Bidirectional Multi-Step Domain Generalization for Visible-Infrared Person Re-Identification

Mahdi Alehdagi[1], Pourya Shamsolmoali[2], Rafael M. O. Cruz[1] and Eric Granger[1]

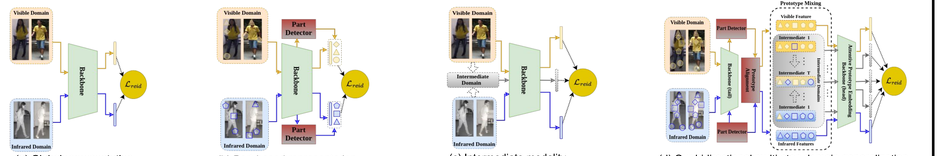
[1]LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada, [2]Dept. of Computer Science, University of York, UK

mahdi.alehdaghi.1@ens.etsmtl.ca, pshams55@gmail.com, {rafael.menelau-cruz, eric.granger}@etsmtl.ca

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Introduction

- V-I ReID is a challenging task due to significant discrepancies across V and I modalities
- State-of-the-art methods use global or part-based representations, but struggle with modality alignment
- Bridging modality gaps effectively requires leveraging intermediate feature spaces

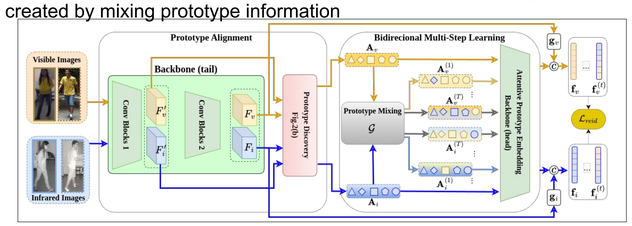


(a) Global representation (b) Part-based representation (c) Intermediate modality (d) Our bidirectional multi-step domain generalization

Proposed Bidirectional Multi-Step Domain Generalization

BMDG Training Architecture:

- **Prototype Alignment:** extracts body part prototype representations from V and I images
- **Bidirectional Multi-Step Learning:** extracts discriminant features using multiple intermediate domains created by mixing prototype information

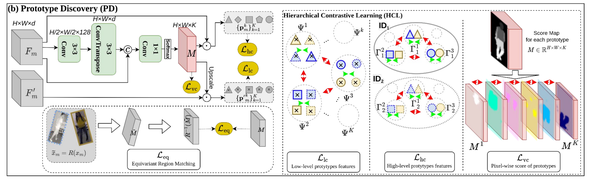


- **Goal**: maximize the similarity of extracted features of same person

$$\max\{S(\mathbf{f}_v^n, \mathbf{f}_v^{l,(t)}) + S(\mathbf{f}_i^n, \mathbf{f}_i^{l,(t)}) \cdot o\} \text{ where } o = 1 \text{ if } y_v^n = y_i^l \text{ else } o = -1 \text{ and } t \in \{1...T\} \text{ is step.}$$

Prototype Alignment

- **Prototype discovery (PD):** mines prototypes from spatial features
- **Hierarchical contrastive learning (HCL):** encourages the prototypes to focus on similar semantics for all individuals without losing ID-discriminative information.



- **Losses:** $\mathcal{L}_{\text{lc}} = -\sum_{k=1}^{K}\sum_{\hat{\mathbf{p}}'\in\Psi^k}\log\frac{e^{\mathbf{p}'^k\cdot\hat{\mathbf{p}}'/\tau}}{e^{\mathbf{p}'^k\cdot\hat{\mathbf{p}}'/\tau}+\sum_{q\neq k}e^{\mathbf{p}'^k\cdot\mathbf{p}'^q/\tau}}$, $\mathcal{L}_{\text{hc}} = -\sum_{y=1}^{C_y}\sum_{k=1}^{K}\sum_{\hat{\mathbf{p}}\in\Gamma_y^k}\log\frac{e^{\mathbf{p}^k\cdot\hat{\mathbf{p}}/\tau}}{e^{\mathbf{p}^k\cdot\hat{\mathbf{p}}/\tau}+\sum_{q\neq k}e^{\mathbf{p}^k\cdot\mathbf{p}^q/\tau}}$.

$$\mathcal{L}_{\text{eq}} = \sum_{k=1}^{K}\|M_m^k - R^{-1}(\tilde{M}_m^k)\| \qquad \mathcal{L}_{\text{vc}} = \sum_{k=1}^{K}\sum_{q=k+1}^{K}\sum_{u\in U}[\|M_m^k - M_m^q\|]_u,$$

Bidirectional Multi-step Learning

- **Mixing Prototypes**: creates intermediate feature representations for each modality by gradually increasing mixing rate

$$\mathbf{A}_m^{(t)} = \mathcal{G}(\mathbf{A}_m, \mathbf{A}_{\tilde{m}}, t) \text{ where } \mathbf{A}_m = [\mathbf{p}_m^1; \mathbf{p}_m^2; \dots; \mathbf{p}_m^K],$$

$$\mathcal{G}(\mathbf{A}_m, \mathbf{A}_{\tilde{m}}, t) = [\mathbf{p}_{r(m,\tilde{m},t)}^1, \dots, \mathbf{p}_{r(m,\tilde{m},t)}^K] \text{ where } r(m,\tilde{m},t) = \begin{cases} m & t/T \leq \mathcal{U}(0,1) \\ \tilde{m} & \text{else} \end{cases}$$

- **Joint feature representation:** $\mathbf{f}_m = [\mathcal{F}(\mathbf{A}_m); \mathbf{g}_m]$ and $\mathbf{f}_m^{(t)} = [\mathcal{F}(\mathbf{A}_m^{(t)}); \mathbf{g}_m]$
- **ReID Losses:**

$$\mathcal{L}_{\text{bcc}} = \mathcal{L}_{\text{cc}}(\mathbf{f}_v, \mathbf{f}_v^{(t)}) + \mathcal{L}_{\text{cc}}(\mathbf{f}_i, \mathbf{f}_i^{(t)})$$

$$\mathcal{L}_{\text{bce}} = \mathcal{L}_{\text{ce}}(\mathbf{f}_v) + \mathcal{L}_{\text{ce}}(\mathbf{f}_v^{(t)}) + \mathcal{L}_{\text{ce}}(\mathbf{f}_i) + \mathcal{L}_{\text{ce}}(\mathbf{f}_i^{(t)})$$

$$\mathcal{L}_{\text{re}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{bcc}}$$

Comparison with State-of-Art Methods

Table 1. Accuracy of the proposed BMDG and state-of-the-art methods on the SYSU-MM01 (single-shot setting) and RegDB datasets.

Family			SYSU-MM01				RegDB			
			All Search		Indoor Search		Visible → Infrared		Infrared → Visible	
	Method	Venue	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Global	AGW	TPAMI'20	47.50	47.65	54.17	62.97	70.05	50.19	70.49	65.90
	CAJ	ICCV'21	69.88	66.89	76.26	80.37	85.03	79.14	84.75	77.82
	RAPV-T	ESA'23	63.97	62.33	69.00	75.41	86.81	81.02	86.60	80.52
	G2DA	PR'23	63.94	60.73	71.06	76.01	-	-	-	-
Part-based	DDAG	ECCV'20	54.74	53.02	61.02	67.98	69.34	63.46	68.06	90.31
	SAAI	ICCV'23	75.90	77.03	82.20	80.01	91.07	91.45	92.09	92.01
	CAL	ICCV'23	74.66	71.73	79.69	83.68	94.51	88.67	93.64	87.61
	PartMix	CVPR'23	77.78	74.62	81.52	84.83	84.93	82.52	85.66	82.27
Intermediate	SMCL	ICCV'21	67.39	61.78	68.84	75.56	83.93	79.83	83.05	78.57
	RPIG	ECCVw'22	71.08	67.56	82.35	82.73	87.95	82.73	86.80	81.26
	G2DA	PR'23	63.94	60.73	71.06	76.01	-	-	-	-
	SEFL	CVPR'23	75.18	70.12	78.40	81.20	91.07	85.23	92.18	86.59
	BMDG	-	**78.08**	**78.22**	**83.59**	**86.35**	**94.76**	**92.21**	**94.56**	**93.07**

- *BMDG outperforms state-of-the-art methods in various cross-modal settings*
- *It improves robustness and accuracy by gradually reducing the modality gap*

Ablation Studies

Table 2. Impact on rank-1 and mAP accuracy using different settings for multi-step domain generalization.

Settings	Number of part prototypes (K)							
	4		5		6		7	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Single step	69.2	65.9	69.4	66.1	70.2	66.3	70.4	66.5
One (V → I)	71.0	66.5	71.5	67.1	72.3	67.5	71.6	66.8
One (I → V)	70.8	66.3	71.2	67.0	72.6	67.6	71.0	66.7
Bidirectional	**74.0**	**71.8**	**74.1**	**71.9**	**75.4**	**72.8**	**73.5**	**70.2**

Our bi-directional approach avoids forgetting source modalities since it is not biased to a specific modality

Table 3. Accuracy of part-based ReID methods with BMDG on the SYSU-MM01.

Method	R1 (%)	mAP (%)
DDAG	53.62	52.71
DDAG with BMDG	**55.36** (+1.74↑)	**54.05** (+1.34↑)
MPANet	66.24	62.89
MPANet with BMDG	**68.74** (+2.50↑)	**64.25** (+1.36↑)
SAAI	71.87	68.16
SAAI with BMDG	**73.69** (+1.82↑)	**70.08** (+1.92↑)

BMDG can be integrated into any part-based V-I ReID method to improve performance

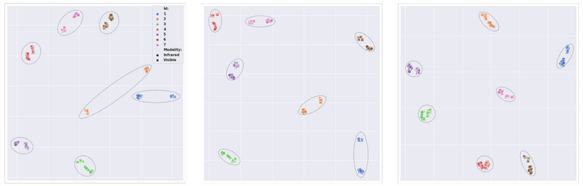
Table 4. mAP accuracy of BMDG for different values of K and T.

T	Number of part prototypes (K)					
	3	4	5	6	7	10
0	65.98	67.03	67.23	68.11	67.55	65.66
1	67.42	68.72	69.28	69.46	68.32	69.28
2	69.51	70.08	70.72	71.14	69.97	71.25
3	71.44	71.69	71.82	72.02	71.06	71.67
4	-	71.98	72.15	**72.86**	71.19	69.00
6	-	-	-	72.40	71.17	69.54
10	-	-	-	-	-	69.46

Accuracy grows with T for a number of prototypes K

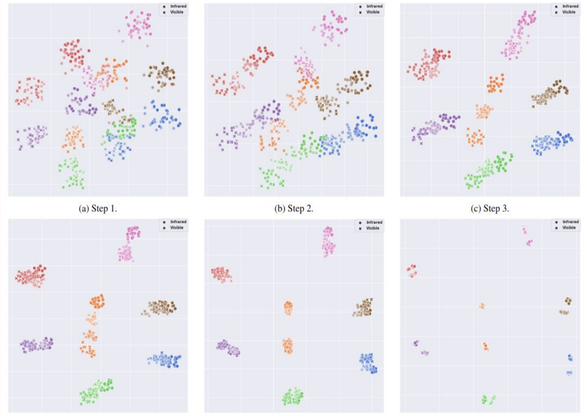
Visual Results (t-SNE Plots)

- Comparing BMDG against One-Step and Baseline approaches:



(a) Baseline. (b) One Step. (c) BMDG.

- Impact of using our prototype mixing for 6 intermediate steps (T):



(a) Step 1. (b) Step 2. (c) Step 3. (d) Step 4. (e) Step 5. (f) Step 6.

Conclusion

- We propose a BMDG training architecture, allowing to learn the discriminative and complementary set of prototypes from both modalities
- These prototypes create intermediate domains by mixing modalities, reducing the domain gap
- BMDG achieves state-of-the-art performance on cross-modal evaluation settings on several challenging V-I ReID datasets

Paper: https://arxiv.org/abs/2403.10782 **Code:** https://alehdaghi.github.io/BMDG