Patch-Mix Transformer for Unsupervised Domain Adaptation: A Game Perspective

Jinjing Zhu [1*], Haotian Bai [1*], Lin Wang

TUE-AM-339

Visual Learning and Intelligent Systems Lab (VLIS LAB)
Thrust of Artificial Intelligence, Information Hub, GZ Campus
Department of Computer Science and Engineering, CWB Campus
The Hong Kong University of Science and Technology (HKUST)

VLIS LAB

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

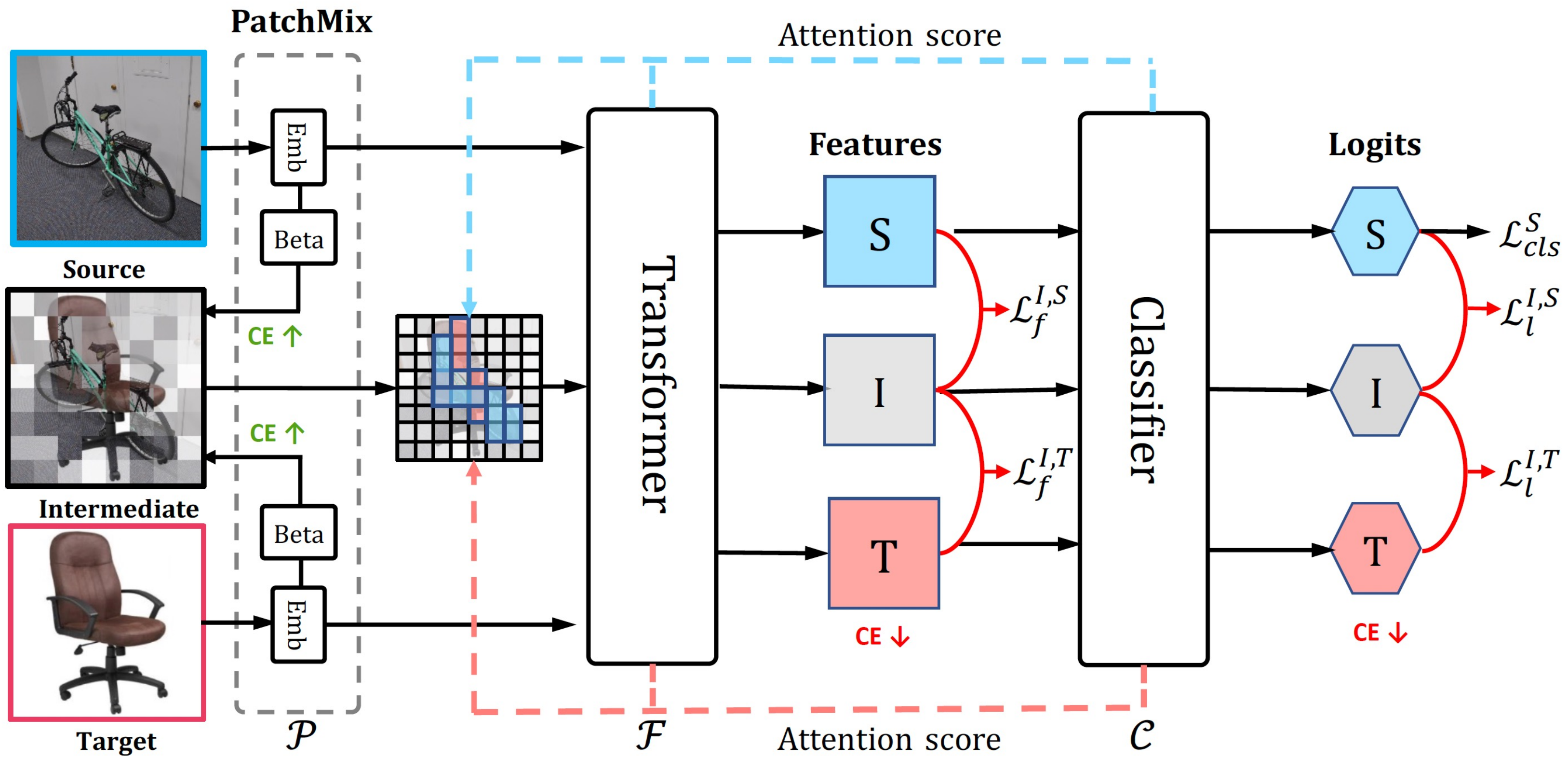
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

PMTrans Overview

Prior works have a a limitation: as the performance of cross-attention highly depends on the quality of pseudo labels, it becomes less effective when the domain gap becomes large.

We probe a new problem for UDA: how to smoothly bridge the source and target domains by constructing an intermediate domain with an effective ViT-based solution?

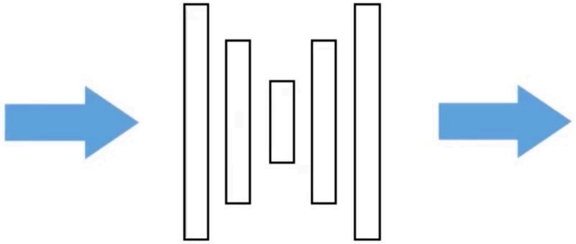


Overview of the proposed PMTrans framework

Research Background

Image

CNN



Classification

Bike?

lack of annotations

expensive

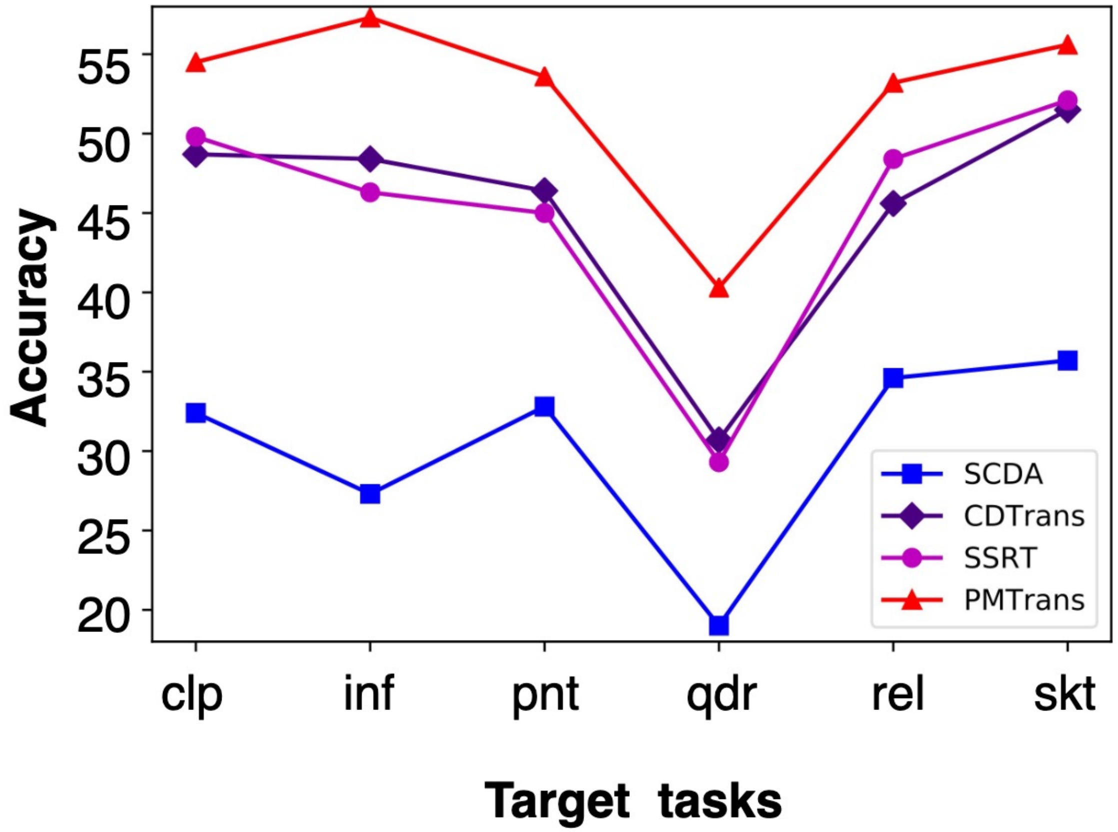
laborious

time-consuming

Research Background

A significant line of solutions reduces the domain gap by producing pseudo labels for target samples.

However, prior method has a distinct limitation: as the performance of cross-attention highly depends on the quality of pseudo labels, it becomes less effective when the domain gap becomes large.



Results of our PMTrans and SOTA methods on DomainNet.

Research Problem

How to smoothly bridge the source and target domains by constructing an intermediate domain with an effective ViT-based solution?

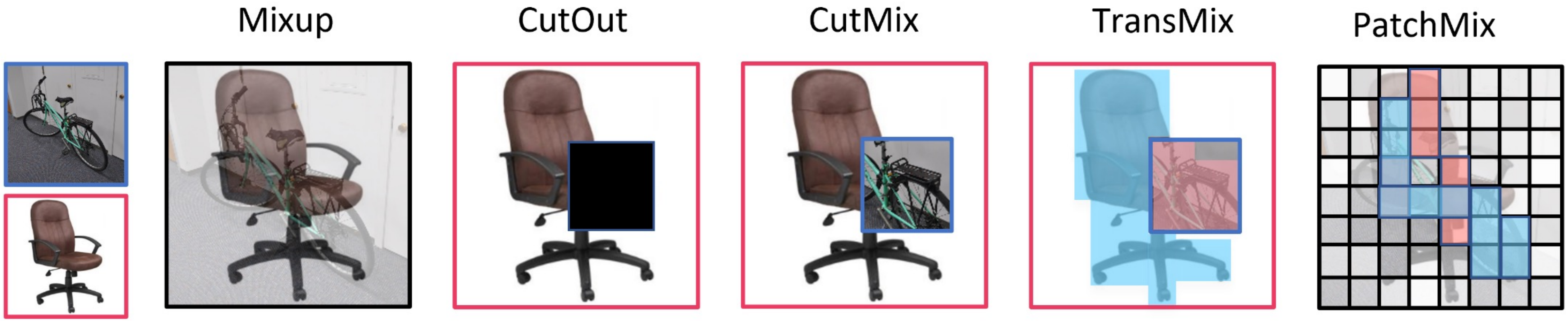
Method

PatchMix

Let $\mathcal{P}_\lambda$ be a linear interpolation operation on two pairs of randomly drawn samples (x^s, y^s) and (x^t, y^t). Then with $\lambda_k \sim Beta(\beta, \gamma)$, it interpolates the **k**-th source patch x_k^s and target patch x_k^t to reconstruct a mixed representation with **n** patches.

$$\boldsymbol{x}^i = \mathcal{P}_\lambda(\boldsymbol{x}^s, \boldsymbol{x}^t),\ \boldsymbol{x}_k^i = \lambda_k \odot \boldsymbol{x}_k^s + (1 - \lambda_k) \odot \boldsymbol{x}_k^t,$$

$$\boldsymbol{y}^i = \mathcal{P}_\lambda(\boldsymbol{y}^s, \boldsymbol{y}^t) = \frac{(\sum_{k=1}^n \lambda_k)\boldsymbol{y}^s + (\sum_{k=1}^n (1 - \lambda_k))\boldsymbol{y}^t}{n}$$



PatchMix and Mixup variants.

Method

A min-max CE game

We interpret UDA as a min-max CE game among three players, namely the feature extractor ($\mathcal{F}$), classifier ($\mathcal{C}$), and PatchMix module ($\mathcal{P}$).

$$
\begin{aligned}
J_{\mathcal{F}}\left(\boldsymbol{\omega}_{\mathcal{F}}, \boldsymbol{\omega}_{-\mathcal{F}}\right) &:= \mathcal{L}_{cls}^{S}(\boldsymbol{\omega}_{\mathcal{F}}, \boldsymbol{\omega}_{\mathcal{C}}) + \alpha \mathrm{CE}_{s,i,t}(\boldsymbol{\omega}), \\
J_{\mathcal{C}}\left(\boldsymbol{\omega}_{\mathcal{C}}, \boldsymbol{\omega}_{-\mathcal{C}}\right) &:= \mathcal{L}_{cls}^{S}(\boldsymbol{\omega}_{\mathcal{F}}, \boldsymbol{\omega}_{\mathcal{C}}) + \alpha \mathrm{CE}_{s,i,t}(\boldsymbol{\omega}), \\
J_{\mathcal{P}}\left(\boldsymbol{\omega}_{\mathcal{P}}, \boldsymbol{\omega}_{-\mathcal{P}}\right) &:= -\alpha \mathrm{CE}_{s,i,t}(\boldsymbol{\omega}),
\end{aligned}
$$

Nash Equilibrium

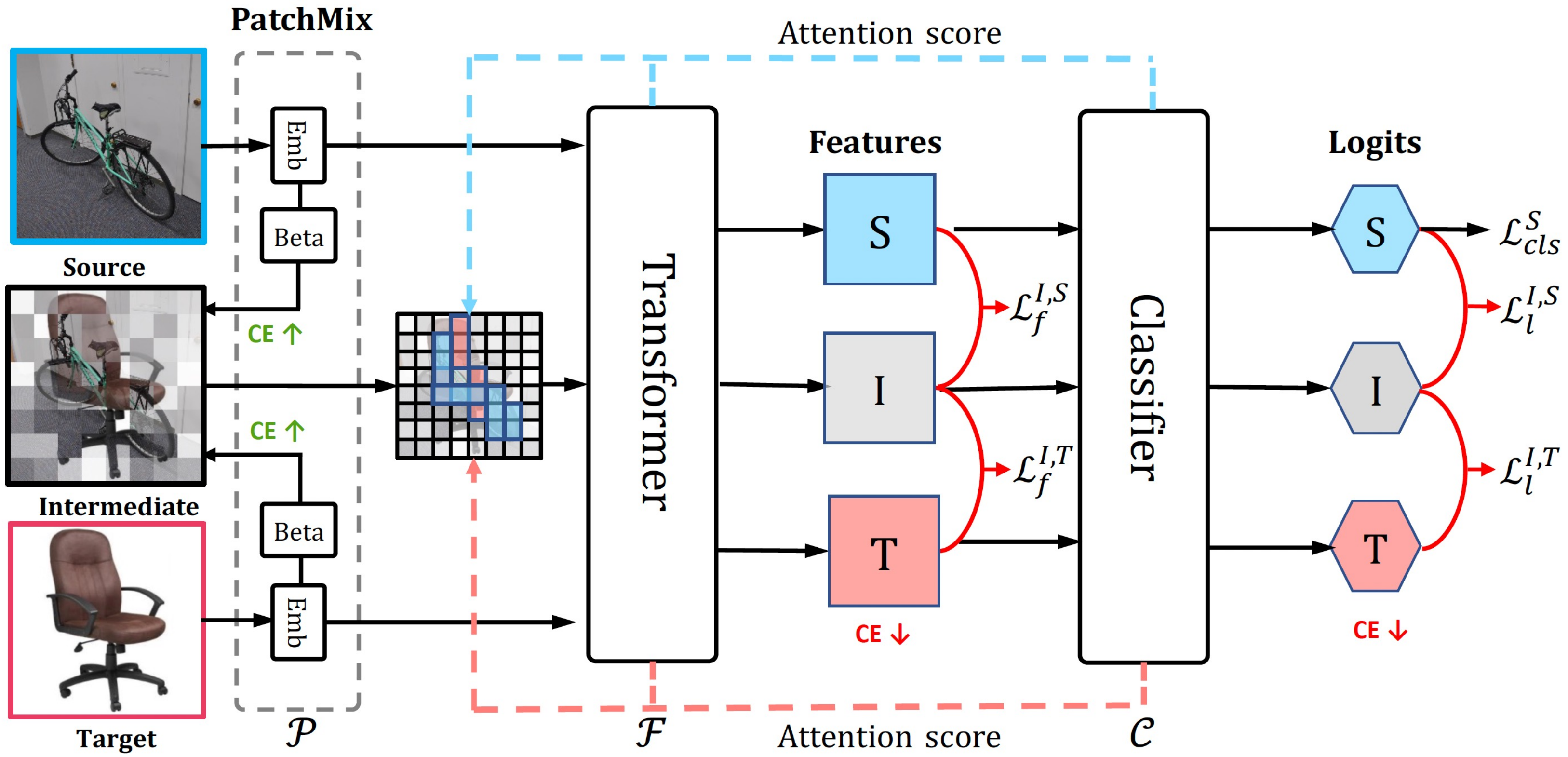
The equilibrium states each player's strategy is the best response to other players. And a point $\omega^* \in \Omega$ is Nash Equilibrium if

$$
\forall \omega_m \in \Omega_i, \forall m \in \{\mathcal{F}, \mathcal{C}, \mathcal{P}\}, s.t. J_m(\omega_m^*, \omega_{-m}^*) \leq J_m(\omega_m, \omega_{-m}^*).
$$

Method

Proposed framework

PMTrans consists of three players: the PatchMix module empowered by a patch embedding (Emb) layer and a learnable Beta distribution (Beta), ViT encoder, and classifier.



Overview of the proposed PMTrans framework.

Method

Semi-supervised mixup loss

In label space:

Use supervised mixup loss in the label space to measure the domain divergence based on CE loss.

$$\mathcal{L}_l^{I,S}(\boldsymbol{\omega}) = \mathbb{E}_{(\boldsymbol{x}^i,\boldsymbol{y}^i)\sim D^i} \lambda^s \ell\left(\mathcal{C}\left(\mathcal{F}\left(\boldsymbol{x}^i\right)\right), \boldsymbol{y}^s\right)$$

$$\mathcal{L}_l^{I,T}(\boldsymbol{\omega}) = \mathbb{E}_{(\boldsymbol{x}^i,\boldsymbol{y}^i)\sim D^i} \lambda^t \ell\left(\mathcal{C}\left(\mathcal{F}\left(\boldsymbol{x}^i\right)\right), \hat{\boldsymbol{y}^t}\right)$$

$$\mathcal{L}_l(\boldsymbol{\omega}) = \mathcal{L}_l^{I,S}(\boldsymbol{\omega}) + \mathcal{L}_l^{I,T}(\boldsymbol{\omega})$$

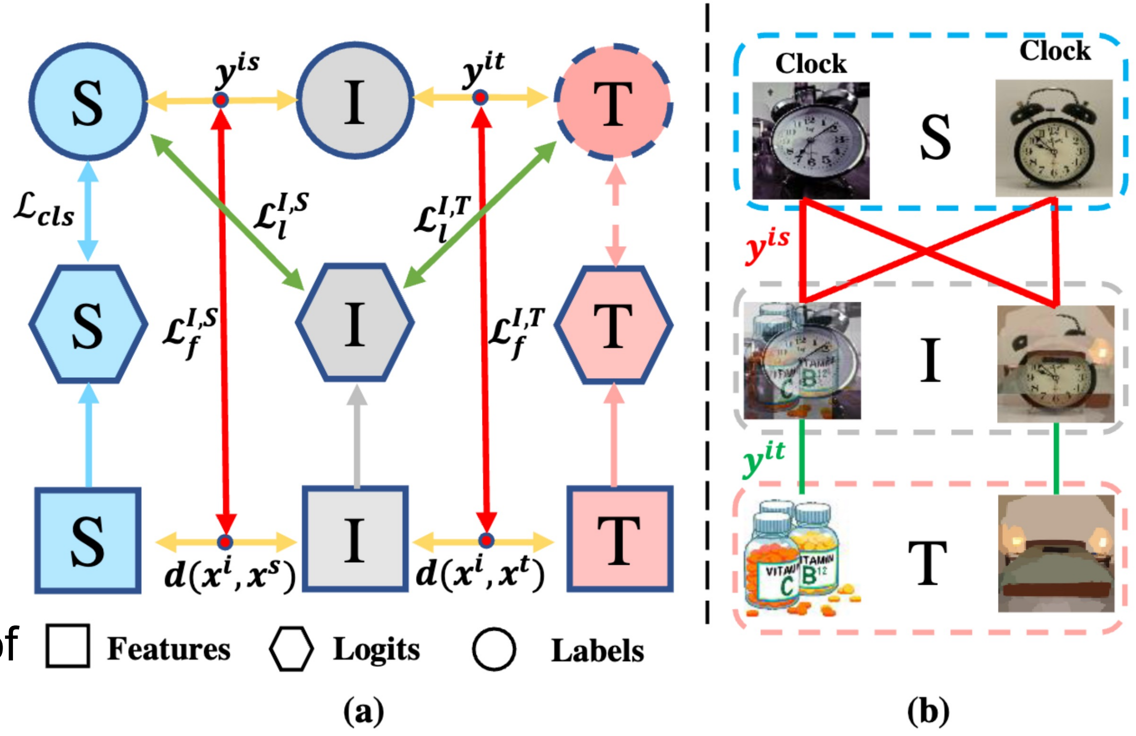
In feature space

Propose to minimize the discrepancy between the similarity of the features and the similarity of labels in the feature space.

$$\mathcal{L}_f^{I,S}(\boldsymbol{\omega}_\mathcal{F}, \boldsymbol{\omega}_\mathcal{P}) = \mathbb{E}_{(\boldsymbol{x}^i,\boldsymbol{y}^i)\sim D^i} \lambda^s \ell\left(d(\boldsymbol{x}^i, \boldsymbol{x}^s), \boldsymbol{y}^{is}\right)$$

$$\mathcal{L}_f^{I,T}(\boldsymbol{\omega}_\mathcal{F}, \boldsymbol{\omega}_\mathcal{P}) = \mathbb{E}_{(\boldsymbol{x}^i,\boldsymbol{y}^i)\sim D^i} \lambda^t \ell\left(d(\boldsymbol{x}^i, \boldsymbol{x}^t), \boldsymbol{y}^{it}\right)$$

$$\mathcal{L}_f(\boldsymbol{\omega}_\mathcal{F}, \boldsymbol{\omega}_\mathcal{P}) = \mathcal{L}_f^{I,S}(\boldsymbol{\omega}_\mathcal{F}, \boldsymbol{\omega}_\mathcal{P}) + \mathcal{L}_f^{I,T}(\boldsymbol{\omega}_\mathcal{F}, \boldsymbol{\omega}_\mathcal{P})$$



(a) The illustration of two proposed semi-supervised losses. (b) Label similarity y^{is} and y^{it}.

Method

A three-player game

The min-max CE game aims to align distributions in the feature and label spaces.

$$\mathrm{CE}_{s,i,t}(\boldsymbol{\omega}) = \mathcal{L}_f(\boldsymbol{\omega}_{\mathcal{F}}, \boldsymbol{\omega}_{\mathcal{P}}) + \mathcal{L}_l(\boldsymbol{\omega})$$

The total objective of PMTrans is defined as:

$$J(\boldsymbol{\omega}) := \mathcal{L}_{cls}^{S}(\boldsymbol{\omega}_{\mathcal{F}}, \boldsymbol{\omega}_{\mathcal{C}}) + \alpha \mathrm{CE}_{s,i,t}(\boldsymbol{\omega})$$

Experiments

Method		A→ C	A→ P	A → R	C → A	C → P	C → R	P→ A	P→ C	P→ R	R→ A	R→ C	R→ P	Avg
ResNet-50	ResNet	44.9	66.3	74.3	51.8	61.9	63.6	52.4	39.1	71.2	63.8	45.9	77.2	59.4
MCD	ResNet	48.9	68.3	74.6	61.3	67.6	68.8	57.0	47.1	75.1	69.1	52.2	79.6	64.1
MDD	ResNet	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
BNM	ResNet	56.7	77.5	81.0	67.3	76.3	77.1	65.3	55.1	82.0	73.6	57.0	84.3	71.1
FixBi	ResNet	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7
TVT	ViT	74.9	86.8	89.5	82.8	88.0	88.3	79.8	71.9	90.1	85.5	74.6	90.6	83.6
Deit-based	ViT	61.8	79.5	84.3	75.4	78.8	81.2	72.8	55.7	84.4	78.3	59.3	86.0	74.8
CDTrans-Deit	ViT	68.8	85.0	86.9	81.5	87.1	87.3	79.6	63.3	88.2	82.0	66.0	90.6	80.5
PMTrans-Deit	ViT	71.8	87.3	88.3	83.0	87.7	87.8	78.5	67.4	89.3	81.7	70.7	92.0	82.1
ViT-based	ViT	67.0	85.7	88.1	80.1	84.1	86.7	79.5	67.0	89.4	83.6	70.2	91.2	81.1
SSRT-ViT	ViT	75.2	89.0	91.1	85.1	88.3	89.9	85.0	74.2	91.2	85.7	78.6	91.8	85.4
PMTrans-ViT	ViT	81.2	91.6	92.4	**88.9**	91.6	93.0	**88.5**	80.0	**93.4**	**89.5**	**82.4**	94.5	88.9
Swin-based	Swin	72.7	87.1	90.6	84.3	87.3	89.3	80.6	68.6	90.3	84.8	69.4	91.3	83.6
PMTrans-Swin	Swin	**81.3**	**92.9**	**92.8**	88.4	**93.4**	**93.2**	87.9	**80.4**	93.0	89.0	80.9	**94.8**	**89.0**

Comparison with SOTA methods on Office-Home.

PMTrans can obtain more robust transferable representations than the CNN-based and ViT-based methods.

Experiments

Method		A → W	D→ W	W→ D	A→ D	D→ A	W→ A	Avg
ResNet-50	ResNet	68.9	68.4	62.5	96.7	60.7	99.3	76.1
BNM	ResNet	91.5	98.5	**100.0**	90.3	70.9	71.6	87.1
MDD	ResNet	94.5	98.4	**100.0**	93.5	74.6	72.2	88.9
SCDA	ResNet	94.2	98.7	99.8	95.2	75.7	76.2	90.0
FixBi	ResNet	96.1	99.3	**100.0**	95.0	78.7	79.4	91.4
TVT	ViT	96.4	99.4	**100.0**	96.4	84.9	86.0	93.9
Deit-based	ViT	89.2	98.9	**100.0**	88.7	80.1	79.8	89.5
CDTrans-Deit	ViT	96.7	99.0	**100.0**	97.0	81.1	81.9	92.6
PMTrans-Deit	ViT	99.0	99.4	**100.0**	96.5	81.4	82.1	93.1
ViT-based	ViT	91.2	99.2	**100.0**	90.4	81.1	80.6	91.1
SSRT-ViT	ViT	97.7	99.2	**100.0**	98.6	83.5	82.2	93.5
PMTrans-ViT	ViT	99.1	**99.6**	**100.0**	99.4	85.7	86.3	95.0
Swin-based	Swin	97.0	99.2	**100.0**	95.8	82.4	81.8	92.7
PMTrans-Swin	Swin	**99.5**	99.4	**100.0**	**99.8**	**86.7**	**86.5**	**95.3**

Comparison with SOTA methods on Office-31.

PMTrans achieves the best performance on each task and outperforms the prior SOTA methods with identical backbones.

Experiments

Method		plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
ResNet-50	ResNet	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
BNM		89.6	61.5	76.9	55.0	89.3	69.1	81.3	65.5	90.0	47.3	89.1	30.1	70.4
MCD		87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
SWD		90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
FixBi		96.1	87.8	90.5	90.3	96.8	95.3	92.8	88.7	97.2	94.2	90.9	25.7	87.2
TVT	ViT	82.9	85.6	77.5	60.5	93.6	98.2	89.4	76.4	93.6	92.0	91.7	55.7	83.1
Deit-based		98.2	73.0	82.5	62.0	97.3	63.5	96.5	29.8	68.7	86.7	96.7	23.6	73.2
CDTrans-Deit		97.1	90.5	82.4	77.5	96.6	96.1	93.6	**88.6**	**97.9**	86.9	90.3	**62.8**	88.4
PMTrans-Deit		98.2	92.2	88.1	77.0	97.4	95.8	94.0	72.1	97.1	95.2	94.6	51.0	87.7
ViT-based		99.1	60.7	70.1	82.7	96.5	73.1	97.1	19.7	64.5	94.7	97.2	15.4	72.6
SSRT-ViT		98.9	87.6	**89.1**	**84.8**	98.3	**98.7**	96.3	81.1	94.8	97.9	94.5	43.1	**88.8**
PMTrans-ViT		98.9	**93.7**	84.5	73.3	**99.0**	98.0	96.2	67.8	94.2	**98.4**	96.6	49.0	87.5
Swin-based	Swin	99.3	63.4	85.9	68.9	95.1	79.6	**97.1**	29.0	81.4	94.2	**97.7**	29.6	76.8
PMTrans-Swin		**99.4**	88.3	88.1	78.9	98.8	98.3	95.8	70.3	94.6	98.3	96.3	48.5	88.0

Comparison with SOTA methods on VisDA-2017.

PMTrans also surpasses the SOTA methods on several sub-categories, such as ”horse” and ”sktbrd”.

Experiments

MCD	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	15.4	25.5	3.3	44.6	31.2	24.0
inf	24.1	-	24.0	1.6	35.2	19.7	20.9
pnt	31.1	14.8	-	1.7	48.1	22.8	23.7
qdr	8.5	2.1	4.6	-	7.9	7.1	6.0
rel	39.4	17.8	41.2	1.5	-	25.2	25.0
skt	37.3	12.6	27.2	4.1	34.5	-	23.1
Avg	28.1	12.5	24.5	2.4	34.1	21.2	20.5

SWD	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	14.7	31.9	10.1	45.3	36.5	27.7
inf	22.9	-	24.2	2.5	33.2	21.3	20.0
pnt	33.6	15.3	-	4.4	46.1	30.7	26.0
qdr	15.5	2.2	6.4	-	11.1	10.2	9.1
real	41.2	18.1	44.2	4.6	-	31.6	27.9
skt	44.2	15.2	37.3	10.3	44.7	-	30.3
Avg	31.5	13.1	28.8	6.4	36.1	26.1	23.6

BNM	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	12.1	33.1	6.2	50.8	40.2	28.5
inf	26.6	-	28.5	2.4	38.5	18.1	22.8
pnt	39.9	12.2	-	3.4	54.5	36.2	29.2
qdr	17.8	1.0	3.6	-	9.2	8.3	8.0
rel	48.6	13.2	49.7	3.6	-	33.9	29.8
skt	54.9	12.8	42.3	5.4	51.3	-	33.3
Avg	37.6	10.3	31.4	4.2	40.9	27.3	25.3

CGDM	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	16.9	35.3	10.8	53.5	36.9	30.7
inf	27.8	-	28.2	4.4	48.2	22.5	26.2
pnt	37.7	14.5	-	4.6	59.4	33.5	30.0
qdr	14.9	1.5	6.2	-	10.9	10.2	8.7
rel	49.4	20.8	47.2	4.8	-	38.2	32.0
skt	50.1	16.5	43.7	11.1	55.6	-	35.4
Avg	36.0	14.0	32.1	7.1	45.5	28.3	27.2

MDD	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	20.5	40.7	6.2	52.5	42.1	32.4
inf	33.0	-	33.8	2.6	46.2	24.5	28.0
pnt	43.7	20.4	-	2.8	51.2	41.7	32.0
qdr	18.4	3.0	8.1	-	12.9	11.8	10.8
rel	52.8	21.6	47.8	4.2	-	41.2	33.5
skt	54.3	17.5	43.1	5.7	54.2	-	35.0
Avg	40.4	16.6	34.7	4.3	43.4	32.3	28.6

SCDA	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	18.6	39.3	5.1	55.0	44.1	32.4
inf	29.6	-	34.0	1.4	46.3	25.4	27.3
pnt	44.1	19.0	-	2.6	56.2	42.0	32.8
qdr	30.0	4.9	15.0	-	25.4	19.8	19.0
rel	54.0	22.5	51.9	2.3	-	42.5	34.6
skt	55.6	18.5	44.7	6.4	53.2	-	35.7
Avg	42.6	16.7	37.0	3.6	47.2	34.8	30.3

CDTrans	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	29.4	57.2	26.0	72.6	58.1	48.7
inf	57.0	-	54.4	12.8	69.5	48.4	48.4
pnt	62.9	27.4	-	15.8	72.1	53.9	46.4
qdr	44.6	8.9	29.0	-	42.6	28.5	30.7
rel	66.2	31.0	61.5	16.2	-	52.9	45.6
skt	69.0	29.6	59.0	27.2	72.5	-	51.5
Avg	59.9	25.3	52.2	19.6	65.9	48.4	45.2

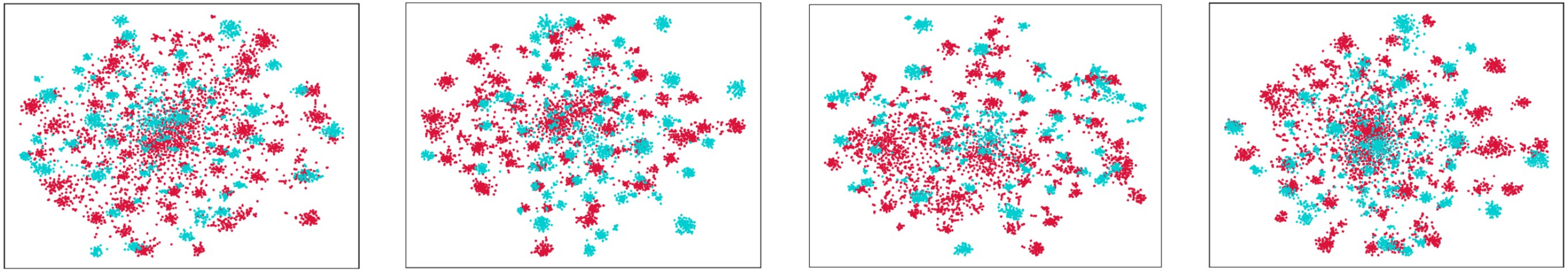
SSRT	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	33.8	60.2	19.4	75.8	59.8	49.8
inf	55.5	-	54.0	9.0	68.2	44.7	46.3
pnt	61.7	28.5	-	8.4	71.4	55.2	45.0
qdr	42.5	8.8	24.2	-	37.6	33.6	29.3
rel	69.9	37.1	66.0	10.1	-	58.9	48.4
skt	70.6	32.8	62.2	21.7	73.2	-	52.1
Avg	60.0	28.2	53.3	13.7	65.3	50.4	45.2

PMTrans	clp	inf	pnt	qdr	rel	skt	Avg
clp	-	34.2	62.7	32.5	79.3	63.7	54.5
inf	67.4	-	61.1	22.2	78.0	57.6	57.3
pnt	69.7	33.5	-	23.9	79.8	61.2	53.6
qdr	54.6	17.4	38.9	-	49.5	41.0	40.3
rel	74.1	35.3	70.0	25.4	-	61.1	53.2
skt	73.8	33.0	62.6	30.9	77.5	-	55.6
Avg	67.9	30.7	59.1	27.0	72.8	56.9	**62.9**

Comparison with SOTA methods on DomainNet.

PMTrans outperforms the SOTA methods by +17.7% accuracy. Incredibly, PMTrans surpasses the SOTA methods in all the 30 subtasks.

T-SNE Visualization



Swin-Base PMTrans-Swin PMTrans-ViT PMTrans-Deit

t-SNE visualizations for task A→C on the Office-Home dataset.

Compared with Swin-based and PMTrans-Swin, our PMTrans model can better align the two domains by constructing the intermediate domain to bridge them.

Ablation Study

Effect of semi-supervised loss

$\mathcal{L}_{cls}^{S}$	$\mathcal{L}_f$	$\mathcal{L}_l$	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
✓			72.7	87.1	90.6	84.3	87.3	89.3	80.6	68.6	90.3	84.8	69.4	91.3	83.6
✓	✓		73.9	87.5	91.0	85.3	87.9	89.9	82.8	72.1	91.2	86.3	74.1	92.4	84.6
✓		✓	79.2	91.8	92.3	88.0	92.6	**93.0**	87.1	77.8	92.5	88.2	78.4	93.9	87.9
✓	✓	✓	**81.3**	**92.9**	**92.8**	**88.4**	**93.4**	**93.2**	**87.9**	**80.4**	**93.0**	**89.0**	**80.9**	**94.8**	**89.0**

Effect of learning parameters

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
Beta(1,1)	79.9	92.0	92.3	88.6	92.6	92.4	86.9	79.0	92.4	88.2	79.3	94.0	88.1
Beta(2,2)	79.9	92.1	92.7	88.4	92.4	92.7	86.9	79.5	92.1	88.1	79.6	94.3	88.2
Learning	**81.3**	**92.9**	**92.8**	**88.4**	**93.4**	**93.2**	**87.9**	**80.4**	**93.0**	**89.0**	**80.9**	**94.8**	**89.0**

Effect of PatchMix

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
Mixup	79.4	92.4	92.6	87.5	92.8	92.4	86.8	**80.3**	92.5	88.2	79.7	**95.4**	88.3
CutMix	79.2	91.2	92.2	87.6	91.8	91.8	86.0	77.8	92.6	88.2	78.4	94.1	87.6
PatchMix	**81.3**	**92.9**	**92.8**	**88.4**	**93.4**	**93.2**	**87.9**	**80.4**	**93.0**	**89.0**	**80.9**	94.8	**89.0**

Conclusion

1. We proposed a novel method, PMTrans, an optimization solution for UDA from a game perspective.

2. PMTrans achieved the SOTA results on three benchmark UDA datasets, outperforming the prior methods by a large margin.

3. We plan to implement our PatchMix and the two semi-supervised mixup losses to solve self-supervised and semi-supervised learning problems.

4. We will also exploit our method to tackle the challenging downstream tasks, e.g., semantic segmentation and object detection.

Thanks !

Email: zhujinjing.hkust@gmail.com

https://vlis2022.github.io/cvpr23/PMTrans