RGB Infrared Cross Modal Person Re-identification

Presented by

Chaitra Jambigi

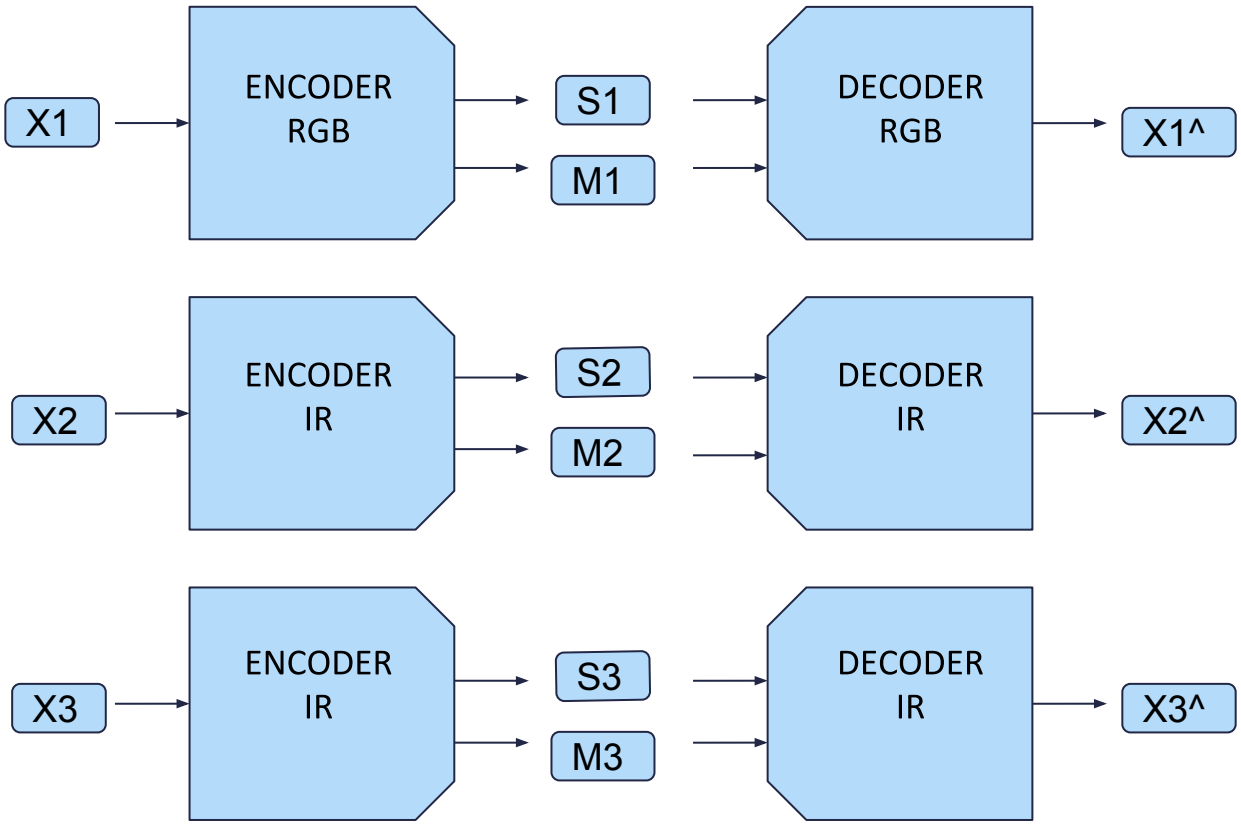
25-05-2020

Overview of Problem

1. Matching Query image with Gallery set
2. Query image - Infrared (7.5-13.5 micrometre)
3. Gallery image - Rgb
4. Used for surveillance
5. Intra modality variations: Pose, camera viewpoint
5. Cross modality variations

Top row - RGB Images
Bottom row- Corresponding IR images

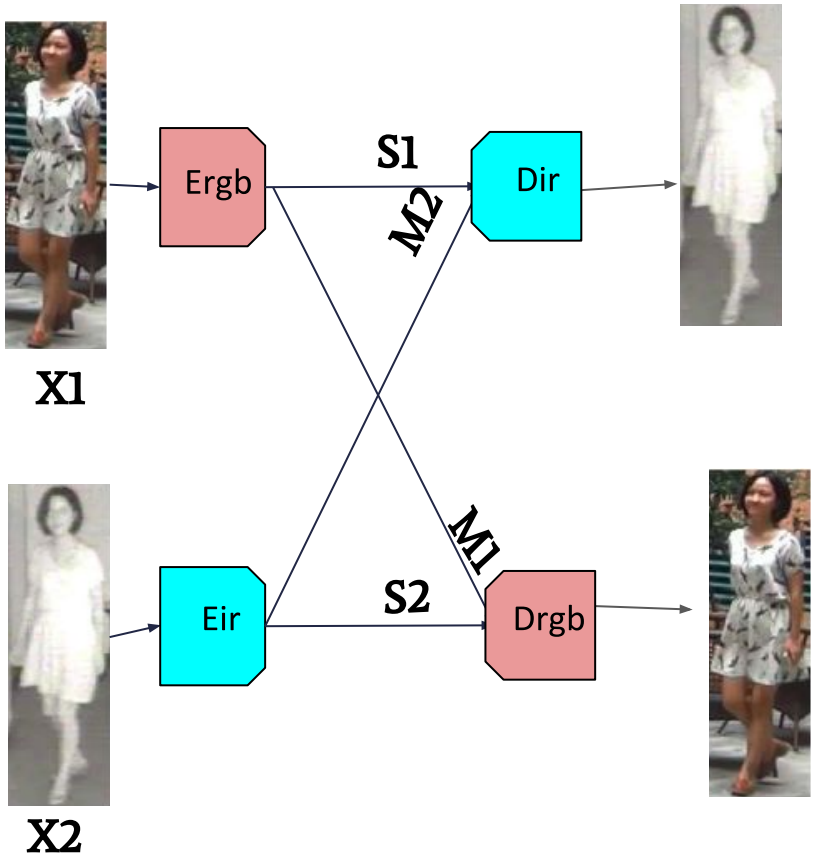
Approach 0 - Architecture



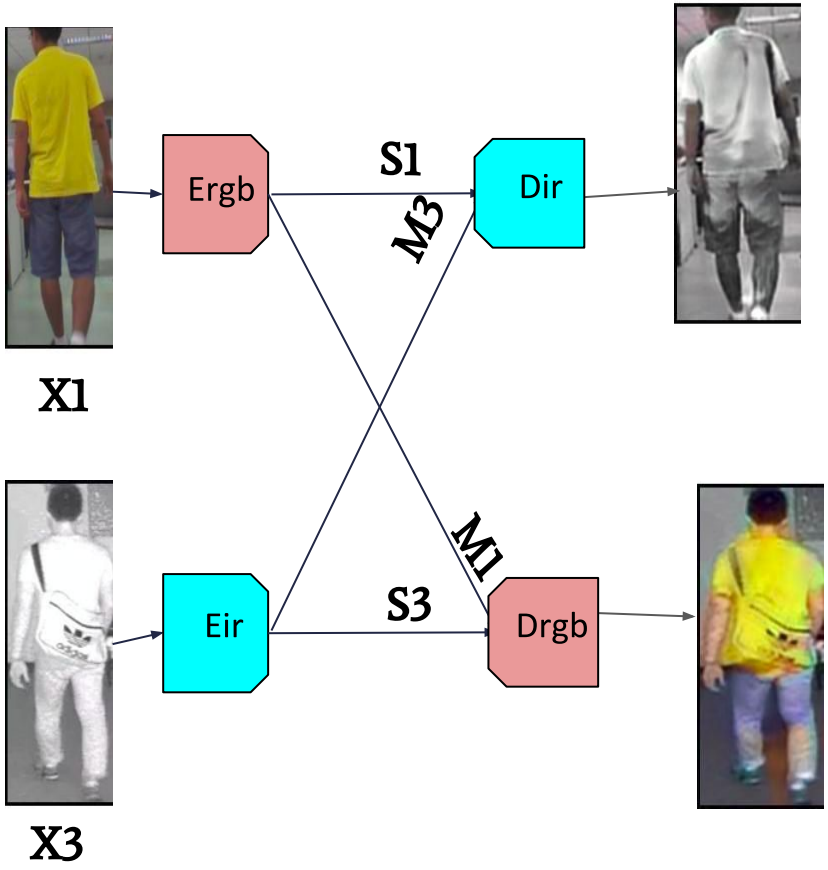
X1- Visible image of ID 1
X2- Infrared Image having same ID as X1
X3- Infrared image having different ID

S- Identity specific component
M- Identity independent component

Approach 0 - Working

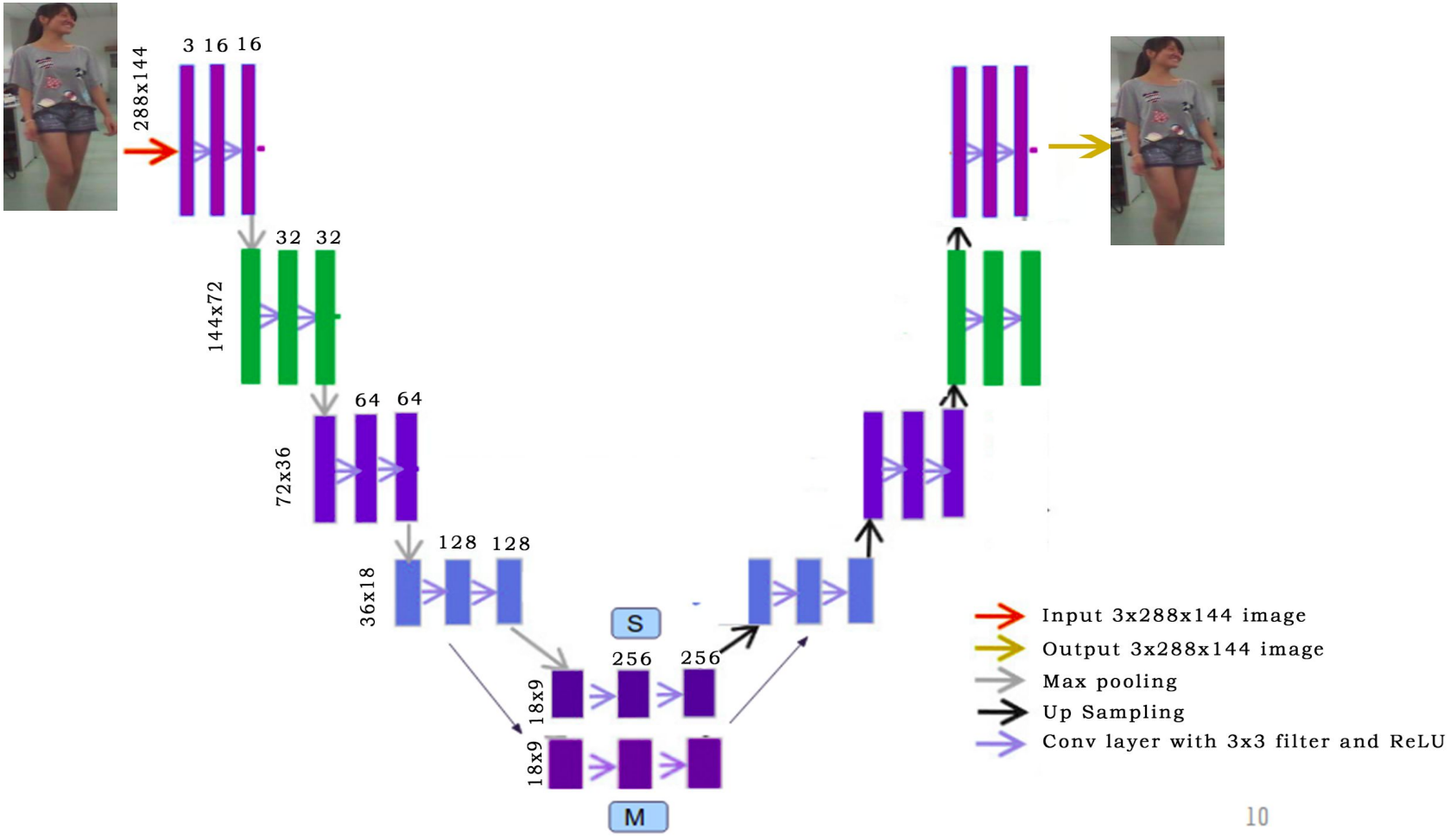


SAME ID GENERATION



DIFFERENT ID GENERATION

Approach 0 - Modified UNet



Approach 0 - Losses

$X_1 - RGB\ image\ of\ Id\ 1\ , X_2 - IR\ image\ having\ same\ Id\ as\ X_1\ , X_3 - IR\ image\ having\ different\ id$

Simple Reconstruction loss:

$$L_{recon} = \left\| X_1 - \widehat{X_1} \right\|^2 + \left\| X_2 - \widehat{X_2} \right\|^2 + \left\| X_3 - \widehat{X_3} \right\|^2$$

Cross reconstruction loss:

$$S_1, M_1 = E_{rgb}(X_1), \quad S_2, M_2 = E_{ir}(X_2), \quad X_1^{'} = D_{rgb}(S_2, M_1), \quad X_2{'} = D_{ir}(S_1, M_2)$$

$$L_{cross} = \left\| X_1 - X_1^{'} \right\|^2 + \left\| X_2 - X_2^{'} \right\|^2$$

Triplet loss:

$$L_{\text{triplet}}\ (X_1, X_2, X_3) = \max\left(0, m + \|f\,(X_1) - f\,(X_2)\|_2^2 - \|f\,(X_1) - f\,(X_3)\|_2^2\right)$$

$$Total\ Loss = L_{recon} + L_{cross} + L_{cycle} + L_{triplet}$$

REMOVE Cyclic loss from equation

Dataset - SYSU-MM01

- Training set - 296 id, Testing - 96 id, Validation - 99 id
- Training RGB images = 22,258, Training IR images = 11909
- Testing Query images = 3803, Testing Gallery = 301

- Evaluation criteria:
- All Search Single shot
- All Search Multi shot
- Indoor Search Single shot
- Indoor Search Multi shot

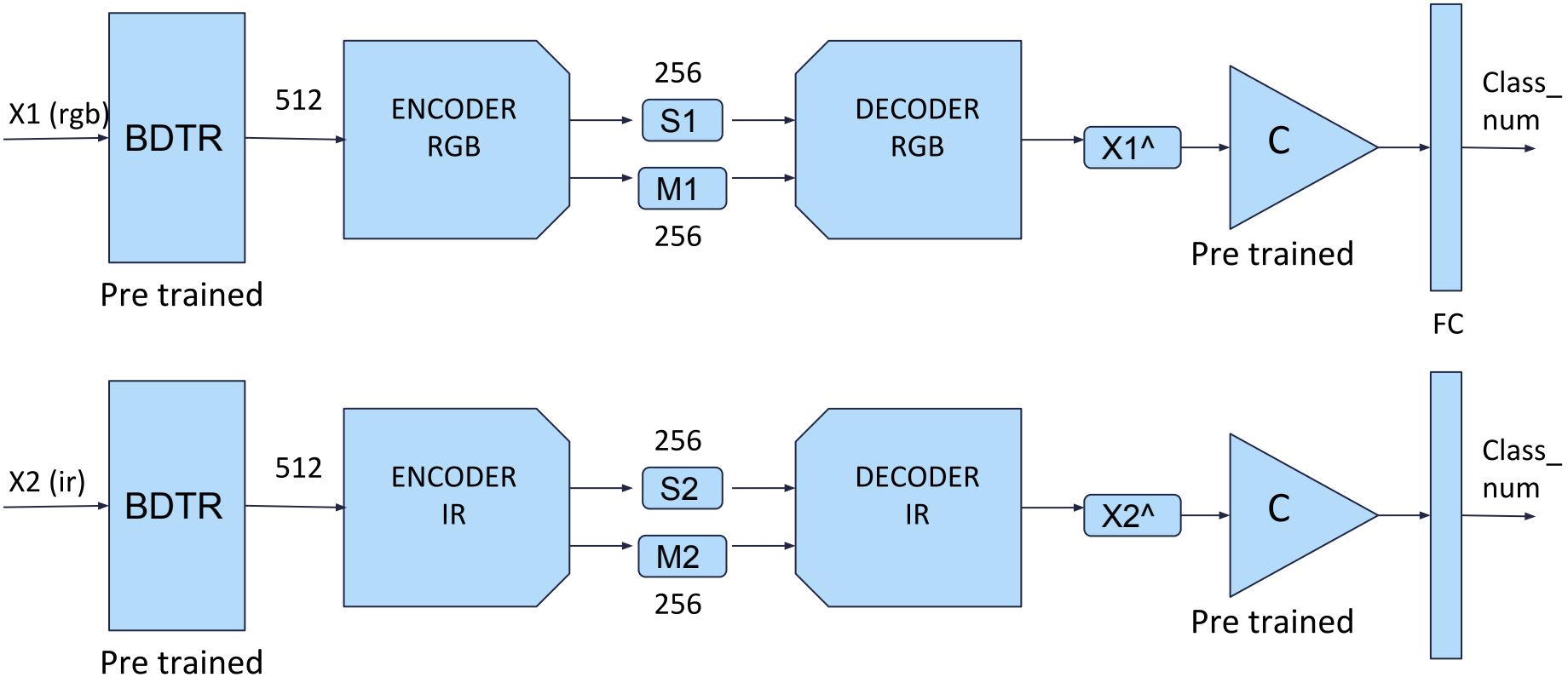
Approach 0 -Results - mAP

Method	All + Single shot	All + Multi shot	Indoor + Single shot	Indoor + Multi shot
Deep zero padding [1]	15.95%	10.89%	26.92%	18.64%
BDTR [2]	19.66%	-	-	-
SDL [3]	**29.01%**	**-**	**39.56%**	**-**
Approach 0	11.03%	4.6%	19.56%	8.8%

[1]: RGB-Infrared Cross-Modality Person Re-Identification
[2] : Visible thermal person re-identification via dual-constrained top-ranking. In IJCAI, 2018
[3] : SDL: Spectrum-Disentangled Representation Learning for Visible-Infrared Person Re-identification

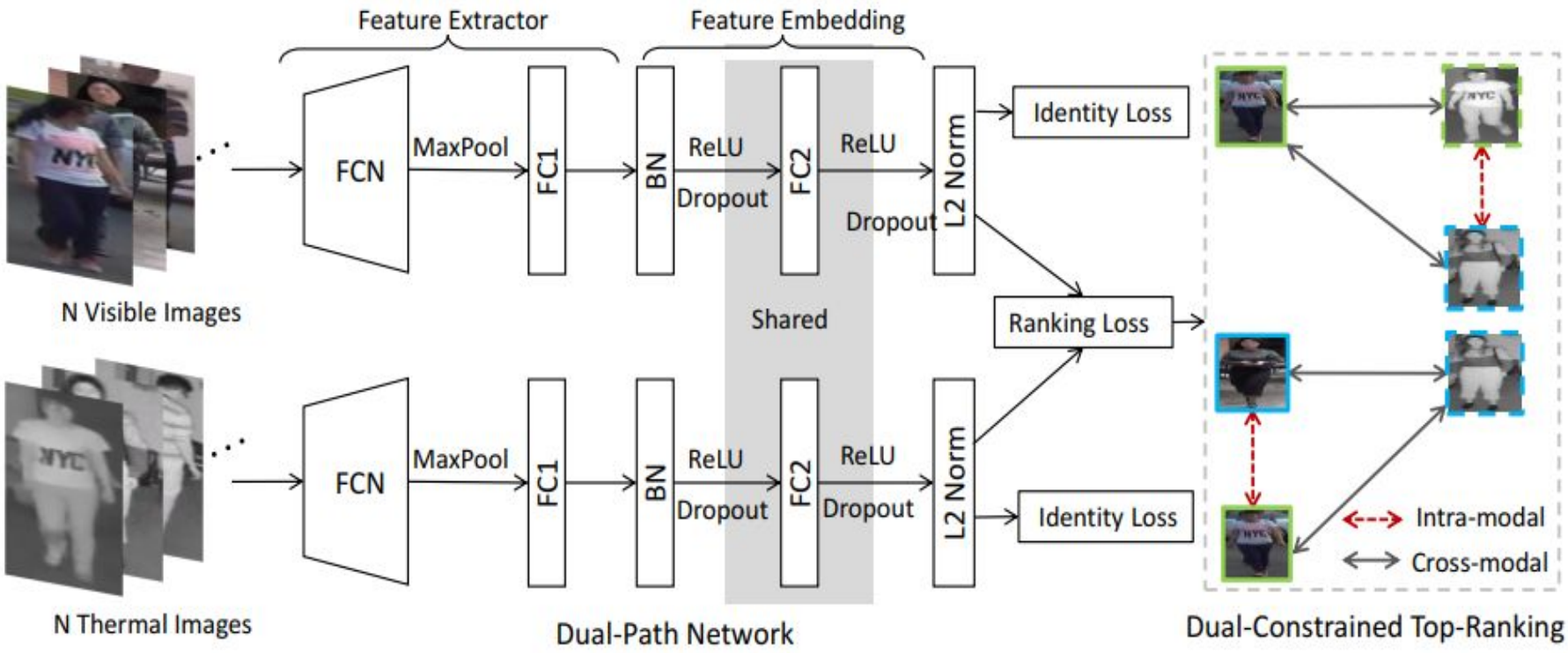
Approach 1

Using Low dimensional embedding space from pretrained BDTR Network



BDTR - Visible thermal person re-identification via dual-constrained top-ranking. In IJCAI, 2018

BDTR - Dual constrained top ranking



Approach 1 - Network

ENCODER-S	ENCODER-M	DECODER	CLASSIFIER
Input- 512	Input- 512	Input - S:256, M:256	Input- 512
1024, ReLU	1024, ReLU	512 ReLU : 512, ReLU	512, ReLU
2048, ReLU	2048, ReLU	2048 ReLU: 2048 ReLU	512, NUM_CLASS (395)
512, ReLU	512, ReLU	512 ReLU: 512 ReLU	
256. ReLU	256, ReLU	Output (S+M) - 512	

Table 5: Network Architecture for training in low dimension

Approach 1 - Loss

$$m_{X1}, s_{X1} = E_{rgb}(X_1), \quad m_{X_2}, s_{X_2} = E_{ir}(x_2), \quad m_{X_3}, s_{X_3} = E_{ir}(X_3),$$

$$L_{simple} = \mathbb{E}\left\|X_1 - D_{rgb}(m_{X_1}, s_{X_1})\right\|_1 + \mathbb{E}\left\|X_2 - D_{ir}(m_{X_2}, s_{X_2})\right\|_1 + \mathbb{E}\left\|X_3 - D_{ir}(m_{X_3}, s_{X_3})\right\|_1$$

$$L_{cross} = \mathbb{E}\left\|X_1 - D_{rgb}(m_{X_1}, s_{X_2})\right\|_1 + \mathbb{E}\left\|X_2 - D_{ir}(m_{X_2}, s_{X_1})\right\|_1$$

$$X_1^3 = D_{rgb}(m_{X_1}, s_{X_3}) \; , X_3^1 = D_{ir}(m_{X_3}, s_{X_1})$$

$$L_{dis} = \mathbb{E}\left[-log(p(y_3|X_1^3))\right] + \mathbb{E}\left[-log(p(y_1|X_3^1))\right]$$

where $p(y_3|X_1^3)$ is the probability that X_1^3 belongs to identity id y_3 and similarly for other term.

$$L_{triplet} = \max(0, m + \left\|s_{X_1} - s_{X_2}\right\|_1 - \left\|s_{X_1} - s_{X_3}\right\|_1)$$

$$L_{total} = L_{simple} + L_{cross} + L_{dis} + L_{triplet}$$

Approach 1 - Results

Results are reported on Indoor Search Single shot

Methods	Indoor Search	
	r1	mAP
$L_{simple} + L_{dis}$	1.33	9.07
$L_{simple} + L_{dis} + L_{cross}$	8.44	20.43
$L_{simple} + L_{dis} + L_{cross} + L_{triplet}$ (random triplet)	14.25	28.85
$L_{simple} + L_{dis} + L_{cross} + L_{triplet}(hardtriplet)$	**18.20**	**34.56**

Table 2: Results on SYSU-MM01 dataset with training on extracted embedding

Observations

- First, difficult to visualise the disentanglement performance since we have only embeddings.
- Second, separation in S and M will be beneficial only if input embedding has separable information.
- Third, mAP at output can be better tha BDTR only if input had redundancy. So our mAP get bounded by mAP of BDTR.

Approach 2 - DG-Net as Baseline

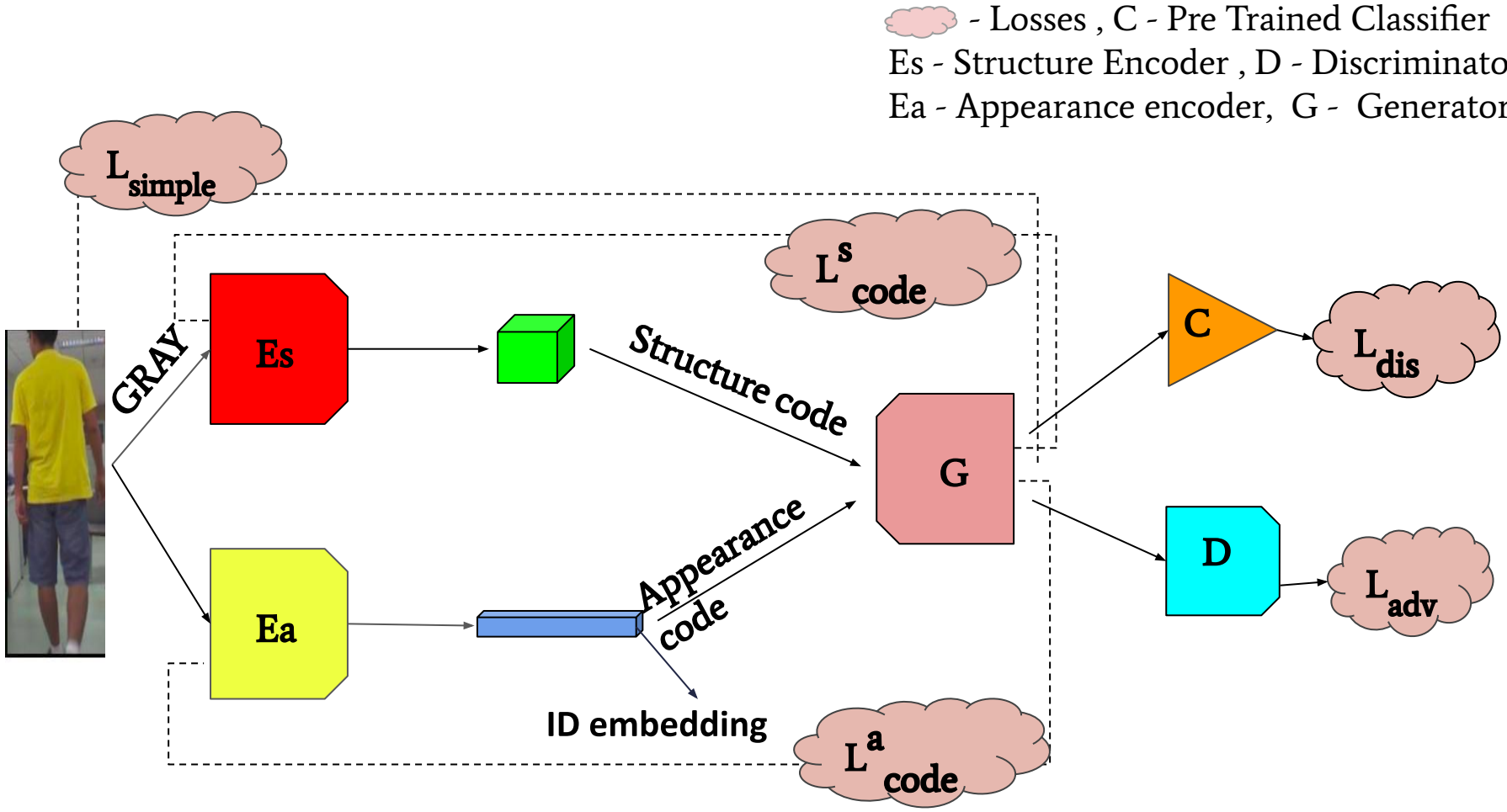
- Fixing what exactly is part of ID specific and ID independent ?
- Broadly, Image of a Pedestrian can have these attributes which we can segregate as present in Appearance space or Structure space.
- Structure space - geometry, position or shape related features
- Appearance space - Color information and other finer details.

Structure pace	Appearance space
Common in RGB, IR-Shape, pose details like hair, bag, etc.	Color of image,shoes, bag, type of clothes eg. Jeans, shorts,etc. face

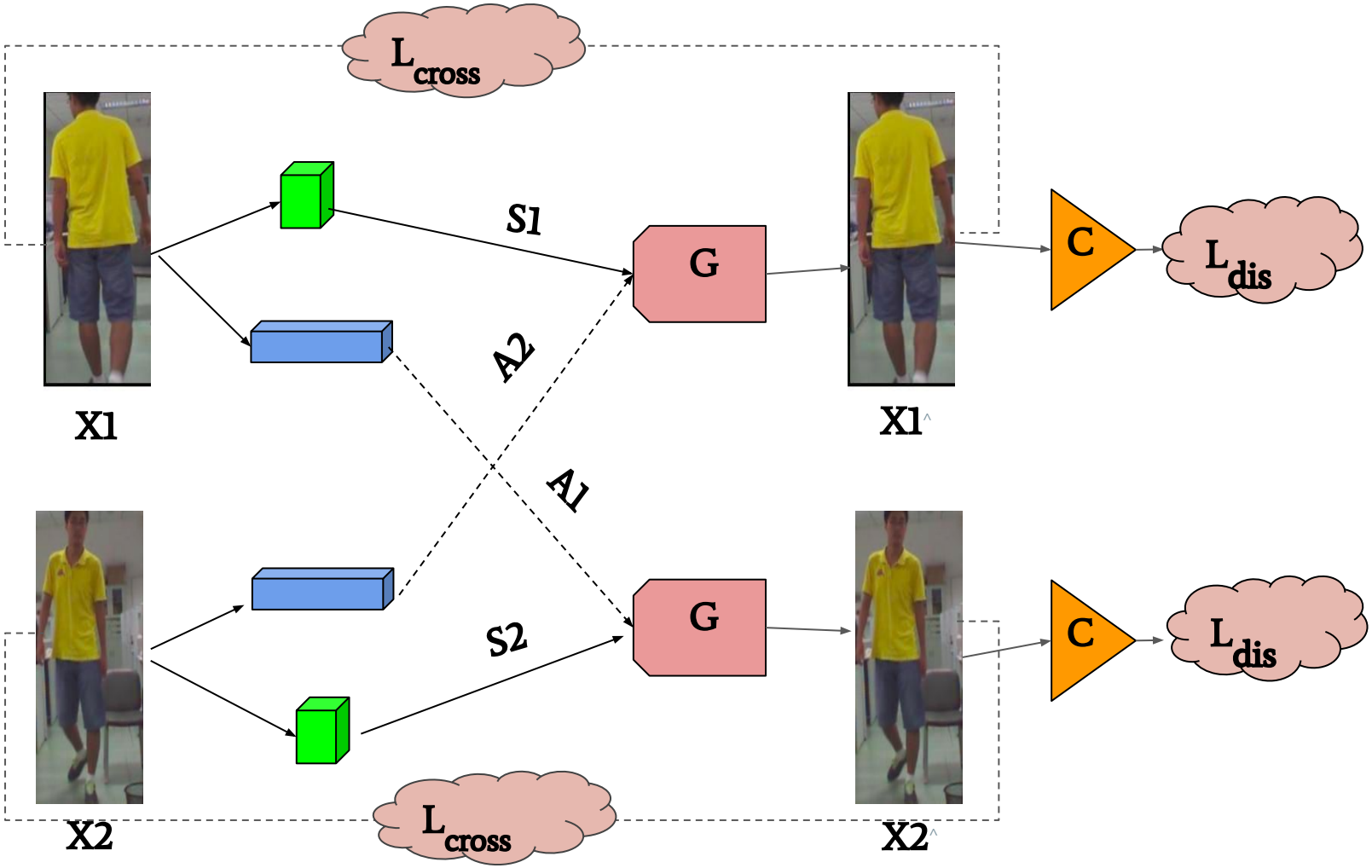
Table 1: Description of the information encoded in Structure space and Appearance space

DG-Net - Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Jointdiscriminative and generative learning for person re-identification. InProceedings of the IEEEConference on Computer Vision and Pattern Recognition, pages 2138–2147, 2019

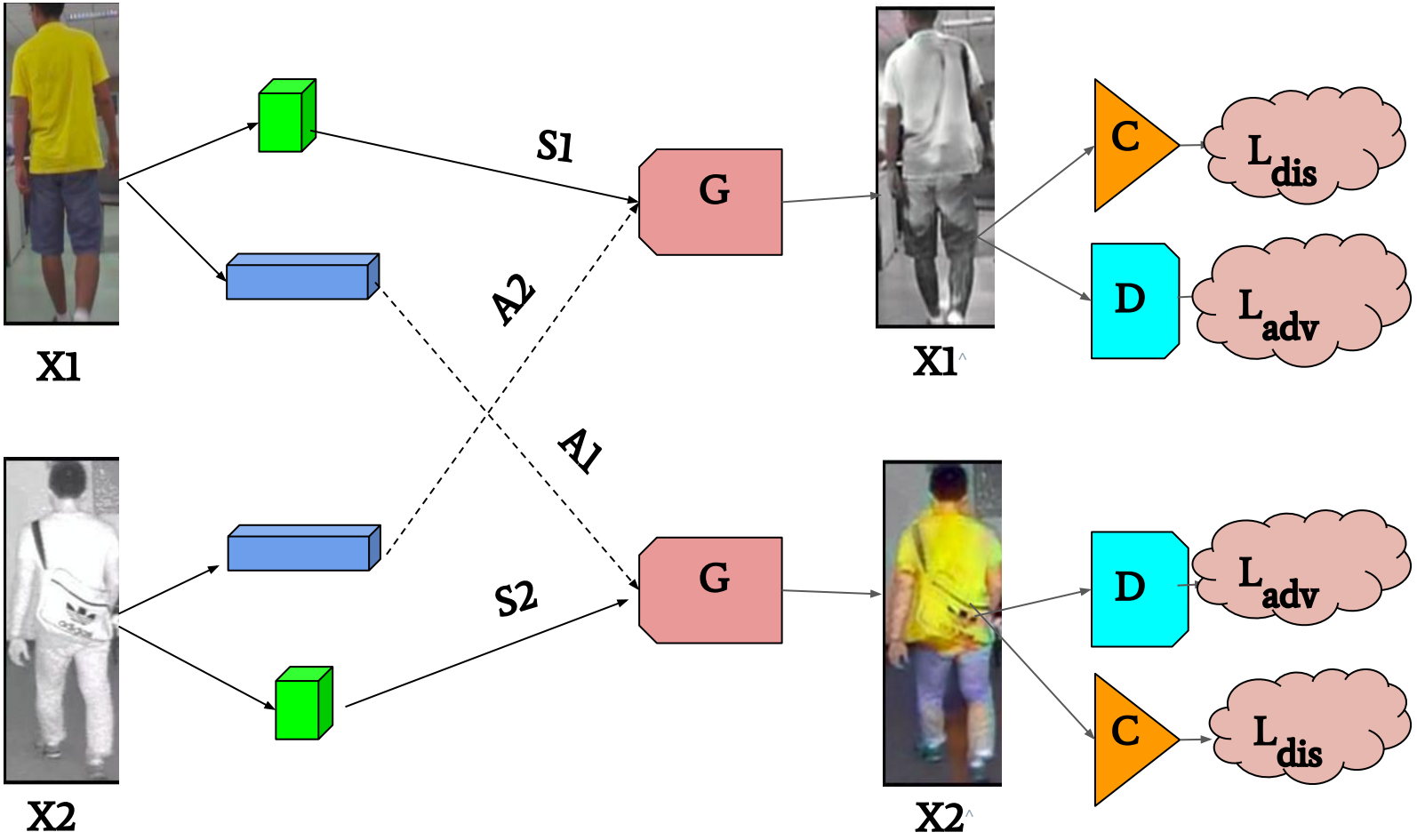
Approach 2 - Architecture



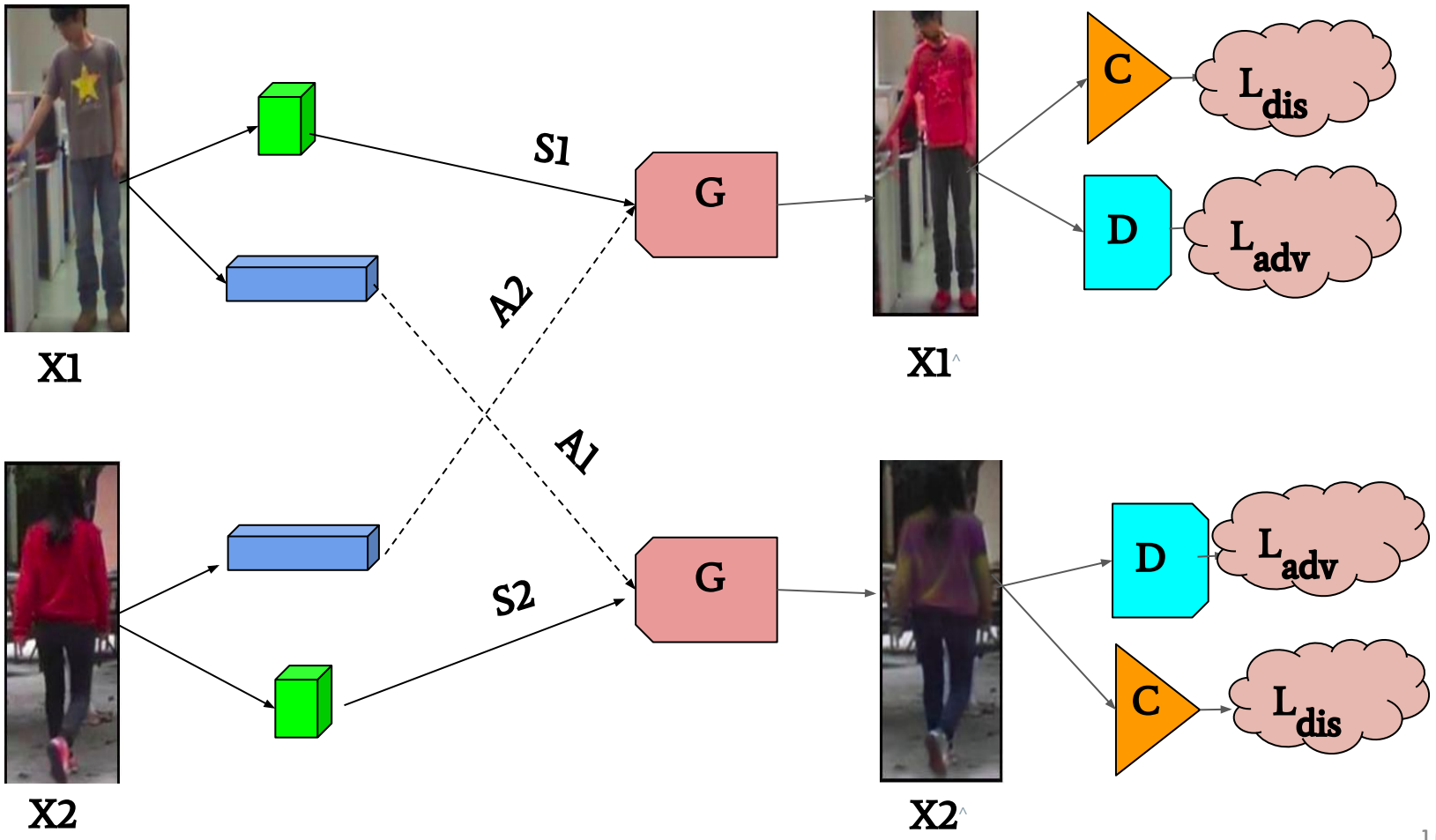
Same identity, same modality



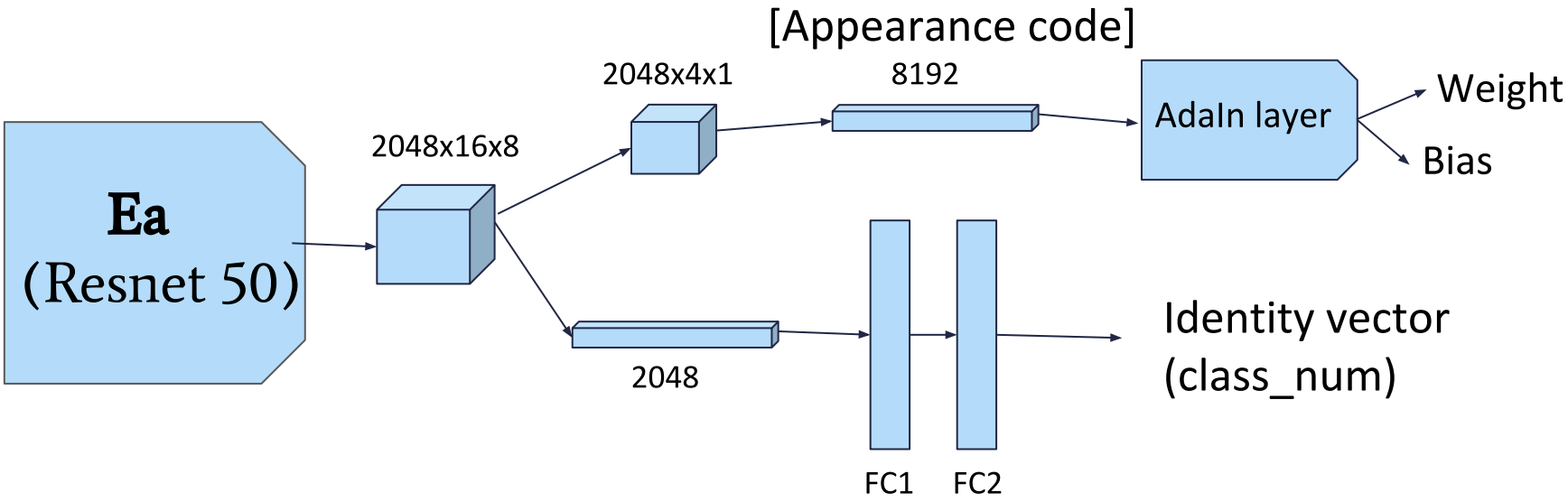
Different identity, Different modality



Different identity, Same modality



Appearance Encoder



Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV),pages 172–189, 2018

Architecture details

Layer	Parameters	Output Size
Input	-	1 x 256 x 128
Conv1	[3 x 3, 16]	16 x 128 x 64
Conv2	[3 x 3, 32]	32 x 128 x 64
Conv3	[3 x 3, 32]	32 x 128 x 64
Conv4	[3 x 3, 64]	64 x 64 x 32
ResBlocks	$\begin{bmatrix} 3X3, 64 \\ 3X3, 64 \end{bmatrix}$ x 4	64 x 64 x 32
ASPP	[1 x 1, 32] $\begin{bmatrix} 3X3, 64 \\ 3X3, 64 \end{bmatrix}$ x 4	64 x 64 x 32
Conv5	[1 x 1, 128]	128 x 64 x 32

Table 3: Architecture of Structure Encoder E_s

Layer	Parameters	Output Size
Input	-	128 x 64 x 32
ResBlocks	$\begin{bmatrix} 3X3, 128 \\ 3X3, 128 \end{bmatrix}$ x 4	128 x 64 x 32
Upsample	-	128 x 128 x 64
Conv1	[5 x 5, 64]	64 x 128 x 64
Upsample	-	64 x x256 x 128
Conv2	[5 x 5, 32]	32 x 256 x 128
Conv3	[3 x 3, 32]	32 x 256 x 128
Conv4	[3 x 3, 32]	32 x 256 x 128
Conv5	[1 x 1, 3]	3 x 256 x 128

Table 4: Architecture of Generator G

Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Jointdiscriminative and generative learning for person re-identification. InProceedings of the IEEEConference on Computer Vision and Pattern Recognition, pages 2138–2147, 2019

Architecture details

Layer	Parameters	Output Size
Input	-	3 x 256 x 128
Conv1	[1 x 1, 32]	32 x 256 x 128
Conv2	[3 x 3, 32]	32 x 256 x 128
Conv3	[3 x 3, 32]	32 x 128 x 64
Conv4	[3 x 3, 32]	32 x 128 x 64
Conv5	[3 x 3, 64]	64 x 64 x 32
ResBlocks	$\begin{bmatrix} 3X3, 64 \\ 3X3, 64 \end{bmatrix}$ x 4	64 x 64 x 32
Conv6	[1 x 1, 1]	1 x 64 x 32

Table 5: Architecture of Discriminator D

Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Jointdiscriminative and generative learning for person re-identification. InProceedings of the IEEEConference on Computer Vision and Pattern Recognition, pages 2138–2147, 2019

Approach 2 - Losses

$$L_{simple} = \mathbb{E}\,\|x_1 - G(a_1, s_1)\|_1$$

$$L_{dis} = \mathbb{E}\,[KL(p(x_1^2)\,\|\,q(x_1^2))]$$

$$L_{cross} = \mathbb{E}\,\|x_1 - G(a_2, s_1)\|_1$$

$$L_{id\_orig} = \mathbb{E}[-log(p(y_1|x_1))]$$

$$x_1^2 = G(a_2, s_1) \text{ and } x_2^1 = G(a_1, s_2)$$

$$L_{code}^s = \mathbb{E}\,\|s_1 - E_s(G(a_2, s_1))\|_1 \quad L_{code}^a = \mathbb{E}\,\|a_1 - E_a(G(a_1, s_2))\|_1$$

$$L_{code} = L_{code}^s + L_{code}^a$$

$$L_{id\_recon} = \mathbb{E}[-log(p(y_2|x_1^2))] \quad L_{id} = L_{id\_orig} + \lambda_{id} L_{id\_recon}$$

$$L_{adv} = \mathbb{E}[log(D(x_1)) + log(1 - D(G(a_1, s_2))]$$

$$L_{total}(E_a, E_s, G, D) = \lambda_{recon}(L_{simple} + L_{cross}) + L_{id} + L_{code} + L_{adv} + L_{dis}$$

DG-Net results

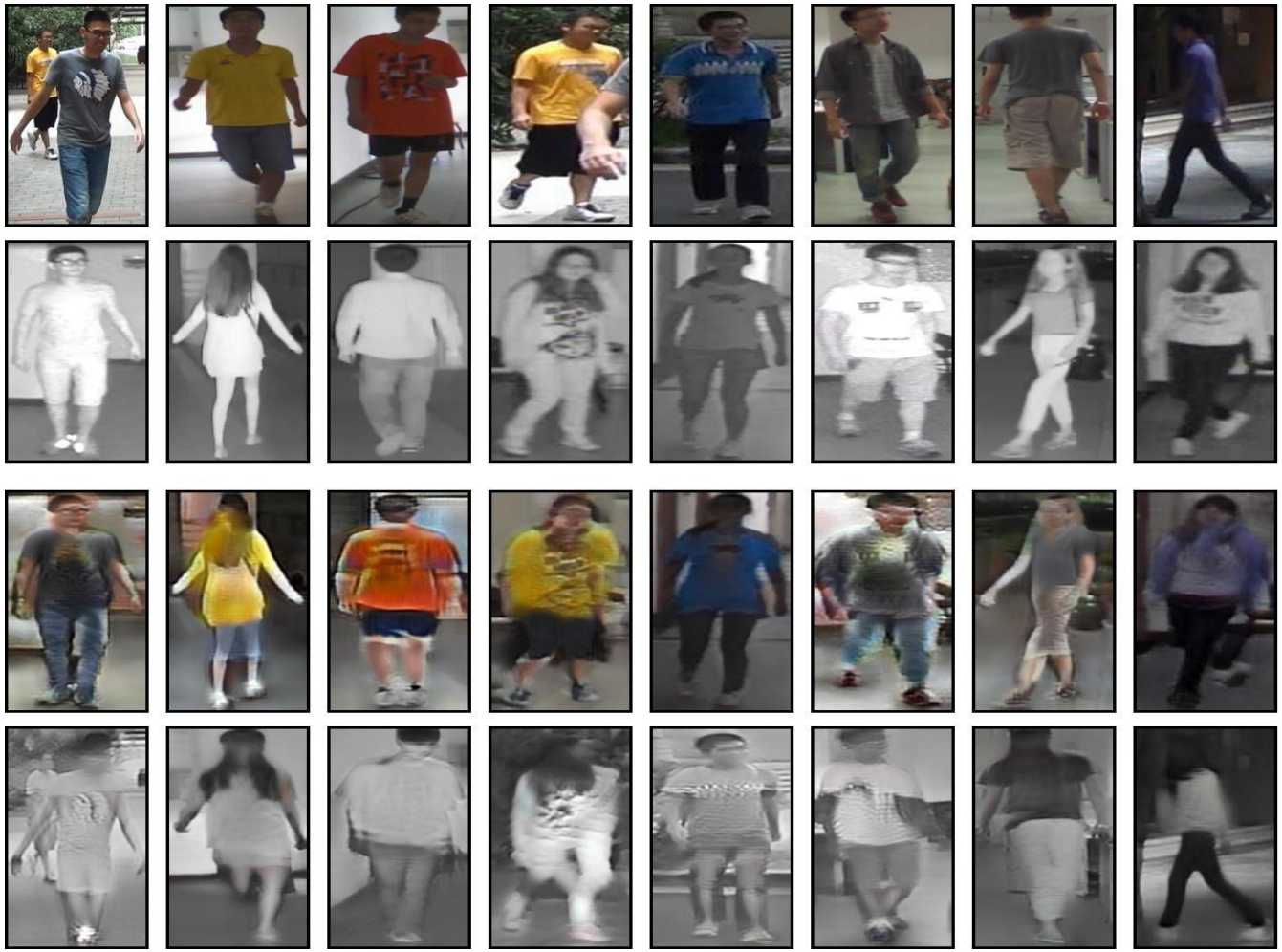
Methods	All Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r12	r10	r20	mAP
BDTR [16]	17.01	55.43	71.96	19.66	-	-	-	-
SDL [8]	28.12	70.23	83.67	29.01	-	-	-	-
cmPIG [13]	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5
Hi-CMD [3]	34.94	77.58	-	35.94	-	-	-	-
CASE-Net [10]	42.9	85.7	94.0	41.5	**52.2**	90.3	96.1	34.5
Baseline [18]	**45.44**	**88.85**	**96.37**	**46.84**	47.73	**90.30**	**96.48**	**40.12**

Table 9: Results on SYSU-MM01 dataset with All Search mode

Methods	Indoor Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r12	r10	r20	mAP
SDL [8]	32.56	80.45	90.67	39.56	-	-	-	-
cmPIG [13]	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
CASE-Net [10]	44.1	87.3	93.7	53.2	**55.0**	90.6	96.8	43.4
Baseline [18]	**51.95**	**95.24**	**98.78**	**63.11**	52.65	**94.10**	**98.35**	**51.22**

Table 10: Results on SYSU-MM01 dataset with Indoor Search mode

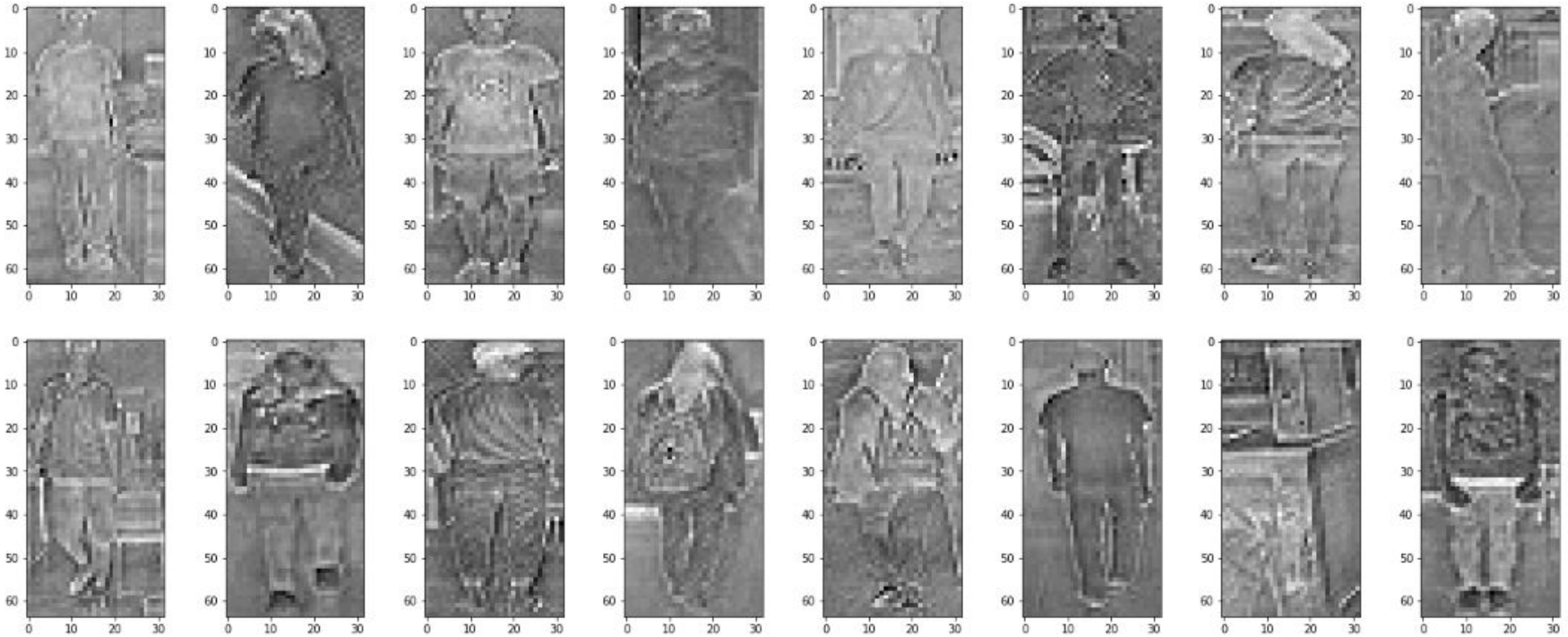
Qualitative result DG-Net



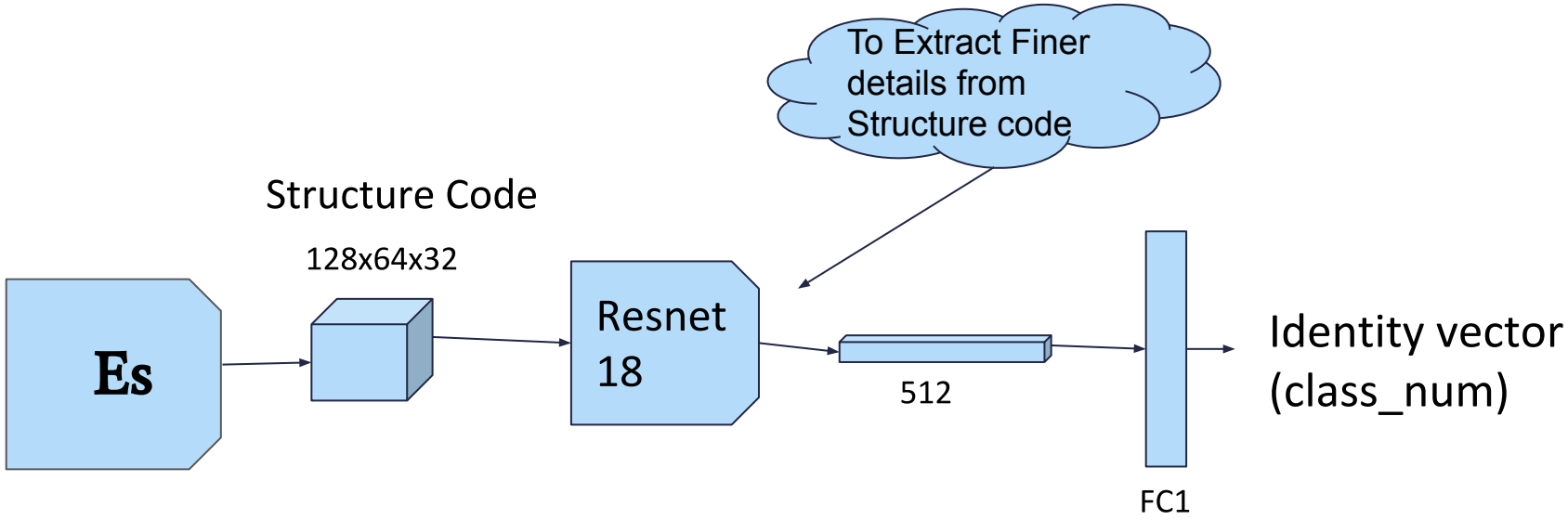
Proposed Modification in DG-Net

- This baseline uses Identity code coming from Appearance encoder which encodes color
- But in RGB-IR, color is not an identity cue.
- So we need to rely more on Structure space information, like shape of body and also make it to encoder finer details
- Taking the identity embedding from Structure space.

Structure space feature maps from baseline



Structure Encoder modifications



Final - Results All search

Methods	All Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r12	r10	r20	mAP
BDTR [16]	17.01	55.43	71.96	19.66	-	-	-	-
SDL [8]	28.12	70.23	83.67	29.01	-	-	-	-
cmPIG [13]	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5
Hi-CMD [3]	34.94	77.58	-	35.94	-	-	-	-
CASE-Net [10]	42.9	85.7	94.0	41.5	**52.2**	90.3	96.1	34.5
Baseline [18]	**45.44**	**88.85**	**96.37**	**46.84**	47.73	**90.30**	**96.48**	**40.12**
Proposed	11.49	46.46	64.66	14.54	10.44	47.78	67.47	10.51

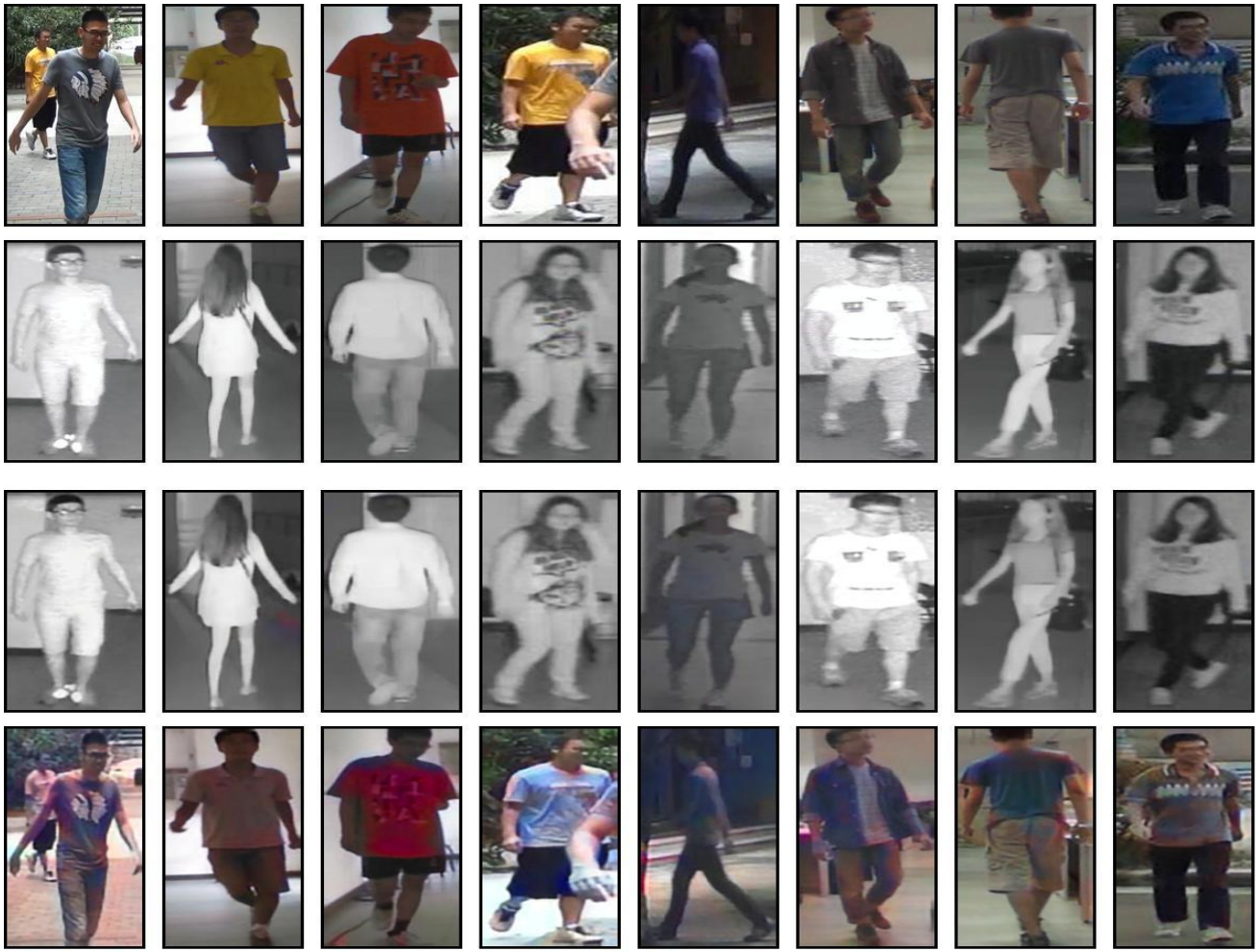
Table 3: Results on SYSU-MM01 dataset with All Search mode

Final - Results Indoor search

Methods	Indoor Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r12	r10	r20	mAP
SDL [8]	32.56	80.45	90.67	39.56	-	-	-	-
cmPIG [13]	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
CASE-Net [10]	44.1	87.3	93.7	53.2	**55.0**	90.6	96.8	43.4
Baseline [18]	**51.95**	**95.24**	**98.78**	**63.11**	52.65	**94.10**	**98.35**	**51.22**
Proposed	13.90	57.70	76.45	23.11	14.58	59.41	78.88	16.40

Table 4: Results on SYSU-MM01 dataset with Indoor Search mode

Qualitative result Proposed



Future Work

- Making structure encoder learn robust shape information (Replacing structure Encoder block with Resnet 50 similar to Appearance encoder to extract finer details)
- Extracting attribute information from the representations like hair, bag etc. using Person Attribute datasets
- Trying with the edges of image to extract identity information, and along with texture information.

Original Image

Edge Image

Any Questions???

Thank You