Object Classification Comparison using Real vs. Synthetic Data

DATASCI 281.3 Spring 2023 with Dr. Allen Yang

Stephen Tan **Victoria Hollingshead**

GitHub Repo

Abstract

The research highlighted in this paper aims to classify objects into 20 different classes given by the PASCAL VOC 2012 dataset and to investigate the utility of synthetic training datasets in training object detection models. This report provides a discussion of Histogram of Oriented Gradients (HOG), Hue Saturation Value color scale (HSV), and Scale-Invariant Feature Transform (SIFT) feature vector extraction for the object classification via Support Vector Machines (SVM). Additional work was done using ResNet-50 to compare classification accuracy. Out of the extracted features, HOG gave the best accuracy at 50% average across all 20 classes. ResNet-50 had an accuracy of 71.45%. Finally, we added synthetic data to both object classification methods to examine if it would have any measurable improvements in bottle object classification. Results showed that for HOG features, there was no change in precision, while there were only 1% improvements in recall and F1-score. For ResNet-50, accuracy increased by 2.07% after transfer learning. Precision, recall, and the F-1 score for the "bottle" object also increased by 2%, 4%, and 3%, respectively. These results support the conclusion that the synthetic dataset had a marginal improvement in the detection of the ITO-EN bottle public domain images. For future work, increasing the synthetic data count or upgrading the synthetic image quality to high-fidelity images (e.g., EPIC game engine) may further improve the accuracy of the ResNet-50 model.

Introduction

Our research proposal is motivated by the work completed by Forensic Architecture[1], a multidisciplinary research group based out of the University of London, and VFRAME[2], a private research group. Both Forensic Architecture and VFRAME contribute to the broader question of whether image classification and object-detection can be used to automate aspects of open source intelligence investigations (OSINT) by human rights practitioners.

In 2019, Forensic Architecture produced "Triple Chaser," an exhibit that highlighted the use of chemical munitions in civilian protests around the world, particularly in Tijuana, Gaza, and later in the United States' Black Lives Matter protests. As a direct result of this work, Safariland, a tear gas manufacturing company responsible for developing such munitions, underwent

[1] *Objects of violence: Synthetic data for practical ML in human rights ...* (n.d.). Retrieved April 20, 2023, from https://www.researchgate.net/publication/337447320_Objects_of_violence_synthetic_data_for_practical_ML_in_human_rights_investigations

[2] VFRAME. (n.d.). Retrieved April 20, 2023, from https://vframe.io/

divestments, as well as the withdrawal of its CEO, Warren Kanders (Pogrebin, 2020). Using similar techniques, human rights researchers from VFRAME uncovered Syrian war crimes, generating 3D-modeled scenes in order to build synthetic training datasets for illegal munitions.

Through their work, Forensic Architecture and VRAME have demonstrated the potential of using machine learning to investigate state violence and uphold human rights worldwide. This final project is an attempt to build upon their work, using comparable techniques in image-generation and classification.

Image 1. *Left* is a public domain image of a man holding two Triple Chaser chemical munitions. *Right* is an image of segmentation masks of cluster munitions. (Forensic Architecture / Praxis Films)

Background

In "Objects of Violence: synthetic data for practical ML in human rights investigations[3]," Kermode et al. trained three models using a synthetic dataset: ResNet-50, U-Net, and MaskRCNN. In their experiment, ResNet-50 AUC hovered at approximately 0.50 for 10 epochs. After applying domain adversarial methods, they found a dramatic increase in model performance, with an AUC of approximately 0.65.

In "Exploring Object-Centric and Scene-Centric CNN Features and their Complementarity for Human Rights Violations Recognition Images[4]," Kalliatakis et al. used public domain data to train ResNet-50, VGG16, and VGG19 models under a number of architectural methods. Using a flatten pool, their ResNet-50 model achieved a Top-1 accuracy of 30.00%.

Likewise, in "Will Large-Scale Generatie Models Corrupt Future Datasets[5]," Hataya et al. used a synthetic dataset to train three models: ResNet-50, SwinTransformer, and ConvNeXt. In their findings, ResNet-50 achieved a test accuracy of 15.7% when using a purely synthetic dataset generated using Stable Diffusion, a state-of-the-art text-to-image generative model.

---

[3] Kermode, L., Freyberg, J., Akturk, A., Trafford, R., Kochetkov, D., Pardinas, R., Weizman, E., & Cornebise, J. (2020, April 1). *Objects of violence: Synthetic data for practical ML in human rights investigations*. arXiv.org. Retrieved April 20, 2023, from https://arxiv.org/abs/2004.01030

[4] Kalliatakis, G., Ehsan, S., Leonardis, A., & McDonald-Maier, K. (2018, May 12). *Exploring object-centric and scene-centric CNN features and their complementarity for Human Rights Violations Recognition in images*. arXiv.org. Retrieved April 20, 2023, from https://arxiv.org/abs/1805.04714

[5] Hataya, R., Bao, H., & Arai, H. (2022, November 15). *Will large-scale generative models corrupt future datasets?* arXiv.org. Retrieved April 20, 2023, from https://arxiv.org/abs/2211.08095

Methods

Task

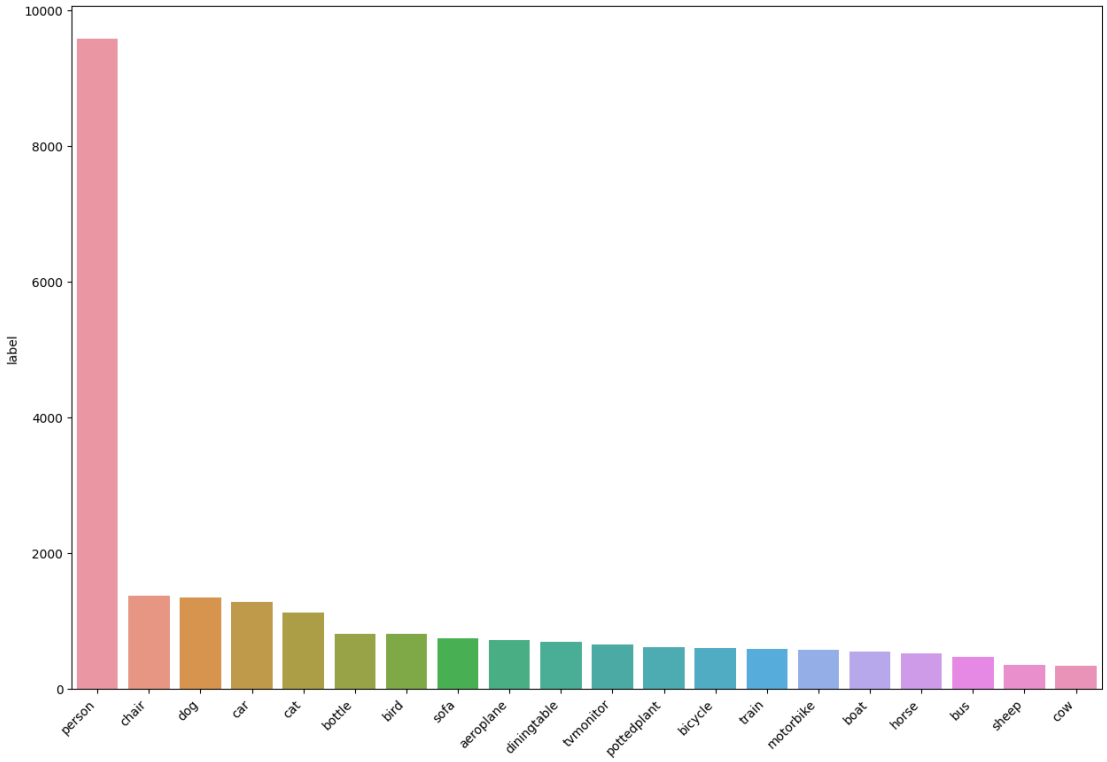
There are two primary objectives to this project. The first objective is to to identify 20 object classes using a custom-built image classifier using Histogram of Oriented Gradients (HOG), Hue Saturation Value (HSV) color scale, Scale-Invariant Feature Transform (SIFT) and ResNet-50. The second objective is to examine the effect of using user-generated synthetic datasets to train object detection models to accurately identify objects of interest in real-world/public datasets.

Data

Our dataset contains three main components. Their role in our research is explained below. Data was split 70-10-20 train-val-test.

PASCAL VOC 2012

This dataset includes approximately 17,000 JPEG images, which include 20 classes that can be grouped into 4 main categories: person (1 class), animal (6 classes), vehicle (7 classes), and indoor objects (6 classes). These images had random numerical file names assigned and were not labeled, but did come with accompanying .xml files that included information about objects in each image, their labels, and their accompanying regions. More information about these annotation files can be found in the Appendix.



Plot 1. PASCAL VOC 2012 label distribution

Synthetic Images Generated via Unity

This dataset includes 8,000 PNG Unity game engine generated images of 500 mL ITO-EN green tea bottles. Variance in object count, rotation, lighting, bo

Image 2. Synthetic images of 500 mL ITO-EN bottles generated using Unity

ITO-EN Public Domain

This dataset includes 50 images of ITO-EN bottles retrieved through Google searches. Further processing, by varying hue, saturation, brightness, and contrast, was conducted to augment the public domain images to increase the count to 516.

Image 3. Public domain images of 500 mL ITO-EN bottles

Preprocessing

Initially, we had conducted our experimentation with the base images offered in the PASCAL VOC dataset. This method was not only time-consuming, but also resulted in terrible classification accuracy of 6%. There was too much information in each of our feature vectors, and further processing was necessary.

In order to remove noise from the images, we first began by looking deeper into the .xml annotation files. We did not notice that images could contain multiple objects and that there were bounding box coordinates for each object within a given image. After extracting that information, we generated new images based off each object crop from its bounding box. Our PASCAL VOC dataset effectively grew from around 17,000 images to almost 24,000 images, post-cropping. We decided not to crop these images to keep them as pseudo-raw images.

Since most of these crops were greater than 128x128 resolution, we padded each image to 128x128 resolution for consistency and for downstream processing.

Image 4. Object extraction from parent image. *Left* image contains two classes – car and motorbike. Each one is extracted and saved as a new cropped image (*middle and right*) that is in 128x128 resolution. Further downstream processing either downsizes or upsizes the image based on crop resolution and then adds white padding to fulfill the 128x128 size.

Evaluation Metrics

In order to evaluate our image classification, we decided to use precision, recall, F1-score, and accuracy. Precision is the ratio of predicted true positives to both predicted true and false positives. Recall is the ratio of predicted true positives to both predicted true positives and false negatives. F1-score gets the best of both worlds where it encompasses the mean of precision and recall. Finally, accuracy is a common metric that evaluates correct predictions based on total predictions.

Strategies and Modeling

Our research is broken down into two phases, which are differentiated by the training dataset. The first phase has two components – A and B – that involves the training set *only* seeing the PASCAL VOC dataset. For Phase 1A, we explored custom feature vectors such as HOG, HSV, and SIFT. For Phase 1B, we explored ResNet-50. The second phase has the same two components, except the training set is composed of both the PASCAL VOC dataset *and* the synthetic image dataset.

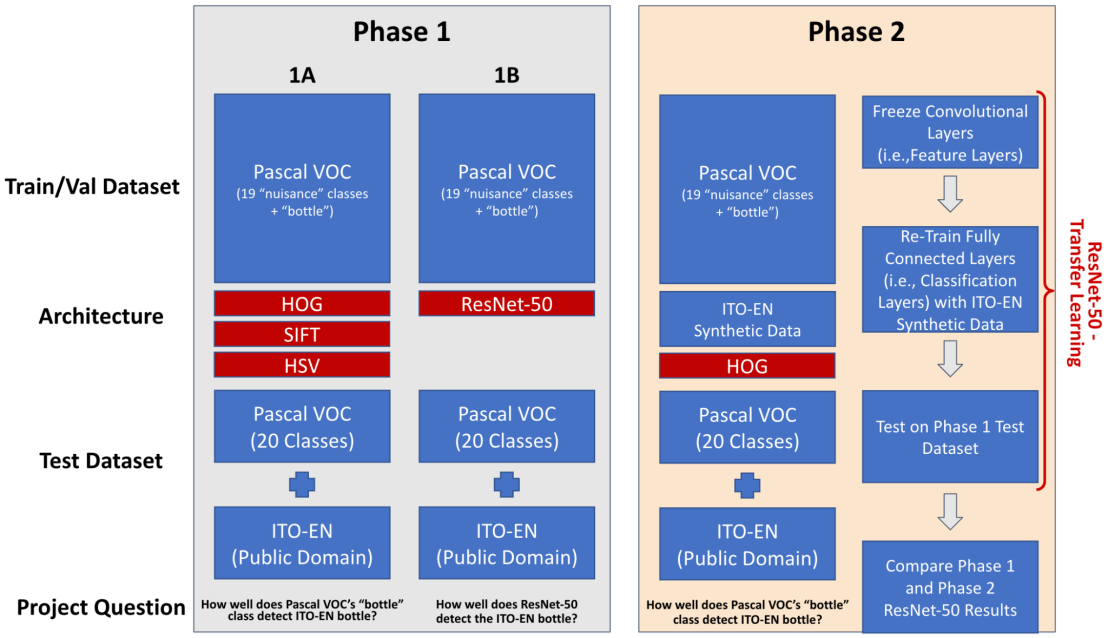


Diagram 1. Experimental approach

Phase 1A

This phase involves the extraction of our custom feature vectors: HOG, HSV, and SIFT. These vectors are then individually fed into a Support Vector Machine (SVM) for object classification.

HOG

Histogram of Gradients is a feature descriptor that characterizes an object by its shape. This descriptor is different from edge features because HOG gives information – oriented gradients – on those edges.

The baseline parameters for HOG feature were chosen to be:

```
blur_value=5, orientations=9, pixels_per_cell=16, cells_per_block=2
```

The resultant length of the baseline HOG vector was 1764 given a 128x128 padded image. We decided to do further optimization on HOG features due to their higher accuracy scores. We modified parameters in two different ways – one for efficiency and another for accuracy.

The HOG parameters optimized for efficiency were chosen to be:

```
blur_value=5, orientations=9, pixels_per_cell=24, cells_per_block=2
```

The resultant length of the baseline HOG vector was 576 given a 128x128 padded image. Since the vector was smaller in length, there is less information. As a result, the SVM classification method effectively has to do fewer calculations, and completes the model in around 4 minutes.

The HOG parameters optimized for accuracy were chosen to be:

```
blur_value=5, orientations=9, pixels_per_cell=8, cells_per_block=2
blur_value=5, orientations=9, pixels_per_cell=12, cells_per_block=2
blur_value=5, orientations=9, pixels_per_cell=20, cells_per_block=2
```

Since the first two vectors were larger in length, there was more information. However, all of these parameter changes actually decreased accuracy for our validation dataset (without public domain ITO-EN images). Scores were 48.34%, 49.09%, 48.97%, respectively. More pixels per cell is not actually better, as it may treat noise as an important feature instead of focusing on the actual object's important characteristics. Doing so does not only cost more computationally but also returns lower accuracy.

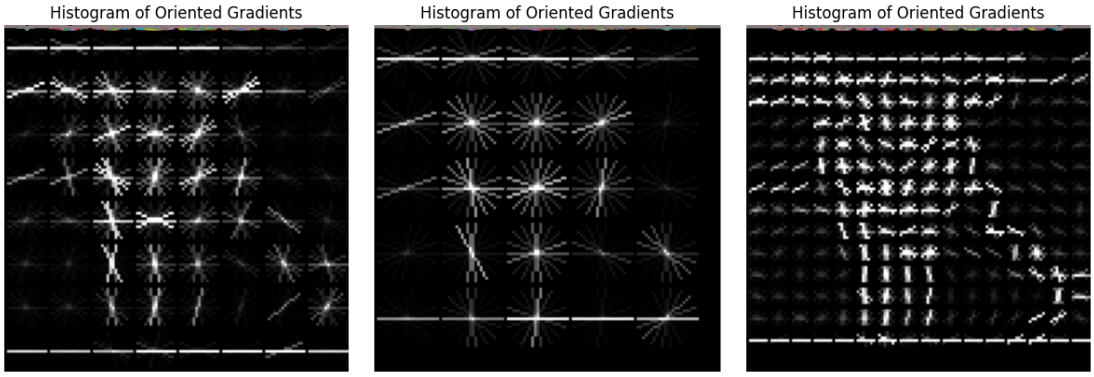


Image 5. HOG optimizations with pixels per cell. More pixels per cell means less information.
From left to right: baseline – 16, efficiency – 24, accuracy – 8

HSV

Hue Saturation Value color scale is one way to identify an object by looking at its specific color characteristics. To process the HSV vector, the native cropped image (unpadded) was converted from RGB to HSV. These images were not padded, since padding would result in a skew of more 255 (white) values appearing from the pad border. Then histograms are calculated for H, S, and V. Values were then flattened to vectorize these values. The resultant length of the baseline HSV vector was 96.

SIFT

Scale-Invariant Feature Transform is a more complex feature vector that includes keypoints and descriptors. While we chose to use it, it is more useful for identifying the same object that has been rotated or scaled. For example, if we had multiple images of the same airplane, SIFT would be beneficial. When extracting these feature vectors, we used a Gaussian blur filter of (3,3) and for efficiency purposes, limited the number of SIFT keypoints to `nfeatures=20`. The resultant length of the baseline SIFT descriptor vector was 6,400.

SVM

Support Vector Machine is the method employed for our object classification. This method was chosen since it is commonly used with HOG and SIFT as shown in Histograms of Oriented Gradients for Human Detection[6]. The extracted features fed into the SVM are numerical, but a visual representation of them before they are vectorized are shown below:

[6] *Histograms of oriented gradients for human detection - INRIA*. (n.d.). Retrieved April 20, 2023, from https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

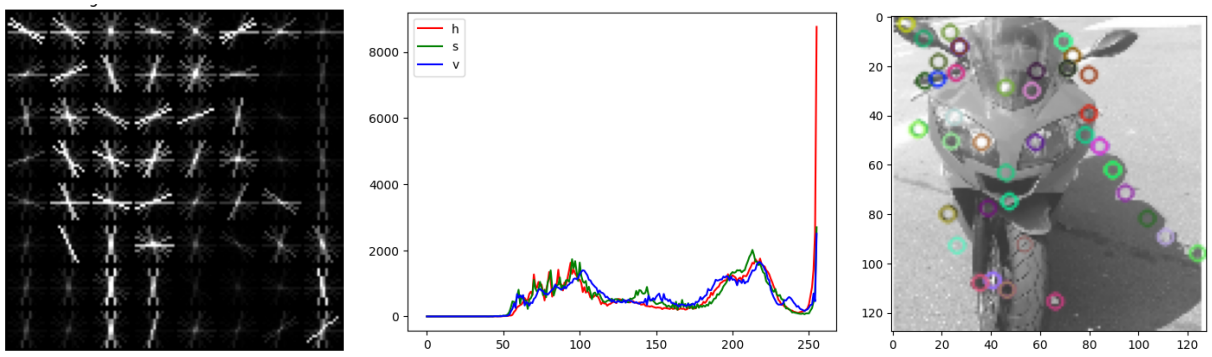


Image 6. Custom Feature Extraction from motorbike example. *From left to right*: HOG, HSV, SIFT

Phase 1B

This phase involves feeding in the resized cropped images into a ResNet-50 architecture for object classification.

ResNet-50

ResNet-50 is a pre-trained deep learning model for image classification using a convolutional neural network (CNN). Trained on over a million ImageNet images, this classifier has been used in a number of publications related to human rights and synthetic dataset research. Using ResNet-50 in our experiment will allow us to make comparisons with literature results.

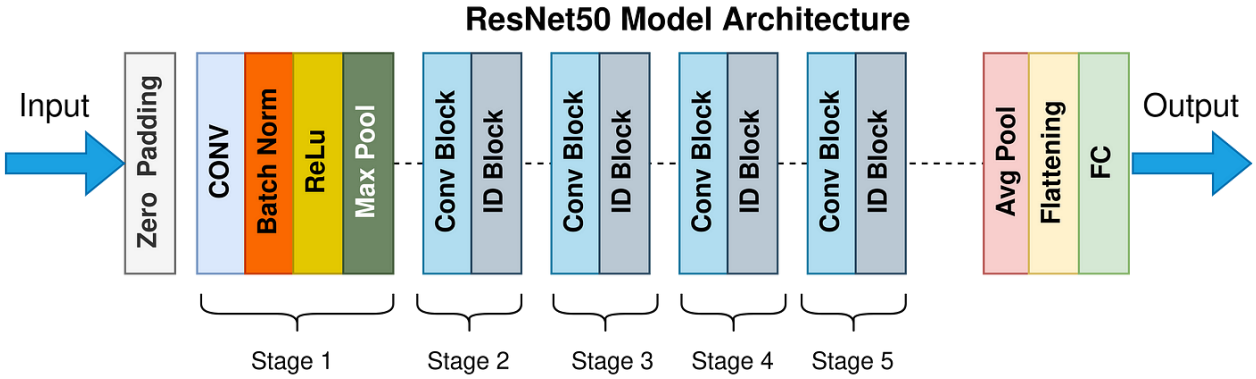


Diagram 2. Outline of ResNet-50 Model Architecture

In Phase 1, we trained our ResNet-50 model with the Pascal VOC dataset while allowing each layer to update. The parameters below were used in the construction of our ResNet-50 model:

```
batch_size = 32, training_epochs = 20, learning rate = 1e-3,
optimizer='adam', loss='categorical_crossentropy'
```

In addition to the parameters above, the pre-trained model was modified with the following dense layers and activation functions, resulting in 23,857,454 total parameters, 269,742 trainable parameters, and 23,587,712 non-trainable parameters:

```
inputs = pretrained_model.input
x = tf.keras.layers.Dense(128, activation='relu')(pretrained_model.output)
x = tf.keras.layers.Dense(50, activation='relu')(x)
outputs = tf.keras.layers.Dense(20, activation='softmax')(x)
```

Phase 2

The extension of our research involved using transfer learning for the Phase 1B ResNet50 model, where we fixed the weights of ResNet-50 (i.e., freeze the convolutional layers) and re-trained the fully connected layers with the 8,000 Unity-generated synthetic dataset of bottle images. We also repeated the same experiment in Phases 1A using the synthetic dataset.

Results and Discussion

A summary of the results are shown below. For each custom feature vector in Phase 1A, along with Phase 1B and 2B, a confusion matrix and a classification report are shown in the Appendix. Additionally, a discussion around the efficiency vs accuracy optimizations of HOG is provided.

Phase 1A and 2A Performance

When used with SVM, the HOG feature descriptors showed better accuracy at 44% than HSV and SIFT. This result makes sense because the general object classes have edge patterns to them albeit some more than others. SIFT feature descriptors had the next best performance at 40%. Color hues, saturations, and values (HSV) showed decent performance at 32%. The reason why HOG performs better than SIFT and HSV is due in part to the images in our dataset. If the images were isolated to different angles and perspectives of similar images such as landmarks or specific brands, SIFT and HSV would likely have better accuracy scores.

When the synthetic images of 500 mL ITO-EN green tea bottles were injected into the training dataset, we saw accuracy increase to 49%. This 5% increase in accuracy can be attributed to the model learning more from the HOG features.

Phase 1B and 2B Performance

Phase 1B ResNet-50 achieved an overall test accuracy of 71.45%, thus outperforming the ResNet-50 models from Kalliatakis et al. and Hataya et al.. After the transfer learning phase, test accuracy jumped by 2.07% to 73.52%, thus outperforming our baseline Phase 1B ResNet-50 model. Due to limited time, we did not calculate the AUC of either model for comparison to the Kermode et al. model, however, precision, recall, and f1 values are outlined in the classification report in the Appendix.

		Compared to Literature		
Literature	**Literature Accuracy (%)**	**Phase 1B Accuracy (%)** (*without* synthetic data)	**Phase 2B Accuracy (%)** (*with* synthetic data)	**Comparison (Phase 2B-1B)**
Kalliatakis et al.	30.0	**+41.45**	**+43.52**	**+2.07**
Hataya et al.	15.7	**+55.75**	**+57.82**	

Table 1. Literature result comparison with and without synthetic training data.

Precision, recall, and the F-1 score for the "bottle" object all increased following transfer learning with the synthetic training dataset. These results support the conclusion that the synthetic dataset had a marginal improvement in the detection of the ITO-EN bottle public domain images.

	"Bottle" Object Performance		
Model	**Precision**	**Recall**	**F-1 Score**
Phase 1B (*without* synthetic data)	0.11	0.08	0.09
Phase 2B (*with* synthetic data)	0.13	0.12	0.12
Comparison (Phase 2B-1B)	**+0.02**	**+0.04**	**+0.03**

Table 2. Bottle object performance with and without synthetic training data.

Limitations

Most of the programming was done on Google Colab. However, due to the size of our dataset and the RAM limitations of the free version of Colab, we had to upgrade to Colab Pro+ in order to handle all of our image data.

Another limitation was that if there were multiple objects of one class in an image, only the last instance of the object would be saved to a cropped image. Most of the images only contained objects of one class and if it contained multiple objects, they would be from different classes, so we felt that this limitation would not significantly affect the results.

Conclusion

For object classification using SVM, it was found that HOG was the best feature descriptor compared to HSV and SIFT. We learned that using HOG features on an entire image with limited preprocessing (i.e. Gaussian blur only), that classification results were terrible at around ~6%. There is simply too much noise and too many feature gradients. However, once objects were further isolated via cropping from their parent images, model accuracy significantly increased to 50%. If future work were to be conducted, it would be beneficial to examine the performance of other object classification methods such as logistic regression and simple perception.

For object classification and transfer learning using ResNet 50, we conclude that the use of synthetic data had a marginal improvement on test accuracy; ultimately increasing the 'bottle' detection precision by 0.02, recall by 0.04, and F-1 score by 0.03. Future teams are encouraged to increase the synthetic data count or upgrade the synthetic image quality to high-fidelity images (e.g., EPIC game engine) in order to further improve the accuracy of the ResNet-50 model.

GitHub Repo

References

Financial Times. "Researchers train AI on ‘synthetic data’ to uncover Syrian war crimes." Financial Times, 11 April 2021, https://www.ft.com/content/8399873e-0dda-4c87-ba59-0e2678166fba

Hataya, Ryuichiro. “Will Large-Scale Generative Models Corrupt Future Datasets?” *N/A*, 15 Nov. 2022, https://arxiv.org/abs/2211.08095 Accessed 21 Mar. 2023.

Kalliatakis, Grigorios. “Exploring Object-Centric and Scene-Centric CNN Features and Their Complementarity for Human Rights Violations Recognition in Images.” *PLOS ONE*, 12 May 2018, https://arxiv.org/abs/1805.04714 Accessed 21 Mar. 2023.

Kermode, Lacklan. “Objects of Violence: Synthetic Data for Practical ML in Human Rights Investigations.” *NeurIPS 2019*, 1 Apr. 2020, https://arxiv.org/abs/2004.01030 Accessed 21 Mar. 2023.

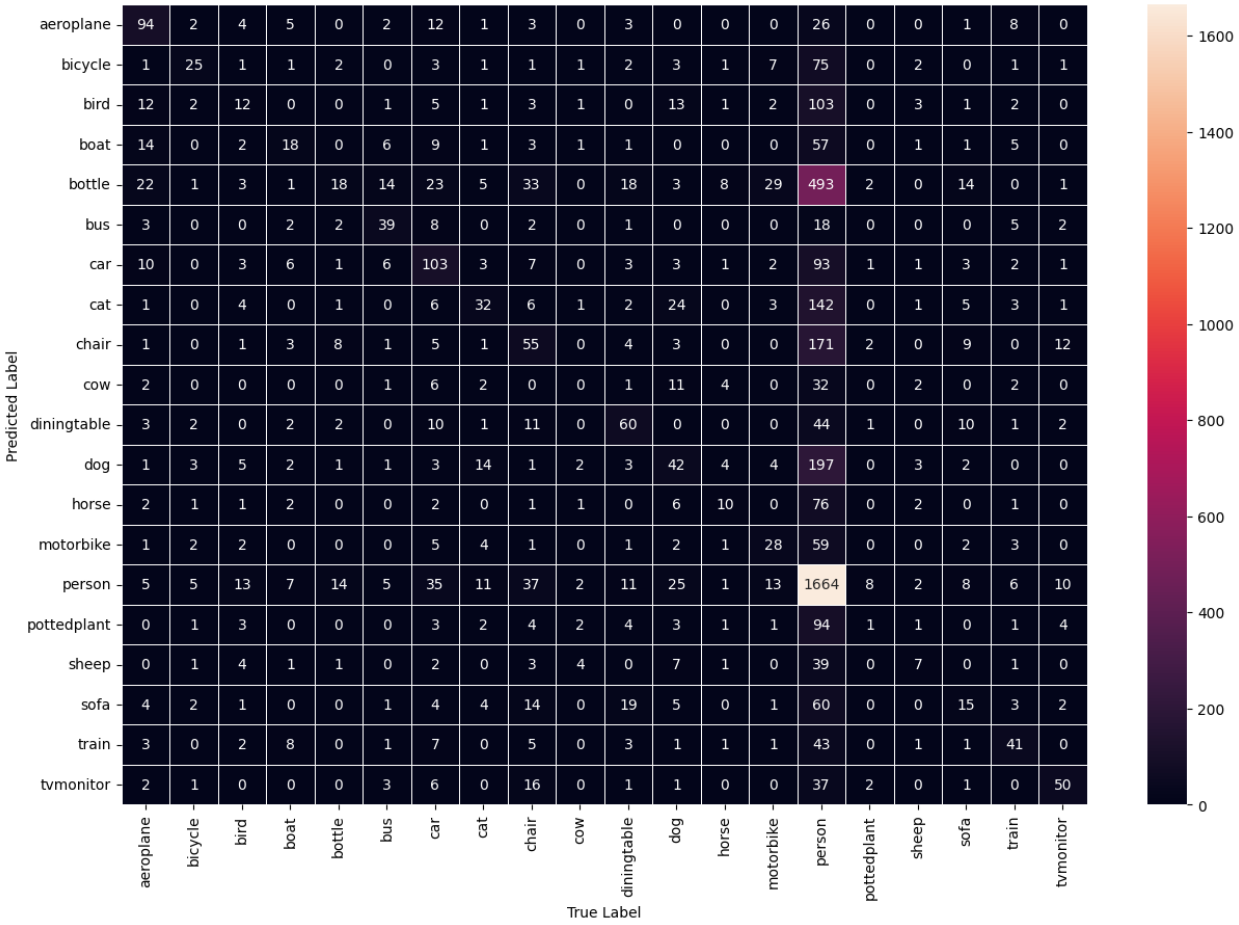
Pogrebin, Robin. 2020. “Warren Kanders Says He Is Getting out of the Tear Gas Business.” The New York Times, June 9, 2020, https://www.nytimes.com/2020/06/09/arts/design/tear-gas-warren-kanders.html?smid=tw-share. Accessed 21 Mar. 2023

Appendix

Train-Val-Test Split Compositions

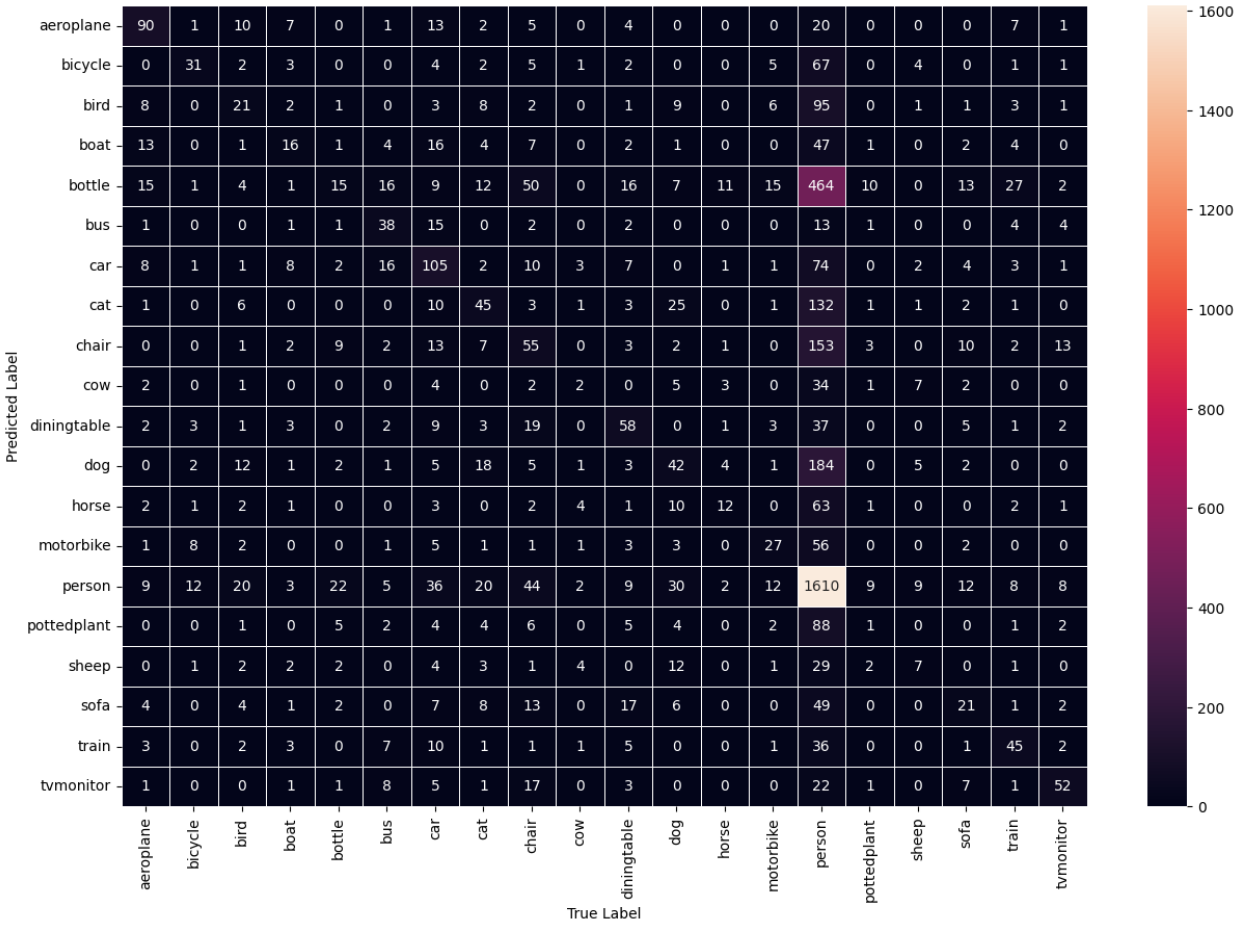
Label	Train %	Validation %	Test %
person	40.2	41.0	40.6
chair	5.8	6.3	5.4
dog	5.7	5.1	5.7
car	5.4	5.5	5.5
cat	4.8	4.7	4.7
bird	3.5	2.9	3.5
bottle	3.5	2.9	3.5
sofa	3.3	2.9	2.6
aeroplane	3.0	3.2	3.2
diningtable	2.9	3.0	2.8
tvmonitor	2.6	3.5	2.7
pottedplant	2.6	2.7	2.7
train	2.5	2.1	2.5
bicycle	2.5	2.7	2.7
motorbike	2.4	2.4	2.6
boat	2.3	2.5	3.2
horse	2.2	2.4	2.0
bus	1.9	1.8	2.2
sheep	1.6	1.0	1.3
cow	1.3	1.9	1.5

HOG Baseline (~25 minute runtime)



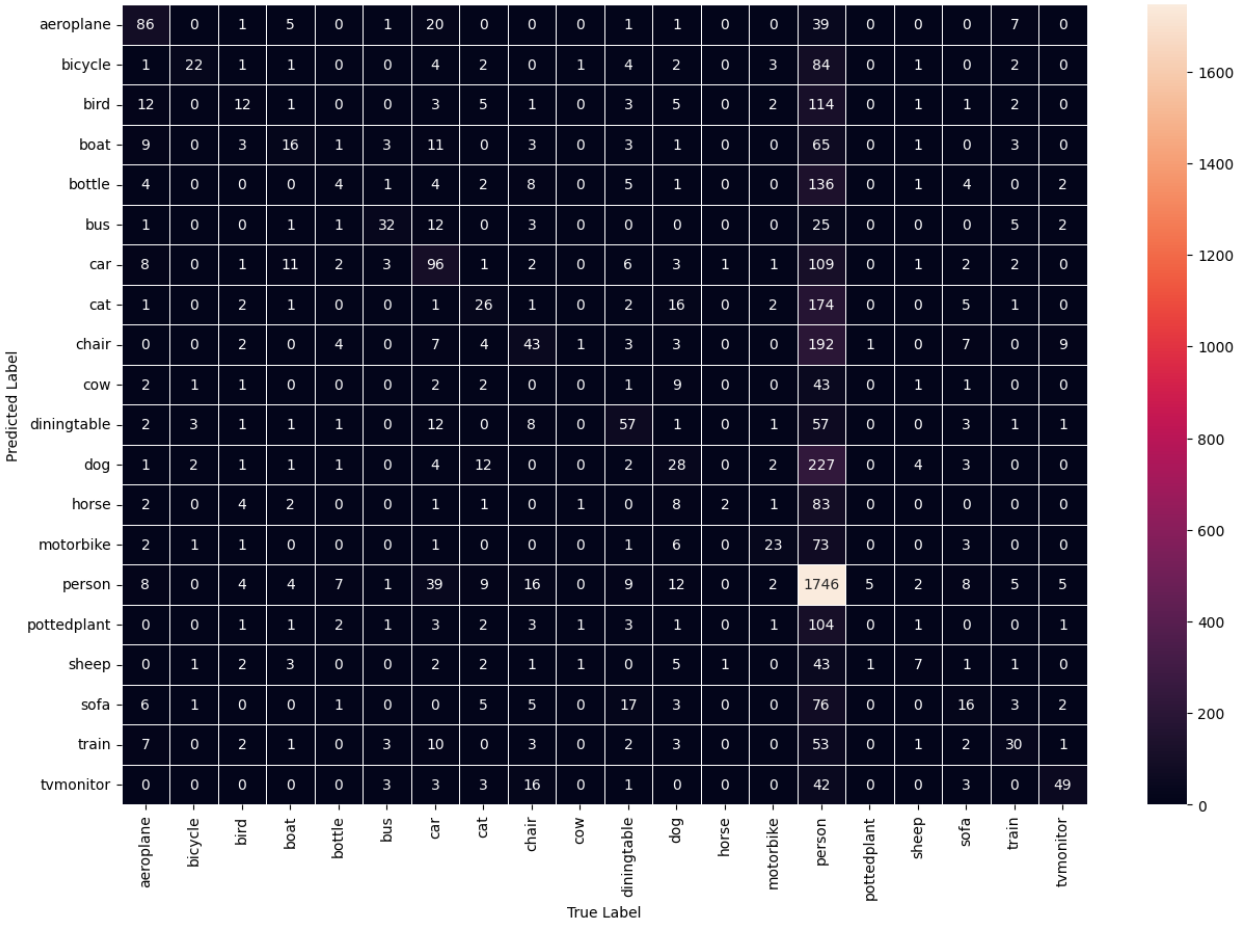
	precision	recall	f1-score	support
aeroplane	0.52	0.58	0.55	161
bicycle	0.52	0.20	0.28	128
bird	0.20	0.07	0.11	162
boat	0.31	0.15	0.20	119
bottle	0.36	0.03	0.05	688
bus	0.48	0.48	0.48	82
car	0.40	0.41	0.41	249
cat	0.39	0.14	0.20	232
chair	0.27	0.20	0.23	276
cow	0.00	0.00	0.00	63
diningtable	0.44	0.40	0.42	149
dog	0.28	0.15	0.19	288
horse	0.29	0.10	0.14	105
motorbike	0.31	0.25	0.28	111
person	0.47	0.88	0.62	1882
pottedplant	0.06	0.01	0.01	125
sheep	0.27	0.10	0.14	71
sofa	0.21	0.11	0.14	135
train	0.48	0.35	0.40	118
tvmonitor	0.58	0.42	0.49	120
accuracy			0.44	5264
macro avg	0.34	0.25	0.27	5264
weighted avg	0.39	0.44	0.36	5264

HOG Optimized for Efficiency (~4 minute runtime)



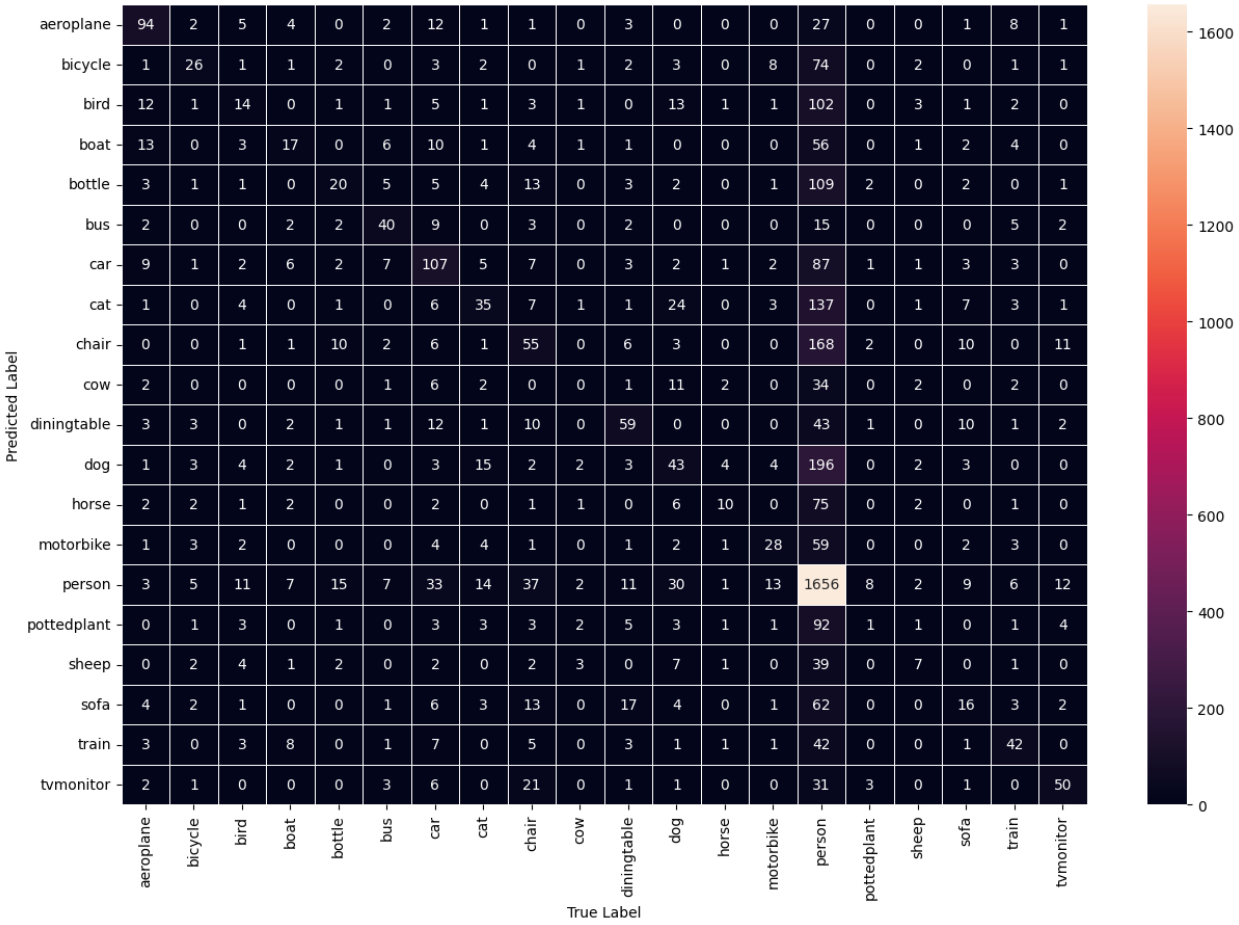
	precision	recall	f1-score	support
aeroplane	0.56	0.56	0.56	161
bicycle	0.51	0.24	0.33	128
bird	0.23	0.13	0.16	162
boat	0.29	0.13	0.18	119
bottle	0.24	0.02	0.04	688
bus	0.37	0.46	0.41	82
car	0.38	0.42	0.40	249
cat	0.32	0.19	0.24	232
chair	0.22	0.20	0.21	276
cow	0.10	0.03	0.05	63
diningtable	0.40	0.39	0.40	149
dog	0.27	0.15	0.19	288
horse	0.34	0.11	0.17	105
motorbike	0.36	0.24	0.29	111
person	0.49	0.86	0.62	1882
pottedplant	0.03	0.01	0.01	125
sheep	0.19	0.10	0.13	71
sofa	0.25	0.16	0.19	135
train	0.40	0.38	0.39	118
tvmonitor	0.57	0.43	0.49	120
accuracy			0.44	5264
macro avg	0.33	0.26	0.27	5264
weighted avg	0.37	0.44	0.37	5264

HOG Optimized for Accuracy (~100 minute runtime)



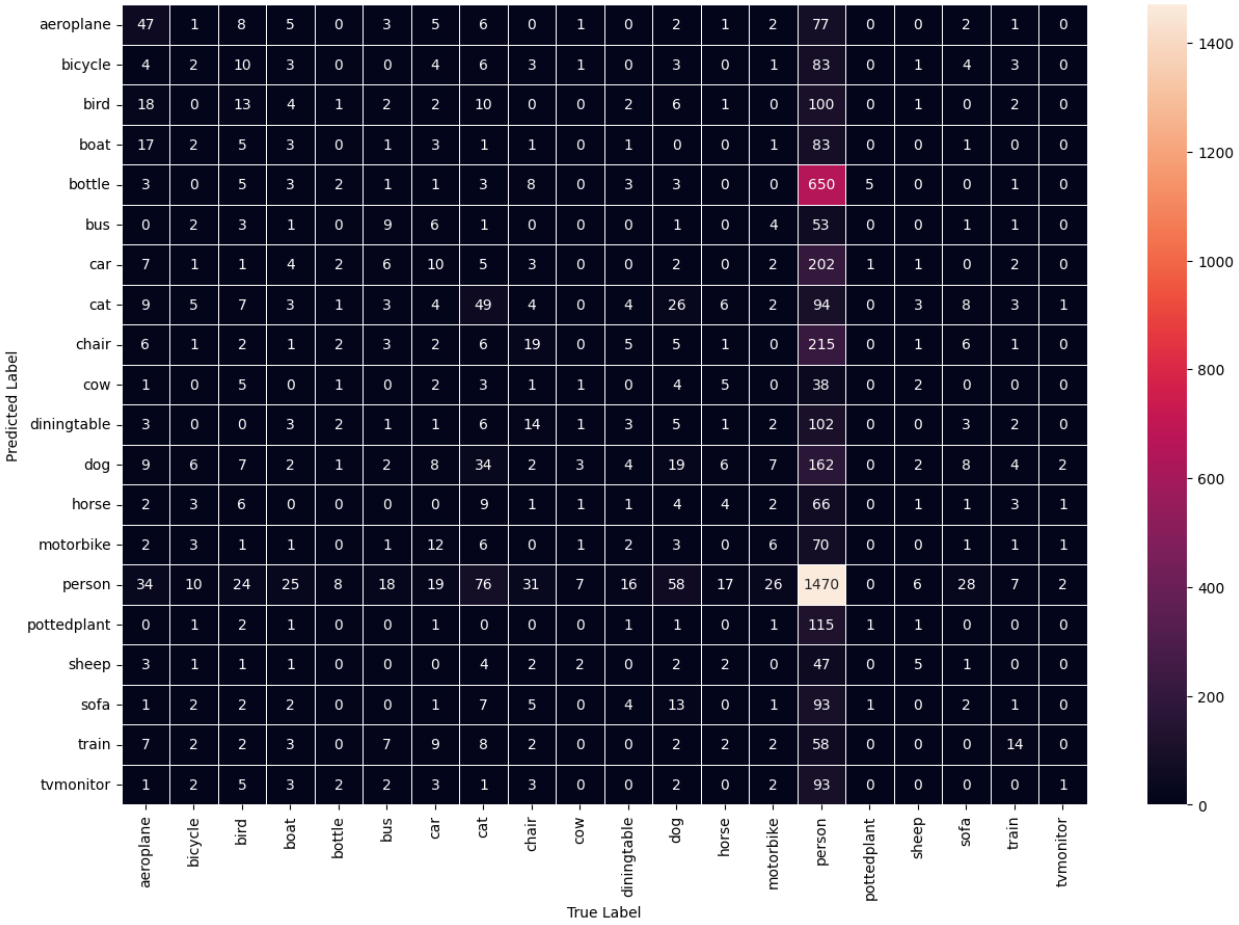
	precision	recall	f1-score	support
aeroplane	0.57	0.53	0.55	161
bicycle	0.71	0.17	0.28	128
bird	0.31	0.07	0.12	162
boat	0.33	0.13	0.19	119
bottle	0.17	0.02	0.04	172
bus	0.67	0.39	0.49	82
car	0.41	0.39	0.40	249
cat	0.34	0.11	0.17	232
chair	0.38	0.16	0.22	276
cow	0.00	0.00	0.00	63
diningtable	0.47	0.38	0.42	149
dog	0.26	0.10	0.14	288
horse	0.50	0.02	0.04	105
motorbike	0.61	0.21	0.31	111
person	0.50	0.93	0.65	1882
pottedplant	0.00	0.00	0.00	125
sheep	0.33	0.10	0.15	71
sofa	0.27	0.12	0.16	135
train	0.48	0.25	0.33	118
tvmonitor	0.68	0.41	0.51	120
accuracy			0.48	4748
macro avg	0.40	0.22	0.26	4748
weighted avg	0.43	0.48	0.40	4748

HOG with Synthetic Data



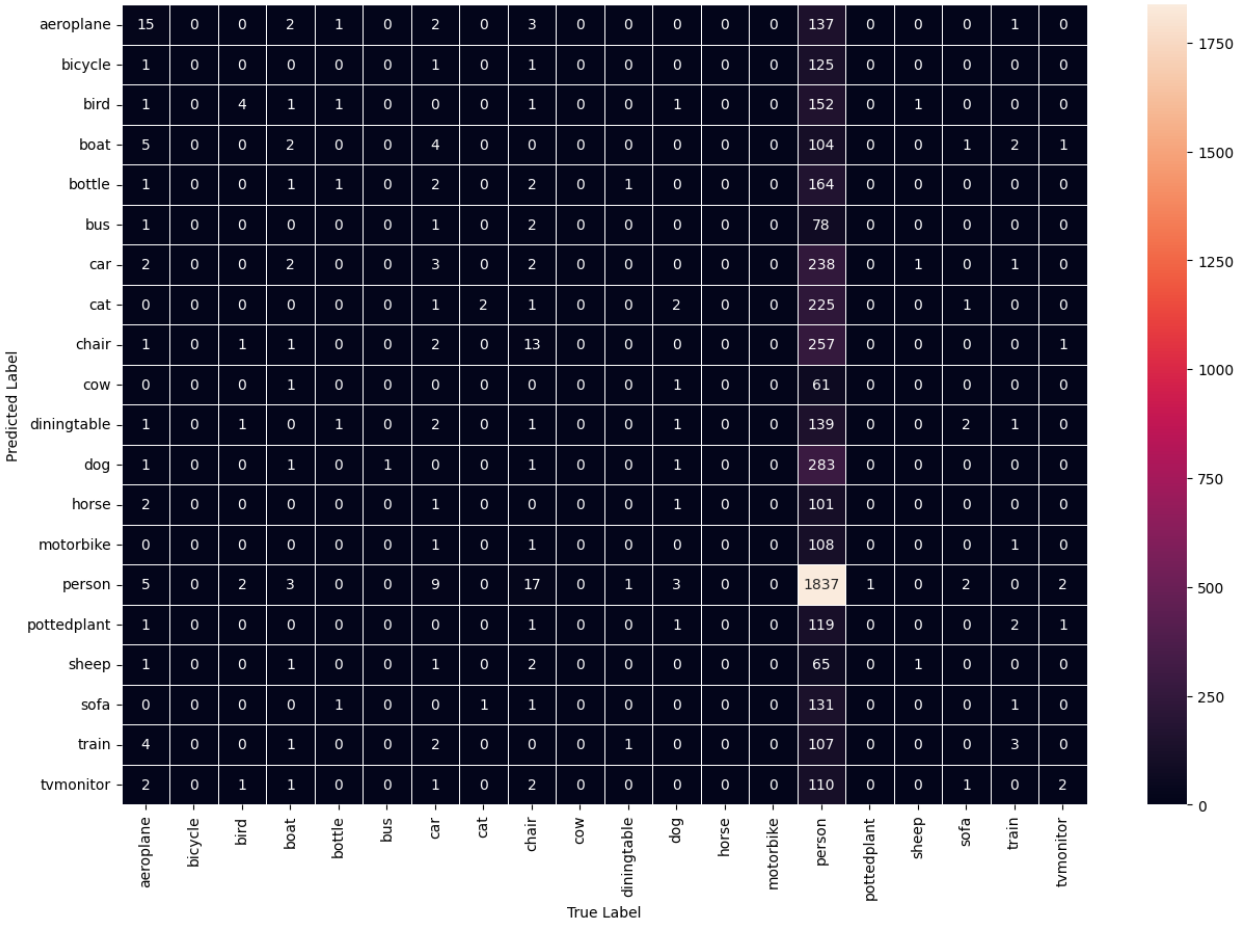
	precision	recall	f1-score	support
aeroplane	0.60	0.58	0.59	161
bicycle	0.49	0.20	0.29	128
bird	0.23	0.09	0.13	162
boat	0.32	0.14	0.20	119
bottle	0.34	0.12	0.17	172
bus	0.52	0.49	0.50	82
car	0.43	0.43	0.43	249
cat	0.38	0.15	0.22	232
chair	0.29	0.20	0.24	276
cow	0.00	0.00	0.00	63
diningtable	0.48	0.40	0.44	149
dog	0.28	0.15	0.19	288
horse	0.43	0.10	0.16	105
motorbike	0.44	0.25	0.32	111
person	0.53	0.88	0.66	1882
pottedplant	0.06	0.01	0.01	125
sheep	0.29	0.10	0.15	71
sofa	0.24	0.12	0.16	135
train	0.49	0.36	0.41	118
tvmonitor	0.57	0.42	0.48	120
accuracy			0.49	4748
macro avg	0.37	0.26	0.29	4748
weighted avg	0.43	0.49	0.43	4748

HSV Baseline



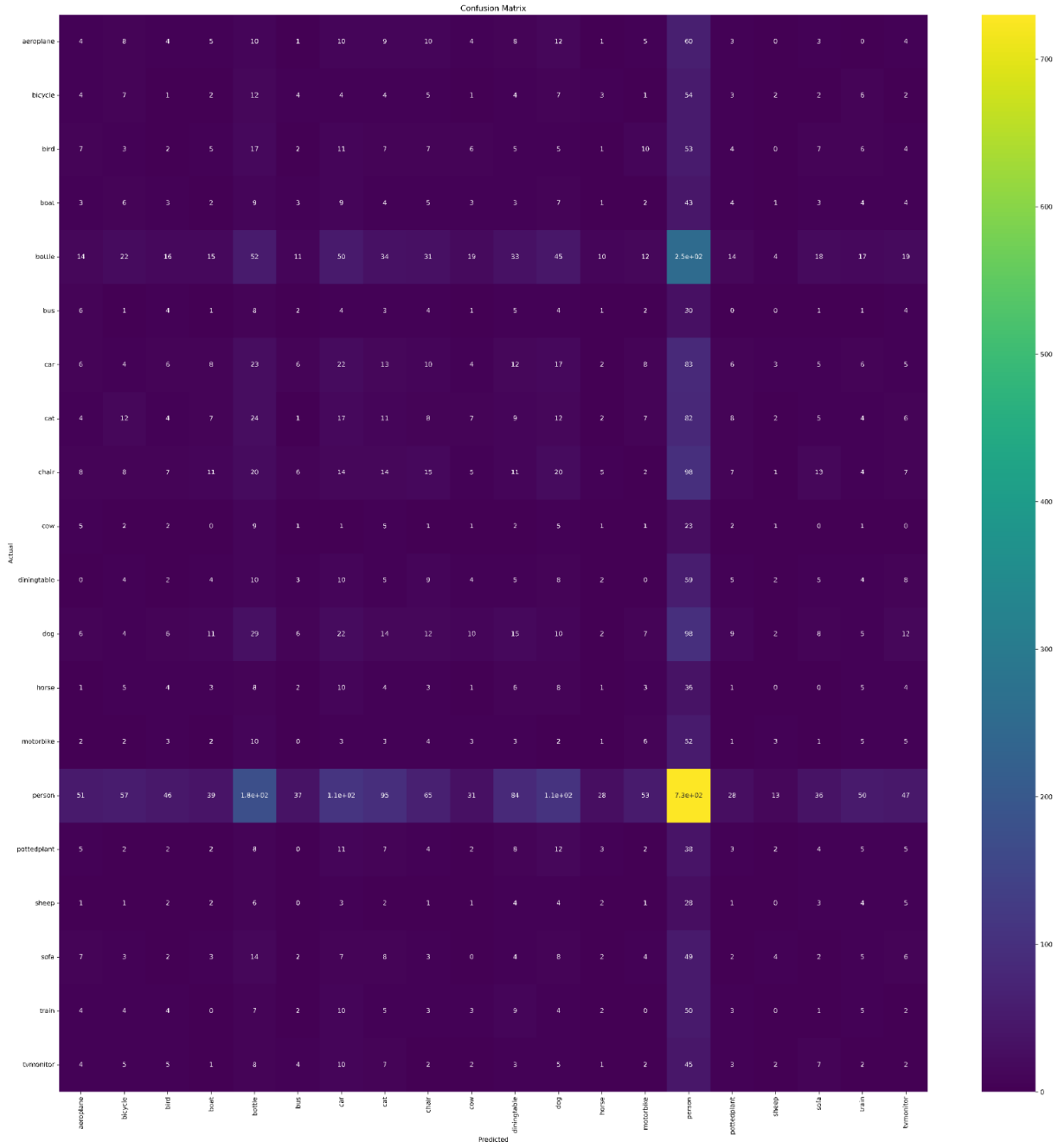
	precision	recall	f1-score	support
aeroplane	0.27	0.29	0.28	161
bicycle	0.05	0.02	0.02	128
bird	0.12	0.08	0.10	162
boat	0.04	0.03	0.03	119
bottle	0.09	0.00	0.01	688
bus	0.15	0.11	0.13	82
car	0.11	0.04	0.06	249
cat	0.20	0.21	0.21	232
chair	0.19	0.07	0.10	276
cow	0.06	0.02	0.02	63
diningtable	0.07	0.02	0.03	149
dog	0.12	0.07	0.08	288
horse	0.09	0.04	0.05	105
motorbike	0.10	0.05	0.07	111
person	0.38	0.78	0.51	1882
pottedplant	0.12	0.01	0.02	125
sheep	0.21	0.07	0.11	71
sofa	0.03	0.01	0.02	135
train	0.30	0.12	0.17	118
tvmonitor	0.12	0.01	0.02	120
accuracy			0.32	5264
macro avg	0.14	0.10	0.10	5264
weighted avg	0.22	0.32	0.23	5264

SIFT Baseline



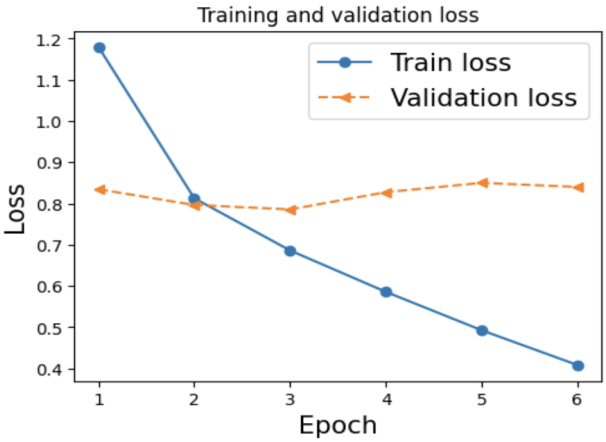
	precision	recall	f1-score	support
aeroplane	0.34	0.09	0.15	161
bicycle	0.00	0.00	0.00	128
bird	0.44	0.02	0.05	162
boat	0.12	0.02	0.03	119
bottle	0.20	0.01	0.01	172
bus	0.00	0.00	0.00	82
car	0.09	0.01	0.02	249
cat	0.67	0.01	0.02	232
chair	0.25	0.05	0.08	276
cow	0.00	0.00	0.00	63
diningtable	0.00	0.00	0.00	149
dog	0.09	0.00	0.01	288
horse	0.00	0.00	0.00	105
motorbike	0.00	0.00	0.00	111
person	0.40	0.98	0.57	1882
pottedplant	0.00	0.00	0.00	125
sheep	0.33	0.01	0.03	71
sofa	0.00	0.00	0.00	135
train	0.25	0.03	0.05	118
tvmonitor	0.29	0.02	0.03	120
accuracy			0.40	4748
macro avg	0.17	0.06	0.05	4748
weighted avg	0.27	0.40	0.24	4748

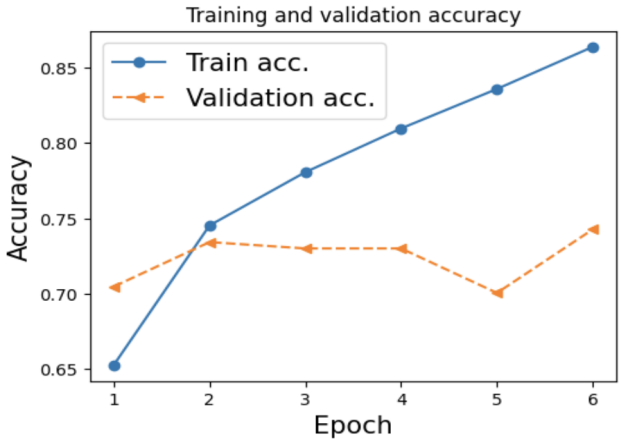
Phase 1B ResNet50



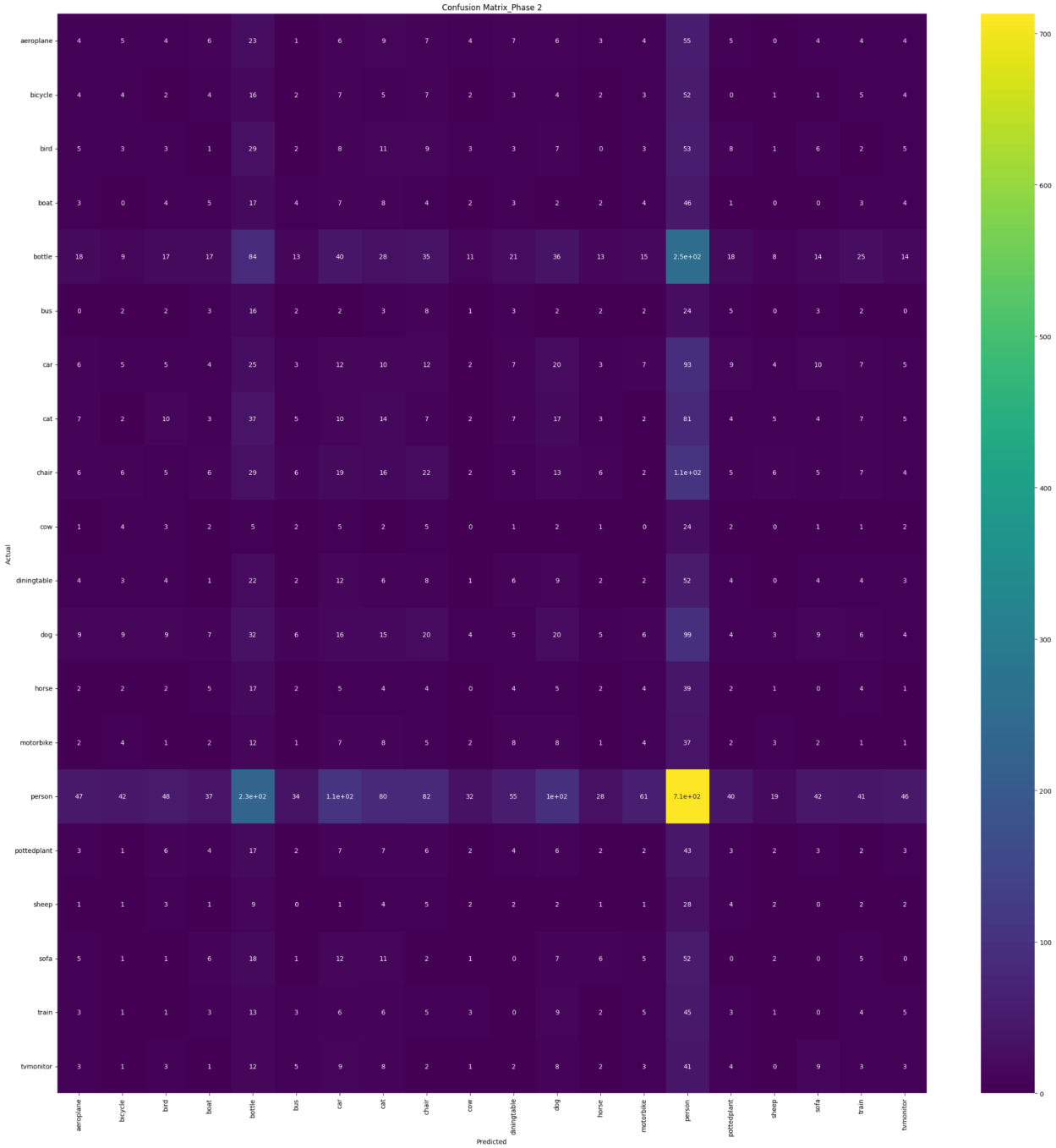
Classification Report:

	precision	recall	f1-score	support
aeroplane	0.03	0.02	0.03	161
bicycle	0.04	0.05	0.05	128
bird	0.02	0.01	0.01	162
boat	0.02	0.02	0.02	119
bottle	0.11	0.08	0.09	688
bus	0.02	0.02	0.02	82
car	0.07	0.09	0.07	249
cat	0.04	0.05	0.05	232
chair	0.07	0.05	0.06	276
cow	0.01	0.02	0.01	63
diningtable	0.02	0.03	0.03	149
dog	0.03	0.03	0.03	288
horse	0.01	0.01	0.01	105
motorbike	0.05	0.05	0.05	111
person	0.37	0.39	0.38	1882
pottedplant	0.03	0.02	0.03	125
sheep	0.00	0.00	0.00	71
sofa	0.02	0.01	0.02	135
train	0.04	0.04	0.04	118
tvmonitor	0.01	0.02	0.01	120
accuracy			0.17	5264
macro avg	0.05	0.05	0.05	5264
weighted avg	0.17	0.17	0.17	5264



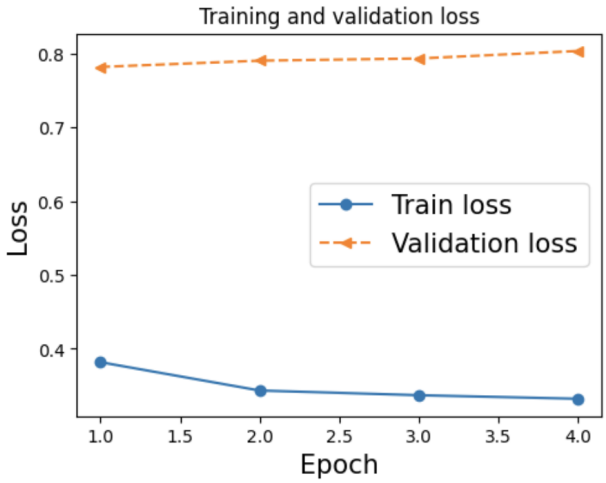


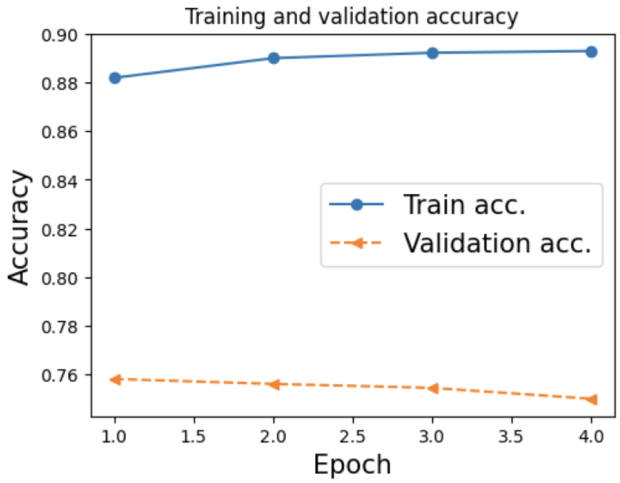
Phase 2B ResNet50



Classification Report:

	precision	recall	f1-score	support
aeroplane	0.03	0.02	0.03	161
bicycle	0.04	0.03	0.03	128
bird	0.02	0.02	0.02	162
boat	0.04	0.04	0.04	119
bottle	0.13	0.12	0.12	688
bus	0.02	0.02	0.02	82
car	0.04	0.05	0.04	249
cat	0.05	0.06	0.06	232
chair	0.09	0.08	0.08	276
cow	0.00	0.00	0.00	63
diningtable	0.04	0.04	0.04	149
dog	0.07	0.07	0.07	288
horse	0.02	0.02	0.02	105
motorbike	0.03	0.04	0.03	111
person	0.37	0.38	0.37	1882
pottedplant	0.02	0.02	0.02	125
sheep	0.03	0.03	0.03	71
sofa	0.00	0.00	0.00	135
train	0.03	0.03	0.03	118
tvmonitor	0.03	0.03	0.03	120
accuracy			0.17	5264
macro avg	0.06	0.06	0.06	5264
weighted avg	0.17	0.17	0.17	5264





.xml Annotation Files

For a given image, its filename, objects within that image (at least one, but can be more), object labels, and their region bounding boxes are listed. An example is below:

```
<annotation>
	<filename>2012_004331.jpg</filename>
	<folder>VOC2012</folder>
	<object>
		<name>person</name>
		<actions>
			<jumping>1</jumping>
			<other>0</other>
			<phoning>0</phoning>
			<playinginstrument>0</playinginstrument>
			<reading>0</reading>
			<ridingbike>0</ridingbike>
			<ridinghorse>0</ridinghorse>
			<running>0</running>
			<takingphoto>0</takingphoto>
			<usingcomputer>0</usingcomputer>
			<walking>0</walking>
		</actions>
		<bndbox>
			<xmax>208</xmax>
			<xmin>102</xmin>
			<ymax>230</ymax>
			<ymin>25</ymin>
		</bndbox>
		<difficult>0</difficult>
		<pose>Unspecified</pose>
		<point>
			<x>155</x>
			<y>119</y>
		</point>
	</object>
	<segmented>0</segmented>
	<size>
		<depth>3</depth>
		<height>375</height>
		<width>500</width>
	</size>
	<source>
		<annotation>PASCAL VOC2012</annotation>
		<database>The VOC2012 Database</database>
		<image>flickr</image>
	</source>
</annotation>
```