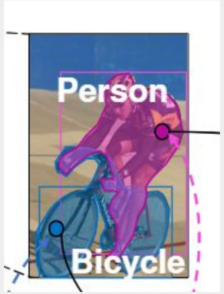
Presenter →  *Seohyun Back (Ph.D. Student)*
Engineer @ Samsung Research

Informative and Consistent Correspondence Mining for Cross-Domain Weakly Supervised Object Detection

***CVPR*, 2021** **(Oral Presentation)**

Luwei Hou[*1,3], Yu Zhang[*†3], Kui Fu[1], Jia Li[†1,2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China [3]SenseTime Research

PURPOSE

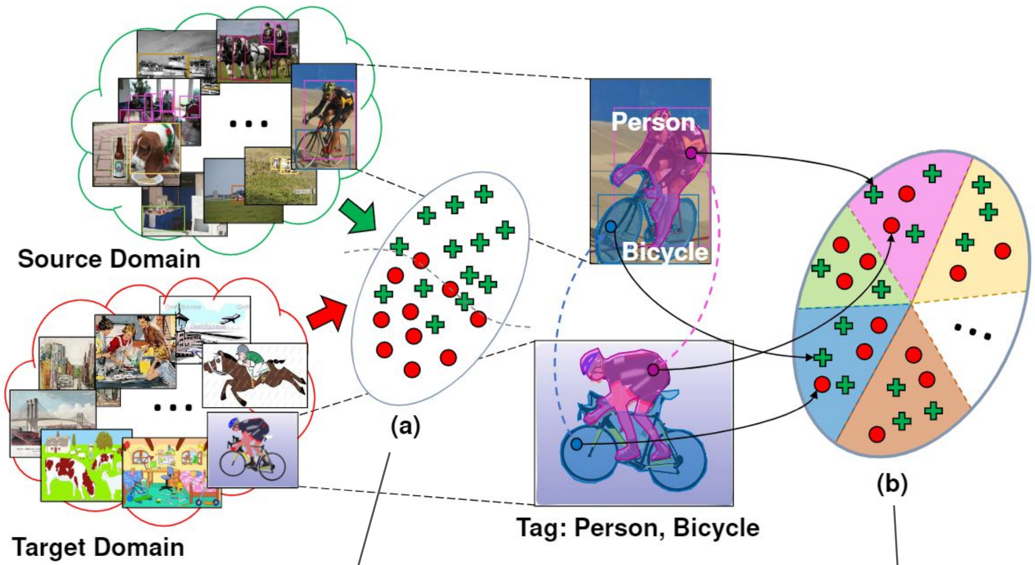


Domain Adaptation by Weakly supervision

Source Domain (real-world)
Fully Annotated: class, bounding-box
(e.g. Pascal-VOC 2007 / 2012)

Target Domain (unreal)
Partially Annotated: presence of classes
(e.g. Clipart1k, Watercolor2k, Comic2k)

Introduction



(a) Conventional approaches (Domain-level)

- Project images from different domains into a unified feature space.
- Adversarially train discriminative classifier not to easily separate them.

(b) Our approach (Pixel-level)

- Explicitly establish pixel-wise correspondence among the semantic regions of cross-domain.
- Form semantic clusters in feature space for well explanation of source domain's region annotation.

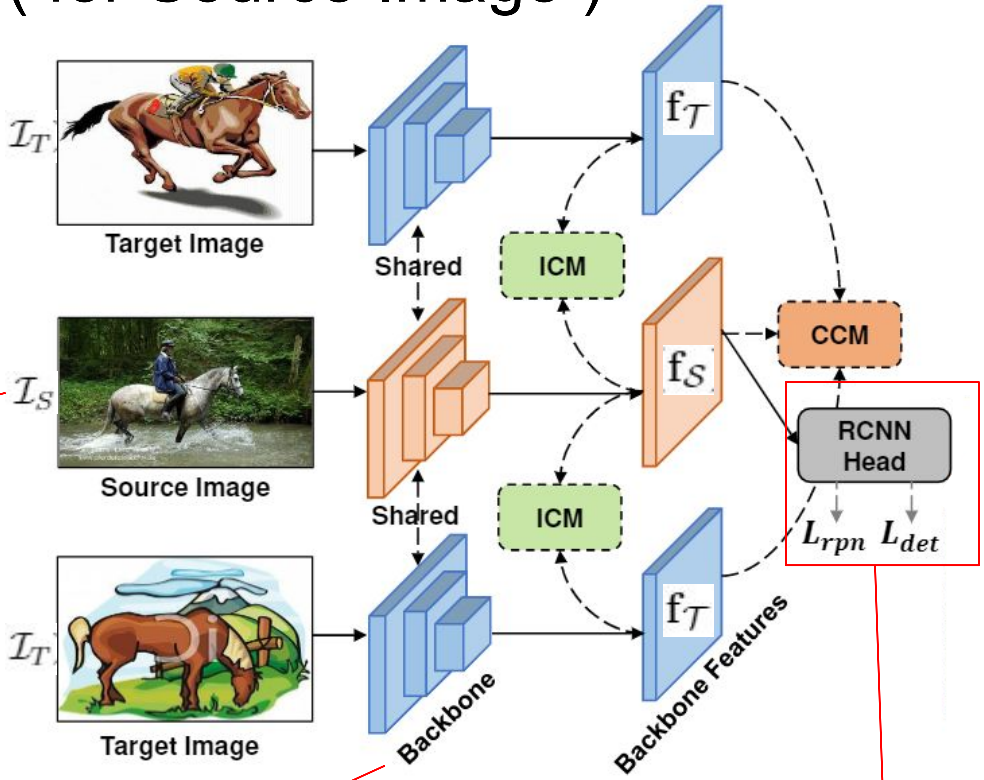
Overview : RCNN Head (for Source Image)

Semantic label

$$\mathbb{Y}_{\mathcal{S}} = \{\mathrm{y}_{\mathcal{B}} \in \{0,1\}^{1\times(|\mathbb{C}|+1)} | \mathcal{B} \in \mathbb{B}\}$$

Bounding-box coordinates

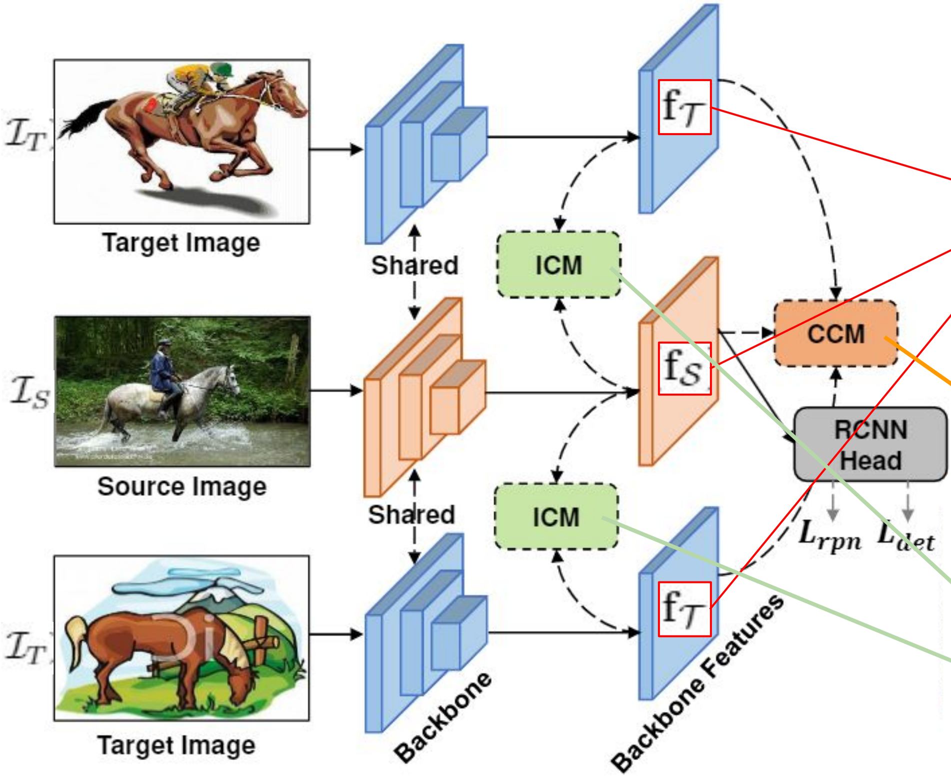
$$\mathbb{P}_{\mathcal{S}} = \{\mathrm{p}_{\mathcal{B}} \in \mathbb{R}^{1\times 4} | \mathcal{B} \in \mathbb{B}\}$$



two−stage
Faster−RCNN

$$L_{\mathcal{S}}(I_{\mathcal{S}}, \mathbb{Y}_{\mathcal{S}}, \mathbb{P}_{\mathcal{S}}) = L_{rpn}(I_{\mathcal{S}}, \mathbb{P}_{\mathcal{S}}) + L_{det}(I_{\mathcal{S}}, \mathbb{Y}_{\mathcal{S}}, \mathbb{P}_{\mathcal{S}})$$

Overview : Contribution Points (Source-target DA)



Previous works

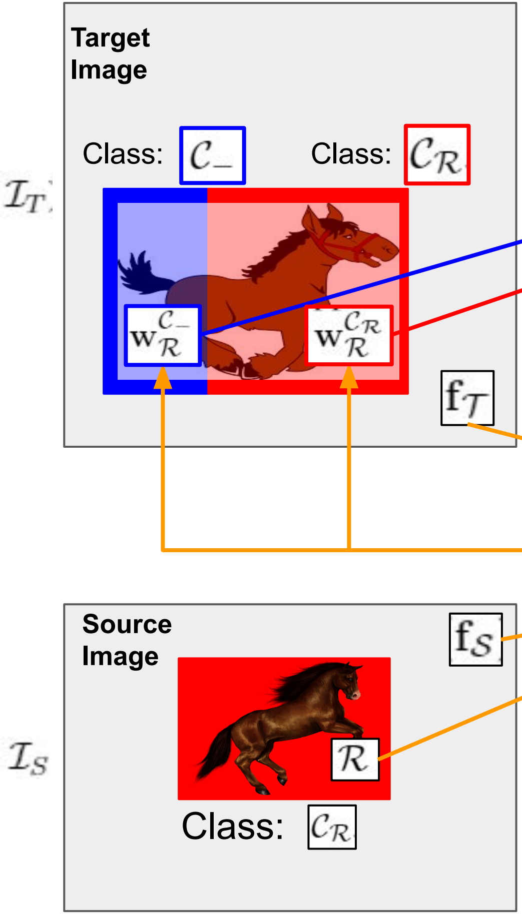
$$L_{\mathcal{T}}\left(I_{\mathcal{S}}, I_{\mathcal{T}}\right)=\mathbb{E}\left[\log \mathcal{D}\left(\mathrm{f}_{\mathcal{S}}\right)\right]+\mathbb{E}\left[\log \left(1-\mathcal{D}\left(\mathrm{f}_{\mathcal{T}}\right)\right)\right]$$

→ Favor in minimizing the most discriminative variance between domains!

Contribution−(2)
Consistent Correspondence Mining (CCM)

Contribution−(1)
Informative Correspondence Mining (ICM)

Informative Correspondence Mining (ICM) - 1/4



"High-level idea : Drop the target partitions, not helping to explain source-domain region"

$$\min_{\Omega} \sum_{\substack{\mathcal{C}_- \in \mathbb{C}_{\mathcal{S}\cap\mathcal{T}} \\ \mathcal{C}_- \neq \mathcal{C}_\mathcal{R}}} I\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}, \mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}} | \mathbb{I} = (\mathcal{I}_\mathcal{S}, \mathcal{I}_\mathcal{T})\right) \quad \boxed{\mathbb{C}_{\mathcal{S}\cap\mathcal{T}} \subseteq \mathbb{C} \cup \{\mathcal{C}_0\}}$$

assume to be constant ($P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbb{I})$ →uniform distribution)

$$\boxed{I(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}, \mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}|\mathbb{I}) = \boxed{H(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbb{I})} - H(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I})} \quad \boxed{\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)}$$

<u>**Minimizing Mutual Information means,**</u>

Minimizing Entropy(Uncertainty) of Target Region for C_
Maximizing Entropy(Uncertainty) of Target Region for C_ given Target Region for C_R

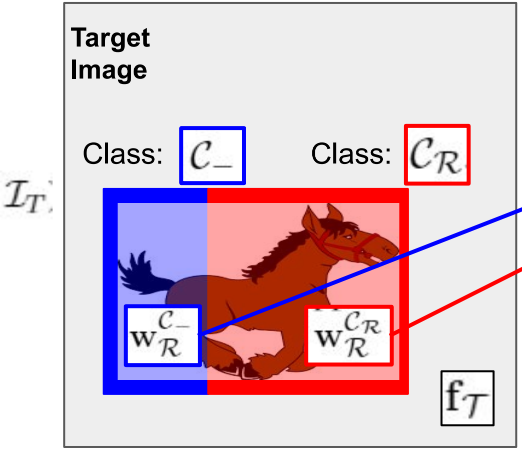
partition generator (Single Conv layer)
$\mathcal{G} : \mathbf{f}_\mathcal{T} \rightarrow (0,1)^{HW \times |\mathbb{C}|}$

Target to Source Attention pooling

$$\mathbf{w}_\mathcal{R}^{\mathcal{C}} = \mathrm{avgpool}\left(\mathrm{norm}\left(\mathbf{m}\left(\Omega_\mathcal{C}\right) \odot \kappa\left(\mathbf{f}_\mathcal{S}, \mathbf{f}_\mathcal{T}\right)\right), \mathcal{R}\right),$$
$$\text{where } \kappa\left(\mathbf{f}_\mathcal{S}, \mathbf{f}_\mathcal{T}\right) = \mathrm{softmax}\left(\mathbf{f}_\mathcal{S}^{\mathrm{T}} \mathbf{W} \mathbf{f}_\mathcal{T}\right).$$

The Problem here is computing posterior for $H(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I}) \rightarrow P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}, \mathbb{I})$

Informative Correspondence Mining (ICM) - 2/4



Solution : Variational approximation !

$$P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}},\mathbb{I}) \approx Q(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}},\mathbb{I}) = \int P(\boxed{\mathbf{a}_\mathcal{R}}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}},\mathbb{I})P(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\boxed{\mathbf{a}_\mathcal{R}},\mathbb{I})d\mathbf{a}_\mathcal{R}, \quad (1)$$

Since $P\left(\mathbf{a}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_+},\mathbb{I}\right) = \dfrac{P\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_+}|\mathbf{a},\mathbb{I}\right)\boxed{P(\mathbf{a}|\mathbb{I})}}{P\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_+}|\mathbb{I}\right)}$ **is delta distribution,**

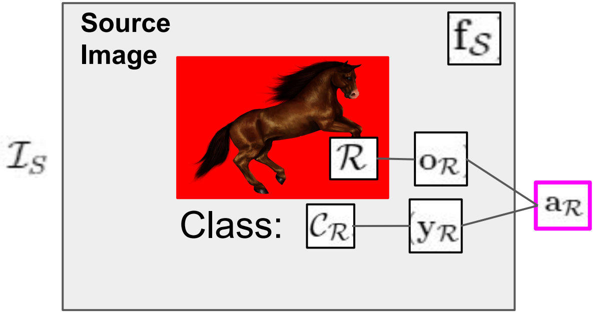
it takes 1 when $\mathbf{a} = \mathbf{a}_\mathcal{R}$ and 0 otherwise.

by marginalizing $\mathbf{a}_\mathcal{R}$**,** $(1) \rightarrow$

$$Q\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}^+},\mathbb{I}\right) = P\left(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_-},\mathbb{I}\right)P\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbb{I}\right) \quad (2)$$
$$\text{s.t. } \mathbf{a}_\mathcal{R} = \arg\max_\mathbf{a} P\left(\mathbf{a}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_+},\mathbb{I}\right).$$

by rule of conditional entropy,

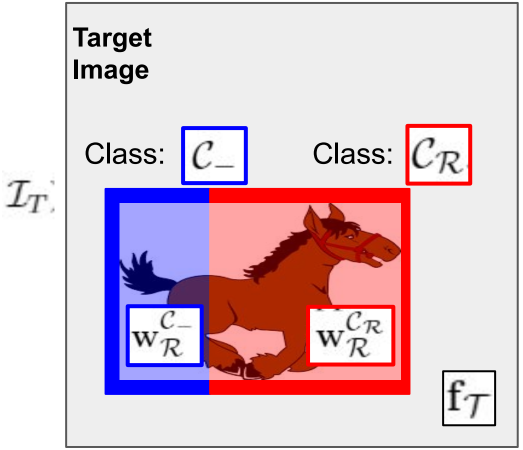
$$H\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_+},\mathbb{I}\right) \approx H_Q\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_+},\mathbb{I}\right)$$
$$= H_Q\left(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_-},\mathbb{I}\right) + H_Q\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbb{I}\right) - H_Q\left(\mathbf{a}_\mathcal{R}|\mathbb{I}\right), \quad (3)$$



Since that $H\left(\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}|\mathbb{I}\right)$ and $H\left(\mathbf{a}_\mathcal{R}|\mathbb{I}\right)$ are constants,

$$\min_\Omega \sum_{\substack{\mathcal{C}_-\in\mathbb{C}_{\mathcal{S}\cap\mathcal{T}}\\ \mathcal{C}_-\neq\mathcal{C}_\mathcal{R}}} I\left(\boxed{\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}}\boxed{\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}}\mathbb{I} = (\mathcal{I}_\mathcal{S},\mathcal{I}_\mathcal{T})\right) \xrightarrow{\textbf{Approximation}} \max_\Omega A(\mathcal{S},\mathcal{T}) = \frac{1}{Z_A}\sum_\mathbb{I}\sum_{\mathcal{R}\in\mathbb{R}_\mathcal{S}}\sum_{\mathcal{C}_-} H\left(\mathbf{a}_\mathcal{R}|\boxed{\mathbf{w}_\mathcal{R}^{\mathcal{C}_-}},\mathbb{I}\right)$$
$$\text{s.t. } \mathbf{a}_\mathcal{R} = \arg\max_\mathbf{a} P\left(\mathbf{a}|\boxed{\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}}},\mathbb{I}\right),$$

Informative Correspondence Mining (ICM) - 3/4



$$\max_{\Omega} A(\mathcal{S},\mathcal{T}) = \frac{1}{Z_A}\sum_{\mathbb{I}}\sum_{\mathcal{R}\in\mathbb{R}_\mathcal{S}}\sum_{\mathcal{C}_-} H\left(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_-},\mathbb{I}\right) \rightarrow$$

Deemed as “adversarial correspondence drop”

$$\text{s.t. } \mathbf{a}_\mathcal{R} = \arg\max_{\mathbf{a}} P\left(\mathbf{a}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}},\mathbb{I}\right),$$

By relaxing the constraints,

$$\boxed{\min_{\mathbf{w}}\left(\lambda N(\mathcal{S},\mathcal{T}) + \max_{\Omega} A(\mathcal{S},\mathcal{T})\right),}$$

$$\mathcal{N}(\mathcal{S},\mathcal{T}) = -\tfrac{1}{Z_N}\textstyle\sum_{\mathbb{I}}\sum_{\mathcal{R}\in\mathbb{R}_\mathcal{S}} \log P(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}_\mathcal{R}},\mathbb{I})$$

Now, what we need is only the posterior $\boxed{P(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}},\mathbb{I})}$

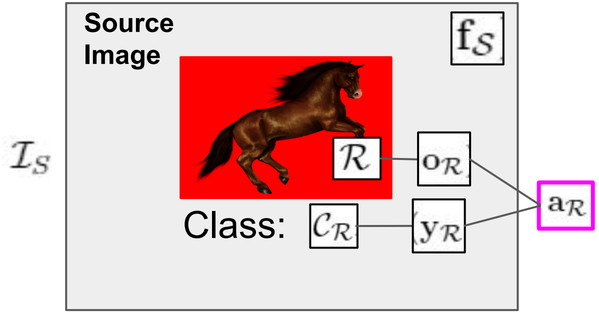
$$P\left(\mathbf{a}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}},\mathbb{I}\right) = P\left(\mathbf{y}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}},\mathbb{I}\right) P\left(\mathbf{o}_\mathcal{R}|\mathbf{w}_\mathcal{R}^{\mathcal{C}},\mathbb{I}\right), \quad \boxed{\mathbf{a}_\mathcal{R}} = (\mathbf{y}_\mathcal{R},\mathbf{o}_\mathcal{R})$$

$$\mathbf{f}_\mathcal{R}^{\mathcal{C}} = \left(\mathbf{w}_\mathcal{R}^{\mathcal{C}}\right)^{\mathrm{T}}\mathbf{f}_\mathcal{T},$$

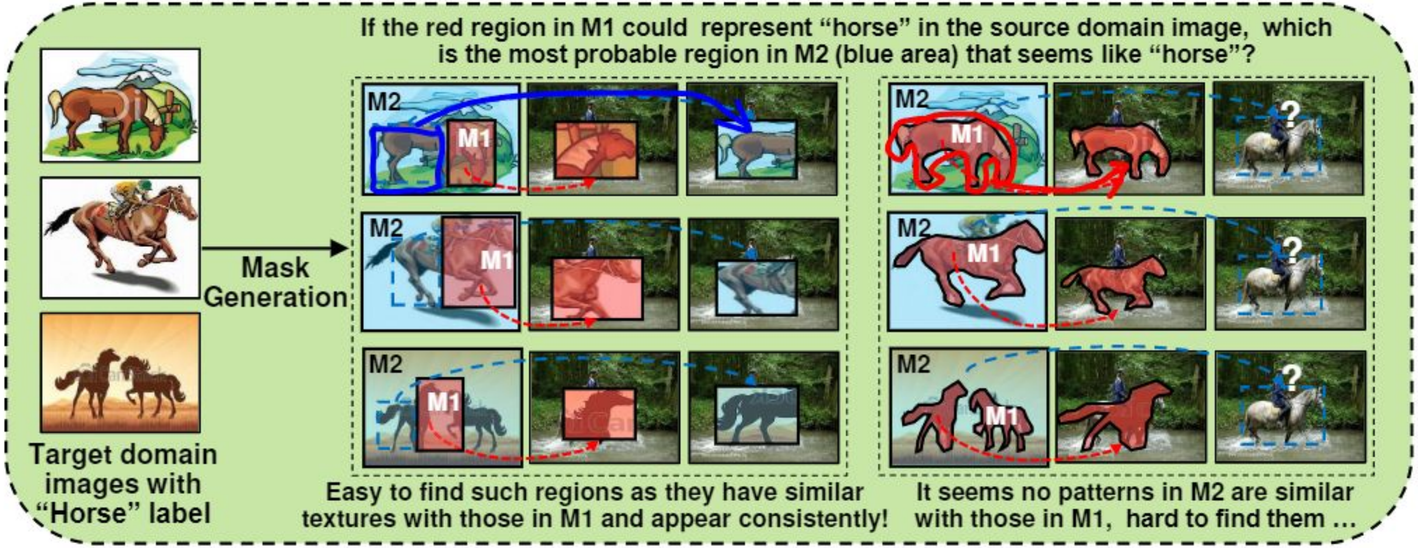
by using 2 separate FC layer,

$$P\left(\mathbf{y}_\mathcal{R}^{\mathcal{C}}|\mathbf{w}_\mathcal{R}^{\mathcal{C}},\mathbb{I}\right) \sim \operatorname{softmax}\left(\mathbf{y}_\mathcal{R}^{\mathcal{C}},\mathcal{F}_c\left(\mathbf{f}_\mathcal{R}^{\mathcal{C}}\right)\right),$$

$$P\left(\mathbf{o}_\mathcal{R}^{\mathcal{C}}|\mathbf{w}_\mathcal{R}^{\mathcal{C}},\mathbb{I}\right) \sim \exp\left(-\frac{\|\mathbf{o}_\mathcal{R}^{\mathcal{C}} - \mathcal{F}_o\left(\mathbf{f}_\mathcal{R}^{\mathcal{C}}\right)\|_1}{\sigma_o^2}\right).$$



Informative Correspondence Mining (ICM) - 4/4

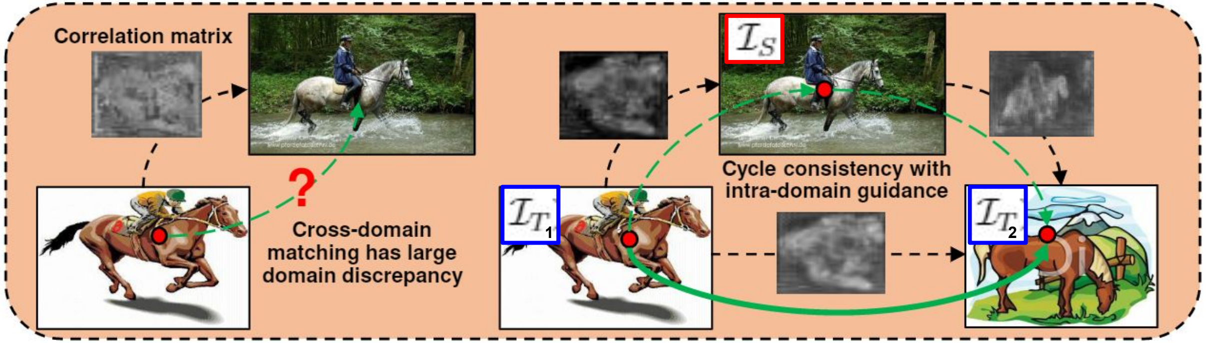


Informative Correspondence Mining (ICM)

By doing this procedure, the model adversarially drops the information on the target image to make the correspondence searching harder.

Consistent Correspondence Mining (CCM) - 1/2

"High-level idea : intra-intra (A→A) domain matching should be equal to intra-inter-intra (A→B→A) domain matching"



Consistent Correspondence Mining (CCM)

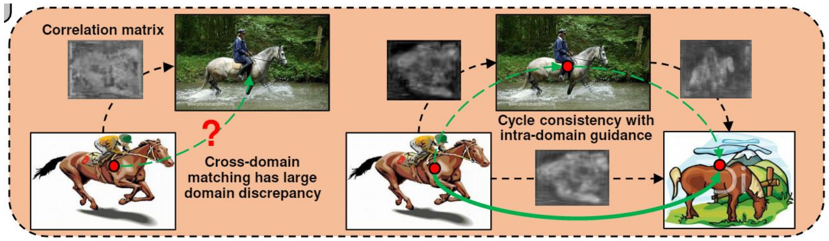
$$C(\mathcal{S},\mathcal{T}) = \frac{1}{Z_C}\sum_{\mathbb{J}} \mathbf{R}_{\mathbb{J}} \left\| \mathbf{K}_{\mathcal{T}_1 \leftarrow \mathcal{T}_2} - \mathbf{K}_{\mathcal{T}_1 \leftarrow \mathcal{S}} \mathbf{K}_{\mathcal{S} \leftarrow \mathcal{T}_2} \right\|_2^2 ,$$

$\boxed{\mathbb{J} = (\mathcal{I}_{\mathcal{S}}, \mathcal{I}_{\mathcal{T}_1}, \mathcal{I}_{\mathcal{T}_2})}$ $\boxed{\mathbf{K}_{\mathcal{B} \leftarrow \mathcal{A}} = \kappa(\mathbf{f}_{\mathcal{A}}, \mathbf{f}_{\mathcal{B}})\mathbf{f}_{\mathcal{B}}}$ $\boxed{\kappa(\mathbf{f}_{\mathcal{S}}, \mathbf{f}_{\mathcal{T}}) = \mathrm{softmax}(\mathbf{f}_{\mathcal{S}}^{\mathrm{T}} \mathbf{W} \mathbf{f}_{\mathcal{T}})}$

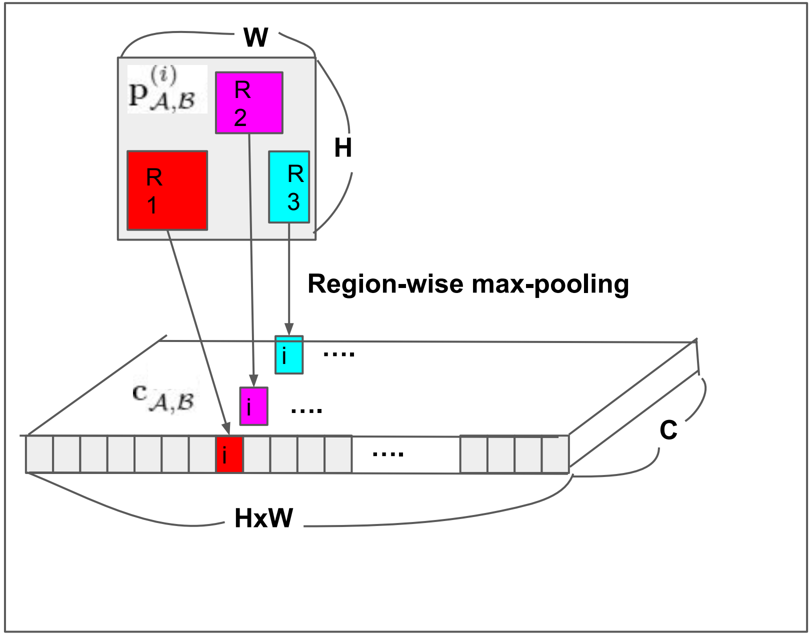
$\mathbf{R}_{\mathbb{J}}$ **is HxW matrix quantifies <u>"transferability"</u> →**

$\mathcal{I}_{\mathcal{T}_1}$ and $\mathcal{I}_{\mathcal{T}_2}$ share a class that is absent in $\mathcal{I}_{\mathcal{S}}$, we cannot expect to reconstruct the warping $\mathcal{T}_1 \leftarrow \mathcal{T}_2$ faithfully everywhere using the immediate warpings $\mathcal{T}_1 \leftarrow \mathcal{S}$ and $\mathcal{S} \leftarrow \mathcal{T}_2$.

Consistent Correspondence Mining (CCM) - 2/2



Consistent Correspondence Mining (CCM)



For $\mathcal{B} \leftarrow \mathcal{A}$ $\boxed{\mathcal{B} \text{ is assumed as source image}}$

$$\mathbf{p}^{(i)}_{\mathcal{A},\mathcal{B}} = \mathrm{softmax}((\mathbf{f}^{(i)}_{\mathcal{A}})^{\mathrm{T}}\mathbf{W}\mathbf{f}_{\mathcal{B}}) \xrightarrow[\text{max-pooling}]{\text{Region-wise}} \mathbf{c}^{(i)}_{\mathcal{A},\mathcal{B}} \in (0,1)^{1\times(|\mathbb{C}|+1)}$$

Transferability $\mathcal{B} \leftarrow \mathcal{A}$ **:** $r^{(i)}_{\mathcal{A},\mathcal{B}} = \exp(-H(\mathbf{c}^{(i)}_{\mathcal{A},\mathcal{B}})).$

if $\mathbf{f}^{(i)}_{\mathcal{A}}$ is a confident match, $\mathbf{c}^{(i)}_{\mathcal{A}}$ tends to have peaks, leading to low uncertainty (high transferability).

$$C(\mathcal{S},\mathcal{T}) = \frac{1}{Z_C}\sum_{\mathbb{J}} \mathbf{R}_{\mathbb{J}} \left\|\mathbf{K}_{\mathcal{T}_1\leftarrow\mathcal{T}_2} - \mathbf{K}_{\mathcal{T}_1\leftarrow\mathcal{S}}\mathbf{K}_{\mathcal{S}\leftarrow\mathcal{T}_2}\right\|_2^2,$$

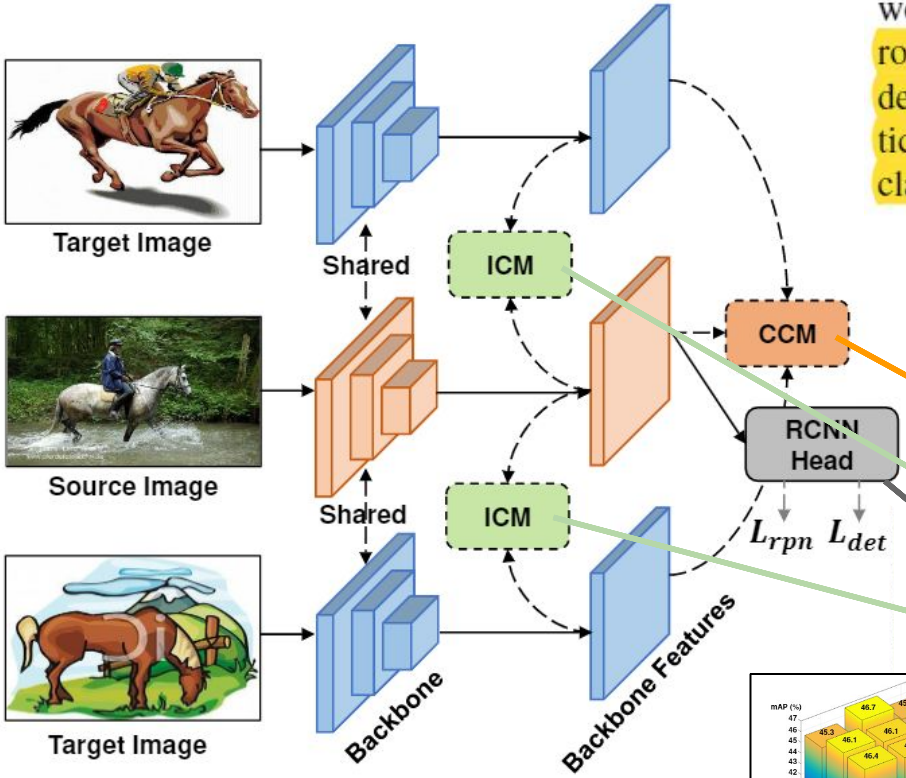
Accumulated Transferability $\mathcal{T}_1 \leftarrow \mathcal{T}_2$ **:** $\mathbf{R}_{\mathbb{J}} = \mathbf{r}_{\mathcal{T}_1,\mathcal{S}} \odot (\mathbf{K}_{\mathcal{T}_1\leftarrow\mathcal{T}_2}\mathbf{r}_{\mathcal{T}_2,\mathcal{S}})$

Detached when training due to gradient instability

Minimizing intra-domain attention-sum with intra-domain attention-sum through inter-domain (Weighted by transferability)

Final training objective

Class-agnostic ICM. The original multi-class version of ICM repeats correspondence learning for every class, which would be slow if the number of classes is large. A walkaround is to unify all the object classes as "foreground", rendering it a binary foreground/background setting. In practice, we do not observe performance degeneration for this class-agnostic setting, yet saving training time significantly.



Used gradient reversal for adversarial training

$$\min_{\theta_0} \Big(\underbrace{L_S}_{\text{OD}} + \alpha(\mathcal{N} + \underbrace{\max_{\theta_\Omega} A}_{\text{ICM}}) + \underbrace{\beta C}_{\text{CCM}} \Big), \quad (7)$$

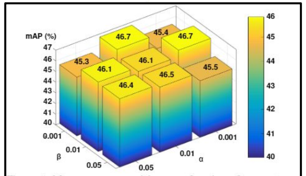
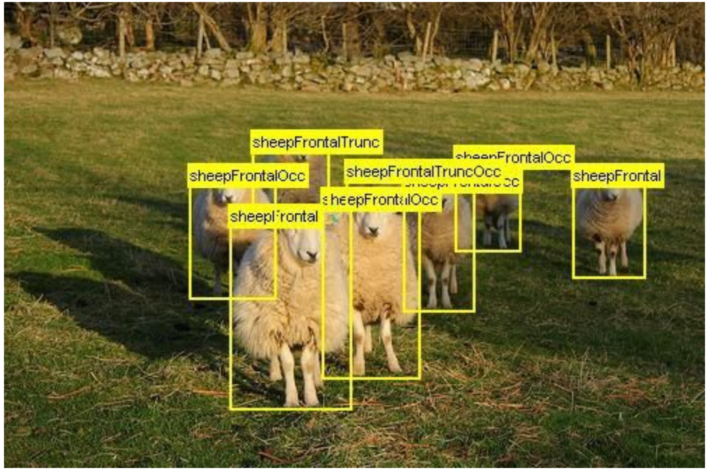


Figure 6. Mean average precision as a function of parameters α and β defined in Eqn. (7), evaluated on the Clipart1k dataset.

α and β in (7) with 0.001 and 0.01,

Datasets

Source domain : Pascal-VOC 2007 / 2012   (16551 real-world photo, 20 classes)
Target domain : Clipart1k (20 classes) , Watercolor2k (6 classes) , Comic2k (6 classes)
→ 1000 for each train/eval split



Pascal-VOC

(a) Clipart1k (b) Watercolor2k (c) Comic2k

Experimental results

WS : Weakly Supervised → w/o Domain Adaptation
UDA : Unsupervised Domain Adaptation → Assumed unlabeled target
CDWS : Cross-Domain Weakly Supervised → Assumed weakly labeled target

Table 1. Average Precisions (AP) and mean AP on Clipart1k. Bold highlights the top place while underline the second place.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Source only	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32	27.8
WL Group → performance degradation may due to style diversity of target domain																					
WSDDN [1]	1.6	3.6	0.6	2.3	0.1	11.7	4.5	0.0	3.2	0.1	2.8	2.3	0.9	0.1	14.4	16.0	4.5	0.7	1.2	18.3	4.4
CLNet [17]	3.2	22.3	2.2	0.7	4.6	4.8	17.5	0.2	4.8	1.6	6.4	0.6	4.7	0.6	12.5	13.1	14.1	4.1	8.0	29.7	7.8
EDRN [34]	2.7	13.5	1.2	4.2	1.8	10.3	25.7	0.4	8.4	0.3	3.2	2.7	1.1	0.7	29.4	17.2	5.2	1.6	2.9	19.1	7.6
PCL [37]	3.4	10.6	2.3	1.7	5.2	3.4	23.3	1.2	5.6	0.4	7.8	3.7	5.6	0.3	24.5	19.7	11.9	3.6	9.2	25.4	8.4
UDA Group → performance degradation may due to inaccurate pseudo labels																					
ADDA [38]	20.1	50.2	20.5	23.6	11.4	40.5	34.9	2.3	39.7	22.3	27.1	10.4	31.7	53.6	46.6	32.1	18.0	21.1	23.6	18.3	27.4
SWDA [32]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	**52.0**	9.3	24.9	54.1	49.1	38.1
STABR [19]	28.0	64.5	23.9	19.0	21.9	64.3	43.5	16.4	42.2	25.9	**30.5**	7.9	25.5	67.6	54.5	36.4	10.3	**31.2**	**57.4**	43.5	35.7
HTD [2]	33.6	58.9	34.0	23.4	**45.6**	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	**51.8**	40.3
CDWS Group																					
CDWSDA [15]	32.0	40.9	29.5	29.3	32.0	**84.7**	38.2	12.4	24.3	54.8	24.7	15.4	36.1	72.1	51.0	41.9	19.0	18.5	47.2	21.4	36.3
Proposed	**39.8**	**66.7**	**37.2**	**42.5**	43.3	48.1	**48.1**	**21.3**	**46.5**	**73.0**	29.0	**29.8**	**57.3**	**78.6**	**67.8**	48.7	**46.3**	19.3	42.8	48.5	**46.7**

Minor poputation categories : train, tv

Experimental results

WS : Weakly Supervised → w/o Domain Adaptation
UDA : Unsupervised Domain Adaptation → Assumed unlabeled target
CDWS : Cross-Domain Weakly Supervised → Assumed weakly labeled target

Table 2. Average Precisions (AP) and mean AP on Watercolor2k. Bold highlights the top place while underline the second place.

Method	bike	bird	car	cat	dog	person	mAP
Source only	68.8	46.8	37.2	32.7	21.3	60.7	44.6
WL Group							
WSDDN [1]	1.5	26.0	14.6	0.4	0.5	33.3	12.7
CLNet [17]	4.5	27.9	19.6	14.3	6.4	31.4	17.4
EDRN [34]	5.2	29.3	15.3	1.4	0.9	34.9	14.5
PCL [37]	6.7	28.8	20.2	9.5	5.4	27.4	16.3
UDA Group							
ADDA [38]	79.9	49.5	39.5	**35.3**	29.4	65.1	49.8
SWDA [32]	82.3	55.9	46.5	32.7	35.5	66.7	53.3
STABR [19]	75.6	45.8	49.3	34.1	30.3	64.1	49.4
HTD [2]	69.2	49.5	49.5	34.9	30.8	61.2	49.2
CDWS Group							
CDWSDA [15]	68.6	46.6	37.7	35.2	36.0	62.5	47.8
Proposed	**86.6**	**64.2**	**52.6**	32.4	**41.2**	**67.4**	**57.4**

Table 3. Average Precisions (AP) and mean AP on Comic2k. Bold highlights the top place while underline the second place.

Method	bike	bird	car	cat	dog	person	mAP
Source only	28.8	13.5	18.6	14.8	15.9	33.9	20.9
WL Group							
WSDDN [1]	1.5	0.1	11.9	6.9	1.4	12.1	5.6
CLNet [17]	0.0	0.0	2.0	4.7	1.2	14.9	3.8
EDRN [34]	1.6	0.5	13.2	7.2	2.5	13.2	6.4
PCL [37]	1.2	0.4	8.9	2.9	2.3	15.6	5.2
UDA Group							
ADDA [38]	39.5	9.8	17.2	12.7	20.4	43.3	23.8
SWDA [32]	30.3	19.6	28.8	15.2	24.9	46.9	27.6
STABR [19]	**50.6**	13.6	31.0	7.5	16.4	41.4	26.8
HTD [2]	35.4	14.8	26.6	13.7	26.9	40.0	26.2
CDWS Group							
CDWSDA [15]	47.0	21.1	30.1	29.0	29.6	40.6	32.9
Proposed	**50.6**	**23.3**	**35.4**	**32.3**	**33.8**	**47.1**	**37.1**

Experimental results

Figure 3. Representative results generated by different approaches (visualized in different rows). Best viewed with zoom in.

Ablation study

$$\boxed{\max_{\Omega} A(\mathcal{S},\mathcal{T}) = \frac{1}{Z_A}\sum_{\mathbb{I}}\sum_{\mathcal{R}\in\mathbb{R}_S}\sum_{\mathcal{C}_-} H\left(\mathbf{a}_{\mathcal{R}}|\mathbf{w}_{\mathcal{R}}^{\mathcal{C}_-},\mathbb{I}\right)}$$

→ **Deemed as “adversarial correspondence drop”**

$$\text{s.t. } \mathbf{a}_{\mathcal{R}} = \arg\max_{\mathbf{a}} P\left(\mathbf{a}|\mathbf{w}_{\mathcal{R}}^{\mathcal{C}_{\mathcal{R}}},\mathbb{I}\right),$$

Table 4. Contributions to the final mAP by different component evaluated on the Clipart1k dataset.

Source	ICM			CCM	mAP
	w/o adv.	full	w/o reg.		
✓					27.8
✓	✓				44.3
✓		✓			45.0
✓	✓			✓	45.7
✓			✓	✓	45.5
✓		✓		✓	46.7

→ demonstrating the advantage of pixel-wise knowledge transfer.

→ demonstrating adversarial mask generation effectiveness. (Full ICM)

→ demonstrating CCM effectiveness (Full ICM + CCM)

→ demonstrating objection position regression in ICM is beneficial for domain adaptation.

Analysis of error reduction

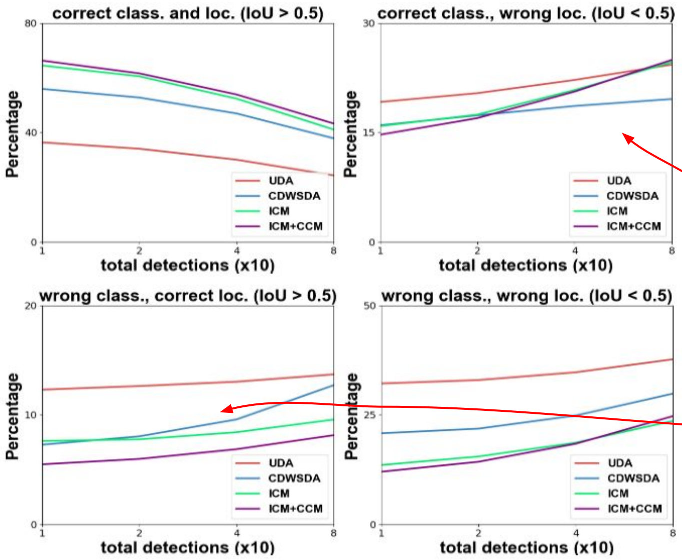
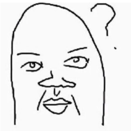


Figure 4. Percentage of detections within each type as a function of the number of detections. Top-left: detections with correct classification and localization. Top-right: classification is correct, but localization is weak ($0.1 < \text{IoU} < 0.5$). Bottom-left: wrong classification, but correct localization (IoU with at least one object exceeds 0.5). Bottom-right: detections with wrong labels and localization ($\text{IoU} < 0.1$). Note that higher percentage is preferred for only the top-left figure, as it counts for true positive detections.

because ICM does not have pseudo labeling process that may introduce undesired labeling noise. (?????)

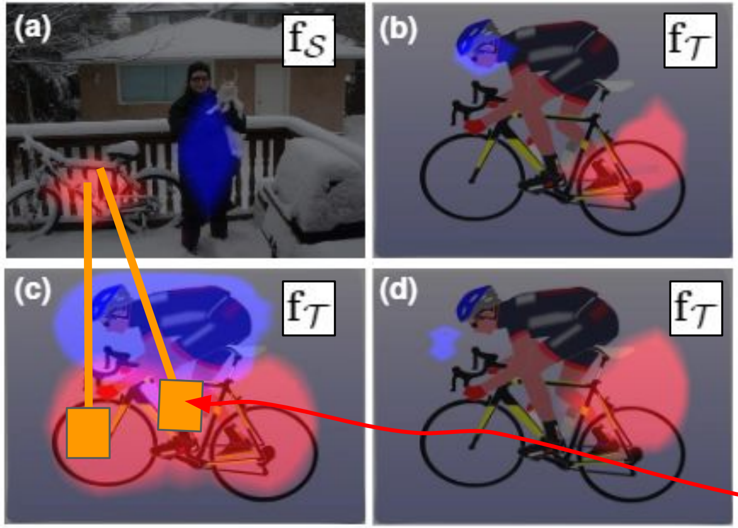


Visualizing the effect of ICM and CCM

$$\boxed{\kappa\left(\mathbf{f}_{\mathcal{S}}, \mathbf{f}_{\mathcal{T}}\right) = \operatorname{softmax}\left(\mathbf{f}_{\mathcal{S}}^{\mathrm{T}} \mathbf{W} \mathbf{f}_{\mathcal{T}}\right)}$$

→ HW x HW (matched position)

seeds are weighted by spatial Gaussian



Accumulated predicted bounding-box with 1 matched position's target feature as input to RCNN head

Figure 5. Visualizing the effect of Informative Correspondence Mining (ICM) and Consistent Correspondence Mining (CCM). (a) Seeds on the source domain image, weighted with a spatial Gaussian. (b) (c) (d): Visualization of the distribution of matched regions, corresponding to: (b) with naive ICM, but without adversarial masking; (c) the full ICM module; (d) without CCM.

Thank you