University of Nottingham
UK | CHINA | MALAYSIA

horizon
CENTRE FOR DOCTORAL TRAINING

Computer Vision Laboratory
cvl.cs.nott.ac.uk

Self-supervised learning of Dynamic Representations for Static Images

Siyang Song
Enrique Sanchez
Linlin Shen
Michel Valstar

Background

Background:

- Temporal information is crucial to human facial behaviour understanding.
- Facial actions are continuous and smooth process
- The same facial actions of different people are similar

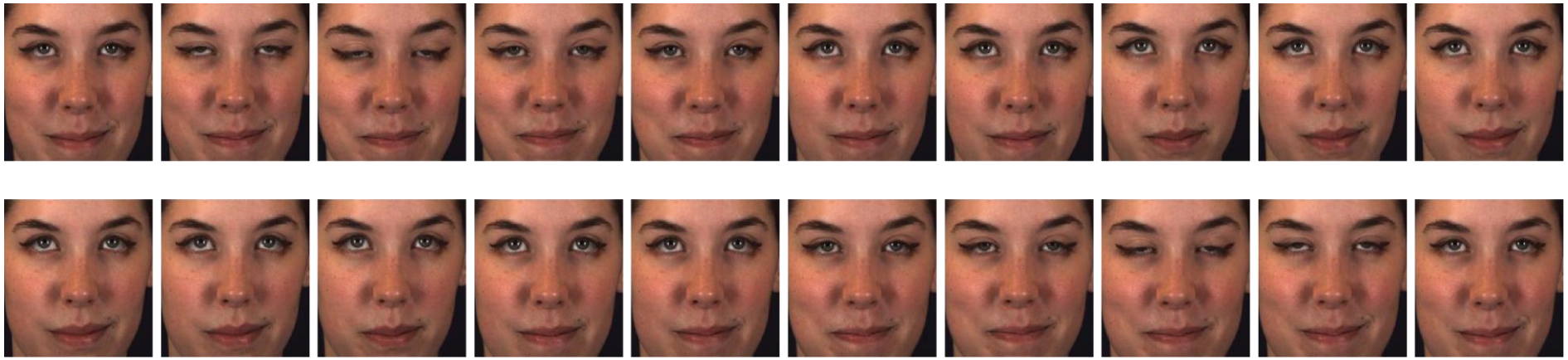
Research Gap:

- In some scenarios, only a still image is available, which means it is almost impossible to use temporal information. As a result, the performance of state-of-the-art methods for facial expression recognition or affect estimation degrades substantially.

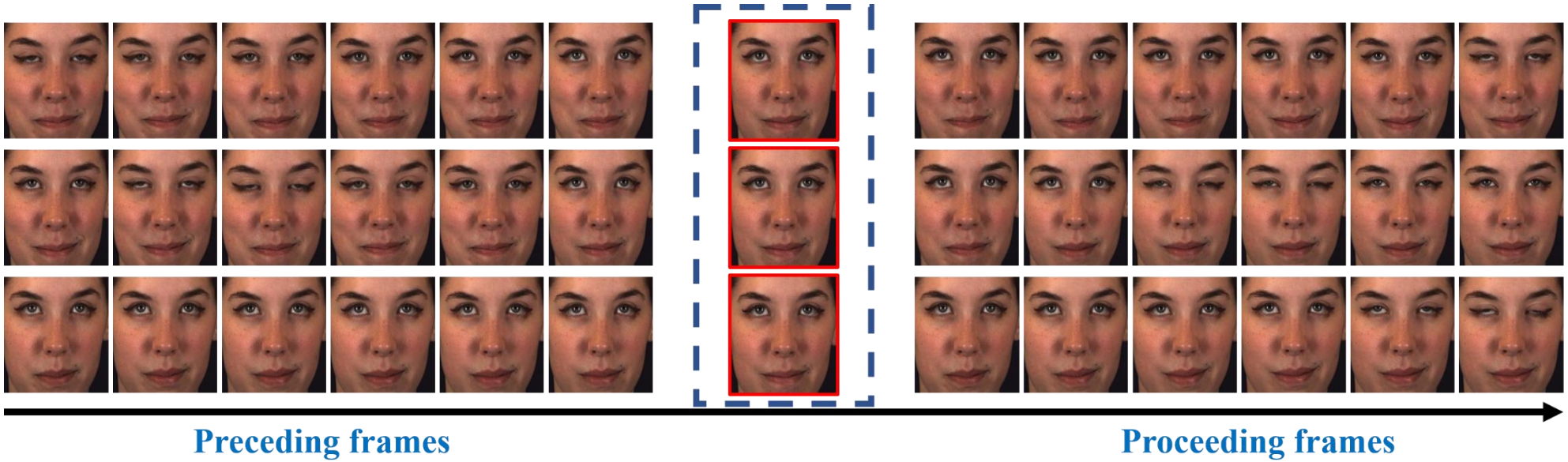
Motivation:

- Propose an approach that can infer generic facial temporal information from a single face image

Short-term facial dynamics encoding



Symmetric pattern



Ambiguous pattern

The proposed approach

Target

$$\delta_{ab}(t) \doteq S(d_t, V_a) - S(d_t, V_b)$$

$$\delta_{ab}(t) > 0 \quad \text{for} \quad \begin{cases} |a-t| < |b-t| \\ (a-t)(b-t) > 0 \end{cases}$$

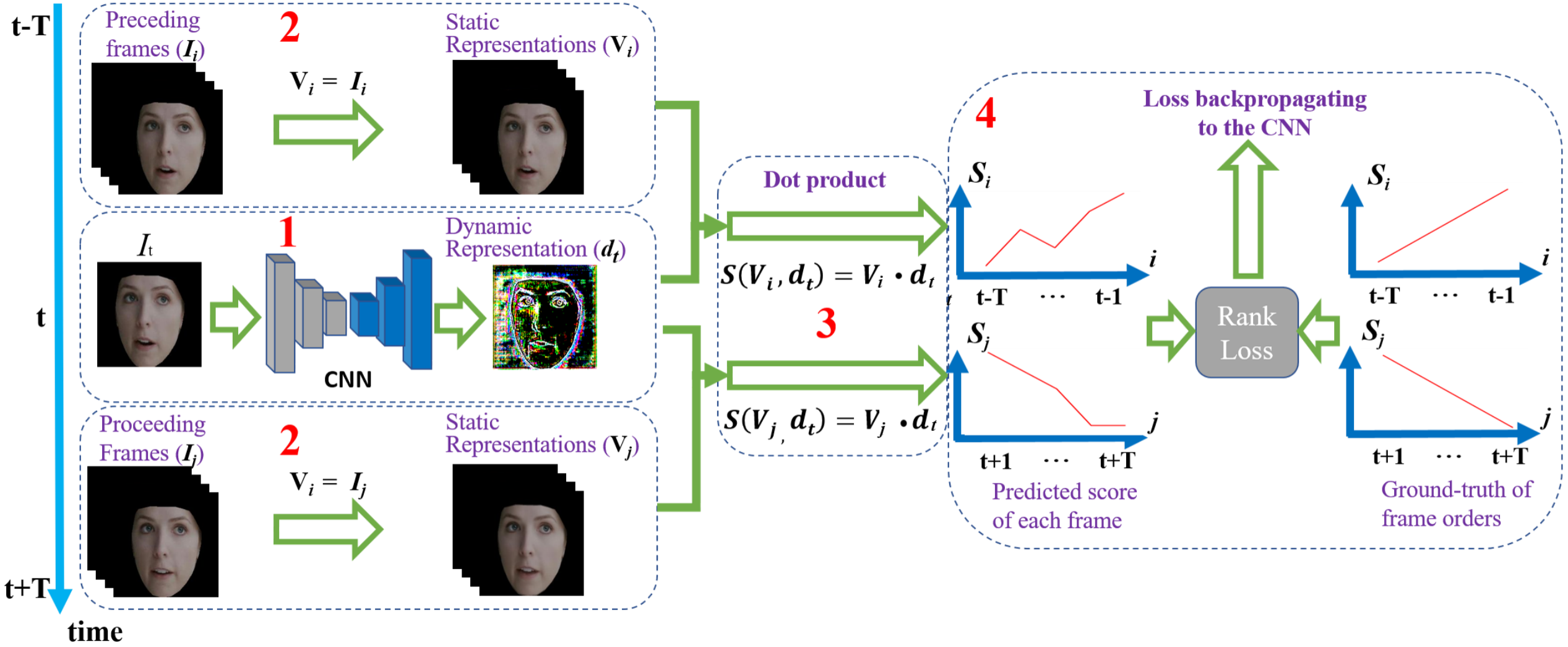
The generated representation is designed to rank both preceding and successive frames of the input frame, based on their temporal distance relative to it.

Rank Loss

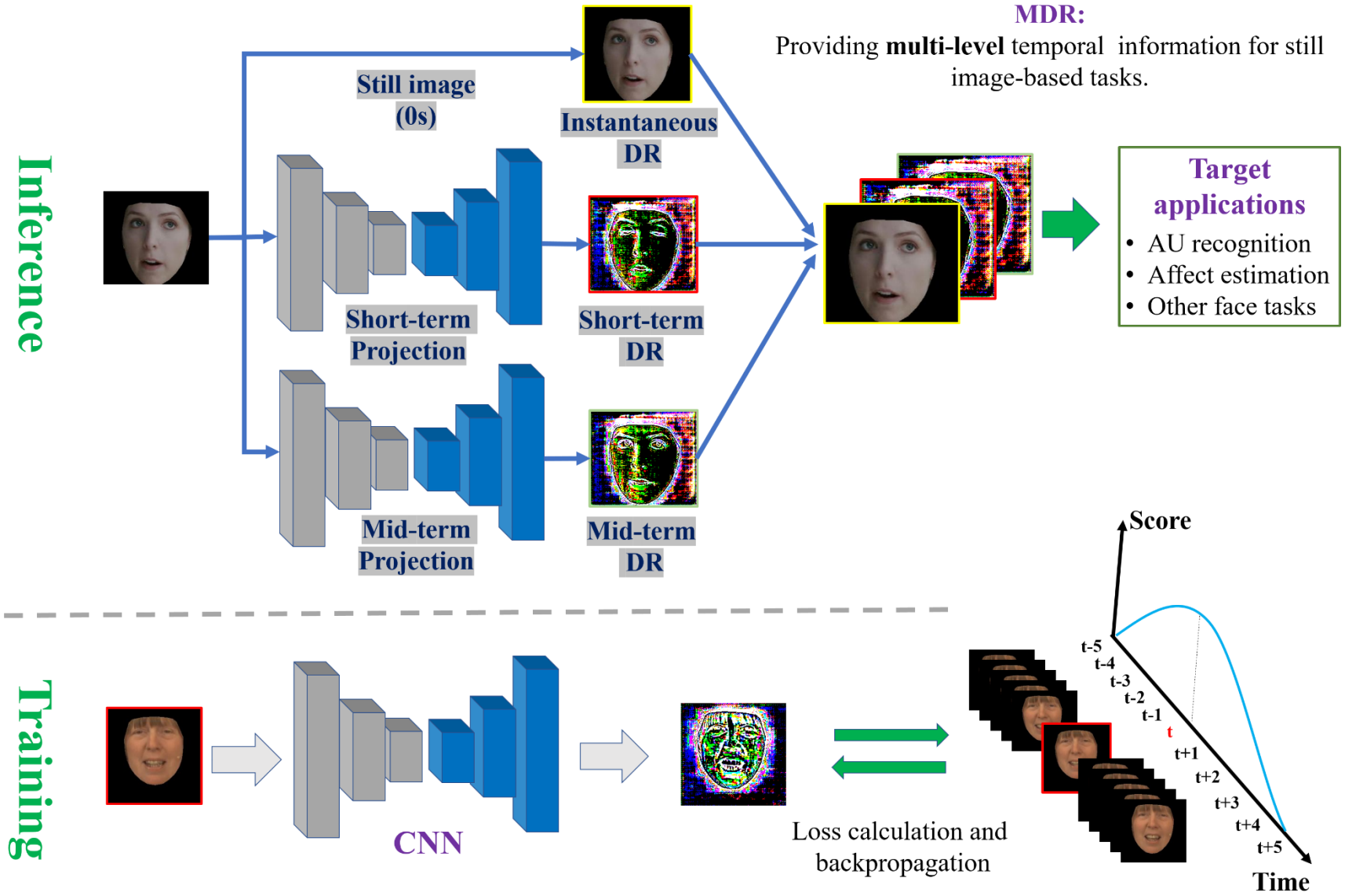
$$L_f(d_t) = \gamma \times \|\mathrm{d_t}\|^2 - \varepsilon + \sum_{b=t-T}^{t-1} \sum_{a=b+1}^{t} \max(0, \theta - \delta_{ab}(t)) + \sum_{a=t}^{t+T-1} \sum_{b=a+1}^{T} \max(0, \theta - \delta_{ab}(t))$$

The proposed approach

Dynamic Representation



The proposed approach



- It encodes facial dynamics in the context of the face.
- The DR can jointly encode both preceding and proceeding dynamics into a single 3-channel raster image.
- It learns generic facial dynamics from a large pool of unlabeled videos and can infer facial dynamics from a single image.

Results

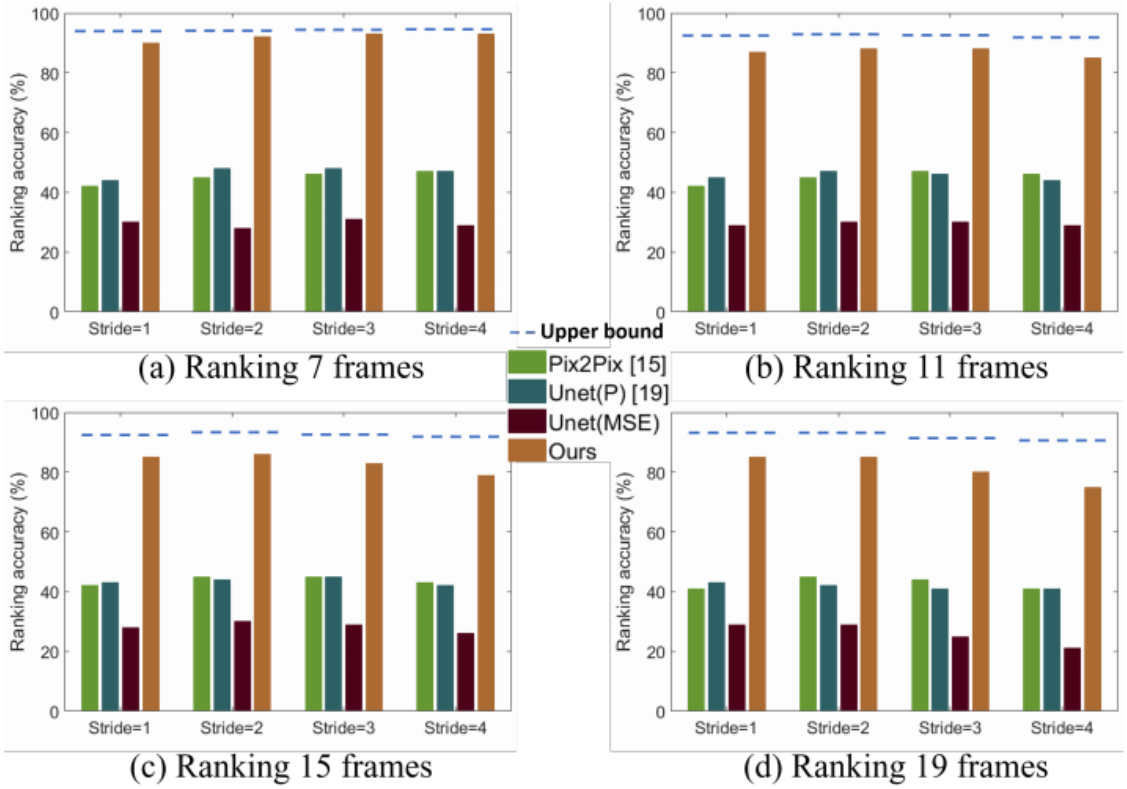
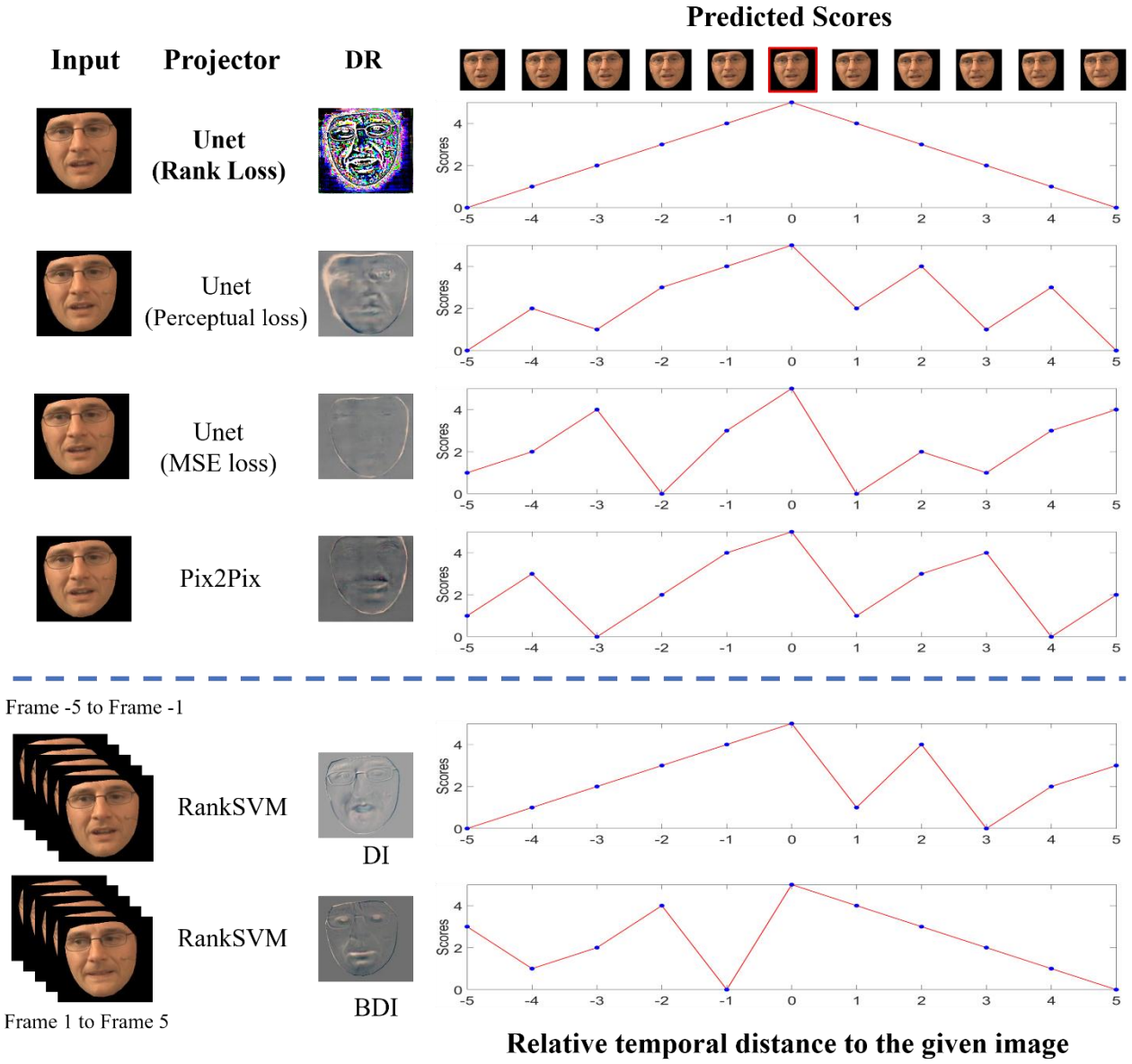


Fig. 3: Average ranking accuracy (%) on two datasets. Four generative models are trained using RECOLA dataset and tested on SEMAINE and BP4D datasets. RankSVM classifiers were trained on SEMAINE and BP4D datasets, and each classifier only rank its training frames. The results obtained by RankSVM are treated as the upper bound.



Results and Conclusion

	AU	6	10	12	14	17	Avg.
ICC	CCNN-IT [44]	0.75	0.69	0.86	0.40	0.45	0.63
	2DC [38]	0.76	0.71	0.85	0.45	**0.53**	0.66
	VGP-AE [45]	0.75	0.66	**0.88**	0.47	0.49	0.65
	HG-HMR [37]	**0.79**	0.80	0.86	0.54	0.43	0.68
	Pix2Pix* [17]	0.59	0.62	0.68	0.29	0.31	0.50
	Unet(P)* [43]	0.55	0.65	0.65	0.30	0.26	0.48
	Unet(MSE)*	0.56	0.63	0.66	0.29	0.26	0.48
	SDR+HG-HMR	0.78	0.80	0.85	0.47	0.45	0.67
	MDR+HG-HMR	0.77	**0.83**	0.87	**0.62**	0.49	**0.72**
MSE	CCNN-IT [44]	1.23	1.69	0.98	2.72	1.17	1.57
	2DC [38]	0.75	1.02	0.66	1.44	0.88	0.95
	VGP-AE [45]	0.82	1.28	0.70	1.43	0.77	1.00
	HG-HMR* [37]	**0.77**	0.92	0.65	1.57	0.77	0.94
	Pix2Pix* [17]	1.22	1.31	0.85	1.90	0.92	1.24
	Unet(P)* [43]	1.53	1.08	1.07	1.62	0.95	1.25
	Unet(MSE)*	1.09	1.55	1.18	2.12	1.15	1.42
	SDR+HG-HMR	0.88	0.84	0.75	1.90	0.60	0.99
	MDR+HG-HMR	0.99	**0.79**	**0.64**	**1.34**	**0.48**	**0.85**

TABLE I: AU intensities estimation results on BP4D dataset.
$*$ denotes results obtained by our own implementation

- The proposed Rank Loss can infer generic facial temporal evolution from previous unseen faces, and generalized better to unseen face images than a model trained using pre-defined representations.
- The proposed method has potential to push up the upper bound for still face image-based emotion and action units recognition.

Thank you

siyang.song@nottingham.ac.uk