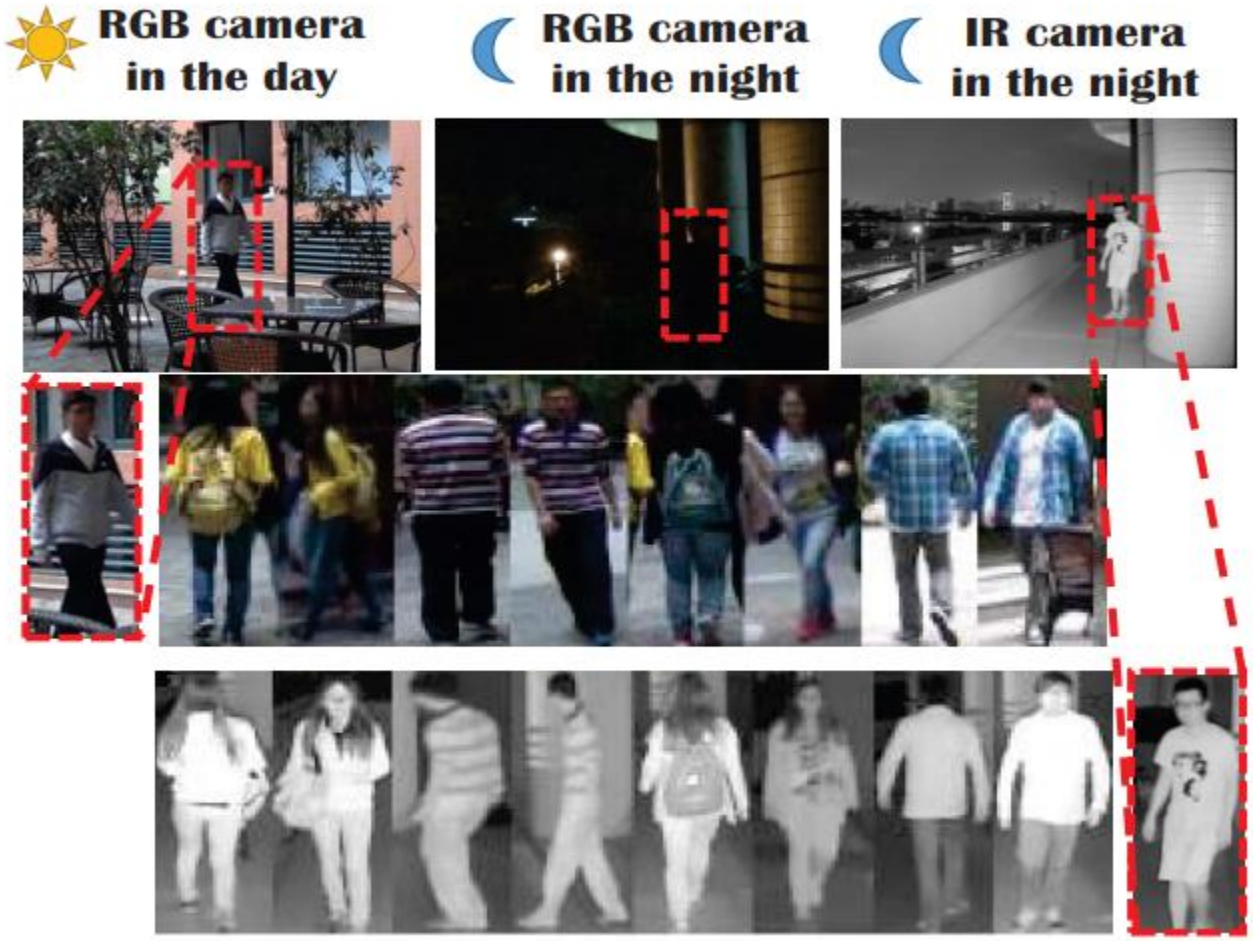
ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

RGB-Infrared Person Re-identification via Image Modality Conversion

Huangpeng Dai, Qing Xie, Yanchun Ma, Yongjian Liu and Shengwu Xiong

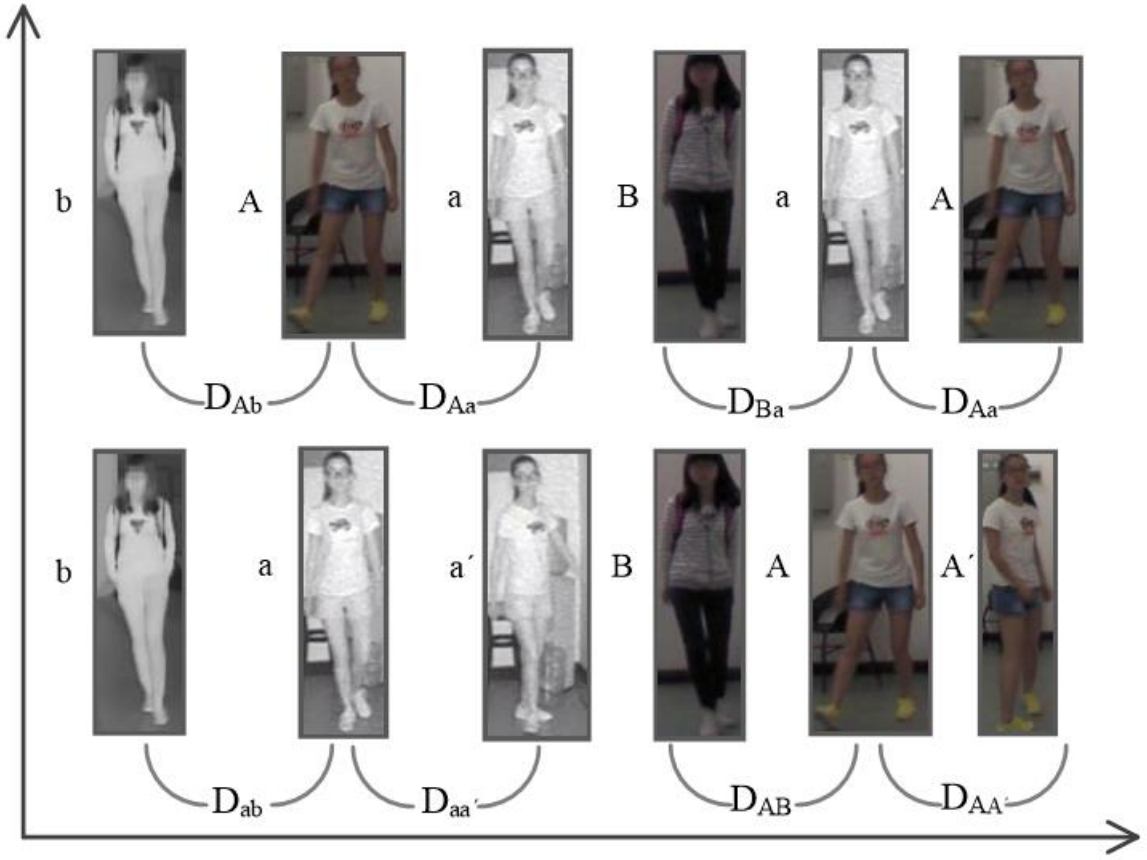
School of Computer Science and Technology, Wuhan University of Technology

Cross-Modality Person Re-Identification



- 1) large cross-modality variations caused by the different cross-camera views and hard negatives D(A, a).
- 2) large intra-modality variations caused by different human poses and viewpoints D(a, b).

it is difficult for even human to recognize whether the persons in the images are the same by only the silhouette of the persons without the help of color information.



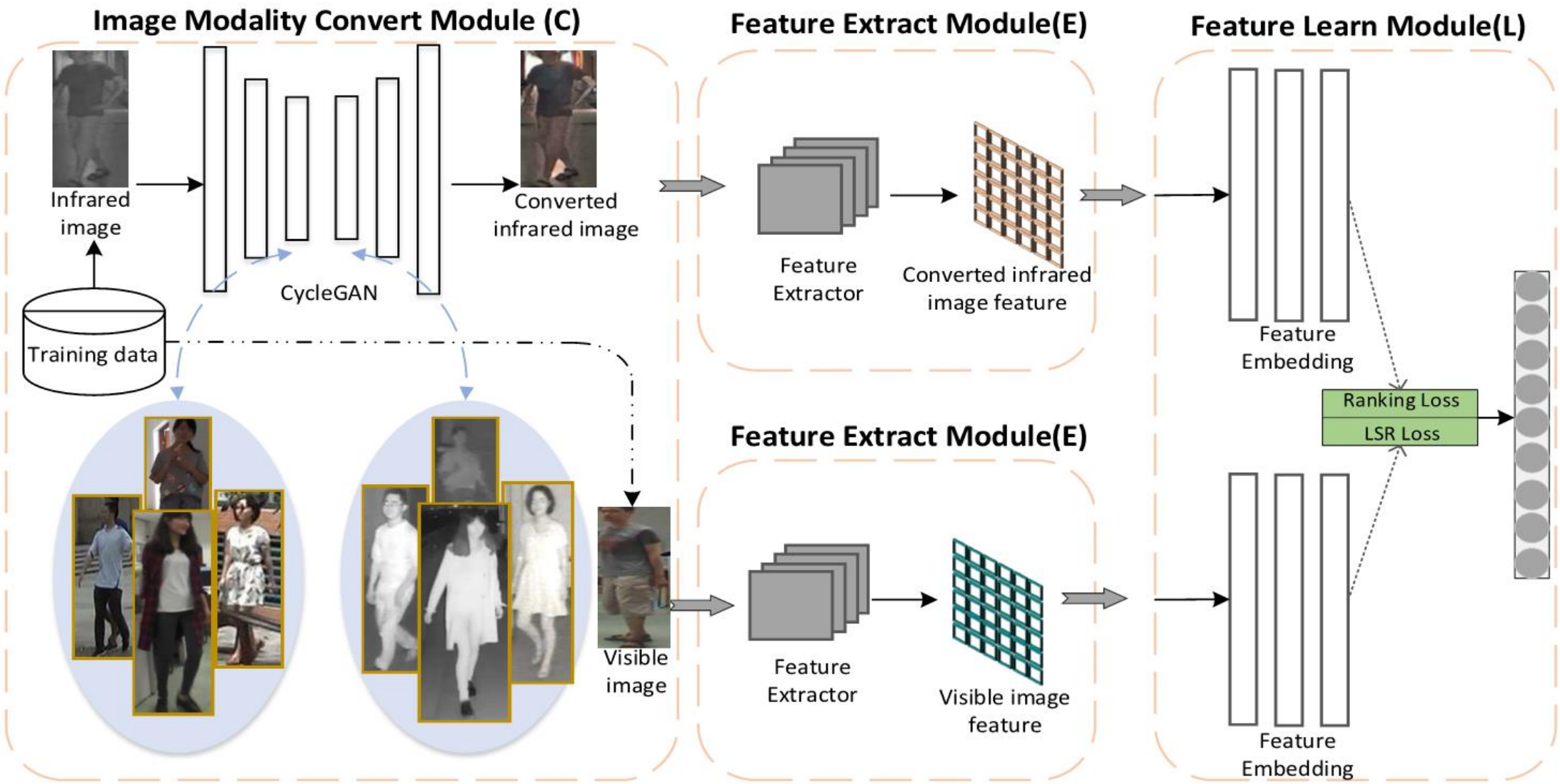
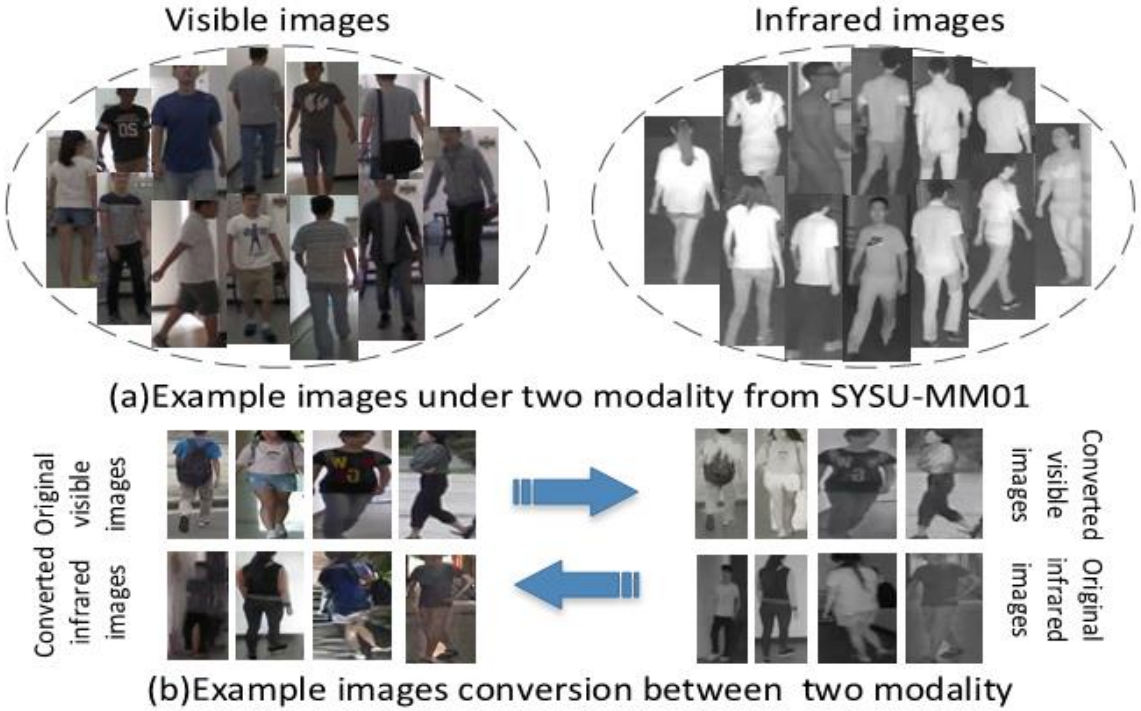


Image Modality Convert Module (C)
Feature Extract Module(E)
Feature Learn Module(L)
Infrared image
CycleGAN
Converted infrared image
Training data
Visible image
Feature Extractor
Converted infrared image feature
Feature Embedding
Ranking Loss
LSR Loss
Feature Extract Module(E)
Feature Extractor
Visible image feature
Feature Embedding

With the learned CycleGAN model, for a modal image captured by a certain camera, it can generate a corresponding image.



(a)Example images from SYSU-MM01. (b)Example images modality conversion between two modality using our proposed method.

- cross-class distances constrained top-ranking loss

$$\ell_{cross} = \sum_{\forall y_i = y_j} max[D(v_i, t_j) - \min_{\forall y_i \neq y_k} D(v_i, t_k) + \rho_1, 0] + \sum_{\forall y_i = y_j} max[D(t_i, v_j) - \min_{\forall y_i \neq y_k} D(t_i, v_k) + \rho_1, 0]$$

- intra-modality constrained loss

$$\ell_{intra} = \sum max[\rho_2 - D(t_j, t_k), 0] + \sum max[\rho_2 - D(v_j, v_k), 0]$$

- LSR loss

$$q_{LSR}(c) = \begin{cases} 1 - \varepsilon + \frac{\epsilon}{C} & c = y \\ \frac{\epsilon}{C} & c \neq y \end{cases} \qquad q(c) = \begin{cases} 1 & c = y \\ 0 & c \neq y \end{cases}$$

- Dataset
  - RegDB is an image dataset consisting of 412 persons while people are moving naturally without any instruction.
  - SYSU-MM01 is a large-scale dataset collected by 6 cameras, including four RGB cameras and IR ones. It contains in total 491 identities and each identity is captured by at least two different cameras.
- Implementation details
  - All images are firstly resized to 256*256, and then randomly cropped to 227*227. Dropout rate and initial learning rate is set as 0.5 and 0.001, respectively.
  - Our model is implemented with GTX 1080Ti, Intel i7 and 256G memory.

Datasets	RegDB				SYSU-MM01			
Methods	r = 1	r = 10	r = 20	mAP	r = 1	r = 10	r = 20	mAP
LOMO [26]	0.85	2.47	4.10	2.28	1.75	14.14	26.63	3.48
MLBP [27]	2.02	7.33	10.90	6.77	2.12	16.23	28.32	3.86
HOG [28]	13.49	33.22	43.66	10.31	2.76	18.25	31.91	4.24
GSM [29]	17.28	34.47	45.26	15.06	5.29	33.71	52.95	8.00
One-stream [3]	13.11	32.98	42.51	14.02	12.04	49.68	66.74	13.67
Two-stream [3]	12.43	30.36	40.96	13.42	11.65	47.99	65.50	12.85
Zero-Padding [3]	17.75	34.21	44.35	18.90	14.80	54.12	71.33	15.95
TONE [2]	16.87	34.03	44.10	14.92	12.52	50.72	68.60	14.42
HCML [2]	24.44	47.53	56.78	20.80	14.32	53.16	69.17	16.16
D^2RL [5]	43.4	66.1	76.3	44.1	28.9	70.6	82.4	29.2
BaseLine [4]	33.47	58.42	67.52	31.83	17.01	55.43	71.96	19.66
CE^2L	47.50	72.17	79.87	44.21	29.52	69.83	82.49	28.43

- Bridge the gap between infrared and visible images by converting image modality.

- For tacking the problem of noise from converting image modality process, we further apply label smoothing regularization to softly distributed data sample labels.

Thanks for your watching!