WNet: Joint multiple head detection and head pose estimation from a spectator crowd image

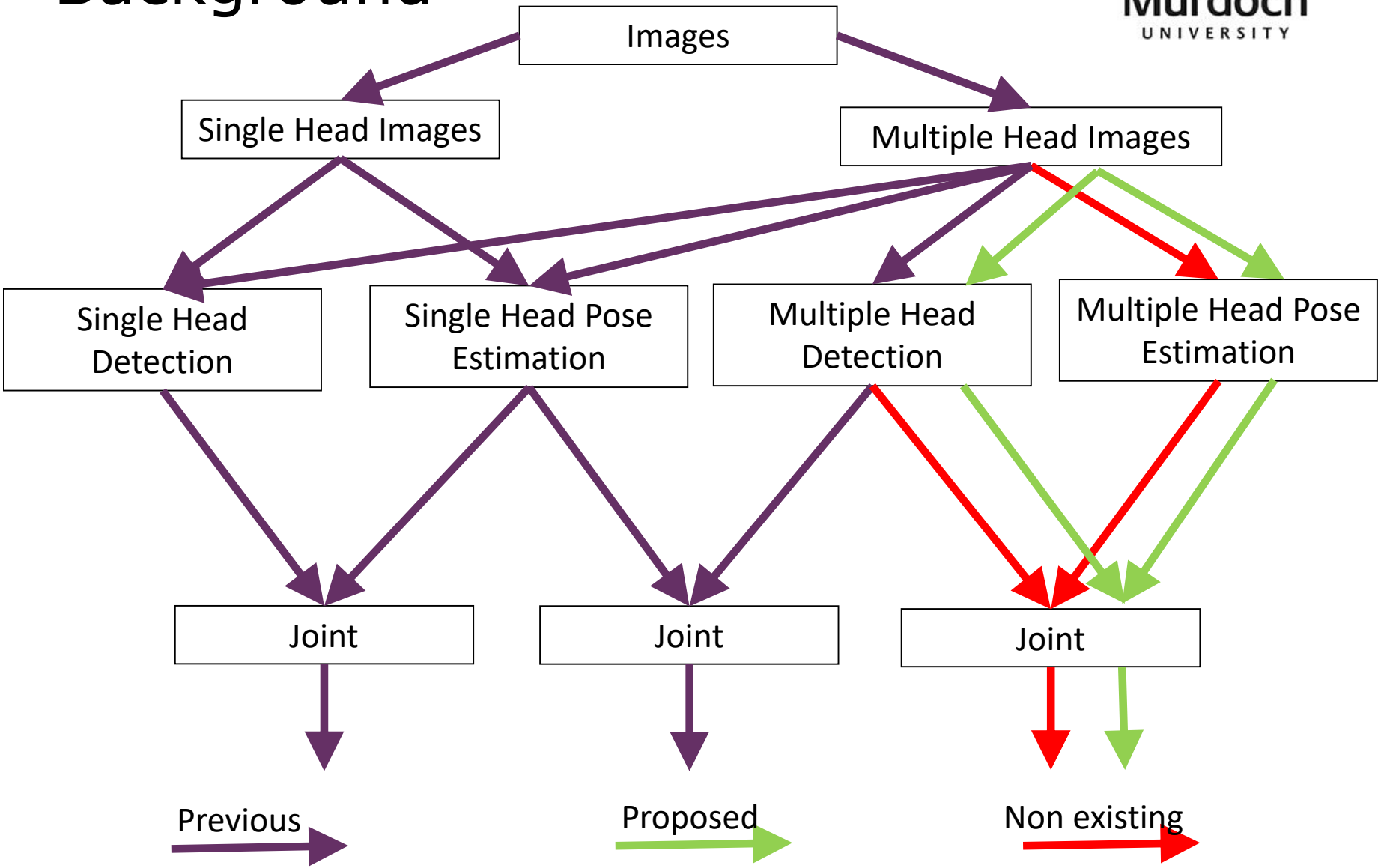
Yasir Jan
Ferdous Sohel
Mohd Fairuz Shiratuddin
Kok Wai Wong

ACCV 2018 / AMV 2018

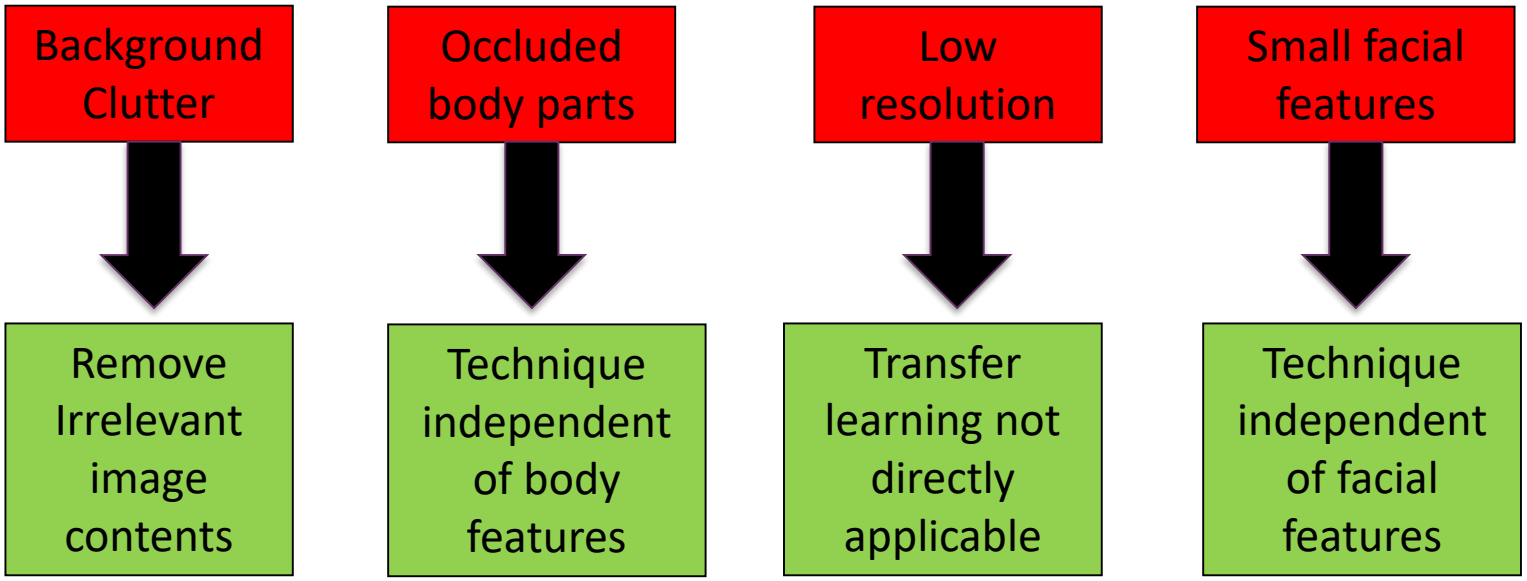
Contents

- Background
- Issues
- Inspiration
- WNet Architecture
- Dataset
- WNet Training & Testing
- Results
- Conclusion

Background

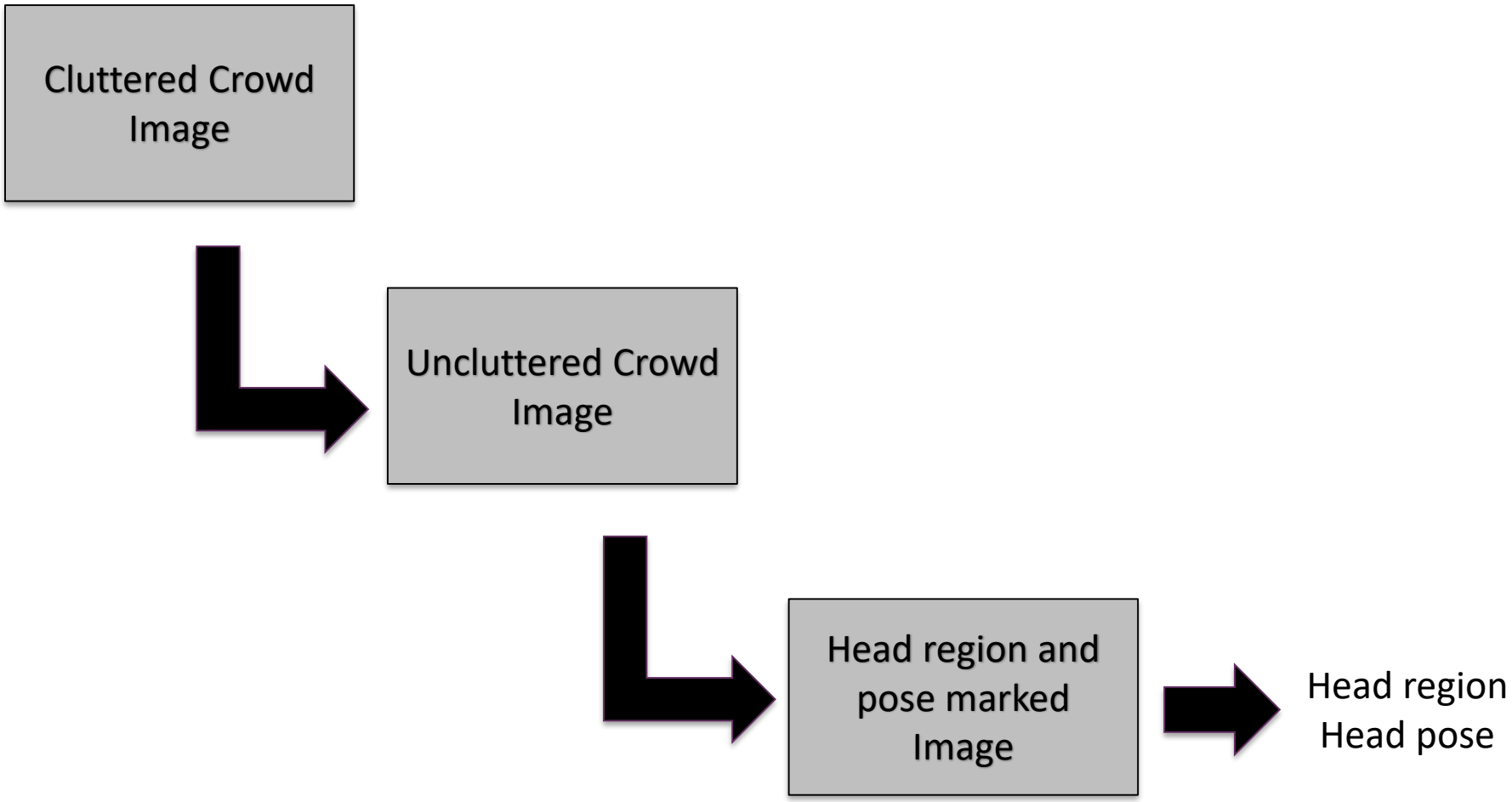

Images
Single Head Images
Multiple Head Images
Single Head Detection
Single Head Pose Estimation
Multiple Head Detection
Multiple Head Pose Estimation
Joint
Joint
Joint
Previous
Proposed
Non existing

Issues


Background Clutter
Occluded body parts
Low resolution
Small facial features
Remove Irrelevant image contents
Technique independent of body features
Transfer learning not directly applicable
Technique independent of facial features

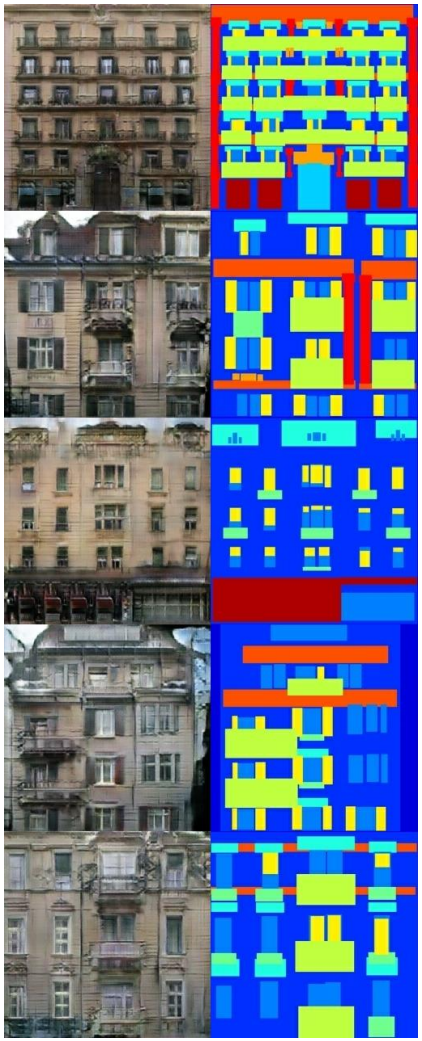
Basic Solution

Murdoch UNIVERSITY



UNet[1] Inspiration

Murdoch UNIVERSITY



Pix2pix Network, based on UNet, can convert complex facade features into simplistic colourful patterns.

https://affinelayer.com/pixsrv/

[1] Image-to-Image Translation with Conditional Adversarial Networks
Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros,
2017 IEEE Conference on Computer Vision and Pattern Recognition

WNet Steps

1) UNet 1: Converts a complex crowd image into less cluttered image, which has only head region left, while other regions are blacked out

Original Input

UNet

Head Region Masked (HRM) Output

WNet Steps (contd …)

2) HRM image is converted into a grayscaled image, so that the region colors don’t effect the next UNet block.

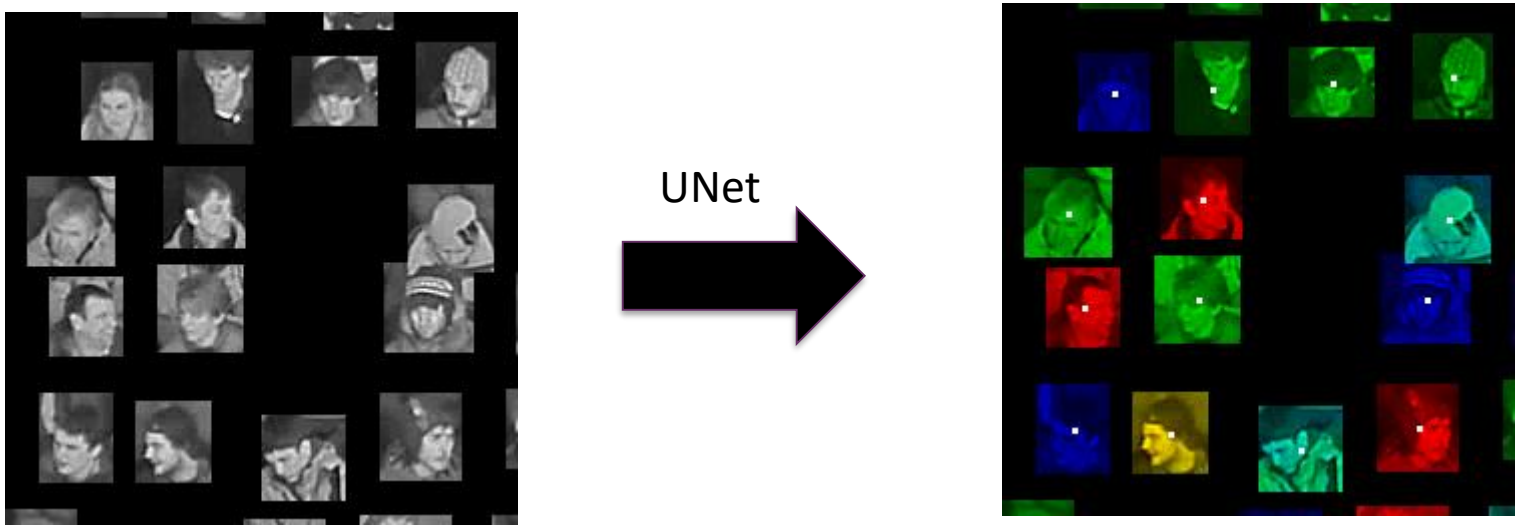
HRM

All channels
Avg of channels

Grayscaled HRM

WNet Steps (contd …)

3) UNet 2: Takes input grayscaled HRM images and generates coloured coded head (CCH) images.



Grayscaled HRM

CCH image

Colours are based on head poses.
Centres are marked with 3x3 white pixels

Colour codes

Murdoch UNIVERSITY

Head pose	Colour code
Left	Red
Right	Green
Front	Blue
Away	Red, Green
Down	Green, Blue

WNet Steps (contd…)

Murdoch UNIVERSITY

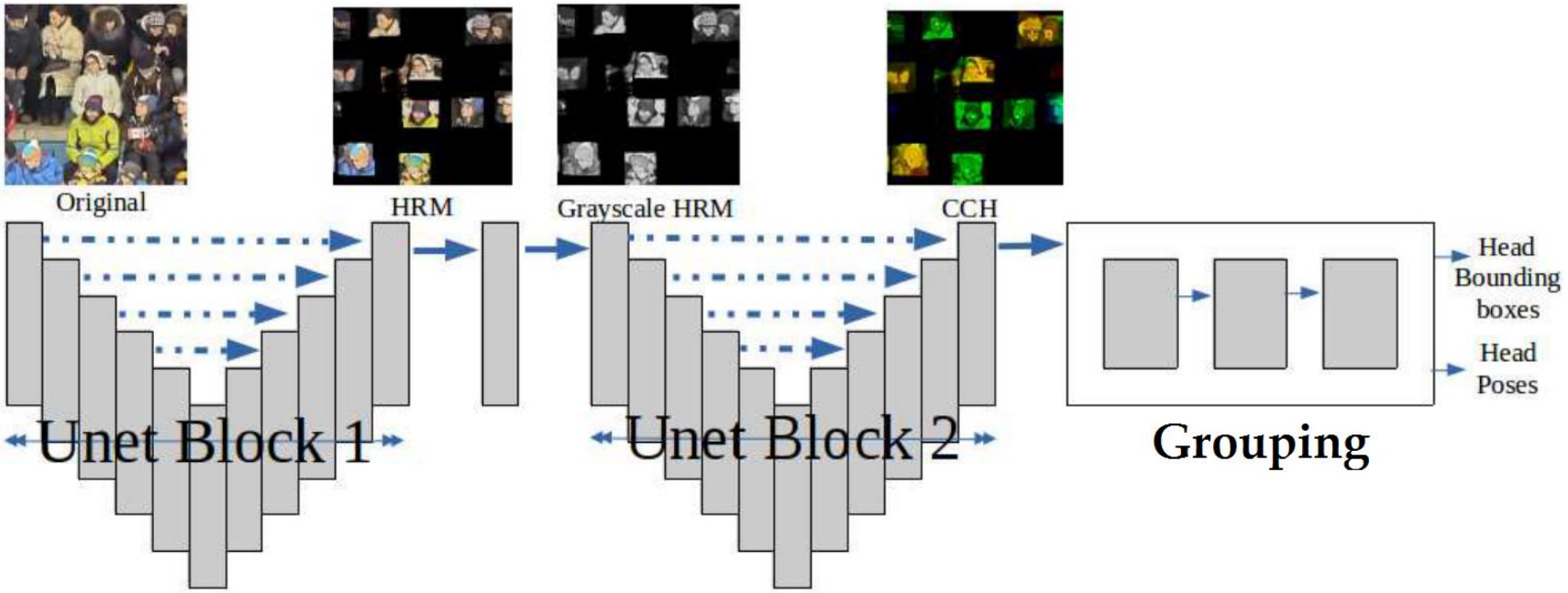
4) Head centres : Locating 3x3 white pixel region

5) Head bounding box : Coloured region around the centre within the head distance range

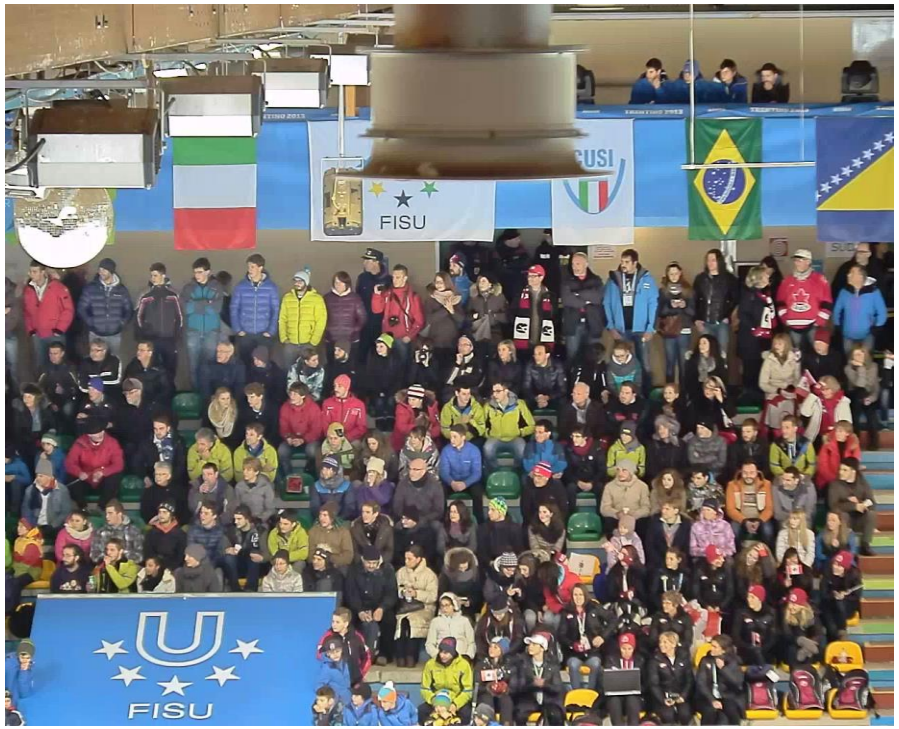
6) Head pose : Maximum colour coding within a head region

WNet Pipeline

Murdoch UNIVERSITY



Dataset
S-HOCK (Spectator Hockey)



S-HOCK dataset frame

The S-Hock dataset: A new benchmark for spectator crowd analysis,
Francesco Setti, Davide Conigliaro, Paolo Rota, Chiara Bassetti, Nicola Conci, Nicu Sebe and Marco Cristani
Computer Vision and Image Understanding, 2017

S-HOCK Videos

Murdoch UNIVERSITY

Hockey Matches

Camera views

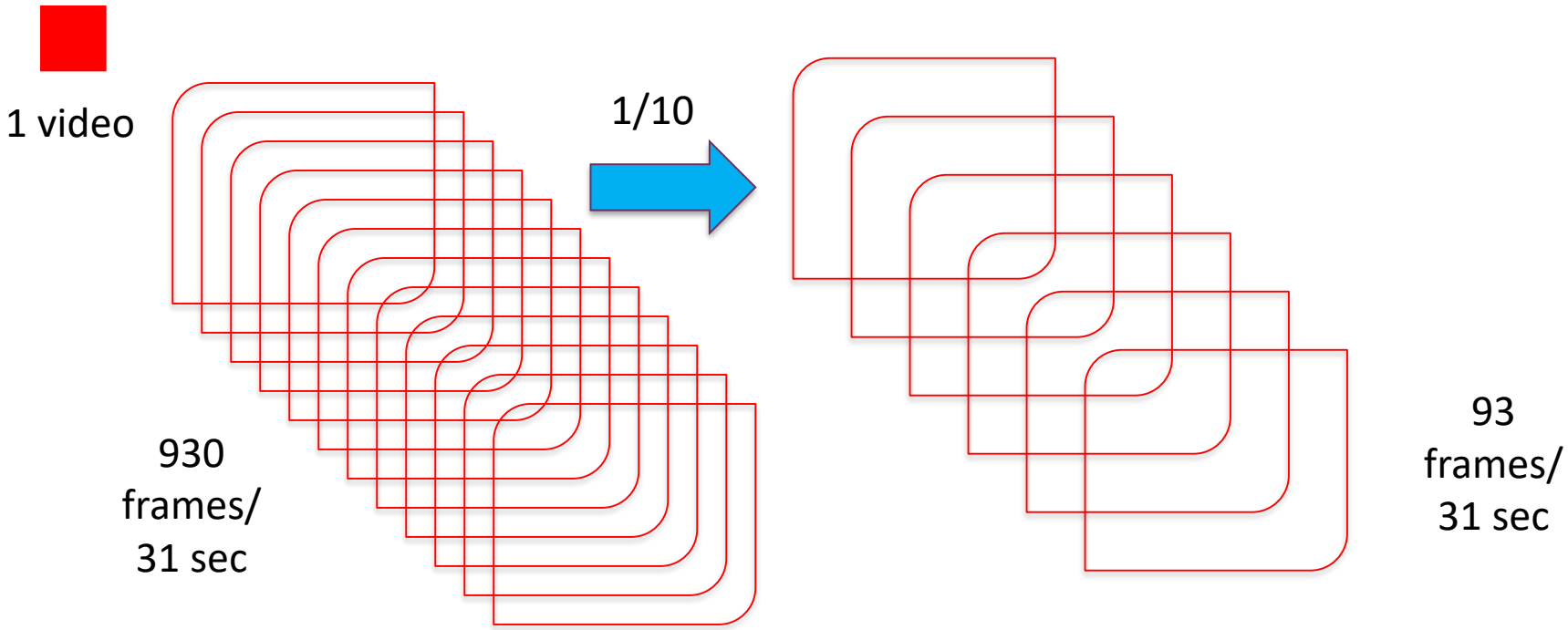
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2			A	N	N	O	T	A	T	E	D				
3															
4															
5															

Training

Validation

Testing

S-HOCK Frames



Resolution: 1280 x 1024 pixels
People Count: Maximum 160 people

WNet Training & Testing

Murdoch UNIVERSITY

UNet 1		UNet 2
• 186 frames : 93 frames x 2 videos • 3720 subframes : 186 x 20 slices • Each slice : 256 x 256 pixels		• Class imbalance is reduced by adding horizontally flipped subframes to the training set. • 372 frames : 93 frames x 2 flips x 2 videos • 7440 subframes : 372 frames x 20 slices • Each slice : 256 x 256 pixels

Testing

- 20460 subframes: 93 frames x 11 videos x 20 slices

Image pairs

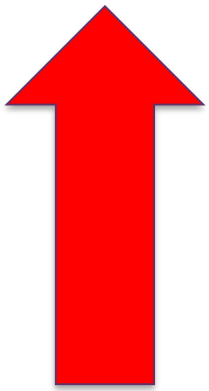
Murdoch UNIVERSITY

WNet Input/Output Pair

Testing Protocol

Murdoch UNIVERSITY

Benchmarks

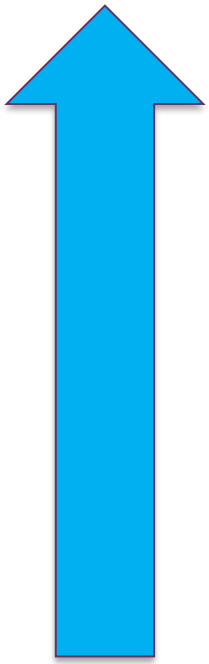


34949 cropped head images

ONLY HPE

WNet

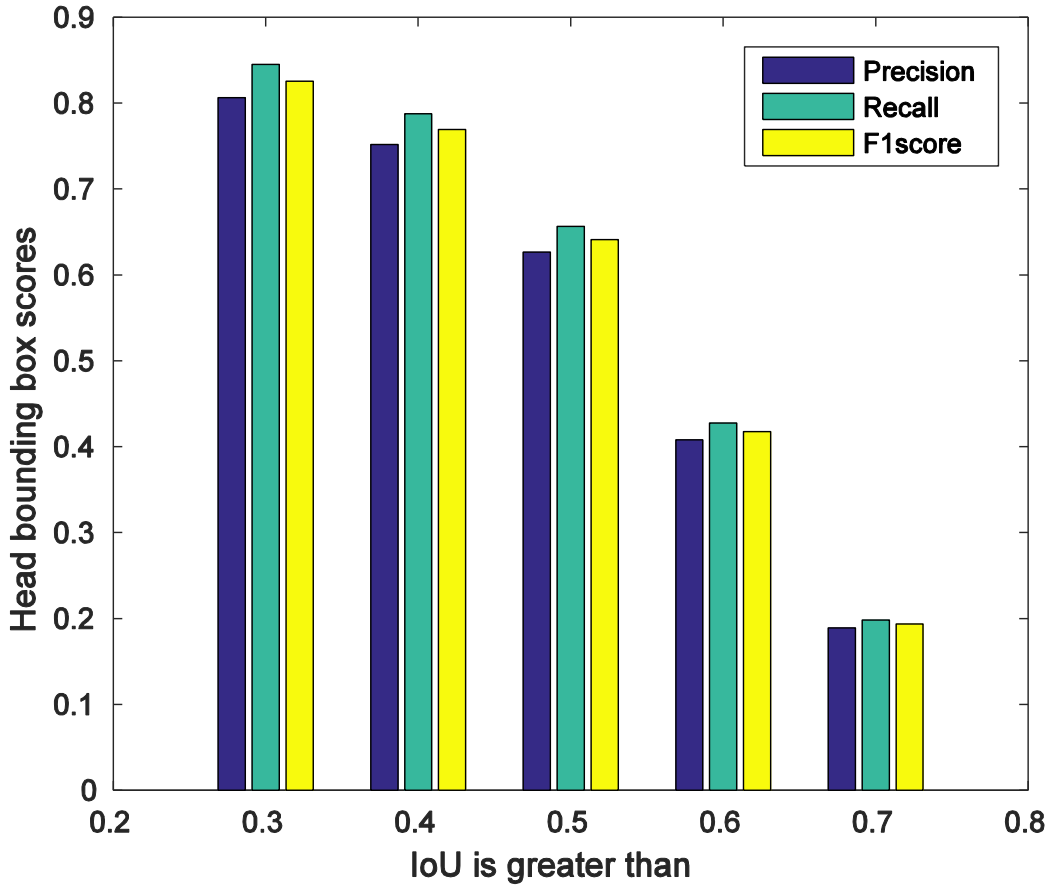
20460 subframes



155085 heads

HD and HPE

Results (Head Detection)



For IoU = (0.3, 0.4, 0.5, 0.6, 0.7) , HD ~= (0.8, 0.7, 0.6, 0.4, 02)

Results (Head Pose Estimation)

Murdoch UNIVERSITY

Method	Avg Accuracy	Acc HPE headcount	Method type
Orozco	0.368	~ 12861	Only single HPE
WArCo	0.376	~ 13140	Only single HPE
CNN	0.346	~ 12092	Only single HPE
SAE	0.348	~ 12162	Only single HPE
WNet (IoU = 0.3)	0.321	**~ 39825**	**Joint multiple HD and HPE**
WNet (IoU = 0.4)	0.323	**~ 35064**	**Joint multiple HD and HPE**
WNet (IoU = 0.5)	0.325	**~ 30241**	**Joint multiple HD and HPE**
WNet (IoU = 0.6)	0.337	**~ 20905**	**Joint multiple HD and HPE**
WNet (IoU = 0.7)	0.34	~ 10545	Joint multiple HD and HPE

HPE Accuracies are ratio to correct HD

Conclusion

Murdoch UNIVERSITY

- Proposed WNet architecture can perform joint head detection and head pose estimation of multiple heads
- WNet converts dense crowd cluttered image into simple color coded head images
- WNet uses lesser number of frames to detect more number of heads and head poses

Thank you

Murdoch
UNIVERSITY