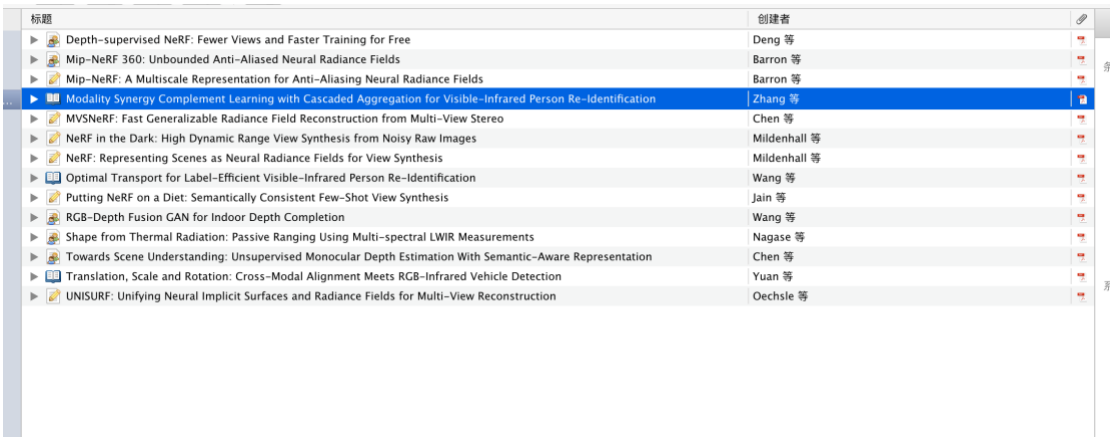
Research Notes

12.05

The first two weeks of paper reading were:



标题	创建者
Depth-supervised NeRF: Fewer Views and Faster Training for Free	Deng 等
Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields	Barron 等
Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields	Barron 等
Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification	Zhang 等
MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo	Chen 等
NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images	Mildenhall 等
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis	Mildenhall 等
Optimal Transport for Label-Efficient Visible-Infrared Person Re-Identification	Wang 等
Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis	Jain 等
RGB-Depth Fusion GAN for Indoor Depth Completion	Wang 等
Shape from Thermal Radiation: Passive Ranging Using Multi-spectral LWIR Measurements	Nagase 等
Towards Scene Understanding: Unsupervised Monocular Depth Estimation With Semantic-Aware Representation	Chen 等
Translation, Scale and Rotation: Cross-Modal Alignment Meets RGB-Infrared Vehicle Detection	Yuan 等
UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction	Oechsle 等

The first week were paper reading & video study of NeRF and the second week focused on pedestrian re-identification. The two papers related to pedestrian re-recognition are discussed and analysed below.

1. Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification

1.1. Research Questions

1) Because VI-ReID faces huge intra-class variations, how to better combine the advantages of Visible and Infrared?
2) Can the synergistic representations be enhanced by the complementary modules of the two modalities?
3) How to solve traditional approaches (hard sample mining and feature aggregation) neglects the comprehensive distribution of all instances?

1.2. Motivation

1) Infrared modal is smoother and represents relatively stable about the same identity and are comparatively immune to noise. But it lacks texture details. The Visible modal contains more discriminative features. Therefore, after the intermediate modality is generated, it needs to be enhanced through the complement module.
2) GAN-based methods usually suffer from computational complexity and noise introduction.
3) If the pursuit of modality-invariance may lead to overlook feature properties of semantic diversity, as well as loss of identity discrimination.

1.3. Framework

Overview

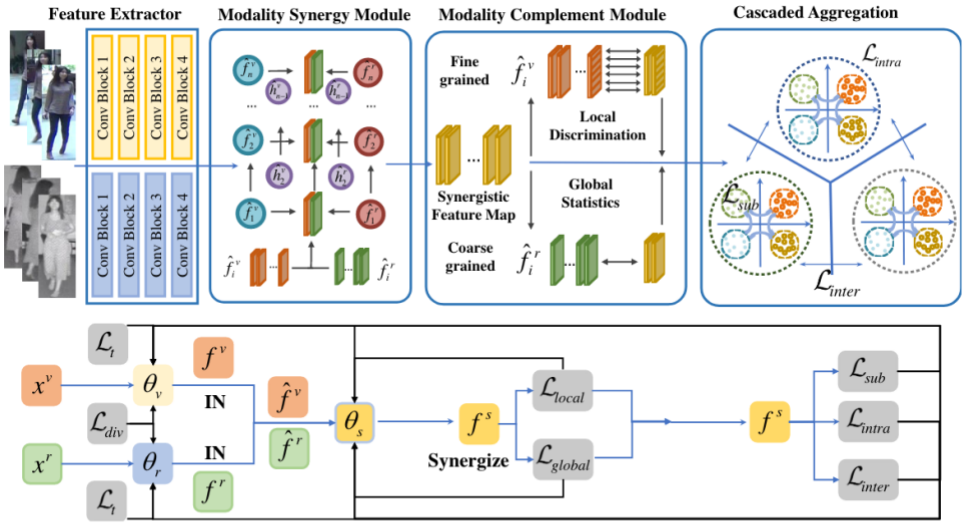


Fig. 3: **Illustration of our MSCLNet.** The images of visible and infrared modalities are fed into convolution blocks for visible and infrared representations. We synergize the single-modal features and complement synergistic features. Then, we design the Cascaded Aggregation strategy to fine-grained and progressively enhance feature embeddings.

First in the Feature Extractor stage, feature representations f^v and f^r are extracted from visible and infrared images. In Modality Synergy Module, MSCLNet constructs synergistic representations f^s by constraining the diversity of the feature distributions between the two modalities. Next, in the Modality Complement Module, the fine-grained discriminative semantics in the visible modality and the corase-grained global semantics in the infrared modality are provided as a complement to the Synergistic Feature Map. In the Cascaded Aggregation section, feature embeddings of the same class are aggregated in three aspects.

Not very understandable:

> the diverse semantics of the two modalities. Given a pair of visible and infrared images $x_i^v \in \mathcal{V}$, $x_i^r \in \mathcal{R}$, the dual-stream network extracts their features f_i^v and f_i^r. With the prerequisite of precise pedestrian re-identification, we concentrate on acquiring the semantic diversity to the largest extent. Features f_i^v and f_i^r are normalized by the following operations.

$$\hat{f}_i^v = \frac{f_i^v - \mathrm{E}\left[f_i^v\right]}{\sqrt{\mathrm{Var}\left[f_i^v\right] + \epsilon^v}} \times \gamma + \beta, \mathrm{E}\left[f_i^v\right] = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} f_{itlm}, \tag{1}$$

> Where $\mathrm{Var}\left[f_i^v\right] = \frac{1}{HW} \sum_{l=1}^{W} \sum_{m=1}^{H} \left(f_{itlm} - \mathrm{E}\left[f_i^v\right]\right)^2$ are calculated per-dimension separately for each instance in a mini-batch. Let $\mathcal{S}(\cdot)$ indicate the Modality Syn-

The mathematical calculations in this part are not understood, and some elements are not explained in the text. (I will explore again in the follow-up code re-implement).

1.4. Experiment

During the training process, a mini-batch is 64 images (8 identities, each identity randomly selects four visible and infrared). Each image will be re-scaled to 288 X 144 in advance.

During training, the learning rate decreases as the number of epochs increases.

and modality synergy module θ_s with SGD optimizer. We set the initial learning rate $\eta = 0.1$, the momentum parameter $p = 0.9$. The learning rate is changed as $\eta = 0.05$ at 21-50 epoch, $\eta = 0.01$ at 51-100 epoch, and $\eta = 0.001$ at 101-200 epoch. The hyper-parameters $\lambda_{div}, \lambda_t, \lambda_{local}, \lambda_{global}$ are set to 0.5, 1.25, 0.8, and

In this way, the model can quickly converge during the training process, and when the number of epochs becomes larger, the learning rate can be reduced to continuously approach the extreme point.

In the ablation study section, the author analyzed the effectiveness of MS, MC, CA on SYSU-MM01 dataset in the all-search mode. We can find that the three modules have significantly improved performance. And the effect of CA and MC modules on performance improvement is more obvious.

In the Comparison with State-of-the-Art Methods section, we found that MSCLNet has achieved the best performance in the SYSU-MM01 dataset, and also achieved satisfied performance in the RegDB dataset.

1.5. The reason of acceptance

1) The question raised is logical and meticulous, and then the three modules proposed are closely related to the meaning of the existence of each module.
2) MSCLNet has achieved the best performance, which can be reflected in the performance comparison part.
3) Modality Complement module and Cascaded Aggregation strategy are very novel, and it can be seen from the ablation study that two modules greatly improve the performance of the model。
4) In Ablation study, visualization analysis, parameters analysis and comparison with other models, author has been done a lot of experiments, which can reflect that the author considered very comprehensively in the experiment。

1.6. Open Questions

First of all, the author did not intuitively tell me what the noise caused by visible modality is, and how to judge whether it is noise or semantic feature. So can we redefine noise?

The author has done a good job of modal complement, but has the modal been fully complement? Can we give full play to the advantages of each modality?

We can clearly see that the features belonging to each identity can be aggregated in the Cascaded Aggregation stage, but when there is a wrong prediction, the author did not consider what to do with the model after the wrong aggregation. Can we optimize it?

1.7. How to relate to our own problems?

When we do Visible-Infrared Person Re-Identification, can we first create a modal noise (and semantic) processing module, and we preprocess the image according to the noise classification (for example, we can set up a detection box to contain the identified Pedestrians are continuously shrunk while deleting redundant background content). After that, we can generate intermediate modality, and then we can also set up a section to see if we can fully utilize the complementary modality feature. We can

also optimize Cascaded Aggregation, judge the wrong aggregation to reduce the wrong prediction among identities.

2. Optimal Transport for Label-Efficient Visible-Infrared Person Re-Identification

2.1. Research Questions

1) How to re-identify and find a person in another modal through a visible/infrared picture?
2) Can we learn a cross-modality model only with one modal supervision or even without supervision?
3) How to take advantage of visible well-annotated without infrared annotations to train a cross-modality model?
4) ow to match the generated pseudo label with the infrared image, and how to reduce the negative effects caused by the wrong pseudo label?

2.2. Motivation

In the VI-ReID dataset, RGB ReID datasets have rich annotation information and annotating infrared images is expensive due to the lack of color information. So unsupervised learning and unsupervised domain adaptation methods are needed to solve the problem of less annotation.

The pseudo label generated by previous methods (such as clustering) may be affected by the color of the clothes and the shape of the human body, and the author hopes to optimize the prediction through a specific module. The current unsupervised framework is often sub-optimal, because the rich annotated visible data is not utilized and the heterogeneous pseudo labels as also not well aligned.

2.3. Framework

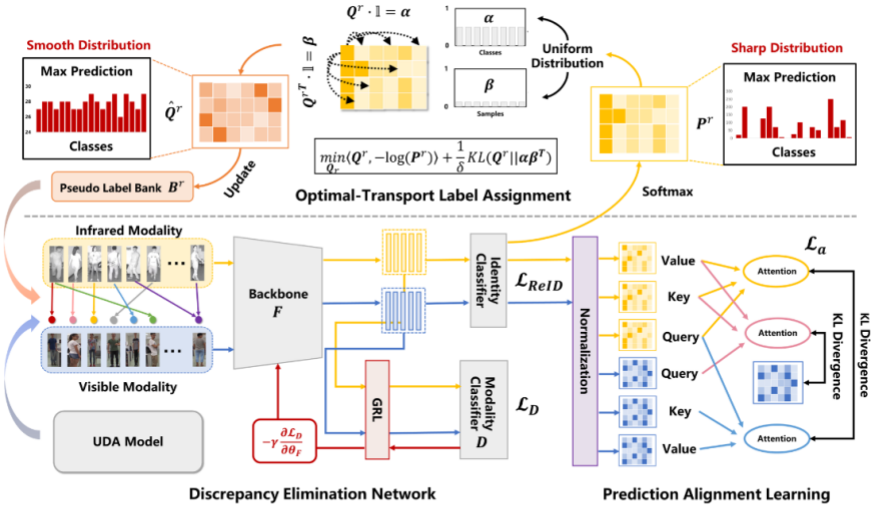


Fig. 2. The pipeline of our framework. Left Bottom: we first use an UDA model to generate the pseudo labels for visible images. Then we take both visible and infrared data into Discrepancy Elimination Network. Upper: The identity prediction of infrared images are sent into Optimal-Transport Label Assignment to assign labels. Right Bottom: The identity predictions are also forwarded into Prediction Alignment Learning to align the mixed predictions, so as to reduce the effects from incorrect pseudo labels.

The author adopts the SOTA clustering based method SpCL to generate the pseudo label, the source domain is RGB dataset, and the target domain is visible-infrared dataset. Then the author roughly divides the overall framework into three modules, namely Discrepancy Elimination Network (DEN), Optimal-Transport Label Assignment module (OTLA), and Prediction Alignment Learning module (PAL). DEN is mainly responsible for feature extractor to reduce the modality gap. OTLA is a relatively innovative idea module, which mainly matches infrared images and visible pseudo classes through class-wised uniform distribution and sample-wised uniform distribution. PAL uses a batch-level self-attention technique to emphasize the truly-related samples while neglecting the incorrect ones, thereby reducing negative effects brought by the inaccurate pseudo labels.

In the implementation stage, the author found through experiments that in the DEN module, the gradient reversal layer (GRL) would degrade the performance in the fully-supervised VI-ReID, but is effective in semi-supervised/unsupervised case. Then this can attract our attention: Is there any other layer or method that does not work well in fully-supervised VI-ReID and will have unexpected gains in semi-supervised/unsupervised.

In the future, I will also read Optimal-Transport related papers, because I think they apply very simple principles, and there should be points for improvement and optimization, which can be further deepened.

2.4. Experiment

This article applies two datasets: SYSU-MM01 and RegDB. In terms of evaluation metrics, cumulative match characteristic (CMC) and mean average precision (mAP) are selected. In the training phase, the author combined the warm-up strategy to let the learning rate decay 10 times at the 20-th and the 50-th epoch. And the input images are randomly flipped and erased with 50% probability, while visible images are extra randomly grayscale with 50% probability (what is the meaning of this processing, this is still not quite understood, and I will figure it out later).

In comparison with other methods, we can find that they have not achieved relatively good performance at present, and we can optimize them on the basis of them to obtain performance improvements.

Settings			SYSU-MM01				RegDB			
			All Search		Indoor Search		Visible2Thermal		Thermal2Visible	
Type	Method	Venue	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
UDA-ReID	SSG‡[9]	ICCV'19	2.3	12.7	-	-	2.2	2.9	-	-
	ECN‡[48]	CVPR'19	8.1	5.0	-	-	1.9	3.2	-	-
	D-MMD†[21]	ECCV'20	12.5	10.4	19.0	15.4	2.2	3.7	2.0	3.6
	MMT†[11]	ICLR'20	13.9	8.4	21.0	15.3	5.3	7.1	11.0	12.1
	SpCL(UDA)†[12]	NIPS'20	15.1	6.5	19.5	12.1	3.3	4.3	8.4	9.5
	GLT†[44]	CVPR'21	7.7	9.5	12.1	18.0	2.9	4.5	6.3	7.6
USL-ReID	BUC†[17]	AAAI'19	8.2	3.2	12.5	6.0	4.7	4.5	8.8	6.0
	SpCL(USL)†[12]	NIPS'20	18.7	11.4	27.1	20.9	20.6	17.3	19.0	16.6
	MetaCam†[36]	CVPR'21	14.7	9.3	23.9	17.1	23.1	17.5	20.9	16.5
	HCD†[46]	ICCV'21	18.0	17.9	24.4	28.8	10.8	12.3	12.4	13.7
SVI-ReID	JSIA-ReID[29]	AAAI'20	38.1	36.9	43.8	52.9	48.5	49.3	48.1	48.9
	Hi-CMD[3]	CVPR'20	34.9	35.9	-	-	70.9	66.0	-	-
	AGW[39]	TPAMI'21	47.5	47.7	54.17	63.0	70.1	66.4	70.5	65.9
	NFS[2]	CVPR'21	56.9	55.5	62.8	69.8	80.5	72.1	78.0	69.8
	LbA[24]	ICCV'21	55.4	54.1	58.5	66.3	74.2	67.6	72.4	65.5
	CAJL[37]	ICCV'21	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8
	MPANet[35]	CVPR'21	70.6	68.2	76.7	81.0	83.7	80.9	82.8	80.7
USVI-ReID	H2H[16]	TIP'21	25.5	25.2	-	-	14.1	12.3	13.9	12.7
	Ours	-	**29.9**	**27.1**	**29.8**	**38.8**	**32.9**	**29.7**	**32.1**	**28.6**
SSVI-ReID	Ours	-	**48.2**	**43.9**	**47.4**	**56.8**	**49.9**	**41.8**	**49.6**	**42.8**

In the ablation study, we can find that the OTLA module has a significant effect on performance improvement.

2.5. The reason of acceptance

1) The OTLA module they set up is very novel, and it is very effective in matching infrared and visible data, and the performance improvement is also obvious.
2) The performance of the model is higher than the performance of pedestrian re-identification using the UDA method before, which also shows the success of the application of the UDA method this time.
3) The author considers the differences between modalities, and uses a batch-level self-attention technique to reduce the negative effects brought by the wrong pseudo label. This optimization of the generated pseudo label shows that the author's thinking is very logical.

2.6. Open Questions

The first thing to think about is why the performance of their current model is not as good as SVI-ReID's. This can be revisited in NFS, LbA, CAJL and MPANet (which is most worth revisiting, this is a better model performance in some aspects than the previous article), and then there may be better insights at the methodological level.

They only consider label- efficient, but not modal complementarity, and could also be optimised in the algorithm for matching pseudo labels, which would reduce the loss from incorrect matches.

2.7. How to relate to our own problems?

First, by using Optimal-Transport to match pseudo labels and infrared images, we can then optimize them to improve label efficiency, and then by reading papers on SVI-ReID such as MPANet, we can analyze why the performance is better than that of UDA and where we can optimize UDA. Some of the layers or methods that do not work well in fully-supervised VI-ReID may have unexpected gains in semi-

supervised/unsupervised.

Apart from that, I am also learning about NeRF and 3D geometry &computer grapichs. From the discussion every Monday, I have a clear understanding of some areas that are not very clear, such as what is shear and 3D Transformations in Homogeneous Coordinates (this is the most rewarding). Afterwards, I would also record my questions while watching the videos and then raise them in the discussion. I feel that the greatest benefit of the discussion every Monday is that the seniors share some of their ideas while watching the videos, so that the discussion among innovative minds can be the source of many ideas for the top articles!