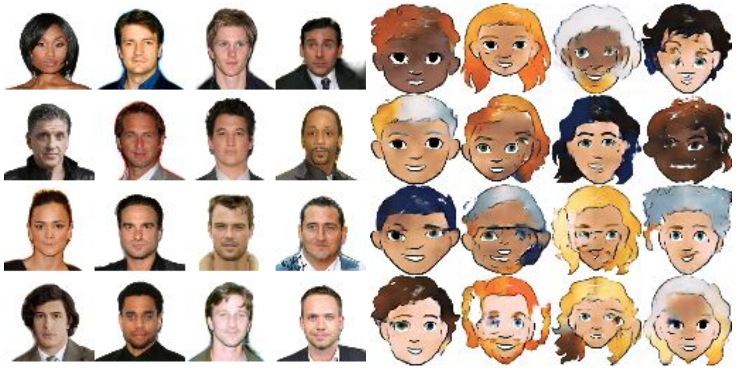
- The CycleGAN setting (Pix2Pix/U-Net architecture) enforces strongly similar pixel structures

Example test samples when transferring Faces to Cartoon with CycleGAN.
With longer training or a deeper Encoder (e.g. Resnet) we obtain better (more cartoon-ish) samples but with no semantic correspondences to the input face.

Baseline 2 - DTN

- The fixed encoder (FaceNet here) cannot bridge the visual shift between the two domains (Face and Cartoon)

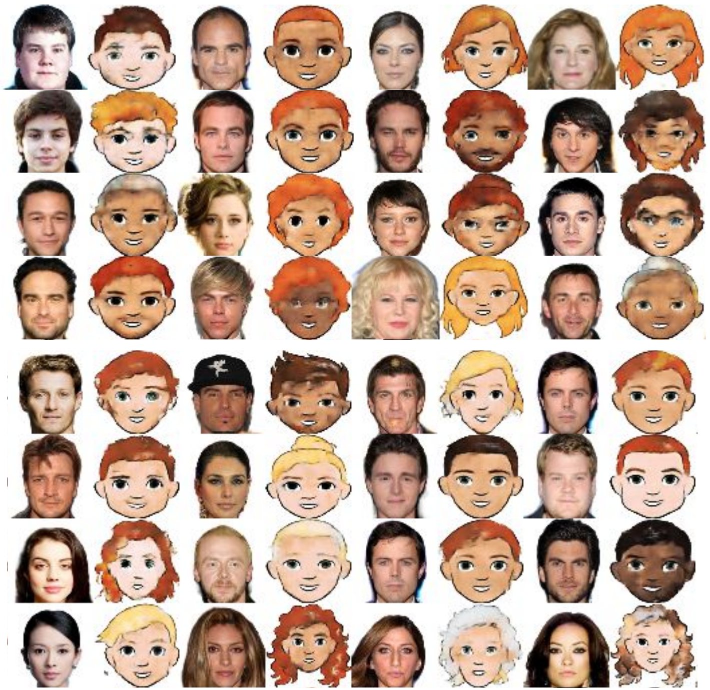
[✓] **SVHN** → **MNIST** (1350 iterations)
The embedding captures the input number's class across the two domains (MNIST *acc ~ 0.7*)



[×] **Face** → **Cartoon** (200k iterations)
The fixed embedding does not generalize well across these two very different domains

Results - XGAN (Source to Target)

64x64 Samples (generated from the test set)



Typical failure cases

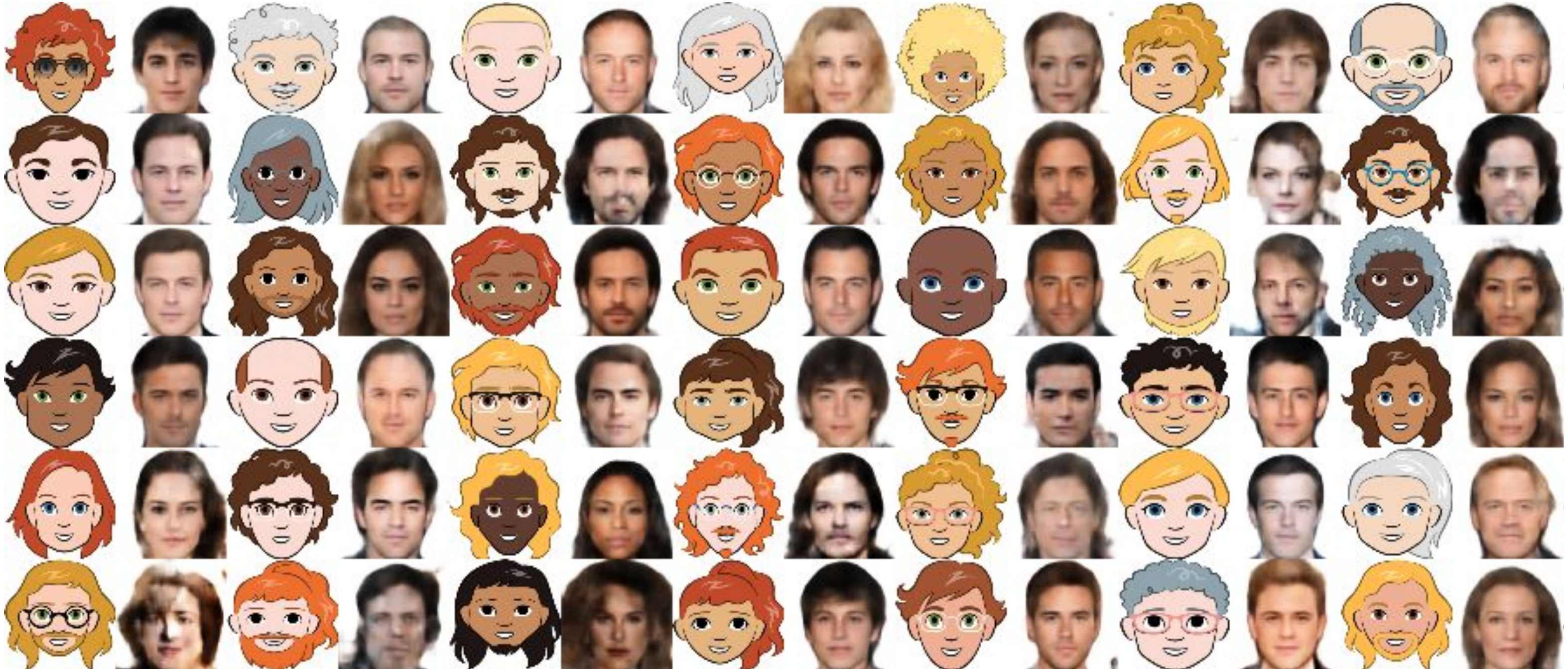
Hair mis-match (e.g., shades of red and grey are over represented in the training set)

Hair hallucinations

wrong skin tone (lighting ?)

Understanding the learned embedding

Source -> Target direction also gives intuitive insights in the model



Experiments (*Active losses*: L_{DA}, L_{Rec})

Failure cases

Low capacity models fail at reconstructing the inputs

DA classifier is too powerful

Necessary for realistic target outputs: preliminary **success criterion**

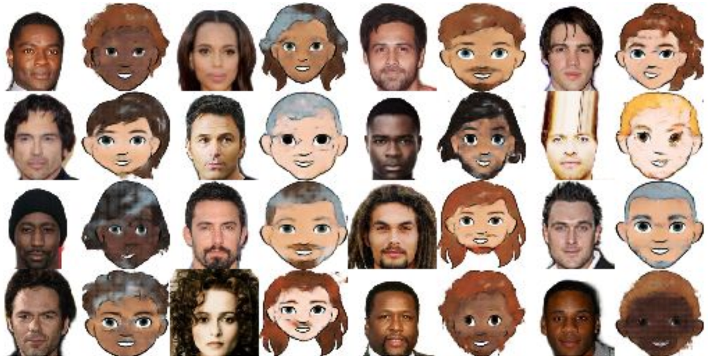
Random samples

In practice, good reconstructions and domain adversarial balance are easy to achieve **without extensive tuning**

Experiments (*ablating the teacher loss*)

Teacher supervision

- Constrain the embedding to more realistic faces
- But harder to tune: High weights lead to lack of variability

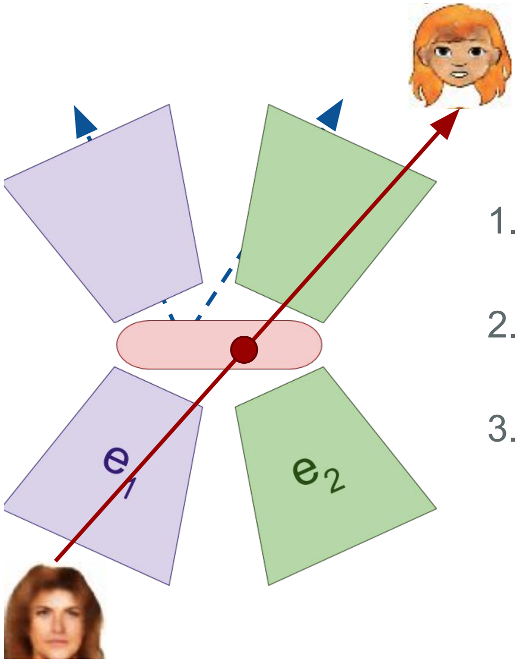


With teacher loss, without semantic consistency

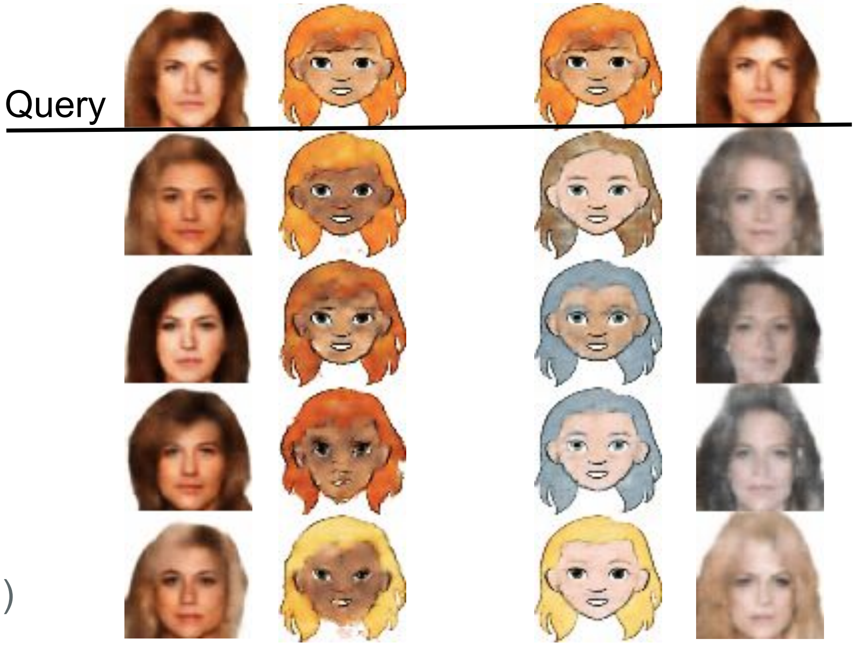
With semantic consistency, without teacher

Understanding the learned embedding

Nearest Neighbor search



1. Compute **query** embedding ●
2. Search **NNs** ● in the embedding space
3. Pass ● through both decoders (visualization)



Top-4 neighbors in $\mathbf{e_1}(\mathbf{D_1})$

Top-4 neighbors in $\mathbf{e_2}(\mathbf{D_2})$

Conclusions

- The **domain adversarial** setting and **semantic consistency** losses contribute to learning an embedding relevant to both domains
- Using a **GAN** framework further improves the sample quality but makes the training unstable
- **Teacher supervision** brings useful supervision at a small cost
- Application to more general domain adaptation framework with quantitative evaluation in future work

Thank you for your attention

Questions ? Suggestions ?

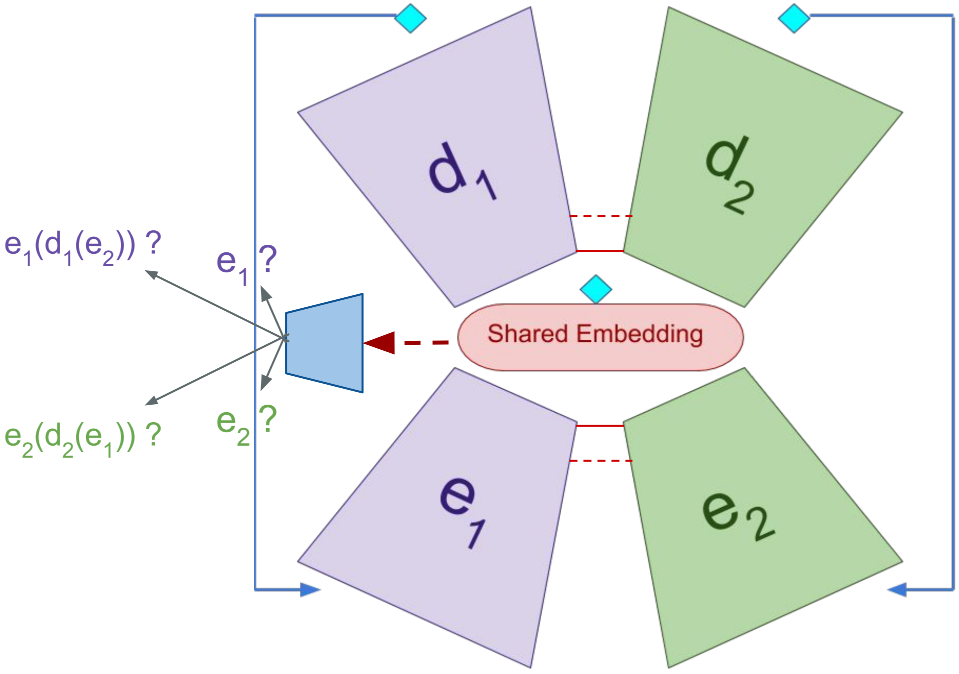
Appendices

Proposed Model - «XGAN»

Additional remark 1: Multi-class DANN

In practice, **4** classes rather than **2**:

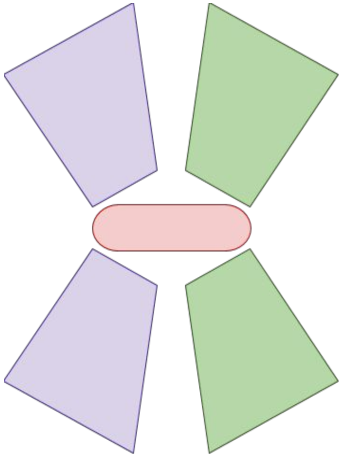
- e_1 // e_2: Shared embedding
- e_1 // $e_1 \circ d_1 \circ e_2$ and e_2 // $e_2 \circ d_2 \circ e_1$: Embeddings after transfer lie in the same subspace ~ Weak semantic consistency



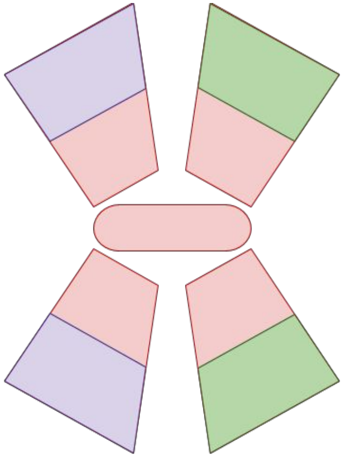
=> Multi-class DANN
(or multiple binary DANNs)

Proposed Model - «XGAN»

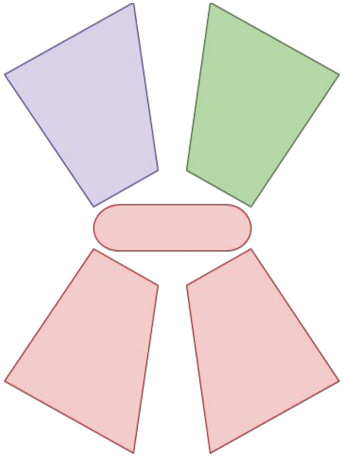
Additional remark 2: Layer Sharing in the Autoencoder



No sharing
Low-capacity



Partial symmetric sharing
More flexibility in the generated samples, but slower to converge to good quality samples

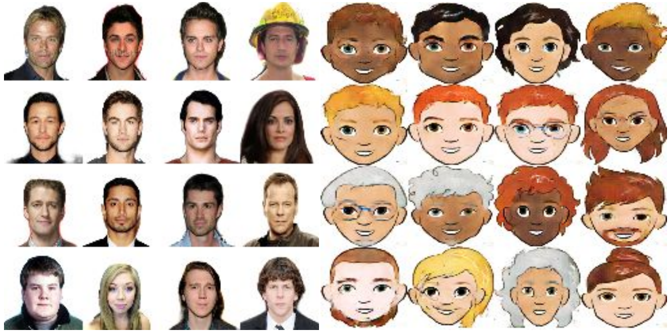


Fully shared encoder
Good quality (crisp) samples but semantics are not always well preserved

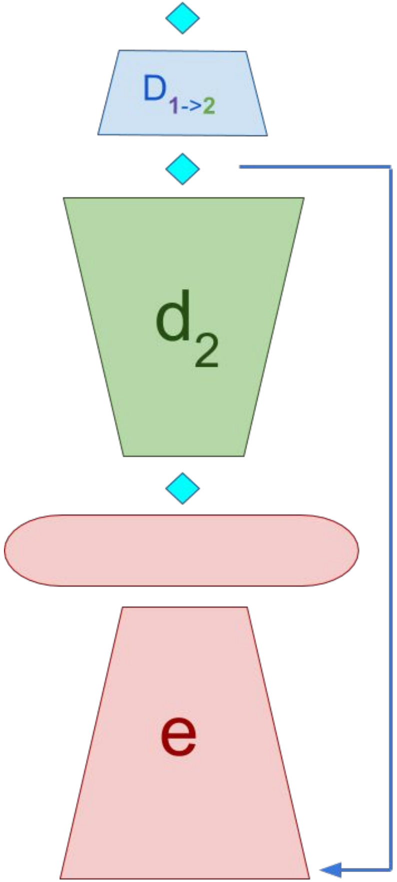
Fine-tuned DTN

- Experiment: Training/fine-tuning the embedding
- Hard to tune, and no control over the initial domain


[✓] **SVHN** → **MNIST** (1350 iterations)
Samples quality is improved (MNIST *acc ~ 0.86*)



[~] **Face** → **Cartoon** (80k iterations)
Some semantic properties are better captured (e.g., gender, skin tone)

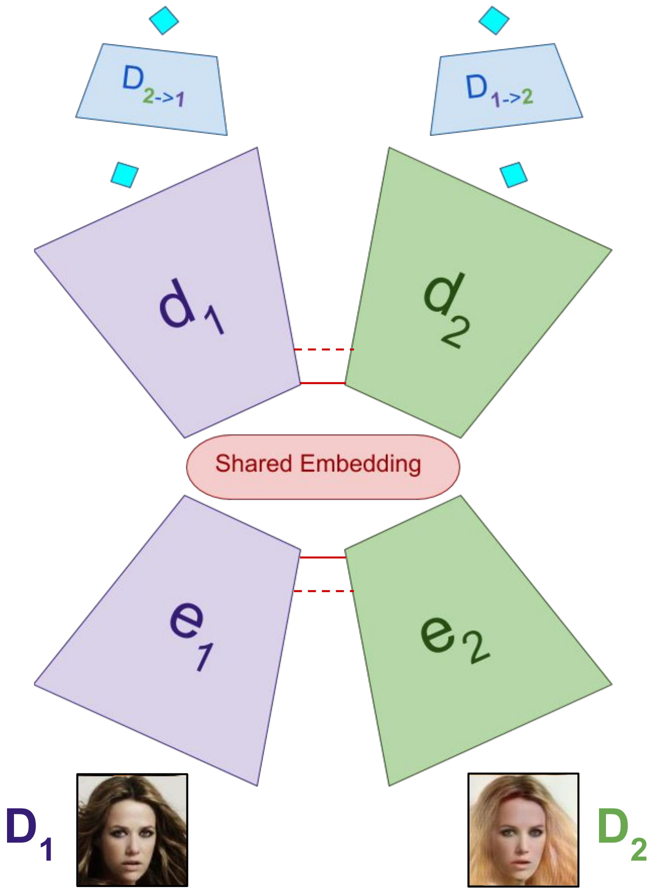


Related Work - UNIT

- The mappings are learned as two VAEGANs with a common representation.
- Two **GAN** objectives
- Two **VAE** objectives (in particular, include reconstruction losses)

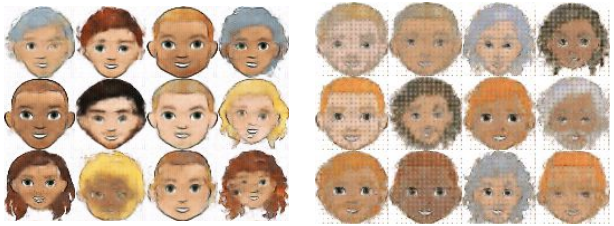
[✅] Pros
- Natural sampling from the VAE framework
- Learned **joint representation** of the two domains

[❌] Cons
- No explicit constraint on the shared embedding
- Pixel-level objective



“Unsupervised Image-to-Image Translation Networks”, Liu et al., arXiv’17

Experiments (*Active losses*: all **+/-** L_{GAN})



Without GAN, the samples look good at first (left) but lack diversity in the long run (right)

Adding the GAN loss (left) and discriminator thresholding (right)

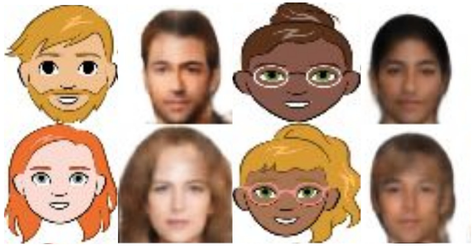
- Reasonable sample quality without discriminator loss but adding the **GAN objective** yields crisper samples
- The discriminator is typically very powerful right from the start → only train if accuracy is below a certain **threshold**

Experiments (*Active losses*: all)

Semantic consistency

- **Both directions** give insight on what the embedding is learning
- Could potentially be used as a criterion for **model selection** (self-supervision)

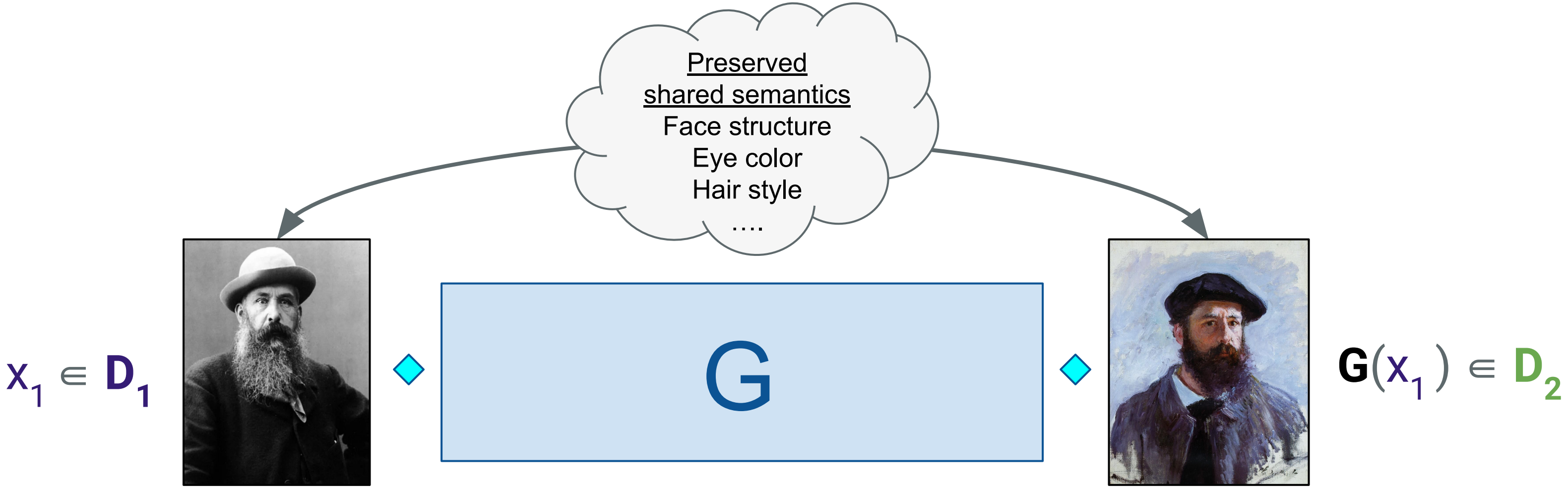
Source to Target



Target to Source

Test samples with lowest (top) and highest (bottom) semantic consistency distance (face → cartoon)

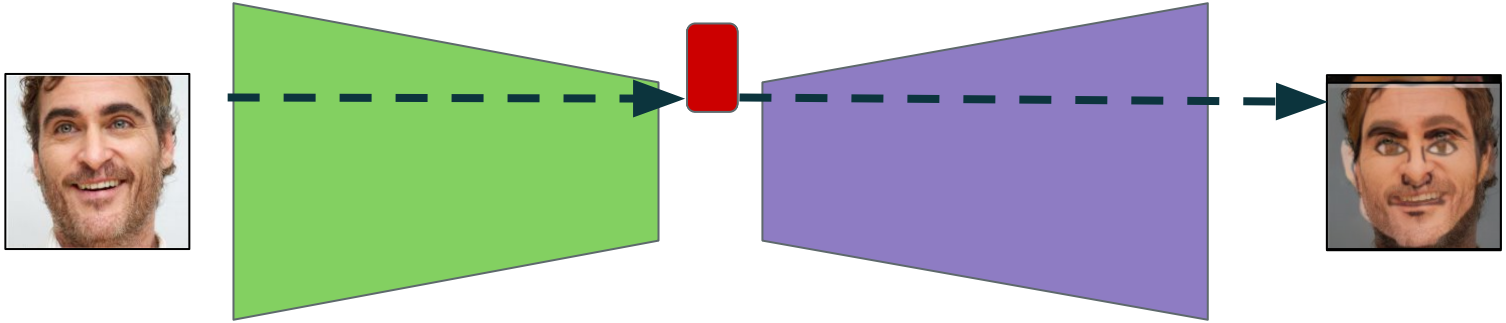
Problem formulation



Domain Transfer Network: Semantic Consistency

"Unsupervised Cross-Domain Image Generation", Taigman et al., ICLR'17

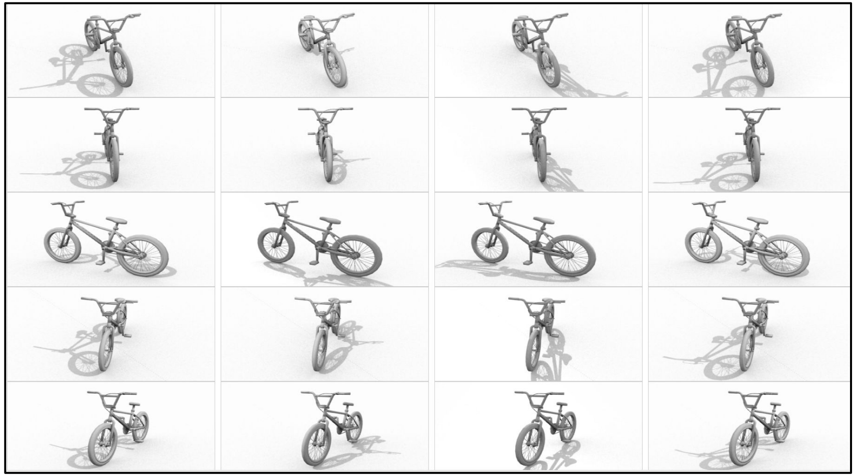
Second loss: semantic loss at the feature-level



The VisDA dataset

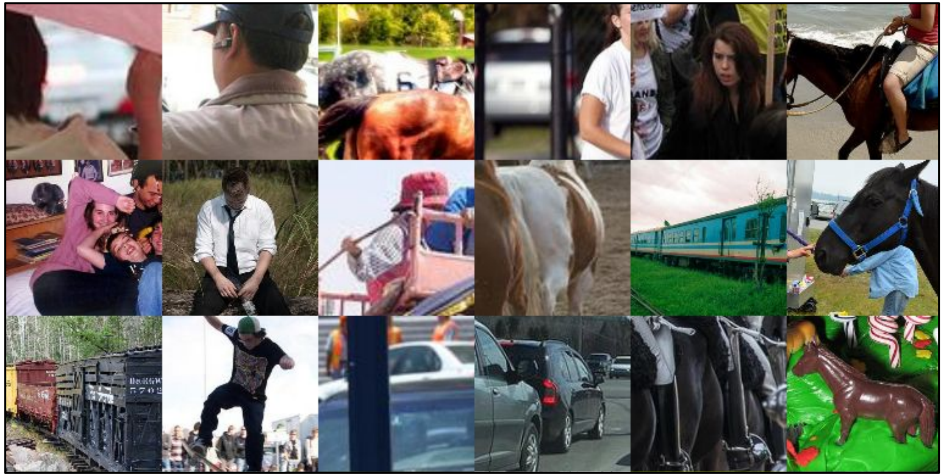
Synthetic Domain (labeled) [source]

12 classes, unbalanced set (~8k per class), grayscale 3D models.

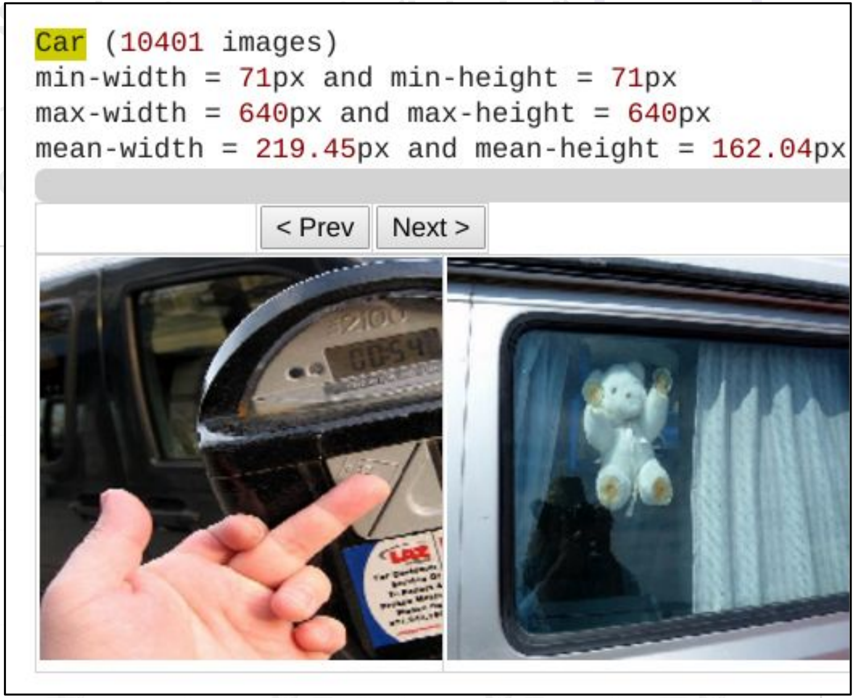


Real Domain (unlabeled) [target]

Varied natural images from the same object classes as the source dataset

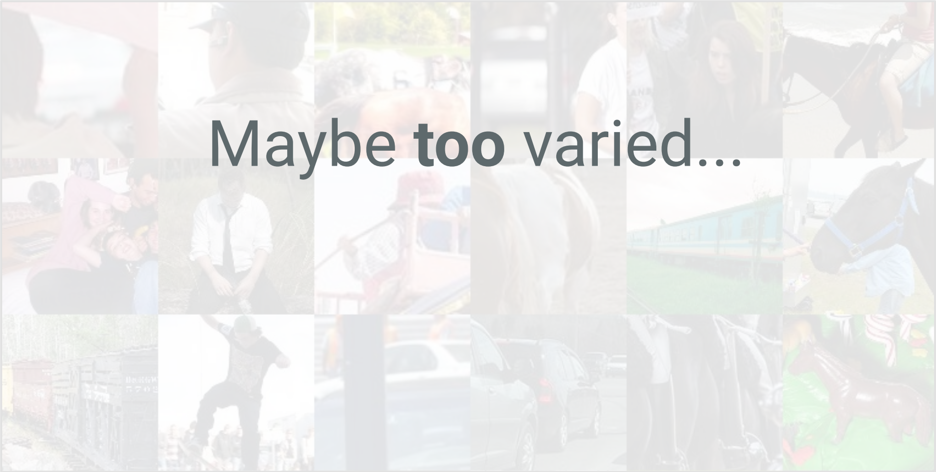


The VisDA dataset



Real Domain (unlabeled) [target]

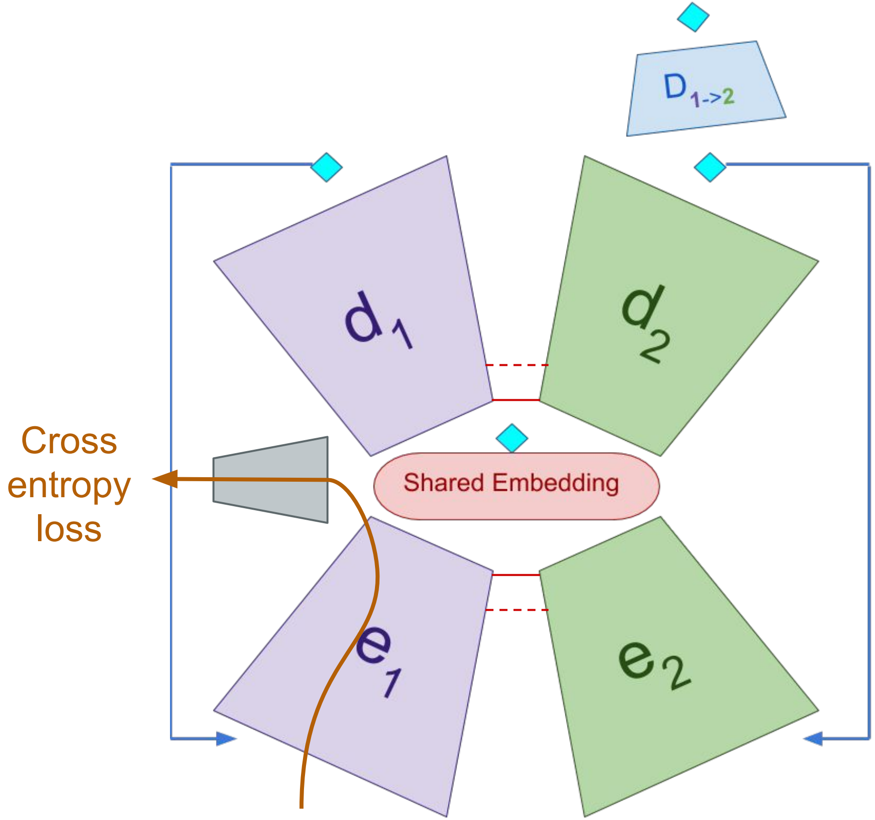
Varied natural images from the same object classes as the source dataset



Adding supervision for the VisDA setting

- Classification “task tower” on top of the embedding for the source labels
- ImageNet pre-trained teacher network on the target domain

→ Two conflicting supervision sources: Alternating training scheme

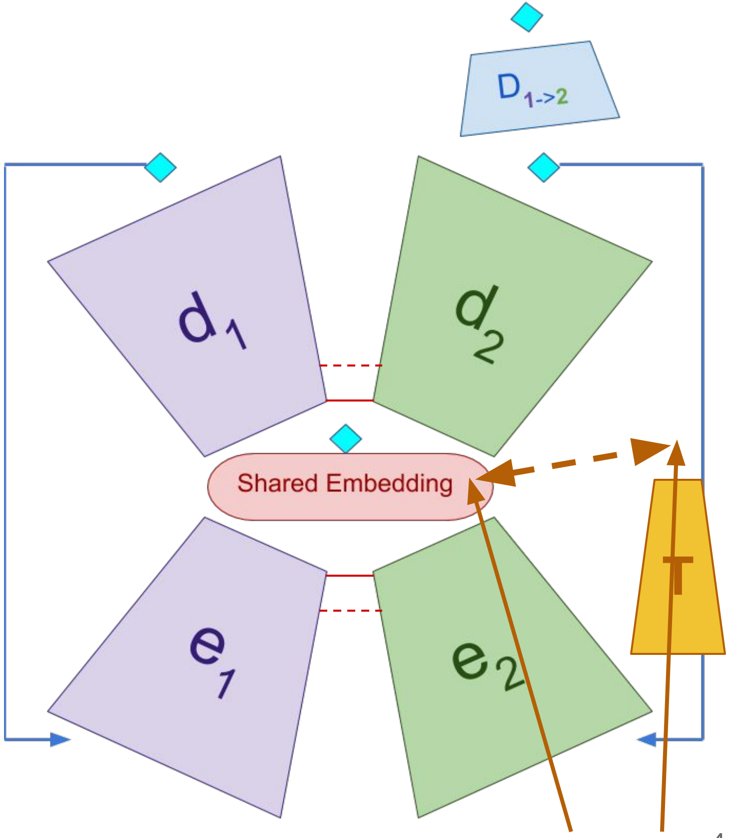


Adding supervision for the VisDA setting

- Classification “task tower” on top of the embedding for the source labels

- **(optional)** ImageNet pre-trained teacher network on the target domain

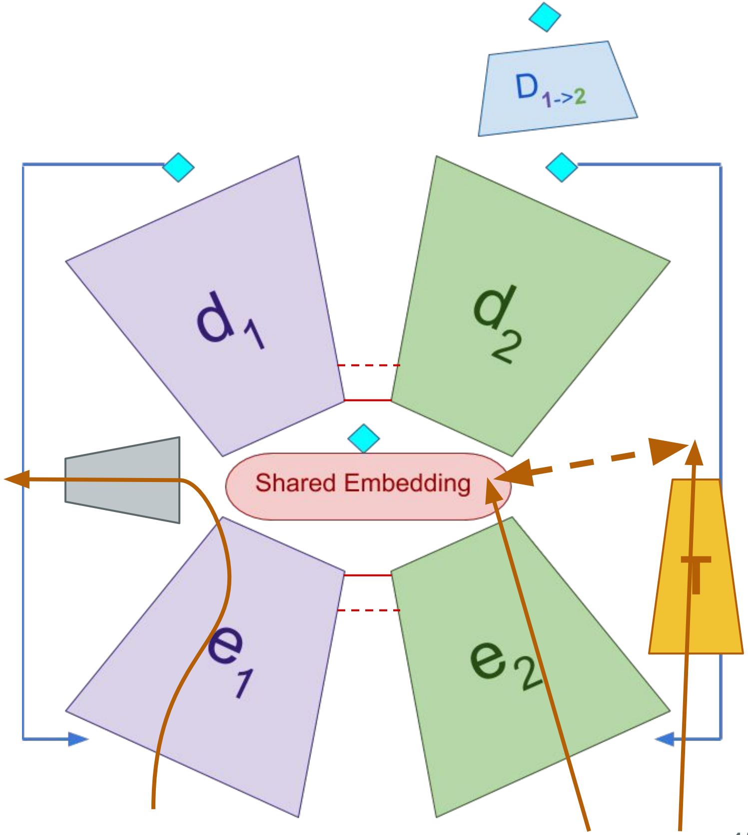
→ Two conflicting supervision sources: Alternating training scheme



Adding supervision for the VisDA setting

- Classification “task tower” on top of the embedding for the source labels
- ImageNet pre-trained teacher network on the target domain

→ Two conflicting supervision sources: Alternating training scheme



Results

- As expected: Classifier overfits to the source dataset
- However: the adaptation losses were not enough to bridge the gap sufficiently (**0.45** acc.)
- The teacher network is mandatory in this setting (**0.2** acc, no other entry, track cancelled…)