Regressive Domain Adaptation for Unsupervised Keypoint Detection

Junguang Jiang[1], Yifei Ji[1], Ximei Wang[1], Yufeng Liu[2], Jianmin Wang[1], Mingsheng Long[1]

[1]Tsinghua University, [2]Kuaishou Technology

JiangJunguang1123@outlook.com

Outline

Motivations

- The annotations of 2D keypoints on real images are expensive and time-consuming to collect while that on synthetic images can be obtained in abundance by CG at a low cost.
- Domain shifts between virtual and real domains will cause significant performance drop, thus domain adaptation is important for this problem.

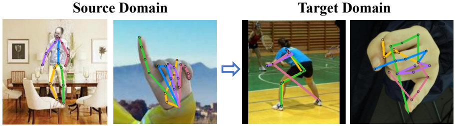


Figure: Confusion of keypoints on the target domain.

- No clear decision boundary exists in regression, thus feature alignment, such as DAN [1] and DANN [2], cannot enlarge the margins of boundaries to generalize the model as done in classification.

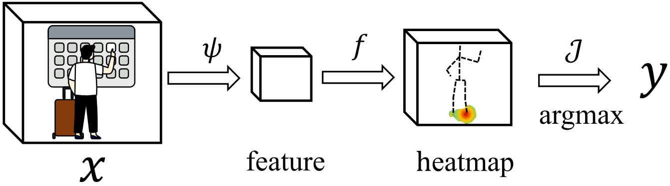
[1] *Long, et al. Learning Transferable Features with Deep Adaptation Networks. ICML 2015.*

[2] *Ganin, et al. Domain-Adversarial Training of Neural Networks. ICML 2015.*

Outline

Learning Setup

In supervised 2D keypoint detection, we have n labeled samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ from $\mathcal{X} \times \mathcal{Y}^K$, where $\mathcal{X} \in \mathcal{R}^{H \times W \times 3}$ is the input space, $\mathcal{Y} \in \mathcal{R}^2$ is the output space and K is the number of keypoints for each input. The goal is to find a regressor $f \in \mathcal{F}$ that has the lowest error rate $\text{err}_D = \mathbb{E}_{(x,y)\sim D} L((f \circ \psi)(\boldsymbol{x}), y)$ on D.



In unsupervised domain adaptation, there exists a labeled source domain $\widehat{P} = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)\}_{i=1}^{n}$ and an unlabeled target domain $\widehat{Q} = \{\boldsymbol{x}_i^t\}_{i=1}^{m}$. The objective is to minimize err_Q.

Disparity Discrepancy (DD)[3]

Definition (Disparity on D)

$\epsilon_D(f_1, f_2) = \mathbb{E}_{(x,y)\sim D}[f_1(x) \neq f_2(x)]$.

Definition (Disparity Discrepancy (DD))

Given a hypothesis space $\mathcal{F}$ and a *specific hypothesis* $h \in \mathcal{F}$, the Disparity Discrepancy (DD) is

$$d_{h,\mathcal{F}}(P, Q) = \sup_{f' \in \mathcal{F}} \left(\mathbb{E}_Q[f' \neq f] - \mathbb{E}_P[f' \neq f]\right) \tag{1}$$

Theorem (Bound with Disparity Discrepancy)

For any $\delta > 0$ and binary classifier $f \in \mathcal{F}$, with probability $1 - 3\delta$, we have

$$\begin{aligned} \mathrm{err}_Q(f) \leq{}& \mathrm{err}_{\widehat{P}}(f) + d_{f,\mathcal{F}}(\widehat{P}, \widehat{Q}) + \epsilon_{ideal} \\ &+ 2\mathfrak{R}_{n,P}(\mathcal{F}\Delta\mathcal{F}) + 2\mathfrak{R}_{n,P}(\mathcal{F}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + 2\mathfrak{R}_{m,Q}(\mathcal{F}\Delta\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \end{aligned} \tag{2}$$

Architecture

- Train the adversarial regressor f' to predict correctly on the source domain, while differ from f as much as possible on the target domain.
- Encourage the feature extractor ψ to output domain-invariant features and deceive f'.
- In keypoint detection where output space is large, it's hard to find the adversarial regressor f' that does poorly *only* on the target domain.

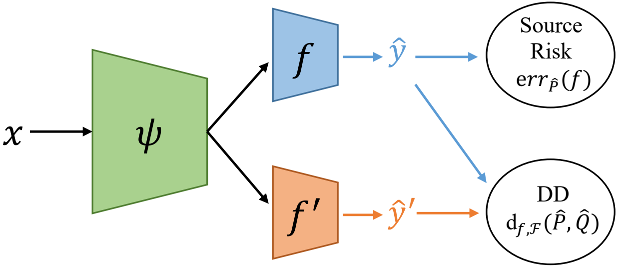
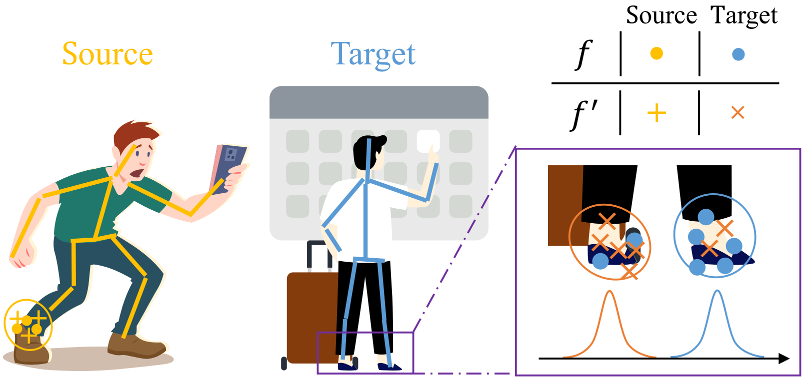


Figure: DD architecture under the keypoint detection setting. ψ: feature generator, f: regressor head, f': adversarial regressor head.

Outline

Sparsity of the Spatial Density

When the position of the right ankle is mistaken, most likely the left ankle is predicted, occasionally other keypoints predicted, and rarely positions on the background are predicted.



Sparsity of the Spatial Density

Denote $\mathcal{H}(\boldsymbol{y}_k) \in R^{H' \times W'}$ as the ground-truth heatmap for keypoint $\boldsymbol{y}_k$. Since wrong predictions are often located at **other** keypoints, we sum up their heatmaps,

$$\mathcal{H}_{\mathrm{F}}(\widehat{\boldsymbol{y}}_k)_{h,w} = \sum_{k' \neq k} \mathcal{H}(\widehat{\boldsymbol{y}}_{k'})_{h,w}, \tag{3}$$

where $\widehat{\boldsymbol{y}}_k'$ is the prediction by the main regressor f. Then we normalize the map $\mathcal{H}_{\mathrm{F}}(\boldsymbol{y}_k)$ into *ground false* distribution,

$$\mathcal{P}_{\mathrm{F}}(\widehat{\boldsymbol{y}}_k)_{h,w} = \frac{\mathcal{H}_{\mathrm{F}}(\widehat{\boldsymbol{y}}_k)_{h,w}}{\sum_{h'=1}^{H'} \sum_{w'=1}^{W'} \mathcal{H}_{\mathrm{F}}(\widehat{\boldsymbol{y}}_k)_{h',w'}}. \tag{4}$$

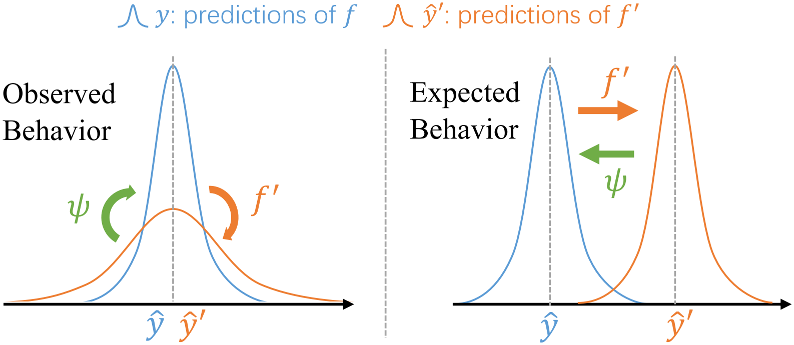
The target disparity is defined as the KL-divergence between the predictions of f' and the *ground false* predictions of f,

$$L_{\mathrm{F}}(\boldsymbol{p}', \boldsymbol{p}) \triangleq \frac{1}{K} \sum_{k}^{K} \mathrm{KL}(\mathcal{P}_{\mathrm{F}}(\mathcal{J}(\boldsymbol{p}))_k || \boldsymbol{p}_k'), \tag{5}$$

where $\boldsymbol{p}' = (f' \circ \psi)(\boldsymbol{x}^t)$ is the prediction of f' and $\boldsymbol{p}$ is the prediction of f.

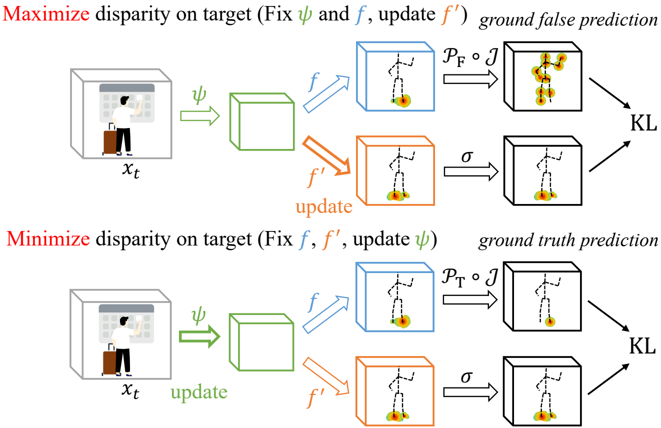
Minimax of Target Disparity

- When maximizing the discrepancy between f' and f, we expect to maximize the mean difference, but often the variance is changed.
- Convert the minimax game to the minimization of two opposite goals.



Overall Objectives

- Supervised training on the source domain.
- Update f' to minimize its KL with *ground false* predictions of f.
- Update ψ to minimize KL between prediction of f' with *ground truth* prediction of f.



Outline

Quantitative Results

- Evaluate the performance of Simple Baseline with ResNet101 as the backbone.
- Percentage of Correct Keypoints (PCK) is used for evaluation.

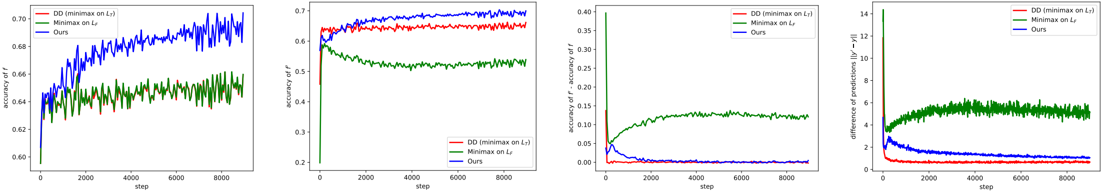
Table: PCK on task *RHD H3D*. Negative transfer happens for all other domain adaptation methods.

Method	MCP	PIP	DIP	Fingertip	Avg
ResNet101	67.4	64.2	63.3	54.8	61.8
DAN	59.0	57.0	56.3	48.4	55.1
DANN	67.3	62.6	60.9	51.2	60.6
MCD	59.1	56.1	54.7	46.9	54.6
DD	72.7	69.6	66.2	54.4	65.2
RegDA	**79.6**	**74.4**	**71.2**	**62.9**	**72.5**
Oracle	97.7	97.2	95.7	92.5	95.8

Ablation

Table: Ablation on the minimax of target disparity.

Method	f'	ψ	MCP	PIP	DIP	Fingertip	Avg
DD	max L_{T}	min L_{T}	72.7	69.6	66.2	54.4	65.2
	min L_{F}	max L_{F}	74.4	71.1	66.9	56.4	66.5
RegDA	min L_{F}	min L_{T}	**79.6**	**74.4**	**71.2**	**62.9**	**72.5**



(a) Accuracy of f (b) Accuracy of f' (c) Accuracy difference (d) Prediction difference

Figure: Empirical statistics during the training process.

Outline

Qualitative results

Source only often confuses different key points, resulting in the predicted skeleton not look real. In contrast, the outputs of **RegDA** look more like a human hand or body.

Figure: *RHD* → *H3D* dataset.

Qualitative results

Source only often confuses different key points, resulting in the predicted skeleton not look real. In contrast, the outputs of **RegDA** look more like a human hand or body.

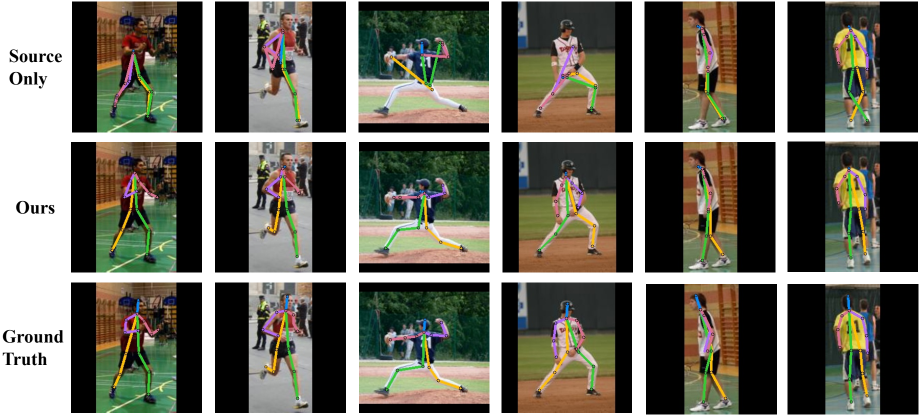
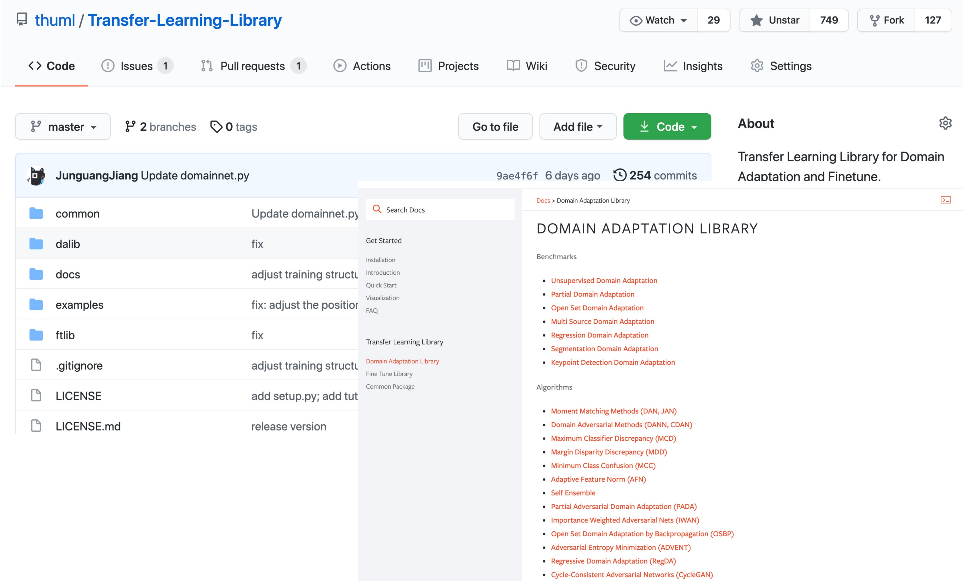


Figure: *SURREAL* $\rightarrow$ *LSP*.

Open-Source Library



Design Patterns

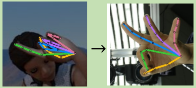
Reproducible | Extendible | Ease of Use | Friendly API | Documentation | TorchVision

Task

Classification
- Close Set DA
- Open Set DA
- Partial DA
- Low-Data Finetune
- ......

DA Segmentation

DA Keypoint Detection



Core

Adaptation
- MDD
- DAN
- CDAN
- ......

Translation
- CycleGAN
- Cycada
- Fourier DA
- ......

Finetune
- DELTA
- Co-Tuning
- BSS
- ......

Modules
- Discriminator
- GradRevLayer
- Kernl
- ......

Platform PYTORCH python ......

https://github.com/thuml/Transfer-Learning-Library