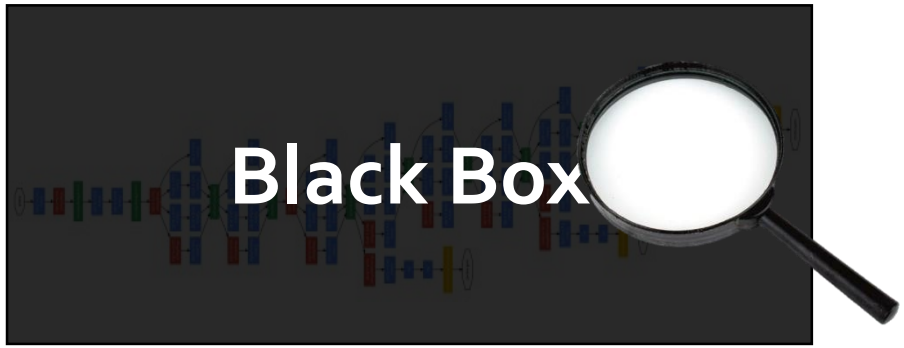
iCaps: An Interpretable Classifier via Disentangled Capsule Networks

Dahuin Jung, Jonghyun Lee, Jihun Yi, and Sungroh Yoon

SEOUL NATIONAL UNIVERSITY

Explainable Artificial Intelligence (XAI)

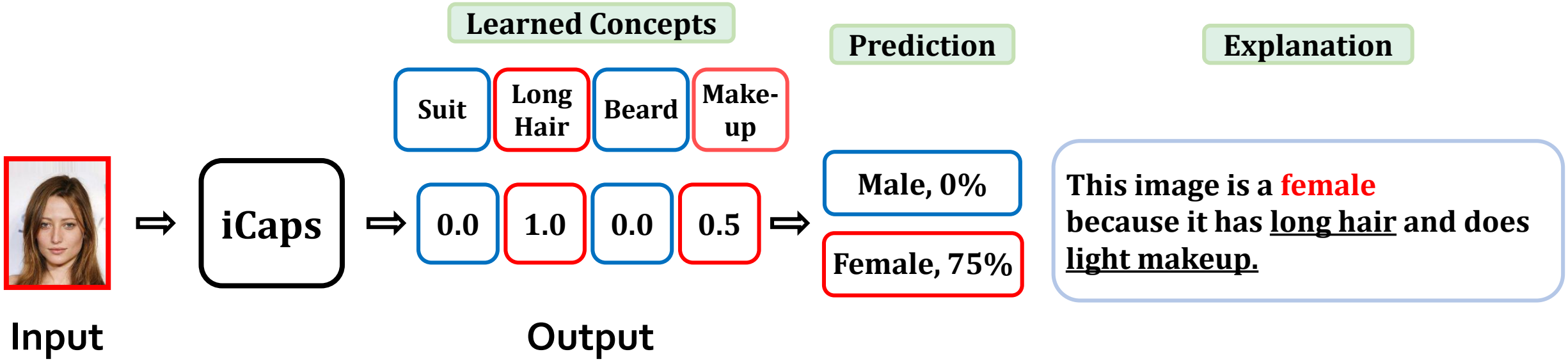
- **Post-hoc Interpretability**



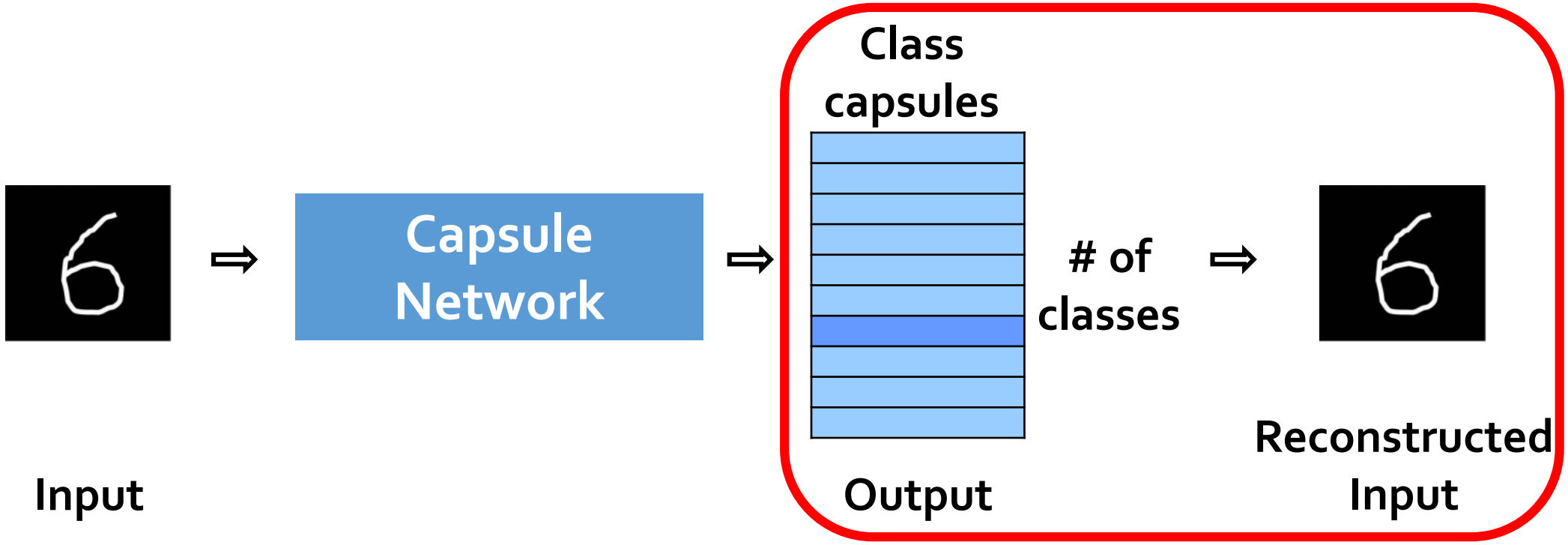
- **Built-in Interpretability**

iCaps

- Design a new built-in interpretable model



Capsule Networks



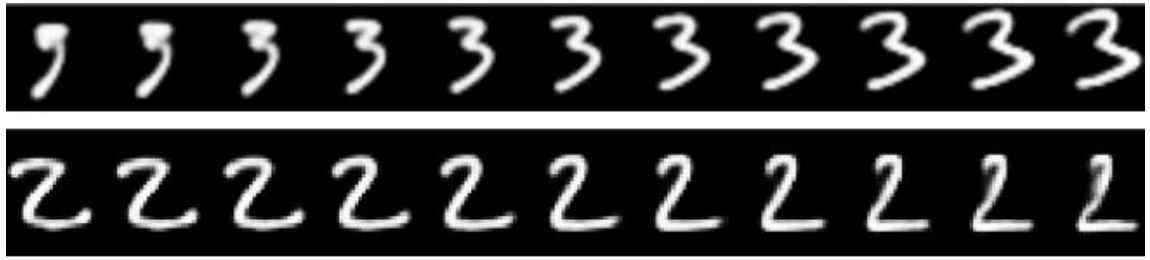
Capsule Networks

1. **Classification-irrelevant concepts**

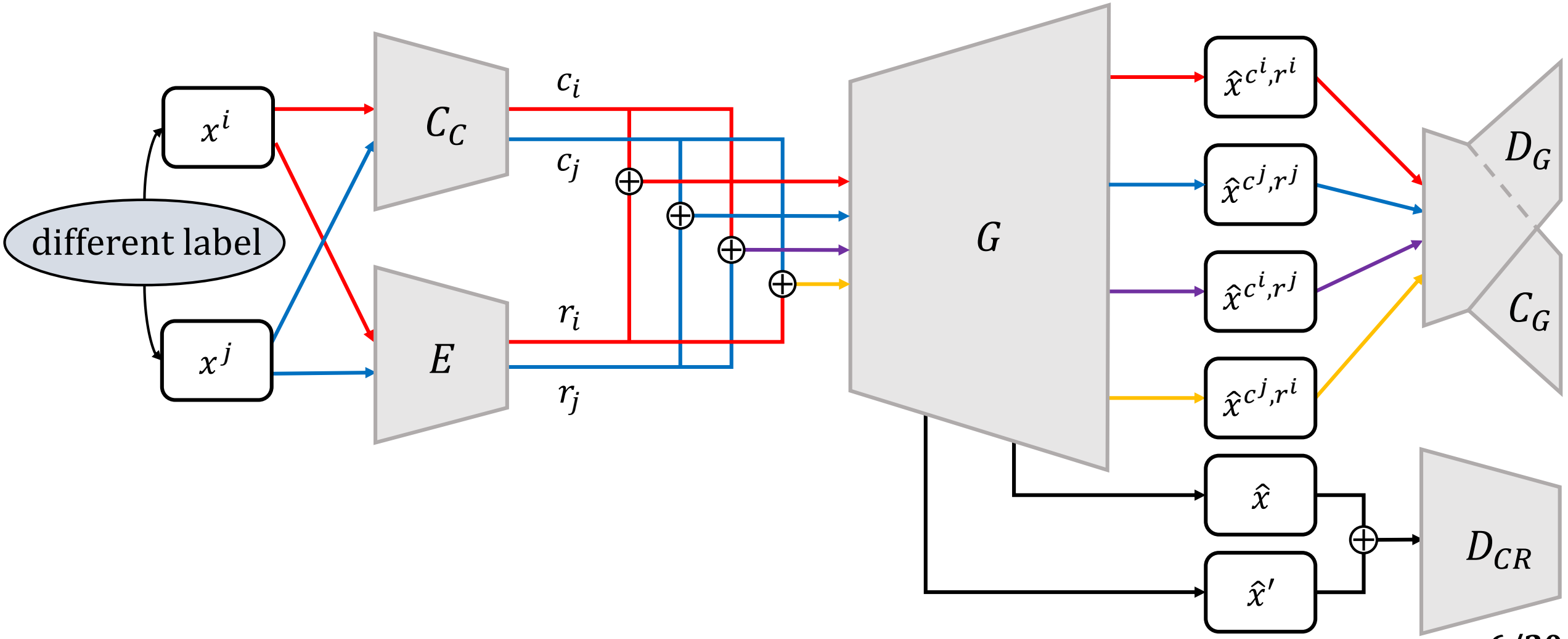
Thickness

2. **Overlap in concepts**

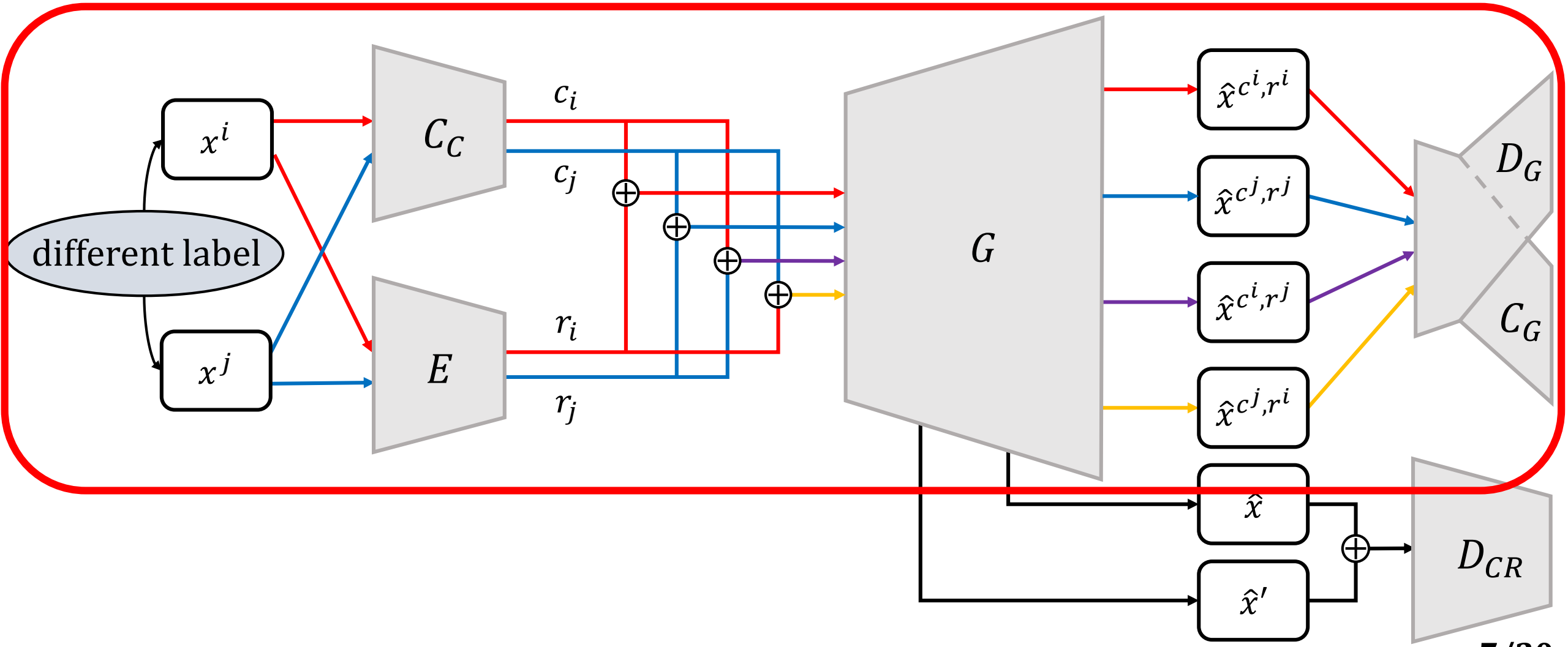
Width



Network Architecture



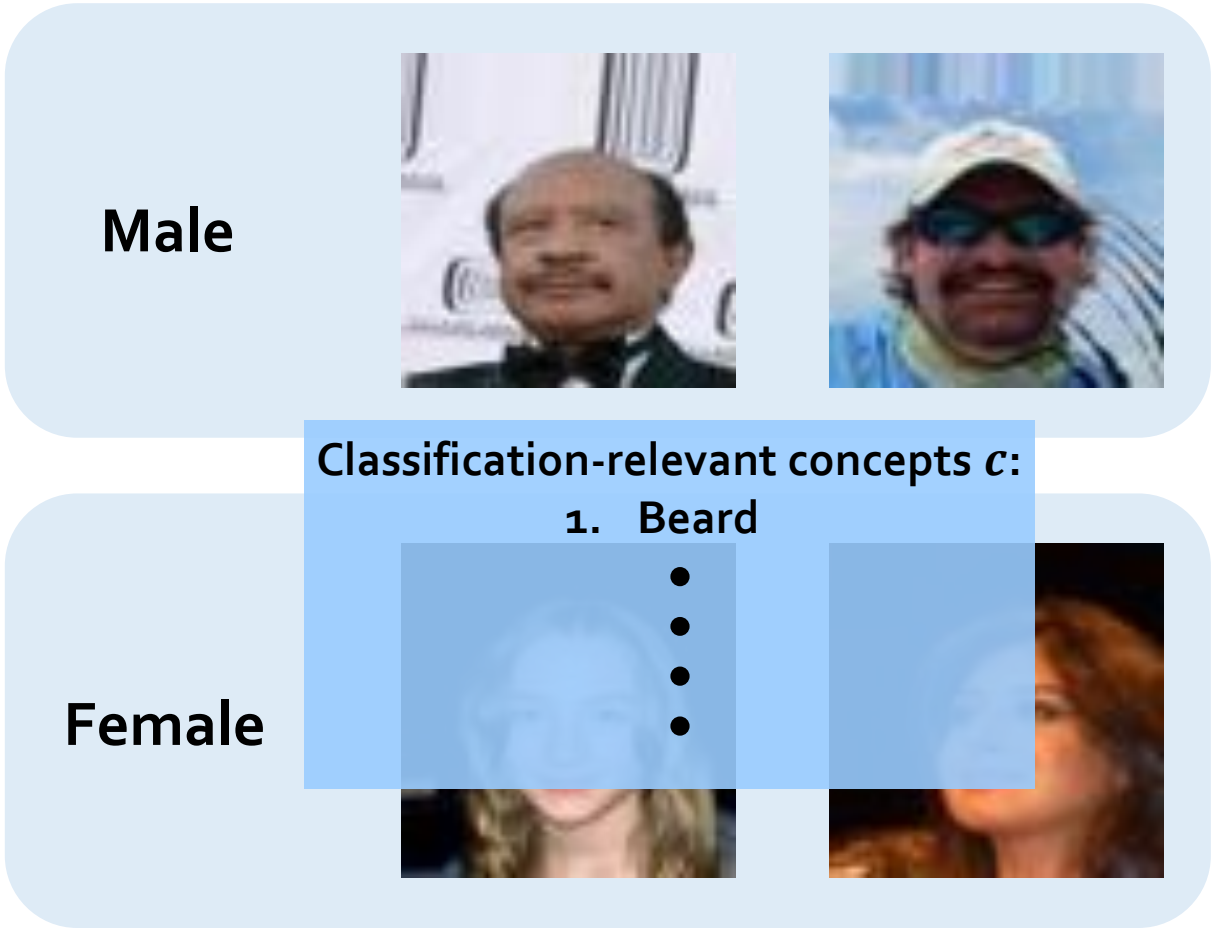
Novel Disentanglement Method



Intra-class Variation

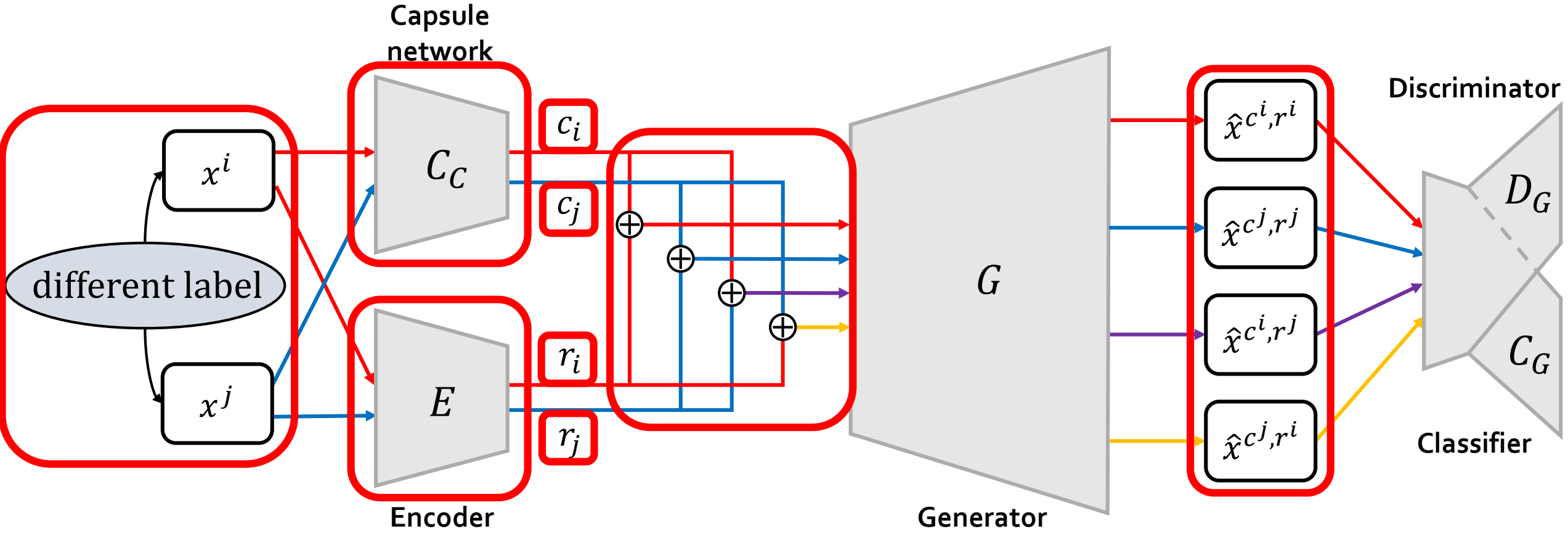
Previous class-supervised disentanglement methods assume

$$\left\| c_i - c_j \right\|_F^2 = 0 \text{ if } y_i = y_j$$



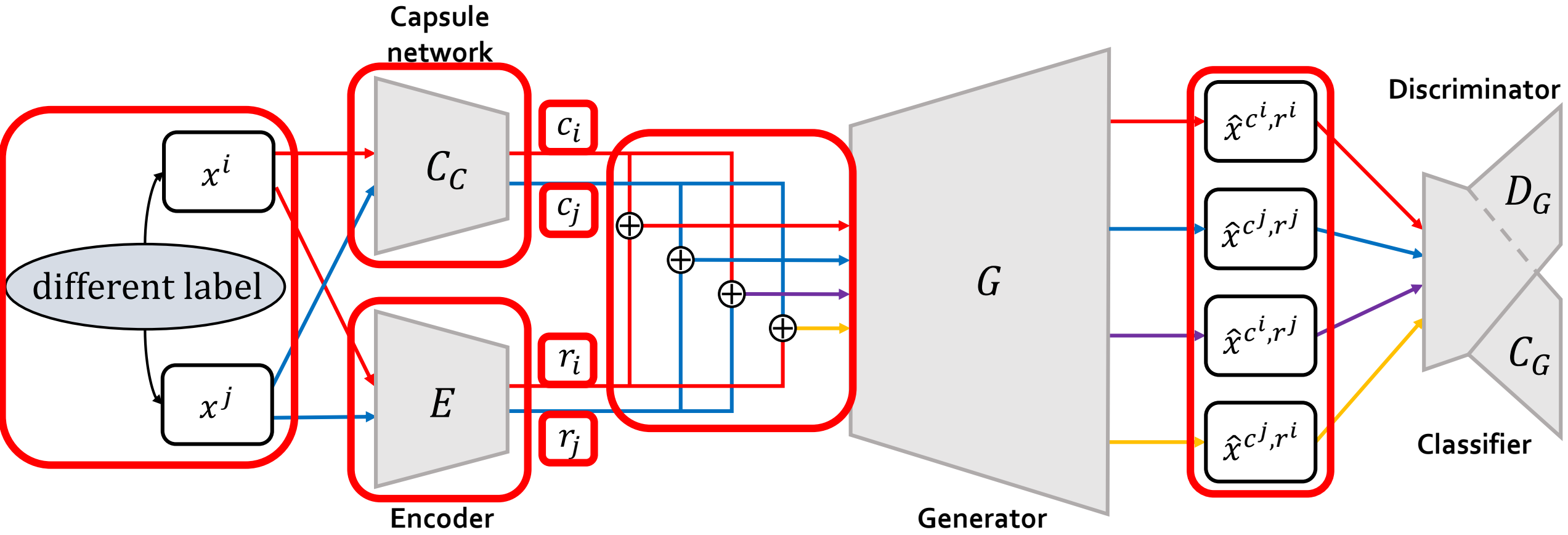
c: class-relevant features, r: class-irrelevant features

Proposed loss: L_{CS}



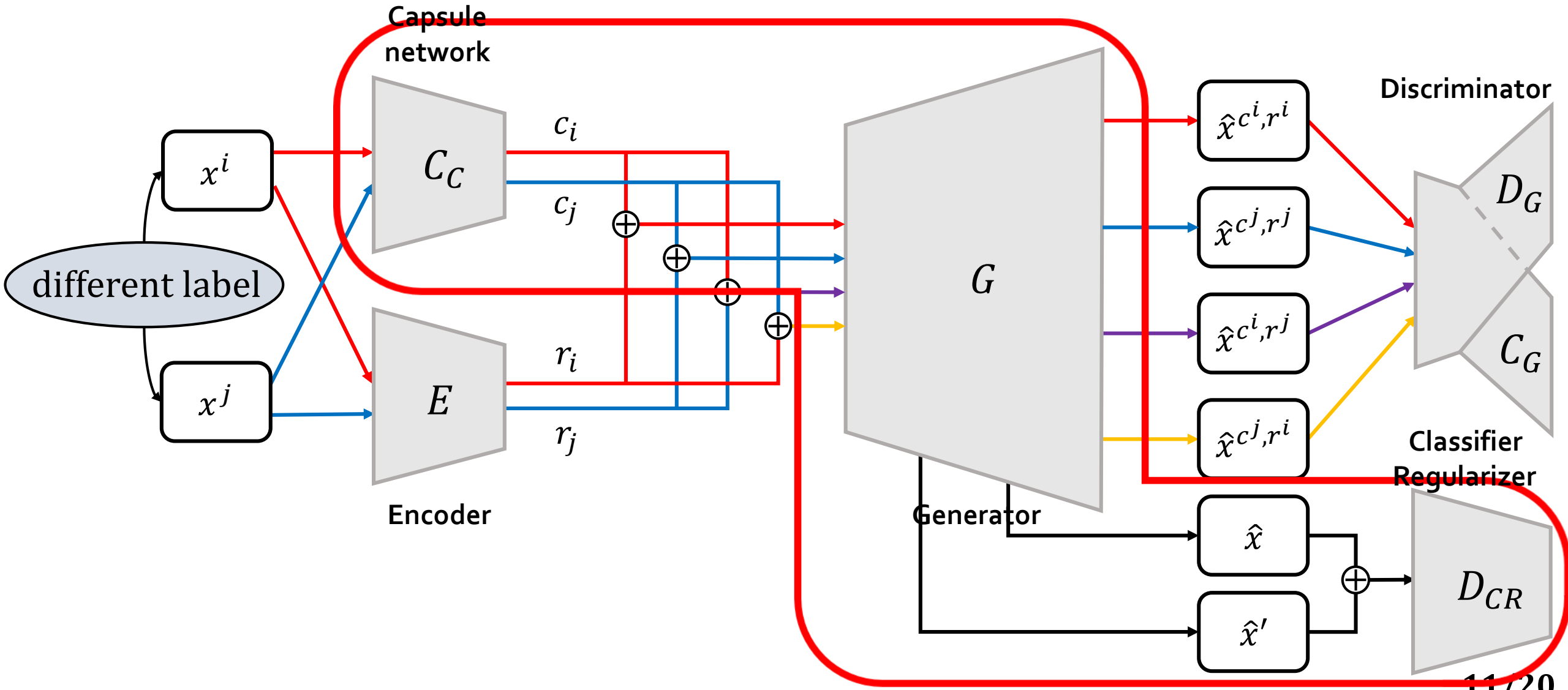
$$\mathcal{L}_{CS} = \lambda_{CS}\,\mathbb{E}\left\| C_G(\hat{x}^{c^i,r^i}) - C_G(\hat{x}^{c^i,r^j}) \right\|_F^2$$

Proposed loss: L_{RS}

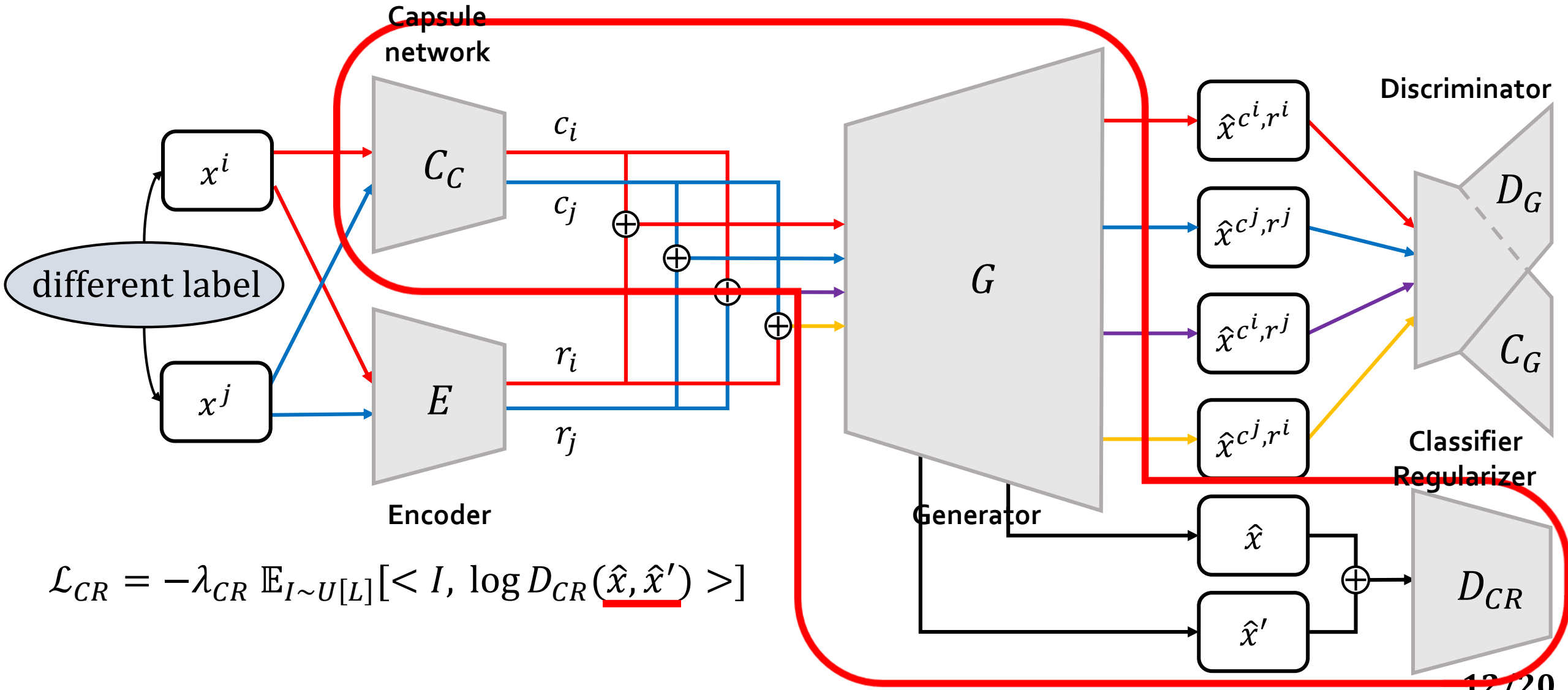


$$\mathcal{L}_{RS} = \lambda_{RS}\,\mathbb{E}\left\|E(x^i) - E(\hat{x}^{c^j,r^i})\right\|_F^2$$

Additional Regularizer



Proposed loss: L_{CR}

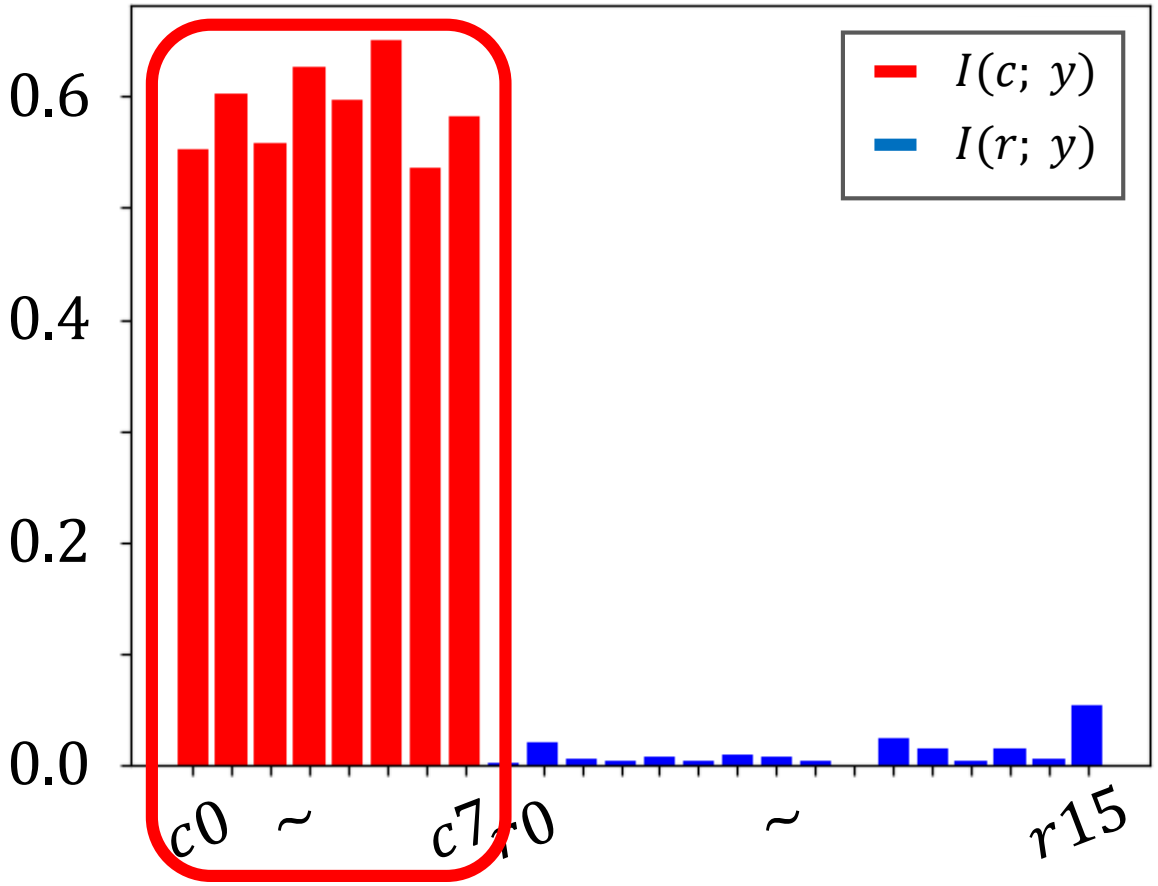


$$\mathcal{L}_{CR} = -\lambda_{CR}\, \mathbb{E}_{I\sim U[L]}[< I,\ \log D_{CR}(\hat{x}, \hat{x}') >]$$

Classification Performance of C_C

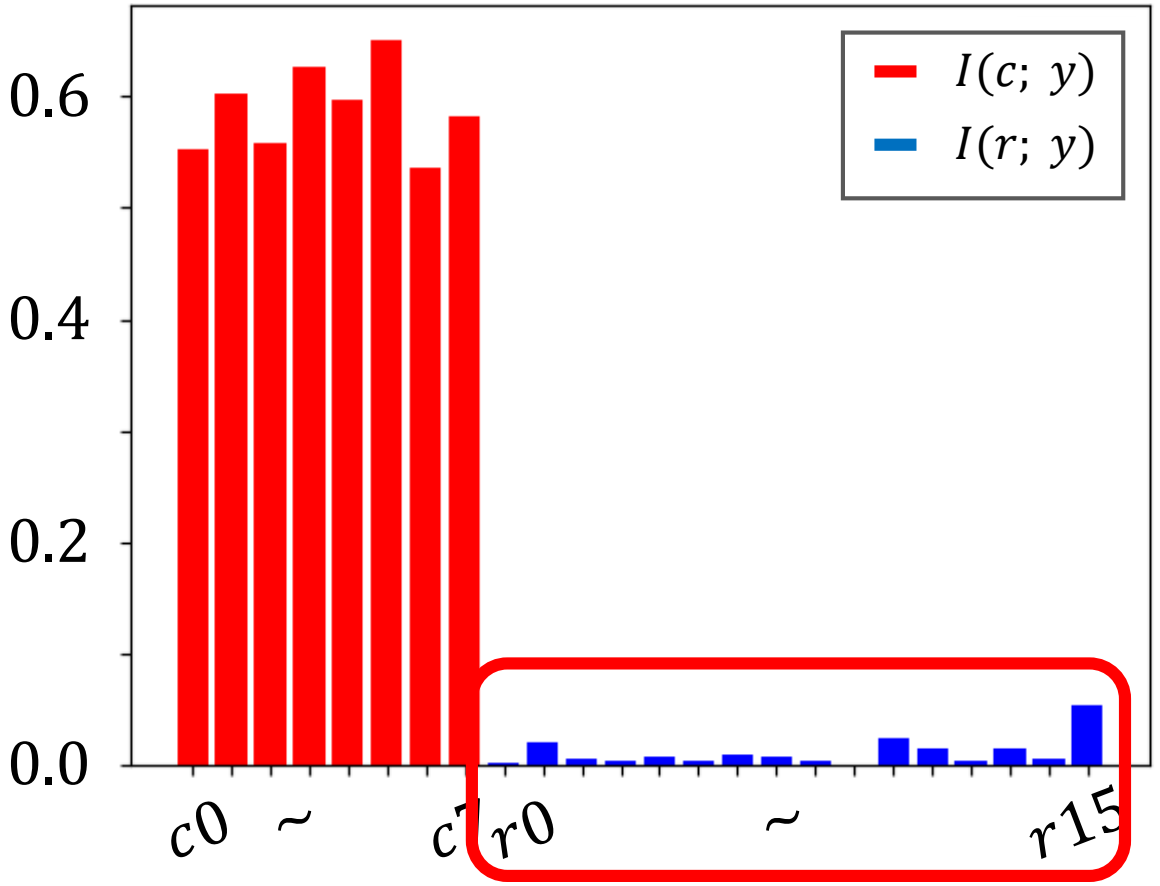
Architecture	MNIST	SVHN	CelebA
ResNet-18	0.992	0.945	0.977
DeepCaps	0.997	0.971	0.974
Ours	0.992	0.920	0.984

Class-supervised Disentanglement



c: class-relevant features, r: class-irrelevant features

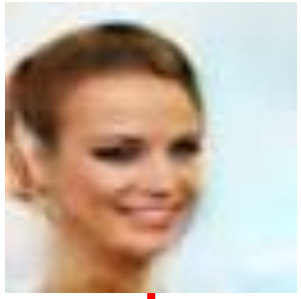
Class-supervised Disentanglement



c: class-relevant features, r: class-irrelevant features

Swapping

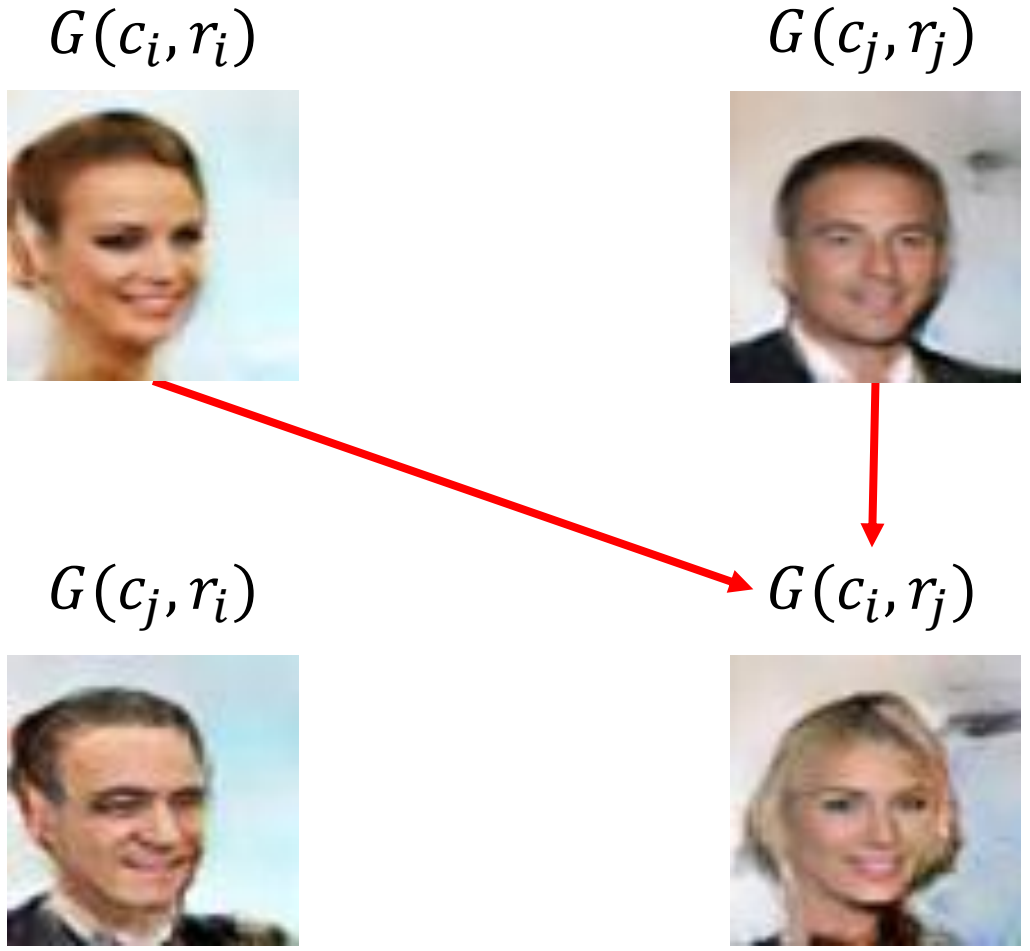
$G(c_i, r_i)$



$G(c_j, r_j)$

$G(c_j, r_i)$

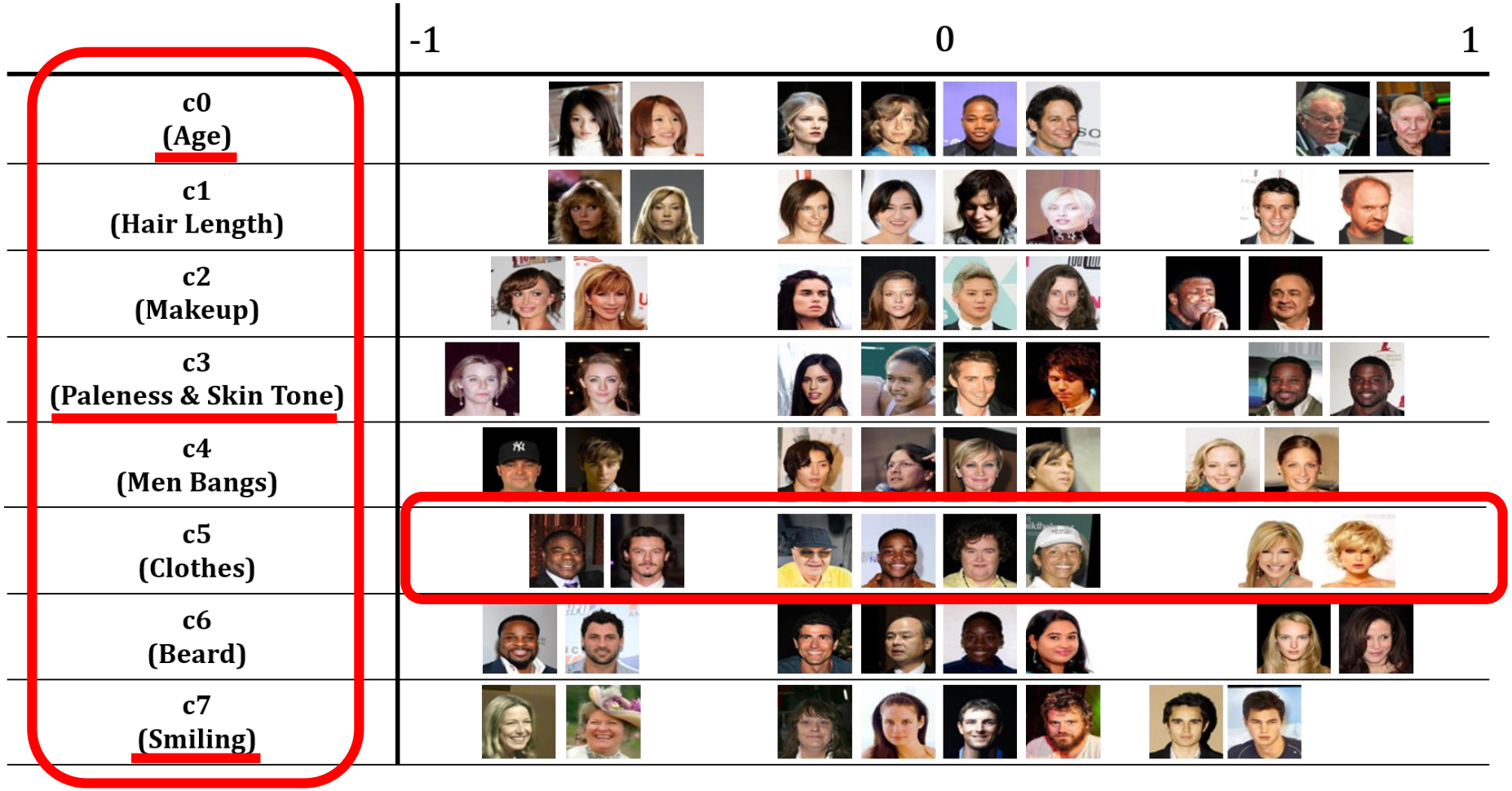
Swapping



Distinctness

	Architecture	FID↓	MI↑	MIG↑
(1)	C_C and G	**101.53**	-	-
(2)	C_C, G, and D_G	**87.00**	**0.02**	-
(3)	C_C, G, D_G, and E	87.02	**0.08**	-
(4)	C_C, G, D_G, E, and C_G	86.83	**0.56**	**0.03**
(5)	C_C, G, D_G, E, C_G, and D_{CR}	86.72	0.56	**0.20**

Learned Concepts for CelebA



			Age	Hair Length	Make up	Skin Tone	Men Bangs	Clothes	Beard	Smiling	Confidence
(a)		F	-0.073	-0.271	-0.364	-0.469	0.306	0.175	0.408	-0.523	0.999
(b)		M	0.439	0.411	0.325	0.300	-0.344	-0.370	-0.410	0.124	0.999
(c)		F	-0.101	-0.261	-0.101	0.006	-0.027	0.073	-0.116	-0.063	0.335
(d)		M	0.043	-0.016	-0.001	0.017	-0.054	0.040	0.001	0.024	0.087

Conclusion

- We propose a novel disentanglement method that the class-relevant subspace contains both class-relevant inter-class and intra-class variation.

- Using the proposed method, we build a new interpretable capsule network that provides explanations of the model's prediction based on only class-relevant concepts.

Thank you

SEOUL
NATIONAL
UNIVERSITY