Detecting Subject-Weapon Visual Relationships

Thomas Truong and Svetlana Yanushkevich
Biometric Technologies Laboratory
Department of Electrical and Computer Engineering
University of Calgary, Canada
{thomas.truong, syanshk}@ucalgary.ca

***Abstract*—Computer vision-based weapon detection methodologies and applications in safety and security have fallen behind when compared to state-of-the-art computer vision applications problems in other areas. In this paper we propose a novel visual relationship detection model trained on the Open Images V6 dataset to detect the visual relationships of "holds" and "wears" between people and objects. We also introduce an application of the proposed model to detect if weapons are being held. Weapons are an unseen object class to the network. The best proposed model achieves an accuracy of 90.01% ± 2.05% on the test set of the Open Images V6 dataset for classifying the "holds" and "wears" visual relationships.**

I. Introduction

Object detection and visual relationship detection datasets and methodologies have progressed to the point where practical applications are of interest to explore for research. Weapon detection and visual relationship detection for weapons has fallen far behind the state-of-the-art methodologies in computer vision research. Applications of these state-of-the-art methodologies is promising in the safety and security field to improve civilian safety and provide assistive technologies for early warning systems. As such, the objective of this paper is to explore state-of-the-art computer vision methodologies and apply it to developing a visual relationship detection model to be eventually applied to safety and security problems.

We seek to present the following novel contributions for computer vision based weapon safety technologies:

- Apply state-of-the-art object detection methodologies to detect weapons.
- Present a novel soft attention mechanism proposed for visual relationship detection.
- Present a proof of concept towards an applicable training methodology for visual relationship detection to be used for detecting held weapons.

Section II reviews relevant literature on recent weapon detection research and identifies the research gaps to be bridged to bring weapon safety focused applications of computer vision to current state-of-the-art. Section III presents the proposed novel methodology and contributions to develop a visual relationship detection model and applying it to detecting held weapons. Section IV summarizes the results of evaluation of the proposed methodology and Section V concludes the paper and presents the future works currently under development for an end-to-end weapon and visual relationship detection model for safety and security applications.

978-1-7281-2547-3/20/$31.00 ©2020 Crown

II. Related Works

Many previous works related to weapon detection have been focused on bounding box detection of weapons such as guns and rifles. The most successful works have come recently and make use of deep convolutional neural networks to do weapon deteciton. Olmos et al. [1] uses a sliding window approach with the VGG-16 [2] backbone architecture and also the Faster R-CNN [3] object detection network to do weapon detection.

Castillo et al. [4] proposes an image brightness adjustment procedure for maximizing weapon detection using deep convolutional neural networks. They apply their procedure to further refine the precision and recall of several object detection networks. [5] and [6] are more recent, shorter works which use networks such as Faster R-CNN [3] and SSD Mobilenet [7] to do their weapon detection.

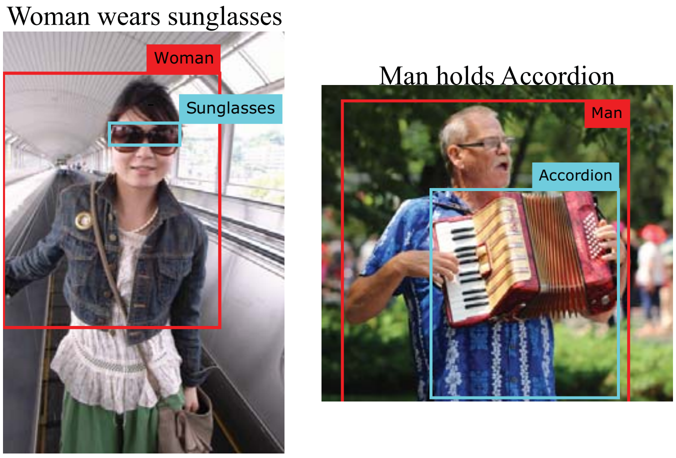
As of writing, no previous published works have focused on visual relationship detection. One of the more successful and more recent developments in visual relationship detection is the Box Attention network [8]. The Box Attention network uses bounding box information to do end-to-end object detection and visual relationship detection. Moreover, instance segmentation networks, such as Mask R-CNN [9] have been developed in recent years and none have been applied to the weapon detection problem. We seek to combine and close these research gaps to bring works on weapon detection and visual relationship detection to current state-of-the-art methodologies.

The visual relationship problem involves detecting subjects, objects, and relationships between subjects and objects in an image. The problem is normally described using the relationship triplet, $\langle subject(S), predicate(P), object(O)\rangle$. In this paper, we focus on the detecting relationships (predicates) between known ground truth subjects and objects, ie. given S and O, what is P.

III. Methodology

This section outlines the procedure and design justification for developing an object detection model for performing weapon and visual relationship detection, Section III-A describes the primary dataset, Open Images V6, and the qualities which make it suitable for this paper. Section III-B provides details on training and evaluating an instance segmentation model for people and weapons. Section III-C details the improvements to the novel soft attention model originally proposed in [10]. This section also describes how the trained instance segmentation model for people and weapons can be used with the visual relationship detection labels from Open Images V6 to make a weapon "holds" detection model.

A. Dataset



(a)

(b)

Fig. 1. Example samples of the "holds" and "wears" visual relationships from the Open Images V6 dataset.

The Open Images V6 dataset [11], [12] was chosen to be the primary dataset to be used for training and testing in this paper for its large number of images containing annotations for bounding boxes, instance segmentations, and visual relationships.

In particular, we are interested in the following object classes with segmentation labels:

- **person**: 149166 training, 269 validation, and 822 test segmentation labels.
- **knife**: 805 training, 79 validation, and 228 test segmentation labels.
- **handgun**: 552 training, 18 validation, and 271 test segmentation labels.
- **baseball bat**: 985 training, 16 validation, and 40 test segmentation labels.
- **hammer**: 122 training, 1 validation, and 0 test segmentation labels.
- **screwdriver**: 69 training, 2 validation, and 6 test segmentation labels.
- **axe**: 134 training, 0 validation, and 2 test segmentation labels.

These object classes were chosen because they are objects which can be (mis)used as a weapon.

Additionally, we are also interested in the following visual relationship labels:

- **wears**: 102,653 training, 1491 validation, and 4668 test visual relationship labels.
- **holds**: 33,018 training, 566 validation, and 1587 test visual relationship labels.

The visual relationship labels contain the $\langle subject(S), predicate(P), object(O)\rangle$ triplet. The subject and object are provided as bounding boxes (and most do not contain a relevant segmentation label) while the predicate is given as the label. Fig. 1 shows samples of the "holds" and "wears" visual relationship These labels generally describe a person holding or wearing everyday object and accessories, like microphones, instruments, hats, and sunglasses. Most visual relationship labels in the Open Images V6 dataset do not contain objects which can be considered as a weapon.

B. Instance Segmentation of People and Weapons

A Mask-RCNN network [9], [13] with a ResNet50 backbone is trained to detect and segment person, knife, handgun, baseball bat, hammer, screwdriver, and axe classes. The Mask-RCNN network was chosen as it is capable of producing high fidelity instance segmentations of these desired classes. The Keras (Tensorflow backend [14]) implementation found at [13] is used for training the network. A Mask R-CNN network that was pre-trained with the COCO dataset [15] is used as the base model for training. Stochastic gradient descent is used as the optimizer and images were resized and padded to 768 × 768 pixels.

Early training of the model is done with a learning rate of 0.001 and a momentum of 0.9 for 10,000 samples with the ResNet50 backbone layer weights frozen to train only the network heads (which were randomly initialized). The entire model then gets trained with a learning rate of 0.0001 and momentum of 0.9 for approximately 400,000 samples. The training was divided into iterations of 2000 samples per iteration. The model was evaluated on the validation set after each iteration and the model was evaluated and the model with the lowest validation loss was selected.

The output of this network provides instance segmentation masks for people and weapons. These instance segmentations can then be used to qualitatively examine our proposed visual relationship detection model. For brevity, this model is not rigorously evaluated in this paper but acts as a proof of concept for developing end-to-end weapon detection and visual relationship detection for future works.

C. Visual Relationship Detection

The visual relationship detection model is based on the proposed soft attention mechanism found in our previous works [10]. This soft attention mechanism was inspired by the Box Attention network [8] and uses the ResNet backbone structure consisting of several bottleneck units. Additional details on the ResNet structure may be found in [16]. The ResNet50 structure will be the focus of this paper as it offers adequate results with a relative short training time when compared to the ResNet101 and ResNet152 counterparts. ResNet is chosen over other popular backbone architectures because it is the same backbone architecture as used in the Mask R-CNN for instance segmentation. Sharing the same

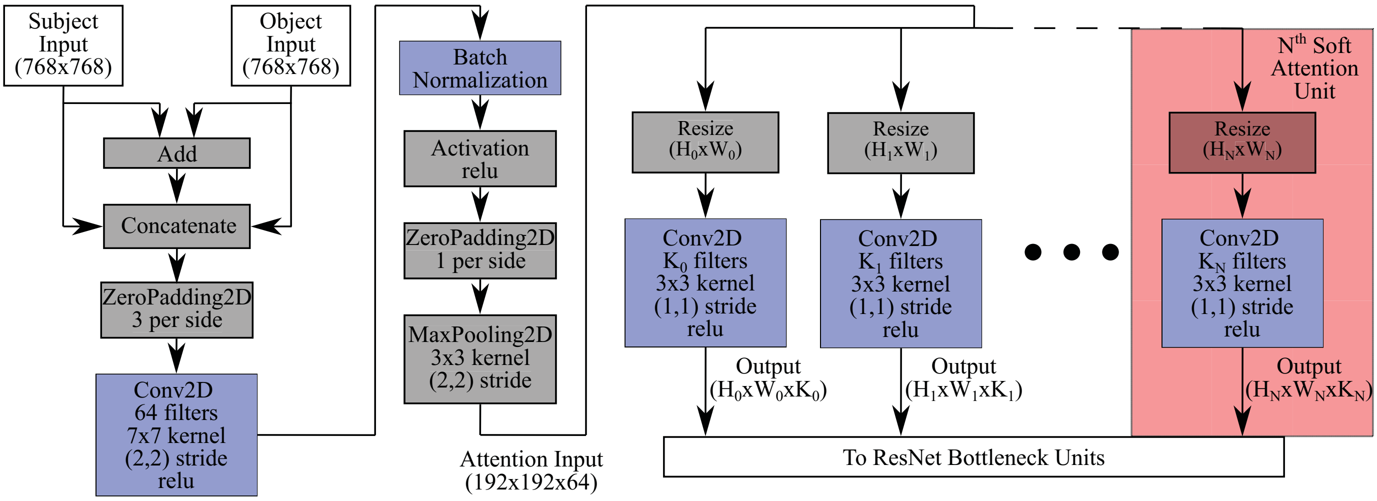

Subject Input (768x768)
Object Input (768x768)
Add
Concatenate
ZeroPadding2D 3 per side
Conv2D 64 filters 7x7 kernel (2,2) stride relu
Batch Normalization
Activation relu
ZeroPadding2D 1 per side
MaxPooling2D 3x3 kernel (2,2) stride
Attention Input (192x192x64)
Resize (H0xW0)
Conv2D K0 filters 3x3 kernel (1,1) stride relu
Output (H0xW0xK0)
Resize (H1xW1)
Conv2D K1 filters 3x3 kernel (1,1) stride relu
Output (H1xW1xK1)
Nth Soft Attention Unit
Resize (HNxWN)
Conv2D KN filters 3x3 kernel (1,1) stride relu
Output (HNxWNxKN)
To ResNet Bottleneck Units

Fig. 2. The proposed improved soft attention mechanism, Soft Attention Unit V2.

backbone architecture as the Mask-RCNN network provides us flexibility implement end-to-end instance segmentation and visual relationship classification in a future work.

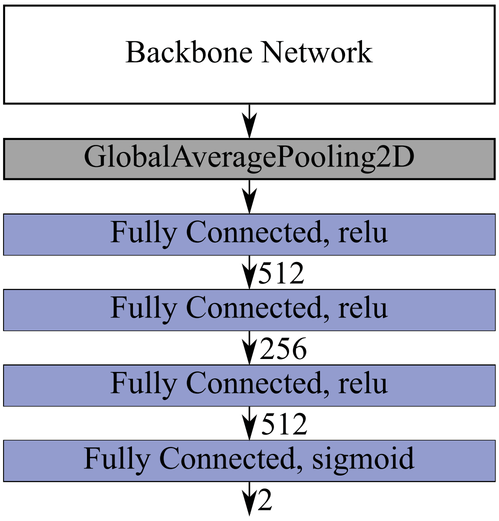
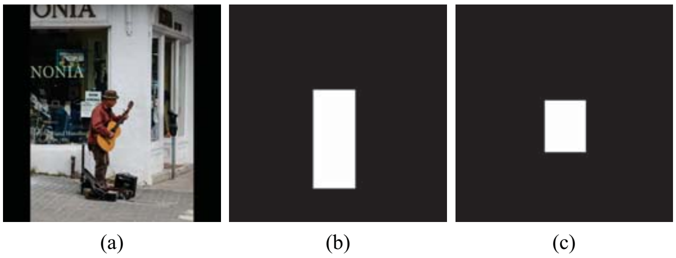


Fig. 3. The classification head used to classify "holds" or "wears".

Each bottleneck unit contains three convolutional blocks in series. The primary function of the second convolutional layer in each bottleneck unit is feature extraction. The soft attention unit as proposed in our previous paper [10] was a trainable unit which was added to the output of the second convolutional layer of each bottleneck unit.

The visual relationship labels as provided by the Open Images V6 dataset are then used to train a visual relationship model using the novel attention unit caled Soft Attention Unit V2 as depicted in Fig. 2. The subject input and object input are gray-scale images of the bounding box for the subject and object, respectively. The inputs are added together and the summation is concatenated with the original inputs. The output of the concatenation is then zero padded before being passed through a convolutional layer with a 7 × 7 kernel, similar to the ResNetV2 input layers. The convolutional layer output is then fed through a batch normalization layer, a relu activation, another zero padding and then a max pooling layer. The output of the max pooling layer is called the Attention Output and is then fed to the soft attention units before being used in the ResNet backbone. Each soft attention unit resizes and then passes resized (H × W) through a convolutional layer with K filters. The output dimensions H × W × K are chosen to match the output of the 3 × 3 convolutional layer of each bottleneck unit. For the ResNet50V2 architecture, there are N = 16 layers. The classification head depicted in Fig. 3 uses the output of the final bottleneck unit to classify "holds" or "wears" with a given image, subject, and object input. Examples of the given image, subject, and object input are depicted in Fig. 4.



(a) (b) (c)

Fig. 4. Example input into the proposed visual relationship network. (a) The resized and padded image, (b) the Subject Input bounding box image which (in this case) is a bounding box for the person, and (c) the Object Input bounding box image which (in this case) is a bounding box for the guitar.

For comparison, we test the new soft attention unit with the previously proposed unit by us in [10], to be retroactively called Soft Attention Unit V1. We also examine hyperparameters such as placement of the soft attention unit and the number of attention units. We have two training methodologies:

- The first methodology uses frozen ResNet50V2 ImageNet weights (provided in Tensorflow) and we only train the soft attention input, soft attention unit, and classification head. We train for 10 epochs and choose the model with the lowest validation loss. This is done to examine the effectiveness of the soft attention unit with a backbone network that has been trained for different purposes (in this case, it was trained to classify images from ImageNet). In the future, this will be useful for developing end-to-end solutions for both instance segmentation and visual relationship detection.

- The second methodology trains the entire soft attention model. The soft attention input, soft attention unit, and classification head is trained for 1 epoch with the ResNetV2 backbone frozen on the ImageNet weights first, then we unfreeze all layers and then train for 10 epochs and choose the model with the lowest validation loss. This is done to maximize the model performance for the task of visual relationship detection.

Te evaluate the models as a multi-label, multi-class problem (even though the Open Images V6 dataset contains only single labels for samples). We justify this decision as the labels for “holds” and “wears” are not necessarily mutually exclusive and future datasets may better describe these relationships in a multi-label problem. We use the standard metrics of accuracy, precision, recall, and specificity in our evaluation on a per-class basis. The equation for each metric is as follows:

$$
\begin{aligned}
\text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{Specificity} &= \frac{\text{TN}}{\text{TN}+\text{FP}}
\end{aligned}
$$

Where TP, FP, TN, FN are the numbers of true positives, false positives, true negativse, and false negatives, respectively. The testing set of the Open Images V6 dataset is imbalanced, containing 1587 samples for “holds” and 4668 samples for “wears”. We report the macro (per-class) average of the metrics, treating each class equally in weighting regardless of the number of samples. Additionally, we divide the test set into 15 equal partitions in order to report the standard deviation of each of the models. The metrics for each partition is calculated sample-wise and then the average and standard deviation of the metrics across the 15 different partitions are calculated and reported. Finally, to compare V1 and V2 soft attention units, we calculate the average performance across all V1 trained models to compare to the proposed V2 soft attention mechanism.

As an early qualitative test and stepping stone for end-to-end subject-relationship-object visual relationship detection, we use the proposed visual relationship detection network with the instance segmentations provided by the people and weapon segmentation network without any additional training.

IV. Results

Fig. 5 shows the results of the instance segmentation model for people and weapons used in combination with the proposed visual relationship model with the ResNet50V2 backbone and 16 units of Soft Attention Unit V2. The visual relationship model confidently labels held weapons in a provided input but struggles with inputs with no relationship. This is likely due to the lack of full negative (subject-object pairs that are neither “held” or “worn”) samples in the training data. Future works will need to address this issue by providing additional full negative samples.

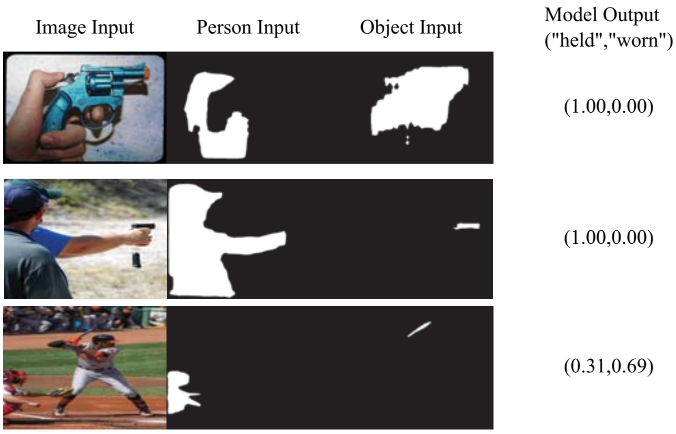


Fig. 5. Example visual relationship detections on unseen data. The person and object inputs are generated using a Mask R-CNN network trained to detect people and weapons.

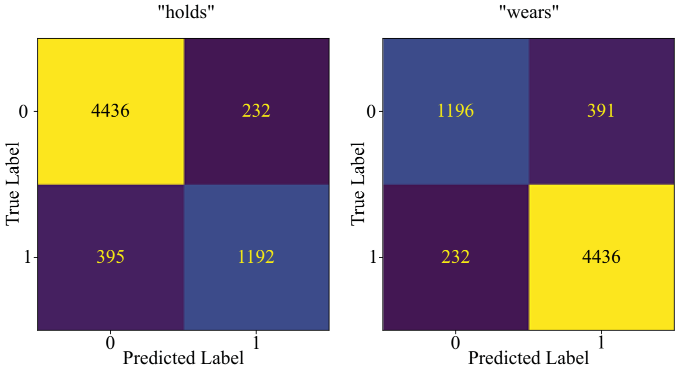


Fig. 6. Confusion matrix for the V2 16 unit model full model training.

Table I shows the results of training only the soft attention input, unit, and classification head with the ResNet50V2 backbone frozen on the ImageNet weights. For convenience in reporting our results, we shorten the ResNet50V2 layer names (as implemented in Tensorflow 2.2) as listed in Table IV. Overall, all models have difficulties with achieving reasonable recall scores. Closer inspection of the macro averaged results indicate that the “holds” class has poor recall performance due to many false negatives/few true positives. Likewise, “wears” class has poor specificity performance due to many false positives/few true negatives, as seen in Fig. 6. This is likely

TABLE I
TEST PERFORMANCE METRICS FOR THE SOFT ATTENTION MODELS USING FROZEN IMAGENET RESNET50V2 WEIGHTS FOR "HOLDS" AND "WEARS" CLASSIFICATION ON THE TEST DATASET FROM OPEN IMAGES V6.

Soft Attention Unit Version	Number of Attention Units	Layers Affected	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
V1	1	c2b1	87.46 ± 2.43	87.43 ± 3.45	78.1 ± 3.53	78.04 ± 3.53
V2	1	c2b1	88.36 ± 2.16	87.87 ± 2.53	80.17 ± 3.07	80.21 ± 3.06
V1	1	c5b3	86.16 ± 2.46	88.24 ± 2.54	74.32 ± 3.83	74.4 ± 3.71
V2	1	c5b3	86.93 ± 2.46	86.57 ± 3.05	77.38 ± 3.8	77.41 ± 4.12
V1	2	c2b1, c5b3	86.16 ± 2.23	87.51 ± 2.3	74.69 ± 3.47	74.8 ± 3.55
V2	2	c2b1, c5b3	88.57 ± 1.83	88.57 ± 2.18	80.16 ± 2.98	80.16 ± 2.98
V1	9	c2b1, c2b3, c3b2, c3b4, c4b2, c4b4, c4b6, c5b2, c5b3	87.52 ± 2.09	87.51 ± 3.04	78.13 ± 3.15	78.33 ± 3.06
V2	9	c2b1, c2b3, c3b2, c3b4, c4b2, c4b4, c4b6, c5b2, c5b3	87.01 ± 2.37	88.80 ± 1.99	76.04 ± 3.74	76.13 ± 3.83
V1	16	all	87.76 ± 1.85	88.37 ± 2.7	78.2 ± 2.59	78.23 ± 2.61
V2	16	all	87.97 ± 2.46	88.76 ± 2.65	78.49 ± 3.47	78.46 ± 3.49

TABLE II
TEST PERFORMANCE METRICS FOR THE SOFT ATTENTION MODELS WITH FULL TRAINING OF THE NETWORK FOR "HOLDS" AND "WEARS" CLASSIFICATION ON THE TEST DATASET FROM OPEN IMAGES V6.

Soft Attention Unit Version	Number of Attention Units	Layers Affected	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
V1	1	c2b1	88.62 ± 2.60	84.64 ± 2.99	85.92 ± 3.04	85.98 ± 3.09
V2	1	c2b1	88.55 ± 2.66	86.24 ± 3.83	82.24 ± 3.38	82.11 ± 3.47
V1	1	c5b3	88.98 ± 2.92	85.47 ± 4.12	85.21 ± 3.27	85.24 ± 3.31
V2	1	c5b3	89.34 ± 2.4	86.36 ± 3.37	85.00 ± 3.5	84.93 ± 3.48
V1	2	c2b1, c5b3	88.97 ± 1.79	85.54 ± 3.32	85.20 ± 2.41	85.24 ± 2.42
V2	2	c2b1, c5b3	89.21 ± 2.36	86.07 ± 3.39	85.17 ± 2.46	85.25 ± 2.57
V1	9	c2b1, c2b3, c3b2, c3b4, c4b2, c4b4, c4b6, c5b2, c5b3	89.32 ± 2.23	85.93 ± 3.43	85.87 ± 2.95	85.89 ± 2.95
V2	9	c2b1, c2b3, c3b2, c3b4, c4b2, c4b4, c4b6, c5b2, c5b3	88.39 ± 1.96	84.46 ± 2.46	85.40 ± 3.03	85.45 ± 3.05
V1	16	all	89.70 ± 2.63	86.67 ± 3.37	85.85 ± 3.31	85.87 ± 3.30
V2	16	all	90.01 ± 2.05	87.80 ± 2.73	85.07 ± 3.27	85.20 ± 3.05

TABLE III
AVERAGE ACROSS ALL TRAINED MODELS METRIC PERFORMANCE COMPARISON BETWEEN THE PREVIOUS V1 SOFT ATTENTION MECHANISM AND THE PROPOSED V2 SOFT ATTENTION MECHANISM.

Soft Attention Unit Version	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
V1	88.06 ± 1.26	86.73 ± 1.28	81.15 ± 4.89	81.20 ± 4.87
V2	88.43 ± 0.97	87.15 ± 1.43	81.51 ± 3.55	81.53 ± 3.55

caused by the dataset imbalance causing the network to favour classifying "wears" over "holds" for more difficult samples. Application of class weights and various resampling methods may be of interest in future works if better recall performance is desired for applications.

Table II shows the results from training the entire soft attention model. We see small performance improvements by training the entire network. We see similar trends as in Table I, with each V2 network outperforming the previous V1 networks. Training of the full network helps the networks overcome the difficulties with the poor recall and specificity performance which was seen previously in Table I. It is interesting to note that the recall performance for the V1 unit actually marginally outperforms the V2 unit when training the full networks at the expense of significantly reduced precision.

Table III shows the average performance of each soft attention unit version across all models for this problem. The proposed V2 soft attention unit, on average, outperforms the previous iteration

V. CONCLUSION AND FUTURE WORKS

In this paper we have developed state-of-the-art training methodologies and network architectures capable of detecting the visual relationships of "holds" and "wears". We improve upon the novel soft attention unit previously proposed in [10] for this problem to achieve 90.01% ± 2.05% accuracy on the test set of the Open Images V6 visual relationships for "holds" and "wears". Additionally, the improved soft attention unit

TABLE IV
ResNet50V2 Tensorflow 2.2 layer names and their shortened form

Layer Name	Shortened Form
conv2_block1_2_conv	c2b1
conv2_block2_2_conv	c2b2
conv2_block3_2_conv	c2b3
conv3_block1_2_conv	c3b1
conv3_block2_2_conv	c3b2
conv3_block3_2_conv	c3b3
conv3_block4_2_conv	c3b4
conv4_block1_2_conv	c4b1
conv4_block2_2_conv	c4b2
conv4_block3_2_conv	c4b3
conv4_block4_2_conv	c4b4
conv4_block5_2_conv	c4b5
conv4_block6_2_conv	c4b6
conv5_block1_2_conv	c5b1
conv5_block2_2_conv	c5b2
conv5_block3_2_conv	c5b3

outperforms, on average, the previous soft attention unit iteration on all major metrics, showing an average improvement of 0.37%, 0.42%, 0.36%, and 0.33% in accuracy, precision, recall, and specificity respectively. We also introduce a practical application of this network by using it on segmented people and weapons to predict whether a particular weapon is being "held".

Immediate future works on developing these visual relationship models include:

- Consolidating works in [10] with the works in this paper for a generalized visual relationship detection network capable of detecting many visual relationship labels.
- Adjusting the soft attention unit to work in parallel with instance segmentation networks to have end-to-end object detection as well as visual relationship detection with minimal extra computational overhead.
- Utilizing the outputs of the developed networks to do risk-analysis and develop assistive technologies for early warning systems in safety and security.

VI. Acknowledgment

This project was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the grant "Biometric-enabled Identity Management and Risk Assessment for Smart Cities" and by the Department of National Defence's Innovation for Defence Excellence and Security (IDEaS) program, Canada.

References

[1] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66–72, 2018.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[4] A. Castillo, S. Tabik, F. Pérez, R. Olmos, and F. Herrera, "Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning," *Neurocomputing*, vol. 330, pp. 151–161, 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.10.076

[5] R. Kakadiya, R. Lemos, S. Mangalan, M. Pillai, and S. Nikam, "AI Based Automatic Robbery/Theft Detection using Smart Surveillance in Banks," *Proc. of the 3rd International Conf. on Electronics and Communication and Aerospace Technology, ICECA 2019*, pp. 201–204, 2019.

[6] S. Xu and K. Hung, "Development of an ai-based system for automatic detection and recognition of weapons in surveillance videos," in *2020 IEEE 10th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, 2020, pp. 48–52.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[8] A. Kolesnikov, C. H. Lampert, and V. Ferrari, "Detecting visual relationships using box attention," *CoRR*, vol. abs/1807.02136, 2018. [Online]. Available: http://arxiv.org/abs/1807.02136

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proc. of the IEEE Int. Conf. on Computer Vision*, vol. 2017-Octob, pp. 2980–2988, 2017.

[10] T. Truong and S. Yanushkevich, "Relatable clothing: Detecting visual relationships between people and clothing," 2020.

[11] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *arXiv:1811.00982*, 2018.

[12] R. Benenson, S. Popov, and V. Ferrari, "Large-scale interactive object segmentation with human annotators," in *CVPR*, 2019.

[13] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," 2017. [Online]. Available: http://github.com/matterport/Mask_RCNN

[14] M. A. et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 740–755.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.