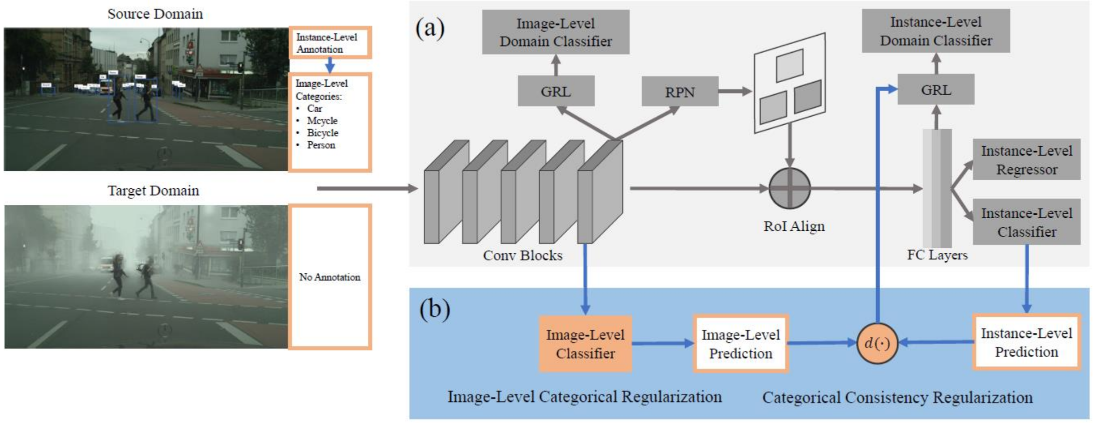
Exploring Categorical Regularization for Domain Adaptive Object Detection

The paper proposes a plug-and-play categorical regularization method for object detection. The main intuition is that image-level classification loss endows discriminative localization ability of CNNs trained on source domain, and suppresses the background activations on source domain. The consistency between image-level and instance-level prediction reflects the importance of target instance proposals for adaptation.

Main idea:

The proposed methods are compatible with other adaptation methods. Do image-level and instance-level regularization.

1. Image level: train a multi-class classifier head for images. The classifier includes all objects of the dataset. Check if the image contains the object.
2. Instance-level: detection heads will give a probability of the instance as class C. The previous multi-class classifier head will also give a probability of if the image includes the object C. penalize the difference between them.



(a) Domain Adaptive Faster R-CNN Series

(b) Proposed Categorical Regularization Framework

Figure 2. Overview of our categorical regularization framework: a plug-and-play component for the Domain Adaptive Faster R-CNN series [1, 28]. Our framework consists of two modules, *i.e.*, image-level categorical regularization (ICR) and categorical consistency regularization (CCR). The ICR module is an image-level multi-label classifier upon the detection backbone, which exploits the weakly localization ability of classification CNNs to obtain crucial regions corresponding to categorical information. The CCR module considers the consistency between the image-level and instance-level predictions as a novel regularization factor, which can be used to automatically hunt for hard aligned instances in the target domain during instance-level alignment.

Cross-Domain Detection via Graph-Induced Prototype Alignment

The paper considers the domain adaptation on object detection. It focuses on two problems:1. The generated proposals of the same object commonly deviate from objects; 2. A single instance cannot represent the multi-modal information. i.e. specific scale or orientation.

Main idea:

Propagate information of adjacent proposals and merge different modal information into prototypes.

1. Generate region proposals
2. Construct relation graph of proposals.
3. Merge features of several proposals of one instance by GCN
3. Merge instance-level feature representations for each object
4. Derive per-category prototypes
5. Category-level alignment by metric learning

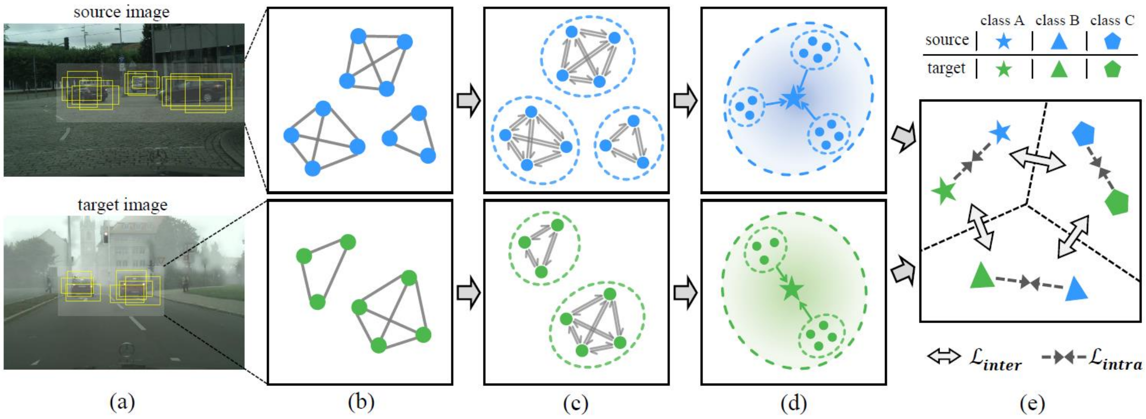


Figure 2. **Framework overview.** (a) Region proposals are generated. (b) Constructing the relation graph on produced region proposals. (c) More accurate instance-level feature representations are obtained through information propagation among proposals belonging to the same instance. (d) Prototype representation of each class is derived via confidence-guided merging. (e) Performing category-level domain alignment through enhancing intra-class compactness and inter-class separability.

Rethinking Classification and Localization for Object Detection

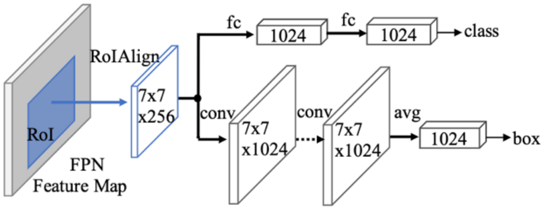
The paper empirically compares the advantages of fully-connected layer and CNN layer in terms of object detection, and find that FC head is better for classification and CNN is better for regression.

FC head has significantly less spatial correlation than CNN head which makes it easier to discover if a proposal covers a complete or partial object but it’s not as robust as CNN for bounding box regression.

Based on this finding, the paper separates the CNN and FC for classification and bounding box regression for the ROI align features. Besides, the paper also does an extension that both heads conduct classification and regression but with an emphasis on their corresponding advantageous tasks.

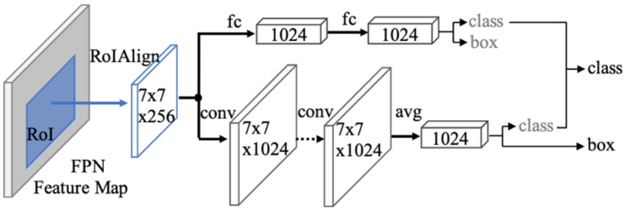
Our approach

Double-Head



$$\mathcal{L} = \omega^{fc}\mathcal{L}^{fc} + \omega^{conv}\mathcal{L}^{conv} + \mathcal{L}^{rpn}$$

Double-Head-Ext



$$\mathcal{L}^{fc} = \lambda^{fc}L^{fc}_{cls} + (1-\lambda^{fc})L^{fc}_{reg}$$

$$\mathcal{L}^{conv} = (1-\lambda^{conv})L^{conv}_{cls} + \lambda^{conv}L^{conv}_{reg}$$

Cross-Modal Cross-Domain Moment Alignment Network for Person Search

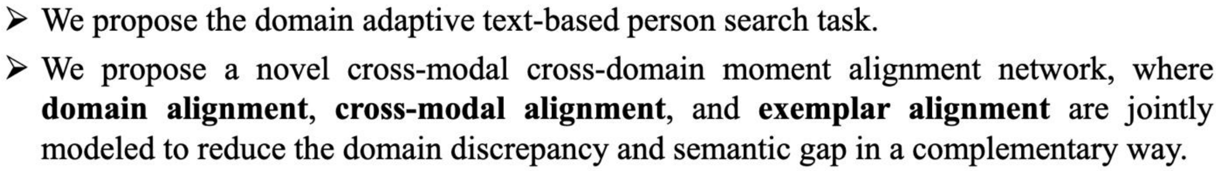
The paper considers the constraint of paired image-text data in text-based person search task. It combines the cross-modal (text-based) person search and cross-domain person search.

Main idea:

Moment alignment (mean and variance) for distributions and penalize the difference of mean and variance of different classes based on Euclidean distance. Pseudo labels of different class is achieved based on self-labeling.

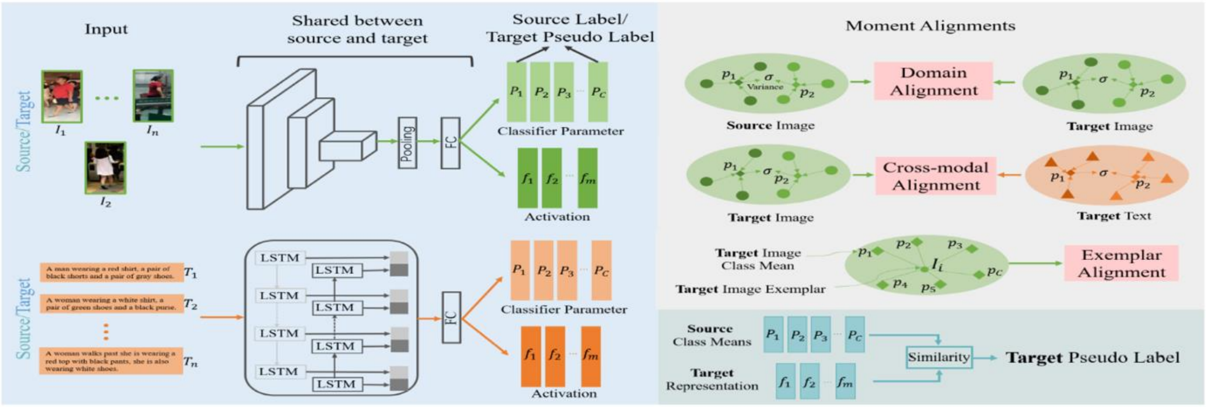
Conduct the alignment on cross-domain and cross-modal. Besides, for cross-domain, do exemplar alignment by enforcing each exemplar to be close to its nearest mean. Distance is measures by cosine similarity. For hard examples, do metric learning by contrastive loss, making positive pair close and negative away.

Experiment results show it achieves SOTA on CUHK-PEDES.



➢ We propose the domain adaptive text-based person search task.

➢ We propose a novel cross-modal cross-domain moment alignment network, where **domain alignment**, **cross-modal alignment**, and **exemplar alignment** are jointly modeled to reduce the domain discrepancy and semantic gap in a complementary way.



Input
Shared between source and target
Source Label/ Target Pseudo Label
Source/Target
I_1
I_n
I_2
Pooling
FC
P_1 P_2 P_3 ··· P_C
Classifier Parameter
f_1 f_2 ··· f_m
Activation
A man wearing a red shirt, a pair of black shorts and a pair of gray shoes. T_1
A woman wearing a white shirt, a pair of green shoes and a black purse. T_2
A woman walks past she is wearing a red top with black pants, she is also wearing white shoes. T_n
LSTM
Moment Alignments
p_1 σ Variance p_2
Domain Alignment
Source Image
Target Image
Cross-modal Alignment
Target Text
Target Image Class Mean
Target Image Exemplar
p_1 p_2 p_3 I_i p_C p_4 p_5
Exemplar Alignment
Source Class Means
Target Representation
Similarity
Target Pseudo Label

Joint Training of Variational Auto-Encoder and Latent Energy-Based Model

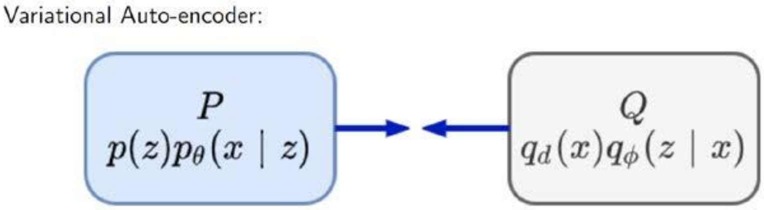
Keywords:

Variational auto-encoder, Energy-based model, Adversarial learning.

Problem:

VAE: directed model, can generate synthesized examples by direct ancestral sampling

Top-down mapping
hidden vector z
$\Downarrow$
signal $x \approx g_\theta(z)$
(a) Generator model
Bottom-up mapping
hidden vector z
$\Uparrow$
observed signal x
(b) Inference model



EBM: undirected model, better approximate the data density; but maximum likelihood learning of EBM require time-consuming Markov chain Monte Carlo (MCMC) sampling.

$$\pi_\alpha(x, z) = \frac{1}{Z(\alpha)} \exp\left[f_\alpha(x, z)\right]$$

Poorer synthesis than GAN: Generator density P seeks to cover all modes of data density Q, and thus it can be overly dispersed.

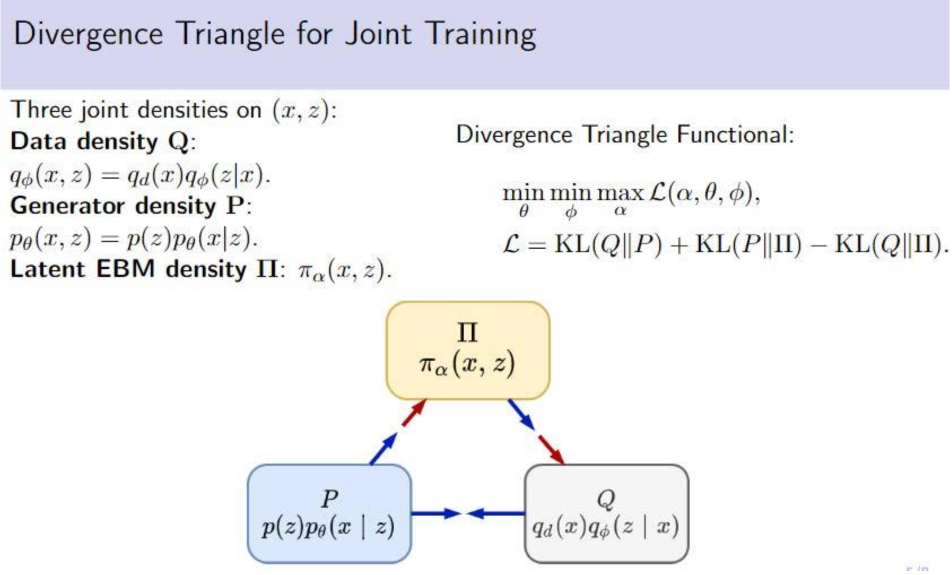
Main idea:

We propose a systematic Divergence Triangle integration of jointly variational learning and adversarial learning, to improve synthesis quality of VAE.

EBM serves as a critic of the generator model by judging it against the data density.

To the generator model, the latent EBM serves as a surrogate of the data density and a target density for the generator model to approximate.

The generator model and the associated inference model, in return, serve as approximate synthesis sampler and inference sampler of the latent EBM, thus relieving the latent EBM of the burden of MCMC sampling.



Divergence Triangle for Joint Training

Three joint densities on (x, z):

Data density $\mathbf{Q}$:

$q_\phi(x, z) = q_d(x) q_\phi(z|x)$.

Generator density $\mathbf{P}$:

$p_\theta(x, z) = p(z) p_\theta(x|z)$.

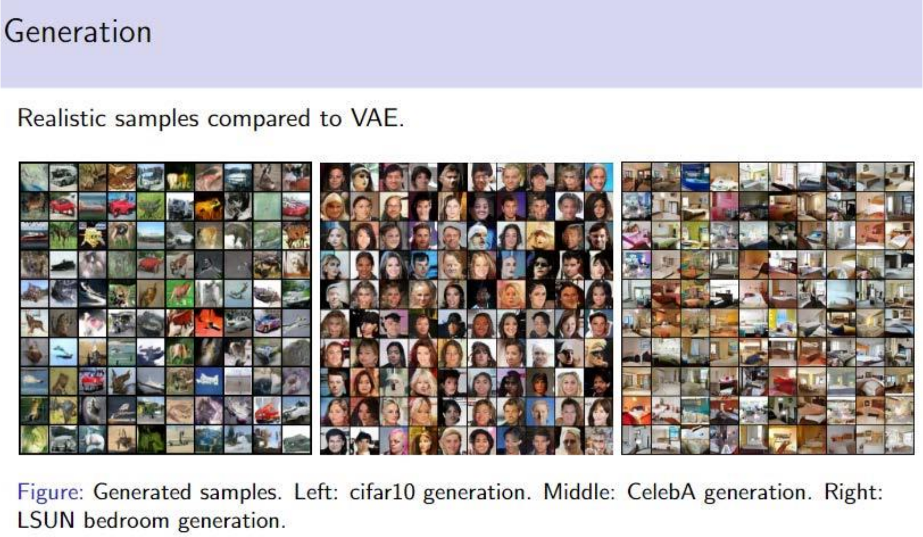
Latent EBM density $\mathbf{\Pi}$: $\pi_\alpha(x, z)$.

Divergence Triangle Functional:

$$\min_\theta \min_\phi \max_\alpha \mathcal{L}(\alpha, \theta, \phi),$$

$$\mathcal{L} = \mathrm{KL}(Q \| P) + \mathrm{KL}(P \| \Pi) - \mathrm{KL}(Q \| \Pi).$$

Results:



Generation

Realistic samples compared to VAE.

Figure: Generated samples. Left: cifar10 generation. Middle: CelebA generation. Right: LSUN bedroom generation.

Guided Variational Autoencoder for Disentanglement Learning

Keywords:

Variational auto-encoder, Multi-task learning, Disentanglement learning, Transfer learning.

Problem:

GANs focus on the generation process and are not aimed at representation learning (without an encoder).

VAEs can model high dimensional data of real-world complexity for dimension reduction.

But it is necessary for VAEs to achieve a high quality reconstruction/synthesis, and make their representation learning more transparent, interpretable, and controllable.

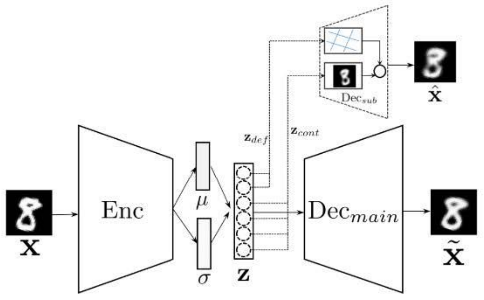
Main idea:

- •Guided-VAE learns a controllable generative model by performing latent representation disentanglement learning.
- •End-to-End, Multi-Task Learning.

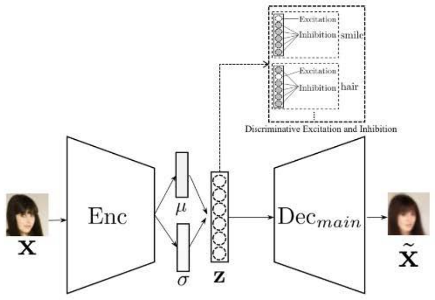
We would like to learn a transparent representation by introducing guidance to the latent variables in a VAE.

Our methods are applicable in a variety of scenarios, including image synthesis, latent disentanglement, transfer learning and few-shot learning.

Guided-VAE makes the VAE latent factors more transparent, interpretable, and controllable



Unsupervised Guided-VAE

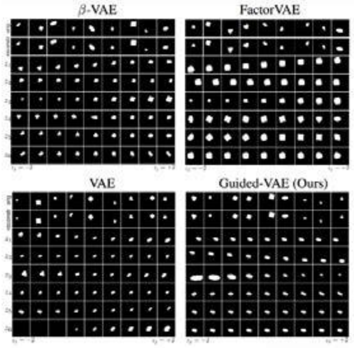


Supervised Guided-VAE

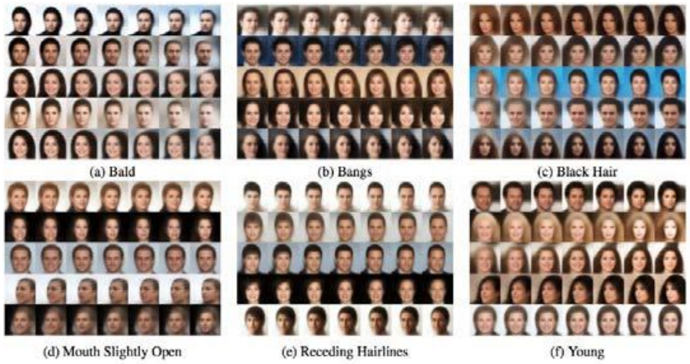
Results:

We apply Guided-VAE to the data modeling and few-shot learning problems.

Our model significantly improves the controllability of the vanilla VAE.



Latent Factors Traversal: z_1, z_2, z_3, z_4 represent traversal results of x-position, y-position, scale and rotation over the unsupervised disentanglement methods on 2D Shapes.


(a) Bald

(b) Bangs

(c) Black Hair

(d) Mouth Slightly Open

(e) Receding Hairlines

(f) Young

Latent Factors Traversal: Traversal results of the supervised Guided-VAE on latent variables representing each attribute on CelebA.

Towards Visually Explaining Variational Autoencoders

Keywords:

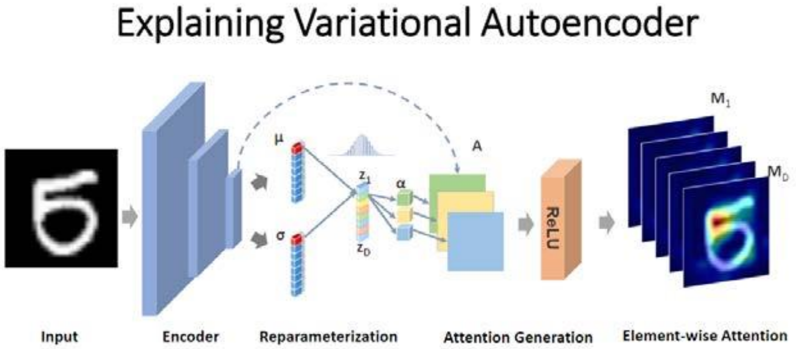
Variational auto-encoder, Explanation for generative models (e.g., VAEs), Anomaly localization, Disentangled attention maps.

Problems:

Understanding the reasoning behind an algorithm's predictions is crucial for safety-critical and consumer-focusing tasks.

Existing techniques need classification to guide model’s explainability. Can we visually explain other models and architectures, such as generative models?

Main idea:



$$\alpha_k = \frac{1}{hw}\sum_{p=1}^{h}\sum_{q=1}^{w}(\frac{\partial z_i}{\partial A_k^{pq}})$$

$$\mathbf{M}_i = \mathrm{ReLU}(\sum_{k=1}^{D}\alpha_k \mathbf{A}_k)$$

We enable visual explanations of generative models (VAE in this paper) using attention maps. We propose to generate gradient-based attentions from the latent space. No classification model needed.

(1) Anomaly localization:

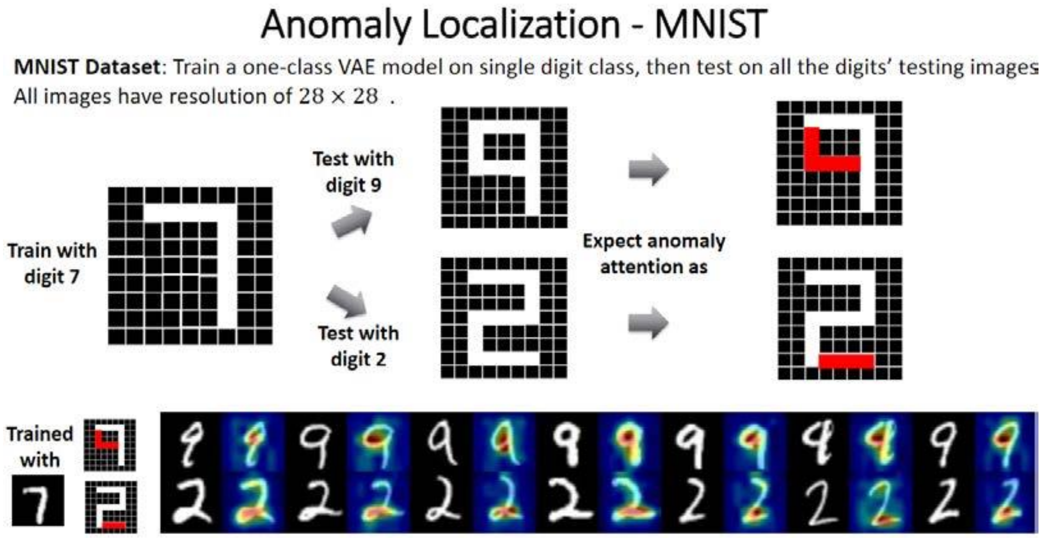
Our VAE attention can localize anomalies, where the input does not follow the standard Gaussian, and we use this difference to compute an attention map to localize anomalies.

(2) Disentangled learning:

Additionally, we use our VAE attention to improve disentangled representation learning. We propose to spatially separate the generated attention maps, achieved as part of our novel attention disentanglement loss, to learn an improved disentangled latent space.

Results:

(1) Anomaly localization:

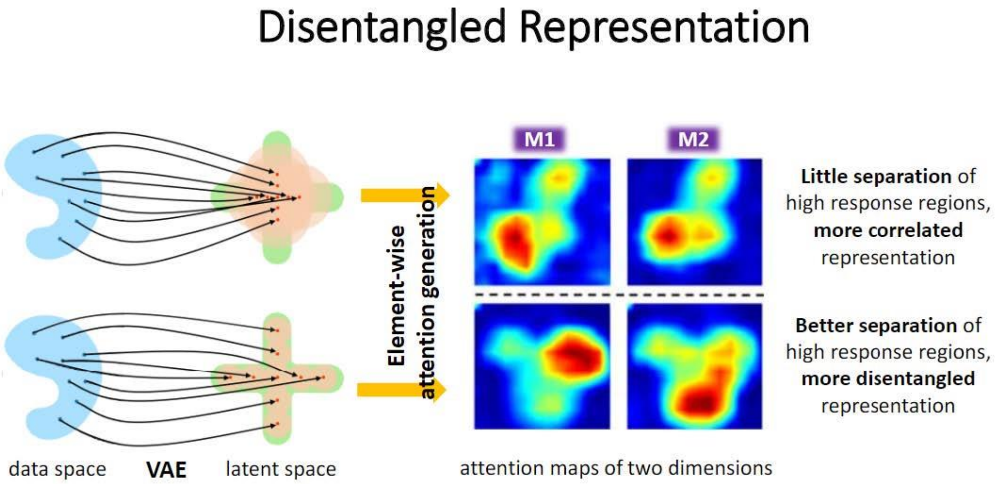


Anomaly Localization - MNIST

MNIST Dataset: Train a one-class VAE model on single digit class, then test on all the digits' testing images

All images have resolution of 28×28 .

(2) Disentangled learning



Disentangled Representation

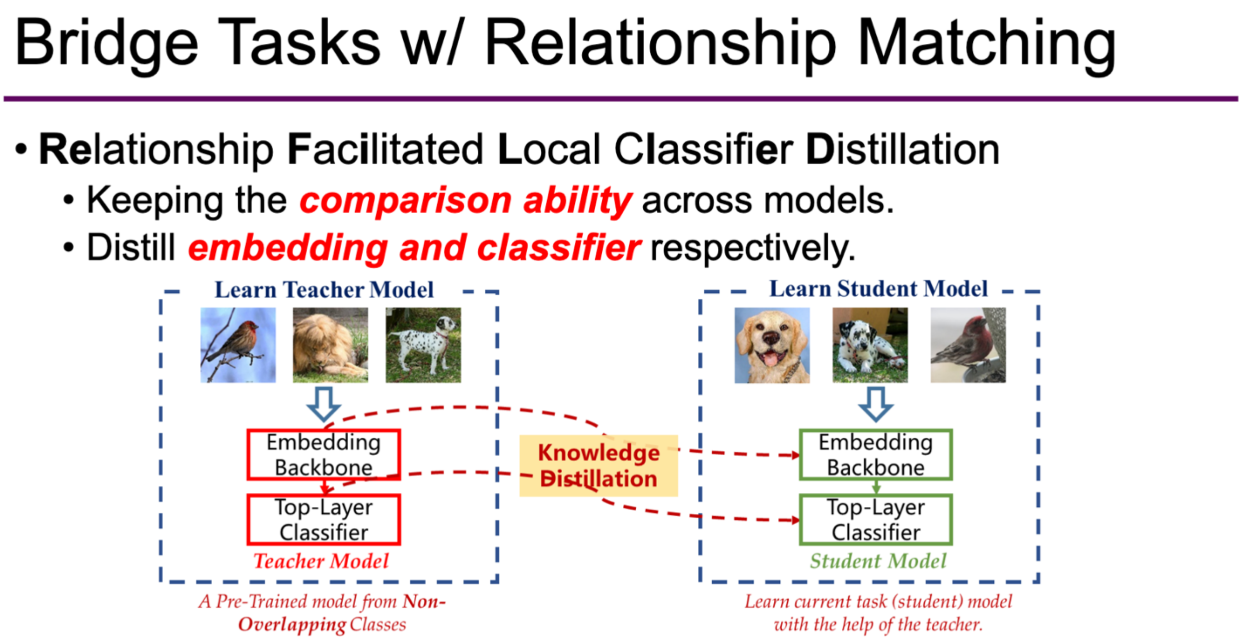
Little separation of high response regions, **more correlated** representation

Better separation of high response regions, **more disentangled** representation

attention maps of two dimensions

Distilling Cross-Task Knowledge via Relationship Matching
The paper uses the Student-Teacher network for cross-task knowledge extraction. However, the teacher and student label space have discrepancy which hinders the process.

Main idea:



Bridge Tasks w/ Relationship Matching

- **Re**lationship **F**ac**i**litated **L**ocal C**l**assifi**e**r **D**istillation
  - Keeping the ***comparison ability*** across models.
  - Distill ***embedding and classifier*** respectively.

A Pre-Trained model from ***Non-Overlapping*** *Classes*

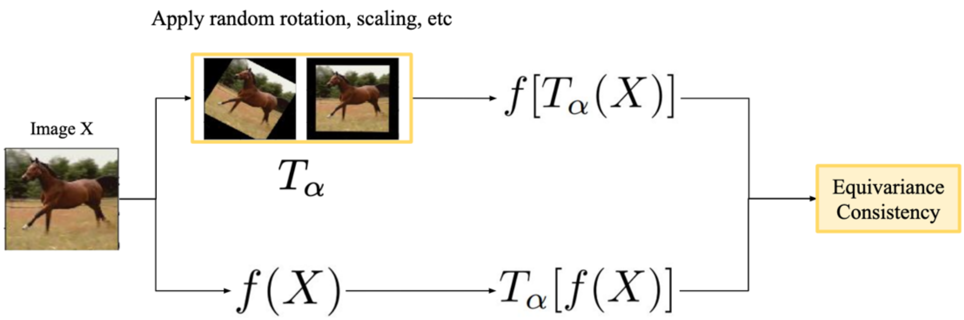
Learn current task (student) model with the help of the teacher.

Learning From Synthetic Animals

The paper discusses the domain adaptation for animal 2D pose estimation. The model is trained on the synthetic data which has low diversity and large domain gap and transferred to real data.

Main idea:
Domain randomization helps to generate more robust pseudo-labels.
1. Train a weak model with labeled source data.
2. Feed the unlabeled target data into the weak model and do self-training. The prediction before or after the manipulation should be the same.

T_α : equivariance transform. Change image and keypoint equally.



T_Δ : Temporal transform. Transform between temporal frames, approximated by optical flow.

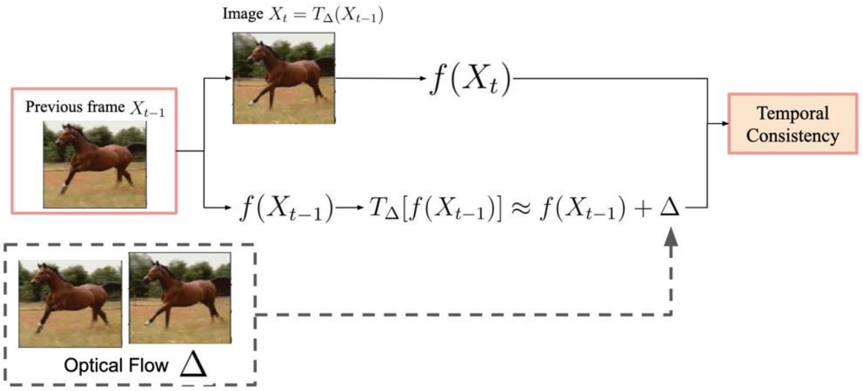
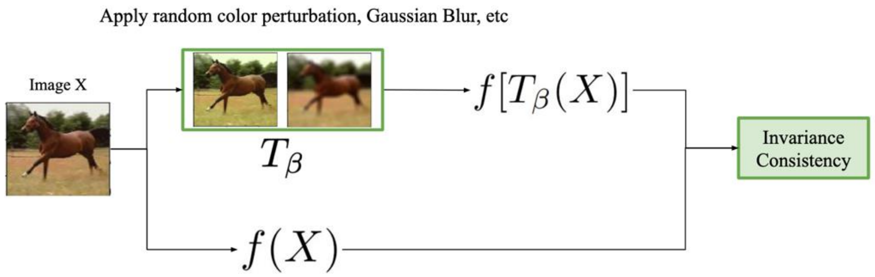


Image $X_t = T_\Delta(X_{t-1})$
$f(X_t)$
Previous frame X_{t-1}
Temporal Consistency
$f(X_{t-1}) \rightarrow T_\Delta[f(X_{t-1})] \approx f(X_{t-1}) + \Delta$
Optical Flow Δ

T_β : invariance transform. Do not change the keypoint locations.


Apply random color perturbation, Gaussian Blur, etc
Image X
$f[T_\beta(X)]$
T_β
Invariance Consistency
$f(X)$