University of Essex

The Human Rights, Big Data and Technology Project

DisplaceNet: Recognising Displaced People from Images by Exploiting Dominance Level

Grigorios Kalliatakis Shoaib Ehsan Maria Fasli Klaus McDonald-Maier
University of Essex, UK

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Code & Trained Models:
https://github.com/GKalliatakis/DisplaceNet

Overview

Objective: Inference of potential displaced people from real-world images.

Motivations:

- **68.5 million** forcibly displaced individuals worldwide - roughly equivalent to the entire UK population being forced to flee their homes.
- Traditional methods for human-rights-related image analysis require **manual labour** by human rights analysts and advocates.
- Computer vision can help **automate parts** of this process and turn recognition of displaced populations into a powerful and cost-effective application that could improve humanitarian responses.

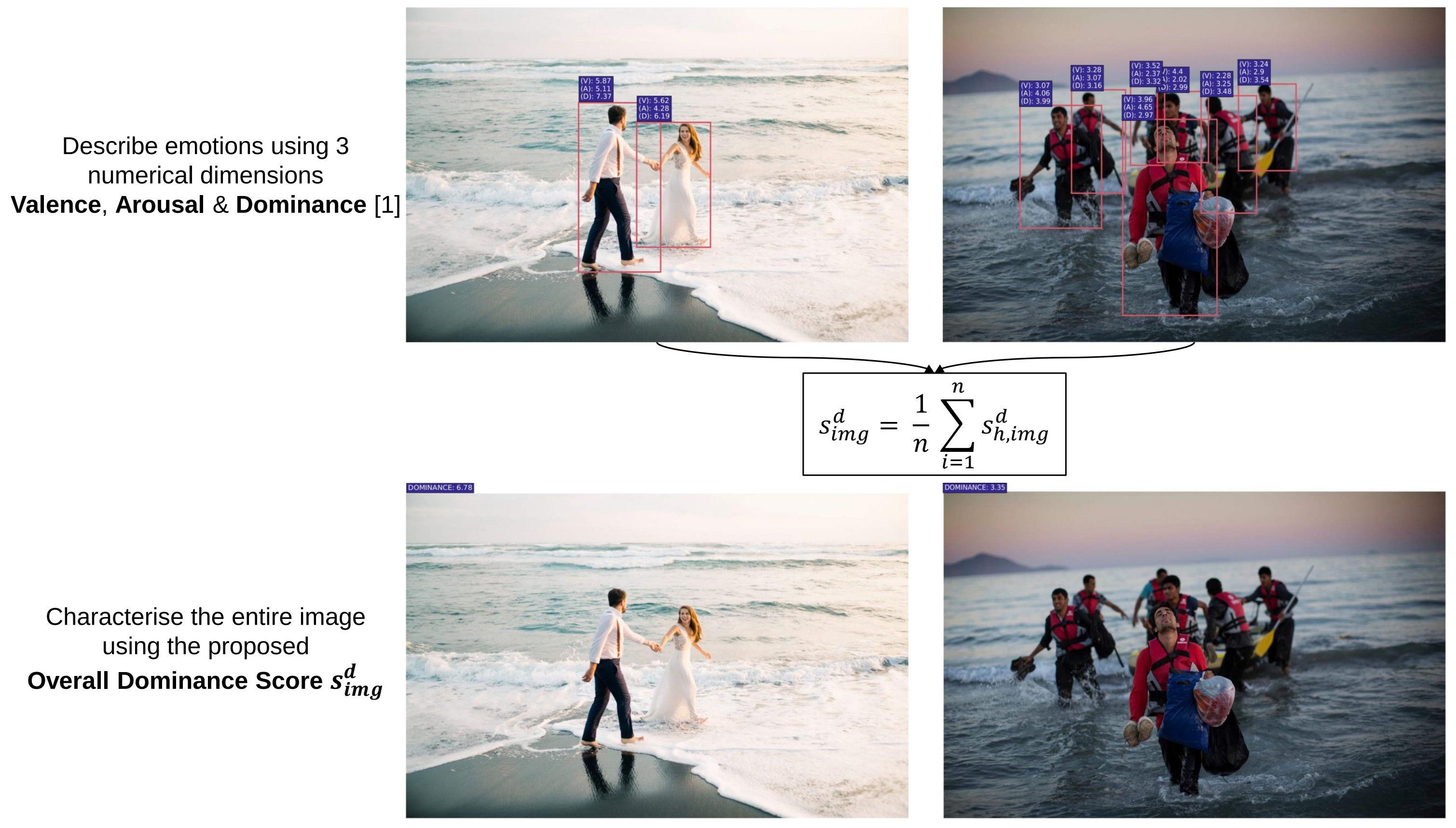
Problem Formulation

Can you label the images below as either displaced people or non-displaced people ?
Try to label them from the inference results of object detection and/or scene recognition.

Main Idea: A person's control level of a situation can be a notifying difference between the encoded visual content of an image that depicts a non-violent situation and the encoded visual content of an image displaying displaced people.

Overall Dominance Score

1. Combine the person bounding box with the information present in scene context similar to [1]–to recognise emotions expressed in continuous dimensions *Valence*, *Arousal* and *Dominance*
2. Introduce the **overall dominance score** that characterises an entire image based on all individuals' control level of the situation



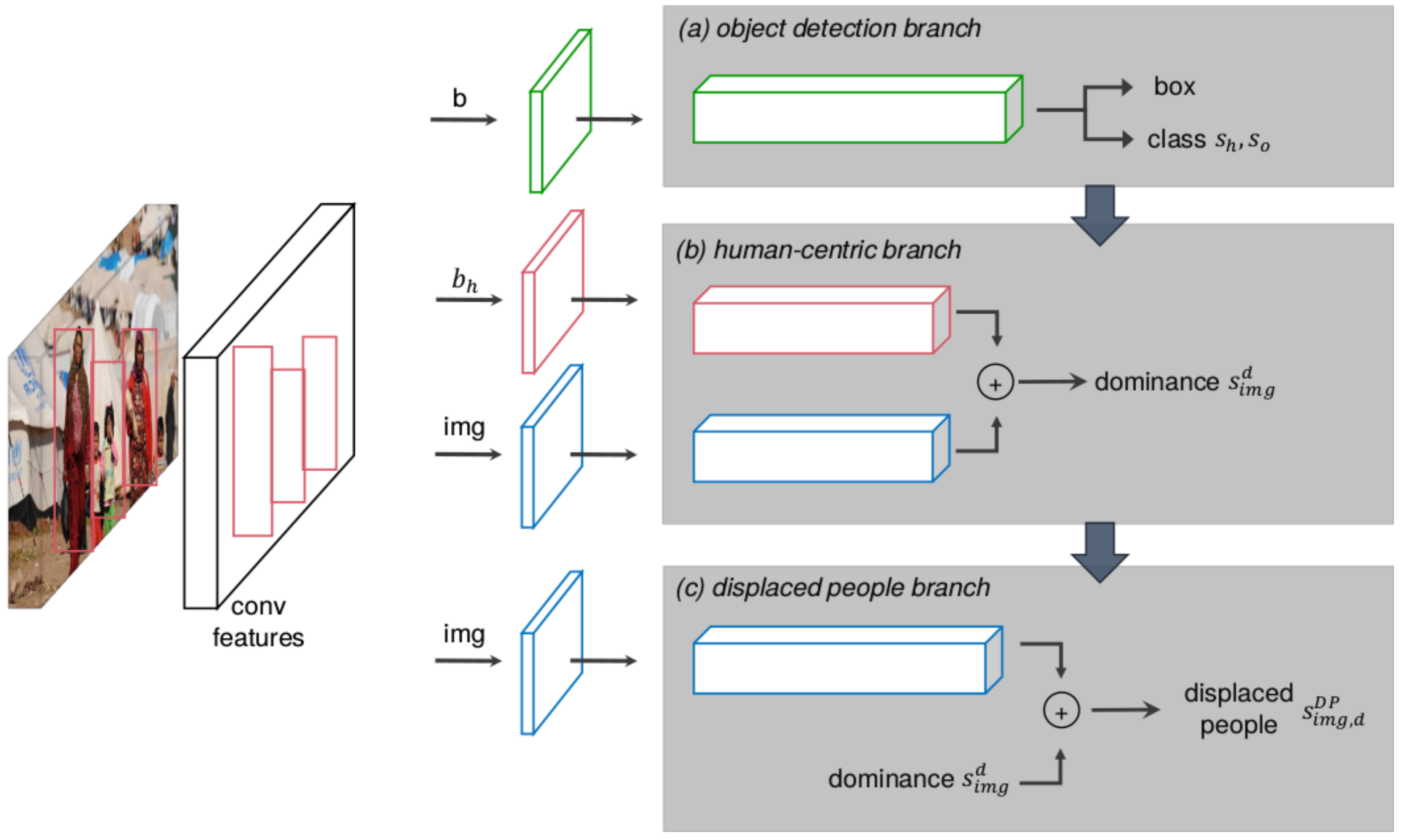
Describe emotions using 3 numerical dimensions **Valence**, **Arousal** & **Dominance** [1]

$$s^d_{img} = \frac{1}{n}\sum_{i=1}^{n} s^d_{h,img}$$

Characterise the entire image using the proposed **Overall Dominance Score** s^d_{img}

Method

Model Architecture:



Proposed Solution: Extend typical image classification by assigning a triplet score $s^{DP}_{img,d}$ to pairs of candidate human boxes b_h and the displaced people category

$$s^{DP}_{img,d} = s_h \cdot s^d_{h,img} \cdot s^{DP}_{img}$$

Components:

Object Detection Branch: Localise the boxes containing a human b_h and the object of interaction b_o using RetinaNet [2].

Human-centric Branch:

- VAD score for each b_h.
- Dominance score s^d_{img} that characterises the **entire** image.

Displaced People Branch:

- Classification score for input image.
- Re-adjust classification score based on s^d_{img}.

Experiments & Results

Quantitative Comparison of vanilla CNNs* & DisplaceNet:

backbone network	layers fine-tuned	vanilla CNN Top-1 acc.	vanilla CNN Coverage	**DisplaceNet** Top-1 acc.	**DisplaceNet Coverage**
VGG16	1	58%	0%	54%	**3%**
VGG19		69%	3%	60%	**6%**
ResNet50		60%	0%	55%	**4%**
VGG16	2	63%	43%	63%	**49%**
VGG19		77%	54%	74%	**58%**
ResNet50		42%	1%	38%	**5%**
mean	-	61.5%	16.83%	57.33%	**20.83%**

Qualitative Results:

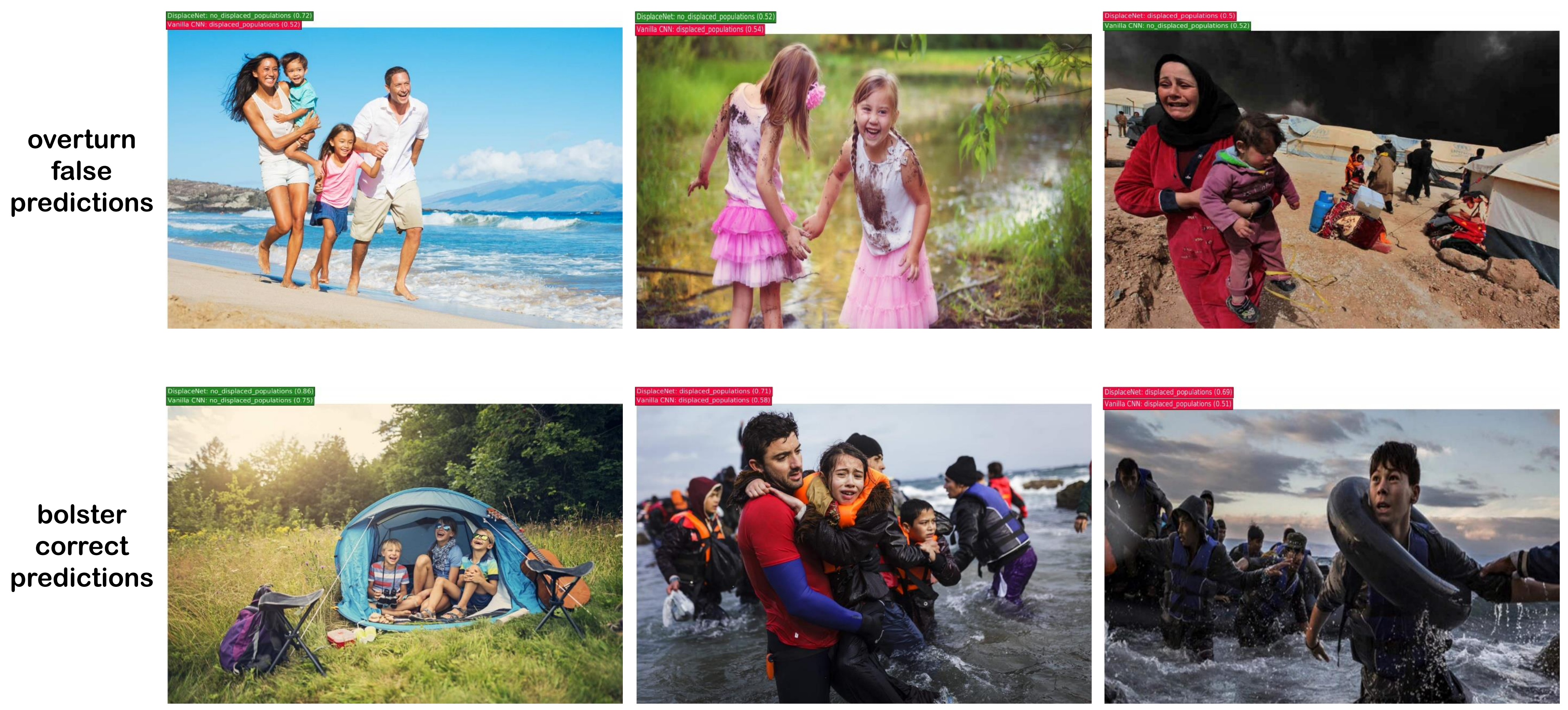

overturn false predictions
DisplaceNet: no_displaced_populations (0.72)
Vanilla CNN: displaced_populations (0.52)
DisplaceNet: no_displaced_populations (0.52)
Vanilla CNN: displaced_populations (0.54)
DisplaceNet: displaced_populations (0.5)
Vanilla CNN: no_displaced_populations (0.52)
bolster correct predictions
DisplaceNet: no_displaced_populations (0.86)
Vanilla CNN: no_displaced_populations (0.75)
DisplaceNet: displaced_populations (0.71)
Vanilla CNN: displaced_populations (0.58)
DisplaceNet: displaced_populations (0.69)
Vanilla CNN: displaced_populations (0.51)

*image classification using solely fine-tuning without any other modification

Dataset & Metrics

- Two-class subset of Human Rights Archive Dataset [3].
- Use of ***coverage***–proportion of a dataset for which a classifier is able to produce a prediction–as a realistic performance metric.
- DisplaceNet refuses to classify an input x, whenever the probability of the output sequence $p(y|x) < t$ for some confidence threshold $t = 0.85$.

References

[1] Emotion Recognition in Context [Kosti *et al.*, CVPR17]

[2] Focal loss for dense object detection [Lin *et al.*, ICCV17]

[3] Exploring Object-Centric and Scene-Centric CNN Features and Their Complementarity for Human Rights Violations Recognition in Images [Kalliatakis *et al.*, IEEE Access 2019]

Acknowledgements

This work is supported by the UK ESRC through grant ES/M010236/1 & EPSRC through grants EP/R02572X/1 and EP/P017487/1