Biometric Technologies Laboratory

UNIVERSITY OF CALGARY

Relatable Clothing: Detecting Visual Relationships between People and Clothing

Thomas Truong, Svetlana Yanushkevich
Department of Electrical and Computer Engineering
Presentation for ICPR 2020

Friday, November 27, 2020

Introduction

- **Motivation:**
  - Datasets for visual relationships related to clothing are lacking.
    - Consequently, detection models for clothing relationships are also lacking.
- **Research Contributions:**
  - To release a large dataset, the Relatable Clothing Dataset, which can be used for detecting visual relationships between people and worn/unworn clothing.
  - To propose and test a novel model architecture for soft attention and visual relationship detection.

Presentation Outline

- Related works
- Visual Relationship Detection
- Relatable Clothing Dataset
- Soft-attention unit
- Results
- Conclusion

Related Works

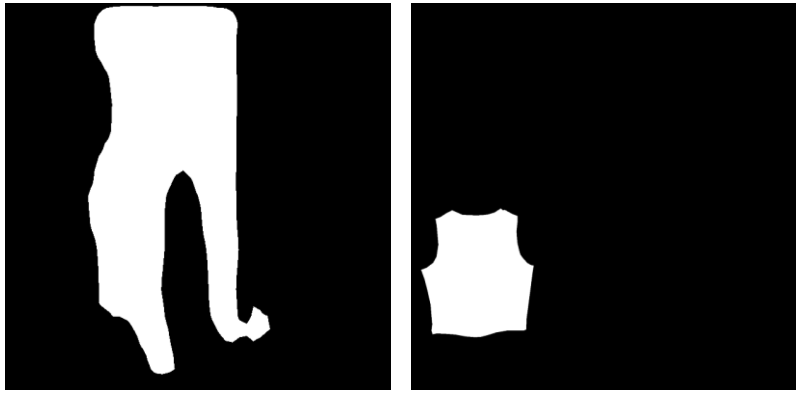
- Verbs in COCO (V-COCO) is the most popular visual relationship detection dataset.
  - Very large dataset but does not contain labels for clothing and whether they are worn/unworn.
- Open Images is another popular visual relationship detection dataset
  - Contains the label “wears” for accessories such as sunglasses, necklaces, purses, and shoes, but no clothing is labeled.
- DeepFashion2 contains labels for worn/unworn clothing.
  - There are a few problems to address first before this dataset is useable to train a worn/unworn clothing detector.

Fig: Example sample from V-COCO

Fig: Example sample from DeepFashion2

Visual Relationship Detection

- Visual Relationship Triplet, <subject, predicate, object> or <S,P,O>.
  - Derived from grammar, subject is the “who/what”, predicate is the “verb” or “relationship”, and object is often a noun which is described in conjunction with the subject and predicate.
- We are concerned with predicate prediction in this paper.
  - The subject will always be a person and the object will always be an article of clothing.

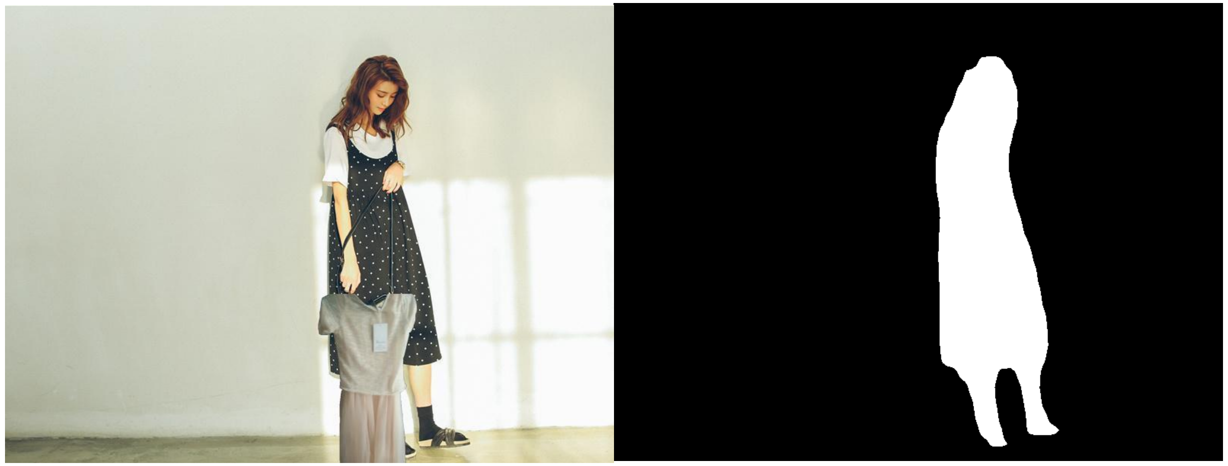


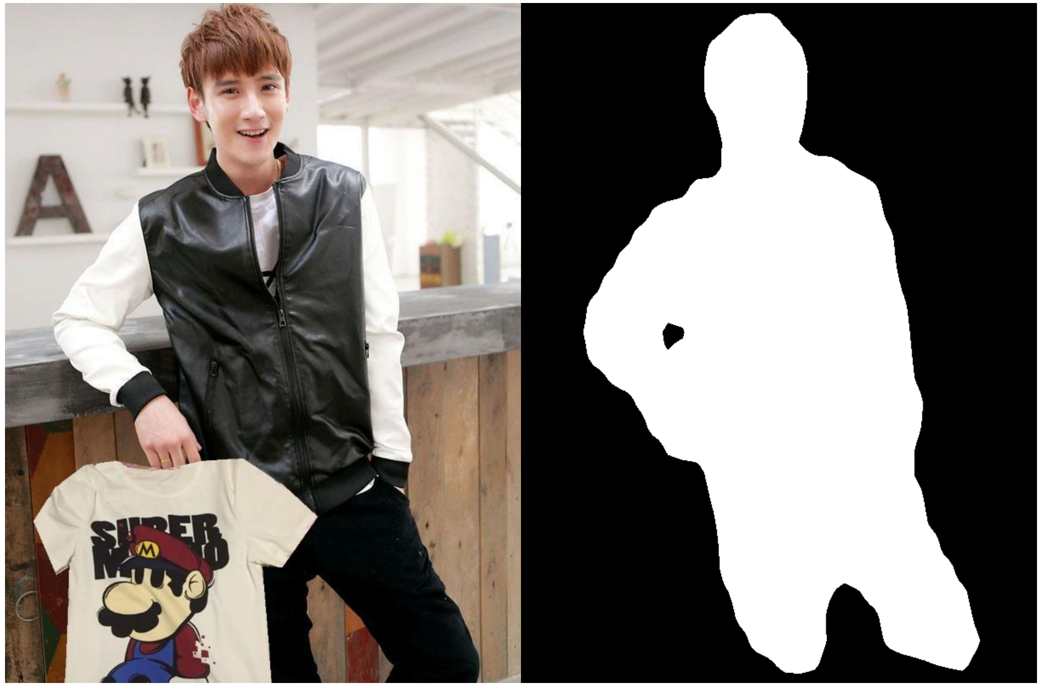
Relatable Clothing Dataset

- DeepFashion2 Dataset lacks two important features that are necessary to be used for visual relationship detection:
  - Subject segmentations. No person is segmented in this dataset.
  - Unworn articles of clothing are close-ups and do not contain any people in the image.
- We propose the Relatable Clothing Dataset for worn/unworn clothing classification problems.
  - A modified subset of the DeepFashion2 Dataset

Relatable Clothing Dataset

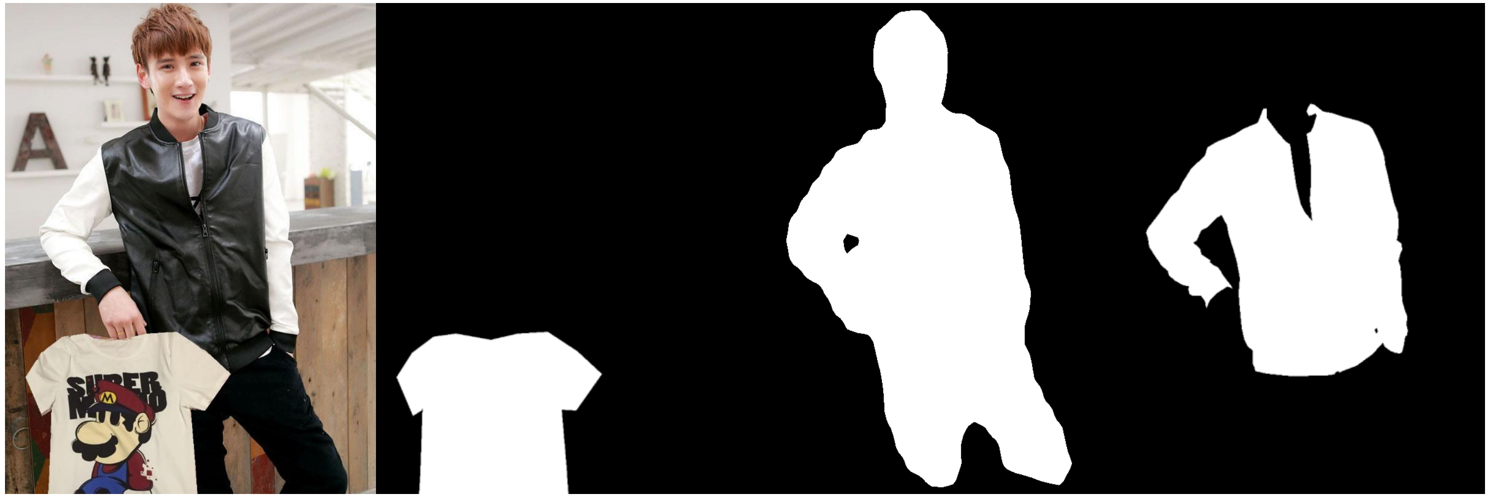
- **Subject segmentations. No person is segmented in this dataset.**
- Unworn articles of clothing are close-ups and do not contain any people in the image.





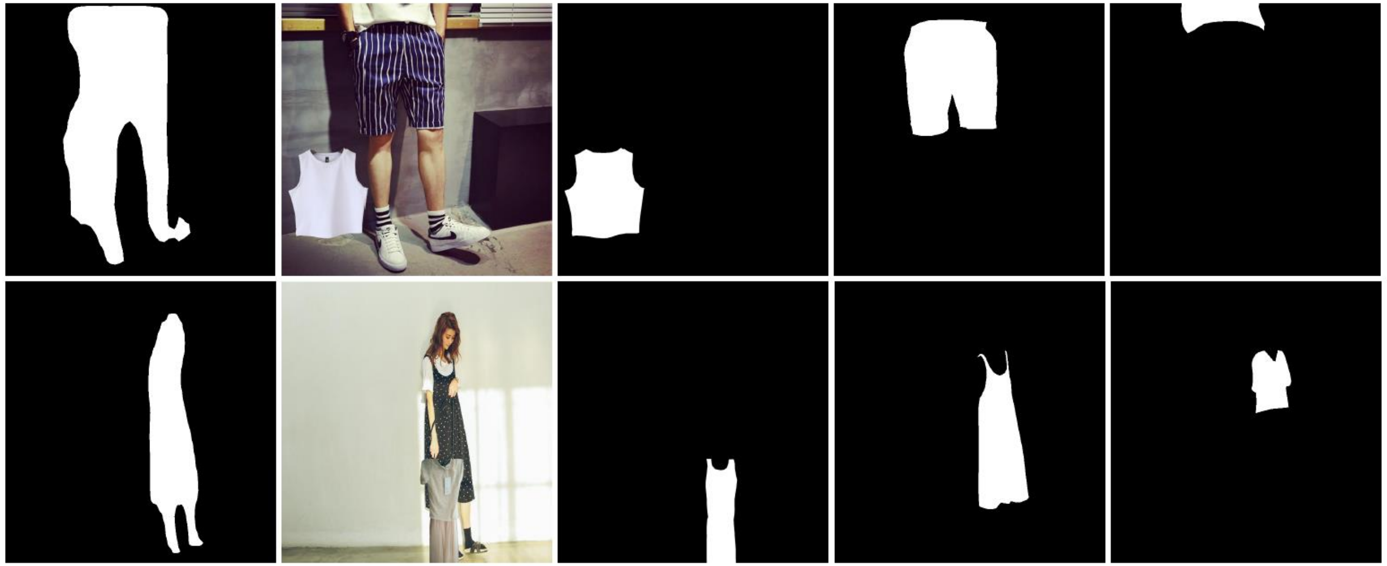
Relatable Clothing Dataset

- Subject segmentations. No person is segmented in this dataset.
- **Unworn articles of clothing are close-ups and do not contain any people in the image.**



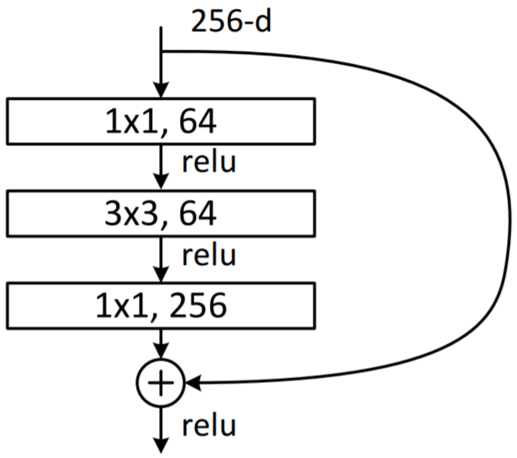
Relatable Clothing Dataset

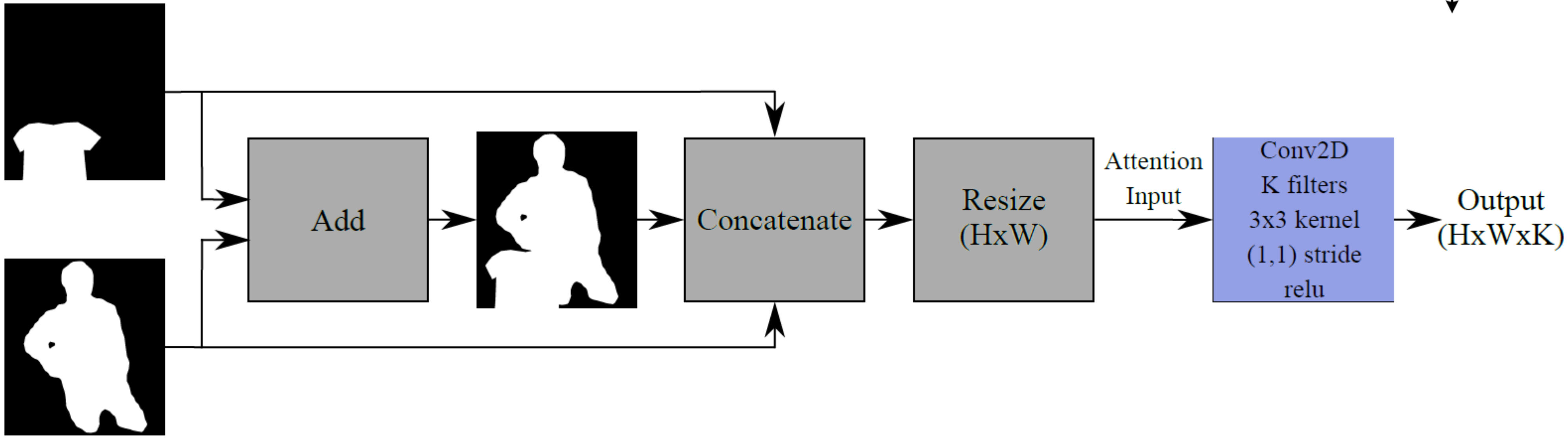
- 29852 person-clothing pairs (18726 “worn” and 11126 “unworn”) available for training
- 5705 person-clothing pairs (3604 “worn” and 2101 “unworn”) for validation and testing



Soft-Attention Unit

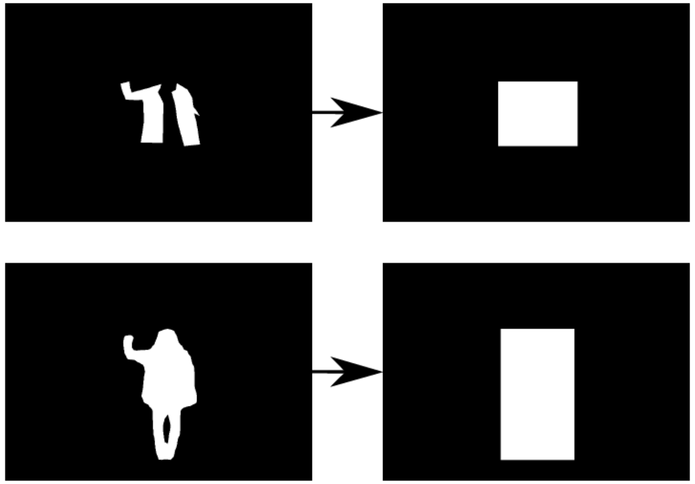
- A trainable unit which guides the "attention" of the network to the areas containing masks.
  - The Output is added to the output of the 3x3 convolutional layer of each bottleneck unit in ResNet.

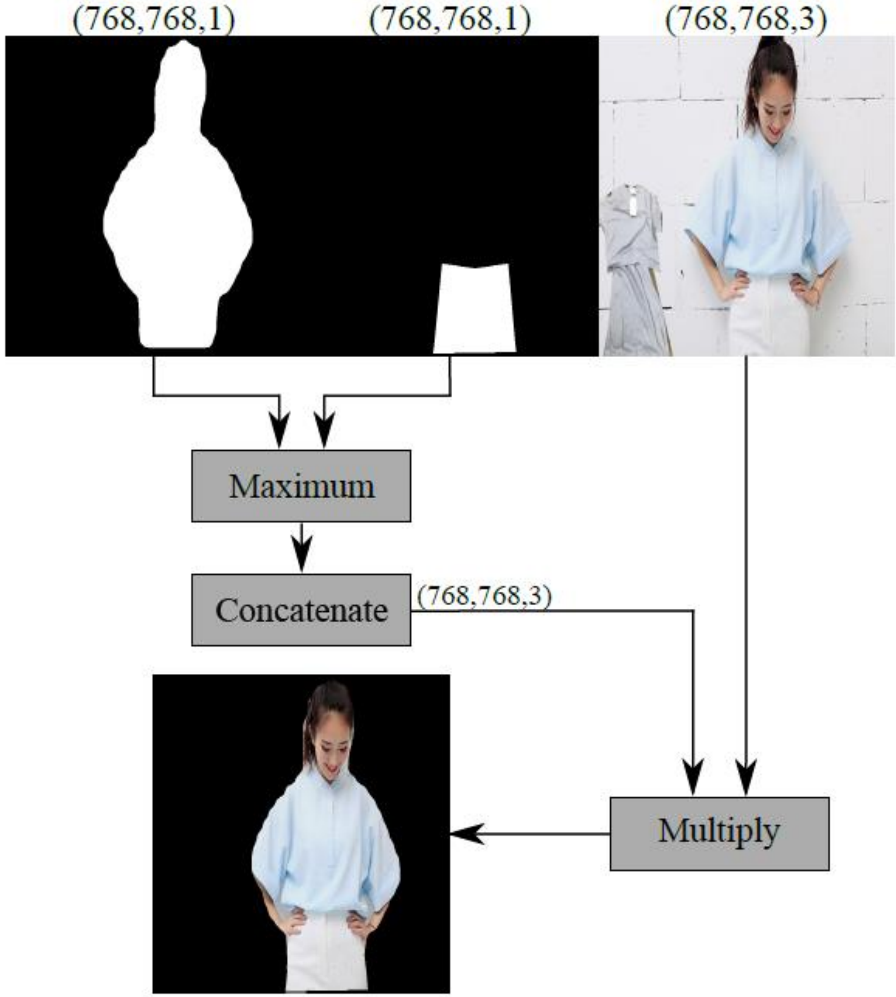




Baseline models

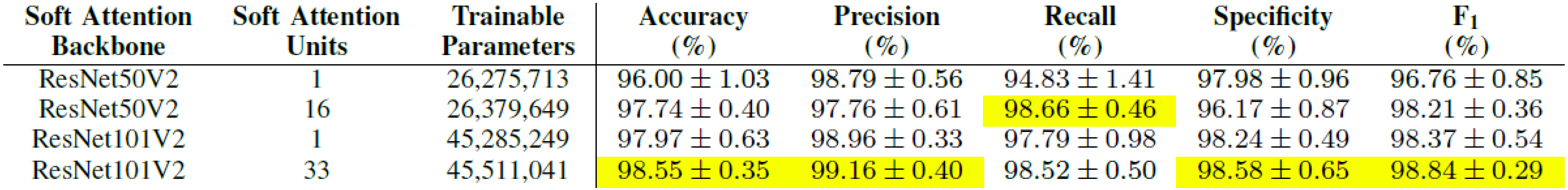
- Hard-attention model
    - Primitive masking of the input image using the masks to provide a basic attention mechanism.
- Box soft-attention model
    - Similar to previous works who use bounding box detections to do visual relationship detection.





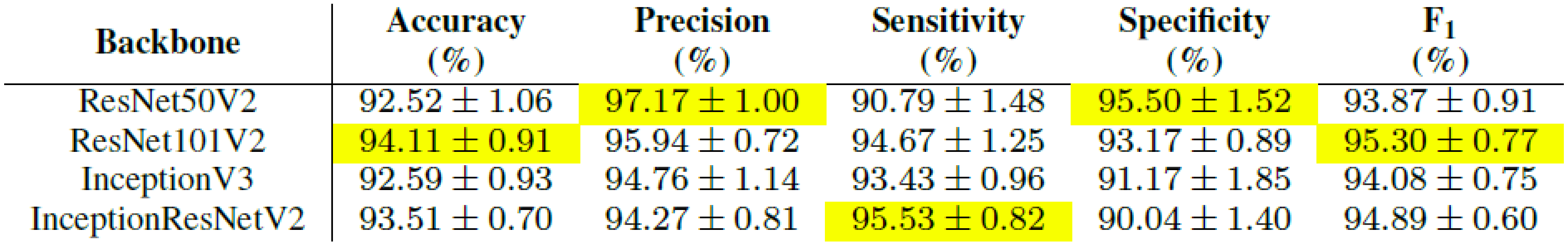
Results

PERFORMANCE METRICS FOR THE PROPOSED SOFT ATTENTION MODELS.



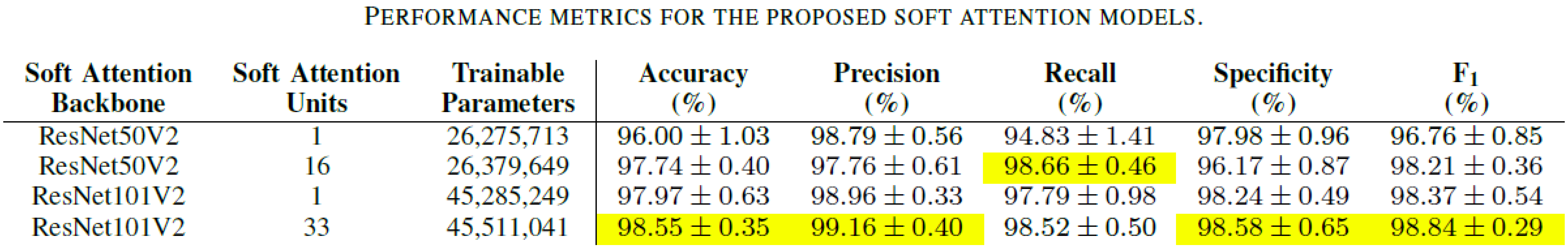
Soft Attention Backbone	Soft Attention Units	Trainable Parameters	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1 (%)
ResNet50V2	1	26,275,713	96.00 ± 1.03	98.79 ± 0.56	94.83 ± 1.41	97.98 ± 0.96	96.76 ± 0.85
ResNet50V2	16	26,379,649	97.74 ± 0.40	97.76 ± 0.61	98.66 ± 0.46	96.17 ± 0.87	98.21 ± 0.36
ResNet101V2	1	45,285,249	97.97 ± 0.63	98.96 ± 0.33	97.79 ± 0.98	98.24 ± 0.49	98.37 ± 0.54
ResNet101V2	33	45,511,041	98.55 ± 0.35	99.16 ± 0.40	98.52 ± 0.50	98.58 ± 0.65	98.84 ± 0.29

PERFORMANCE METRICS FOR THE HARD ATTENTION MODELS.



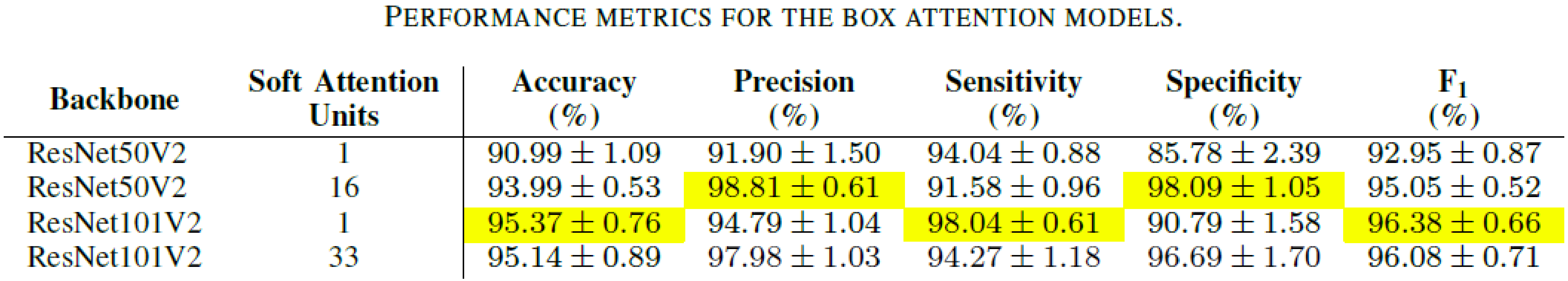
Backbone	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F_1 (%)
ResNet50V2	92.52 ± 1.06	97.17 ± 1.00	90.79 ± 1.48	95.50 ± 1.52	93.87 ± 0.91
ResNet101V2	94.11 ± 0.91	95.94 ± 0.72	94.67 ± 1.25	93.17 ± 0.89	95.30 ± 0.77
InceptionV3	92.59 ± 0.93	94.76 ± 1.14	93.43 ± 0.96	91.17 ± 1.85	94.08 ± 0.75
InceptionResNetV2	93.51 ± 0.70	94.27 ± 0.81	95.53 ± 0.82	90.04 ± 1.40	94.89 ± 0.60

Results



PERFORMANCE METRICS FOR THE PROPOSED SOFT ATTENTION MODELS.

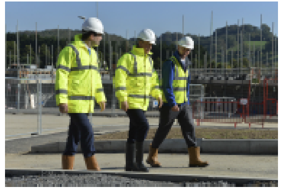
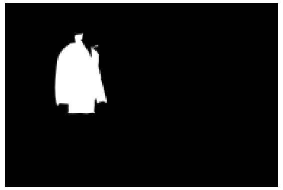
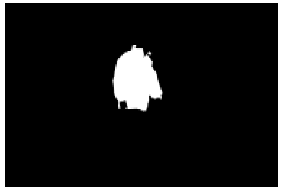
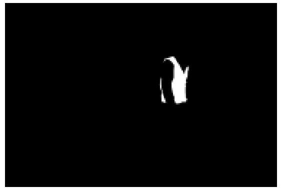
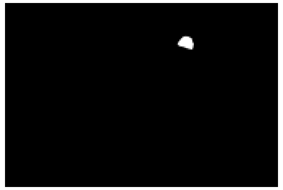
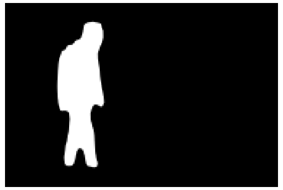
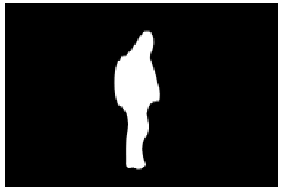
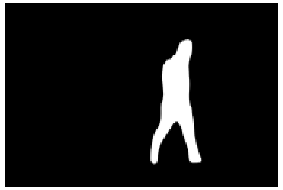
Soft Attention Backbone	Soft Attention Units	Trainable Parameters	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1 (%)
ResNet50V2	1	26,275,713	96.00 ± 1.03	98.79 ± 0.56	94.83 ± 1.41	97.98 ± 0.96	96.76 ± 0.85
ResNet50V2	16	26,379,649	97.74 ± 0.40	97.76 ± 0.61	98.66 ± 0.46	96.17 ± 0.87	98.21 ± 0.36
ResNet101V2	1	45,285,249	97.97 ± 0.63	98.96 ± 0.33	97.79 ± 0.98	98.24 ± 0.49	98.37 ± 0.54
ResNet101V2	33	45,511,041	98.55 ± 0.35	99.16 ± 0.40	98.52 ± 0.50	98.58 ± 0.65	98.84 ± 0.29



PERFORMANCE METRICS FOR THE BOX ATTENTION MODELS.

Backbone	Soft Attention Units	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F_1 (%)
ResNet50V2	1	90.99 ± 1.09	91.90 ± 1.50	94.04 ± 0.88	85.78 ± 2.39	92.95 ± 0.87
ResNet50V2	16	93.99 ± 0.53	98.81 ± 0.61	91.58 ± 0.96	98.09 ± 1.05	95.05 ± 0.52
ResNet101V2	1	95.37 ± 0.76	94.79 ± 1.04	98.04 ± 0.61	90.79 ± 1.58	96.38 ± 0.66
ResNet101V2	33	95.14 ± 0.89	97.98 ± 1.03	94.27 ± 1.18	96.69 ± 1.70	96.08 ± 0.71

Results

Input Image	Person 1 Vest	Person 2 Vest	Person 3 Vest	Person 1 Helmet	Person 2 Helmet	Person 3 Helmet
						
Person 1 	1.000	0.003	0.527	1.000	0.001	0.889
Person 2 	0.100	1.000	0.449	0.848	1.000	0.884
Person 3 	0.069	0.000	1.000	0.697	0.000	1.000

Conclusions and Future Work

- Release of the Relatable Clothing Dataset
  - 29852 person-clothing pairs for training, 5705 person-clothing pairs for validation and testing.
- Proposal of a novel soft-attention unit for visual relationship detection.
  - Demonstrated good performance for worn/unworn clothing detection on the Relatable Clothing Dataset and decent generalizability on unseen articles of clothing.
- Currently extending these works for full end-to-end object detection and visual relationship detection for applications in safety and security.

Acknowledgements

- NSERC “Biometric-enabled Identity Management and Risk Assessment for Smart Cities”
- Department of National Defence’s Innovation for Defence Excellence and Security (IDEaS)