https://doi.org/10.2352/ISSN.2470-1173.2020.13.ERVR-381

WARHOL: Wearable Holographic Object Labeler

Matthew Shreve, Bob Price, Les Nelson, Raja Bala, Jin Sun, Srichiran Kumar
Palo Alto Research Center
Palo Alto, CA, USA

Abstract

Deep learning has significantly improved the accuracy and robustness of computer vision techniques but is fundamentally limited by access to training data. Pretrained networks and public datasets have enabled the building of many applications with minimal data collection. However, these datasets are often biased: they largely contain images with conventional poses of common objects (e.g., cars, furniture, dogs, cats, etc.). In specialized applications such as user assistance for servicing complex equipment, the objects in question are often not represented in popular datasets (e.g., fuser roll assembly in a printer) and require a variety of unusual poses and lighting conditions making the training of these applications expensive and slow. To overcome these limitations, we propose a fast labeling tool using an Augmented Reality (AR) platform that leverages the 3D geometry and tracking afforded by modern AR systems. Our technique, which we call WARHOL, allows a user to mark boundaries of an object once in world coordinates and then automatically project these to an enormous range of poses and conditions automatically. Our experiments show that object labeling using WARHOL achieves 90% of the localization accuracy in object detection tasks with only 5% of the labeling effort compared to manual labeling. Crucially, WARHOL also allows the annotation of objects with parts that have multiple states (e.g., drawers open or closed, removable parts present or not) with minimal extra user effort. WARHOL also improves on typical object detection bounding boxes using a bounding box refinement network to create perspective-aligned bounding boxes that dramatically improve the localization accuracy and interpretability of detections.

Introduction

Deep learning has achieved tremendous success in computer vision in the past decade, spanning a range of applications from image classification to object segmentation at or beyond human level accuracy. However, a key practical challenge that remains is the efficient annotation of the large datasets required to train the underlying systems. Typically, labels are drawn onto images manually by humans. To alleviate this time-consuming and often tedious process, researchers have proposed several semi-automated approaches [13, 20, 19]. Unfortunately, many of these existing approaches impose a number of constraints that limit their applicability, including a fixed camera perspective [13, 20], static scene [23, 14], or smooth motion trajectories of objects in scenes [19]. These constraints are often violated in industrial applications of computer vision. For instance, consider a printer maintenance application that uses image recognition to assist a user performing operations such as opening panels, changing lever positions and removing and replacing components. The application would ideally tailor its advice to the user based on recognizing the printer, its components and states. The objects in the scene are idiosyncratic machine parts that undergo complex articulations: panels are opened to varying degrees, covers are removed, and parts are disassembled. Such objects are not found in standard datasets, much less under the encountered variations in pose and articulation. We propose a fundamentally different approach for rapidly acquiring labeled data for real-world objects which we call **WARHOL**, a **W**e**ar**able **H**olographic **O**bject **L**abeler. The user wears an augmented reality headset with the WARHOL app installed. When the user looks through the AR display, s/he will see the object to be labeled overlaid with a mesh representing the 3D surface of the object. The user can then drop markers onto the surface of the object to indicate significant features such as doors, sub-assemblies and controls (Figure 1). Because markers are anchored to the room's reference frame, WARHOL will maintain their position under changes in user position, ambient lighting, occlusion, or object articulation. The application then projects these markers from the 3D room coordinates back into the user's display perspective in real time to provide a stream of labeled images (Figure 2). The underlying data is stored in the form of raw images and text files that can be directly used for training deep object detectors such as SSD. The ease and speed with which users can collect data with WARHOL enable rapid development of robust vision systems for industrial applications on real-world objects.

Exploratory experiments show that in a typical scenario, WARHOL generates labels 22 times faster than fully manual labeling, at 90% accuracy (IOU). While we designed the method to work on dynamic articulated objects, it is also perfectly suited to quickly collect labeled imagery of regular static objects under a diverse range of illuminations, viewpoints, and cosmetic appearance changes. While our focus has been on rapid training for custom industrial applications, we have also observed the value of the labeler for improving existing public data sets. Since many existing datasets use images off the web, they are largely restricted to conventional views of objects [22]. These views do not necessarily provide sufficient coverage for applications such as tracking objects in the household leading to poor generalization. However, by adding a small number of images from various views on even a few instances of a class, we note a much improved classification accuracy (see Section). We surmise that the web based images provide diversity in style, while WARHOL based images improve diversity in pose.

Currently AR headsets are in the early stages of development and are still expensive. In an industrial setting however, the enormous savings in time and shortening of time to market can compensate for this. The applicability and impact will only increase as the technology matures and economies of scale kick in [1]. The primary focus of this work is to test the hypothesis that the track-

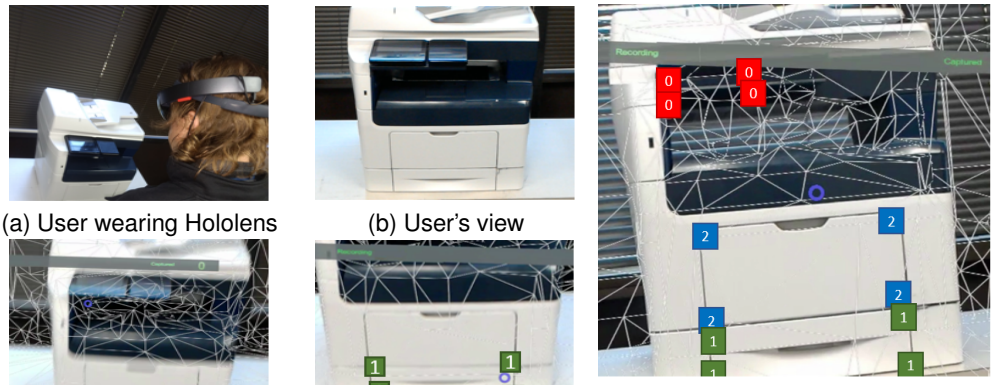


Figure 1. *WARHOL provides an intuitive way for users to quickly label a variety of movable parts on a device for training detection models.*

ing and mapping capabilities in current AR headsets are reliable enough to generate thousands of image labels with a single set of annotations. Since WARHOL is designed to work on any headset, it will leverage any technical advances made with future headsets: for example, more robust environmental mapping and tracking.

We will release both our Unity-based [4] labeling application and datasets (25,000+ images with annotated bounding boxes) to the community.

Related Work

The gold standard of data labeling is fully manual annotation with a visual user interface. In its basic form, a user clicks on an image to indicate landmark points or object boundaries. The VGG Image Annotator (VIA) [2] is one such tool that has been used to label data for object detection and semantic segmentation. With crowdsourcing platforms such as Amazon Mechanical Turks, large amounts of labeled data can be acquired quickly and inexpensively as exemplified by the ImageNet [18] project that comprises millions of images, classification labels and segmentation masks. One potential drawback of using crowdsourcing platforms is that additional protocols are often required to ensure quality and consistency of annotation; this is by itself an active research topic [21]. Also crowd-sourced approaches are often not applicable in scenarios where annotation requires expert knowledge (e.g., fuser-roller in a printer, condenser in an air conditioner, cancer node in radiograph etc.).

Several methods have been proposed introducing some level of automation to ease the labeling burden. ViPER [13, 20] uses key-frame animation style predictions to estimate (strictly linear) trajectories of objects moving in videos. LabelMe [19] interpolates future locations of bounding boxes based on feature-based tracking. Both approaches have limitations. The first relies on a fixed camera perspective with objects moving at a relatively constant rate throughout the scene. The second approach does handle some shift in perspective, however the bounding boxes remain axis aligned and do not track accurately when the object undergoes rapid translations or its parts articulate. Polygon RNN [7] guesses an initial set of polygon vertices defining the boundary of an object, which is then refined by a human operator. While this substantially reduces human effort, manual intervention is still needed for each and every object to be annotated. Similarly, there have been methods that identify relevant segmentation masks from image descriptions [26], scribbles [11], as well as a single point [5]; however, none of these methods have been shown to generate ground truth level results [7].

Recently several methods have been proposed to collect semantic segmentation labels that leverage 3D information to speed up the labeling process [23, 14, 8]. SemanticPaint [23] is an interactive VR approach that allows users to paint the surface of 3D reconstructed objects and scenes using a hand gesture that triggers a pixel-level label propagation algorithm. This system is designed to be fully online as a user provides live feedback of the labeling. Another interactive 3D labeling approach can be found in [14], wherein an initial 3D segmentation of the scene is performed using a combination of Markov Random Fields (MRF) and object localization, followed by refinement by a user. [8] uses a depth sensor and state-of-art algorithms to reconstruct a 3D indoor scene. Crowdsourced workers then annotate objects in the reconstructed 3D scenes. One major limitation of each of these aforementioned 3D methods is that they assume that the objects being labeled are static; in other words, if a single part has been articulated, the reference crowdsourced or painted label for the part is no longer valid, and the entire mesh must be reconstructed again. Given objects with multiple part articulations and the combinatorial complexity with which the overall object's appearance can change as a result (vehicles, furniture with drawers, industrial machines, etc.), this quickly becomes an intractable process and it is clear that an alternative approach is needed.

WARHOL Interface

We now describe WARHOL in greater detail. We begin with the overall design and critical components that we found necessary to create an effective and efficient annotation tool.

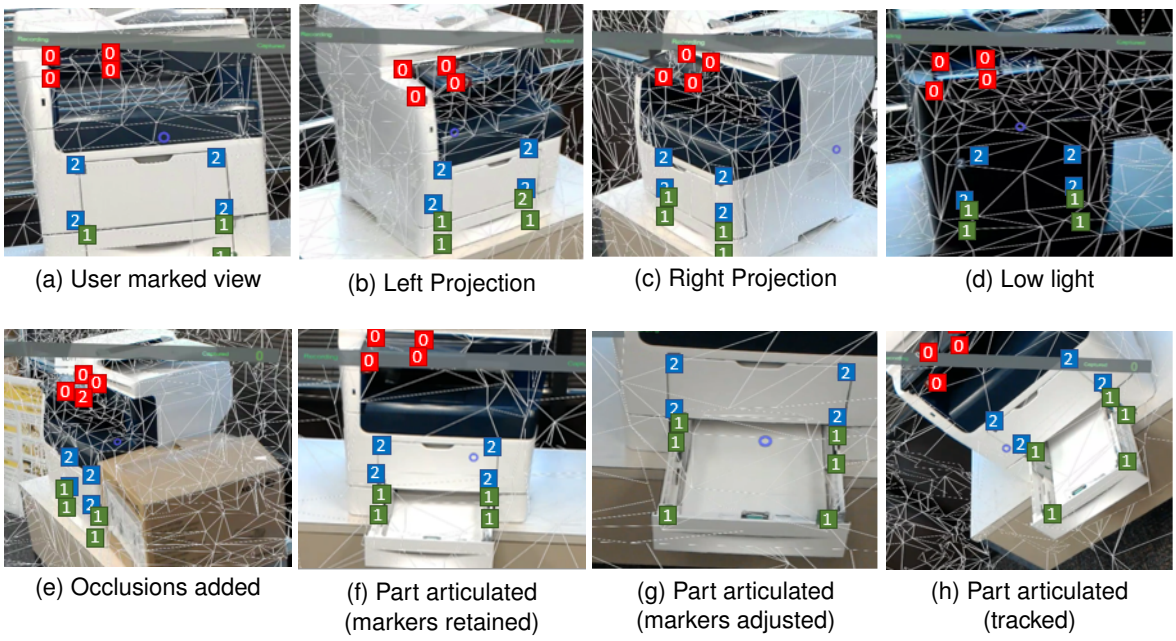


Figure 2. *WARHOLexploits the Hololens 3D reference frame to project markers from a single human labeling into video frames. This allows the user to generate views under an enormous range of conditions at 30 frames a second. The user can also articulate parts of the device and generate new labeled images with minimal changes (Views captured from WARHOL running on the Microsoft Hololens)*

Design Requirements

Intuitive Interactions. The majority of annotation tools available in the literature are restricted to a keyboard and mouse and annotate images on a computer screen. However with WARHOL, we can label objects directly in the world using the heads up display and leverage voice and gesture modes that are fully integrated with most commercially available AR devices. Our goal is to minimize the effort required by the annotator in terms of hand and gaze movements.

Low Cognitive Load Data. To reduce the cognitive burden on the annotator, there should be no difficulty disambiguating annotations once placed. We address this by leveraging depth perception and using easily distinguishable marker colors.

Flexible Object Boundary Labeling. As a general tool, we want to be able to label not only bounding boxes, but also keypoints and the detailed boundaries of target objects. WARHOL leverages the surface meshes that are produced by the AR device to initially guide placement of boundary points. These individual markers can then be adjusted using simple voice commands and/or gestures.

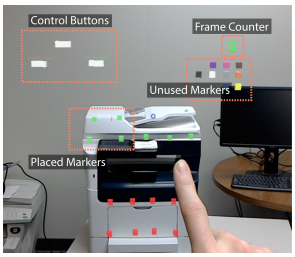
Interface Components.

As shown in (Figure) WARHOL is a virtual environment embedded in the real world from a user's point of view. It builds upon the visual components and functional modules provided by typical AR devices.

Virtual Markers. These are the unit-length cubes that can be rendered onto the real world's surfaces. Depending on the number of markers being used, it is possible to annotate bounding boxes, keypoints, or complete outlines of an object. Markers are grouped by eleven visually distinct colors based on the color naming theory of Berlin and Kay [6]. Each named color group represents a consistent visual concept: e.g., the outline of a car v. outline of a door. An unlimited number of markers can be created one by one by voice command "New #COLOR".

World Surface Mesh Map. This is obtained from the AR device. The mesh is an estimate of the $[X,Y,Z]$ coordinates of solid surfaces in the real world, and is used as a guidance to let the user quickly attach virtual markers onto real world objects. Because these meshes are produced with 3D sensors, they are robust to environmental changes in lighting and/or cosmetic appearance changes made to objects. The mesh over objects is limited in resolution, but generally tracks flat surfaces very well. Users can readily drop a marker cube on the mesh and manually refine the location. Manual adjustment is typically only required for a small number of markers where the mesh is ambiguous. The tracking accuracy of placed cubes varies with the degree of surface texture, occlusions and lighting conditions, however, reports from researchers under realistic conditions indicate that the hololens can track to within about 5mm over extended periods of time [3]. In our experience, for commonly encountered objects in the 30 to 300 cm scale, the subjectivity of the manual bounding box placement (inter-annotator agreement) is greater than the error caused by the Hololens localization. For example, 5mm error on 30cm is less than 2% error. This is discussed in more detail in Section .

Interaction Mode. We use the user's gaze (indicated by a white dot for the Hololens) to guide and attach markers onto the surface mesh. To improve marker placement accuracy after markers are initially attached, the user can use hand gestures to adjust markers freely in three dimensions, without the constraint that they lie on a mesh.



hr WARHOL interface with control buttons, virtual markers, and the capture counter.

Controls. Once markers have been placed, there are three virtual buttons for controlling the data collection procedure: start, stop, and reset. Each of these actions can also be invoked using voice commands.

Geometric Transformation. The coordinates of the markers in 3D are first transformed to camera coordinates, and then converted to the 2D image plane with the AR camera's projection matrix. In addition, a transform is stored for every captured frame based on the placement of an orientation cube in the scene.

Visibility Test. After placement, markers might be hidden in a particular view due to being out-of-sight or occluded. We check a marker's location in the 2D image plane to determine if it is out-of-sight using the surface mesh of the object.

Data Collection Workflow

We now describe the data collection process using WARHOL. A video demonstration can be found in the supplementary material.

The user begins by identifying the target object(s) to be annotated. Walking around the scene may be required for the AR device to estimate and register the various surfaces of both the object(s) and the scene. Next, the user selects a marker color by voice command and begins placing virtual markers on the target object(s). To label a single object or a part of an object, the user places markers of the same color around its extent. To label multiple objects in a room, the user can use colors or numbers (e.g., "red", "blue", "green" or "1", "2", "3" etc.) as labels for different object categories or parts. A file in csv format is used to provide human readable descriptions for each marker.

Once the initial labeling is completed, the user starts the recording process by voice command and simply walks around the scene as WARHOL takes photos from various viewing angles and lighting conditions (day, night, flashlight, etc.). Photos are captured at a preset frame rate, and all of the markers' positions for each frame are automatically calculated and stored. The process can be interrupted and resumed over an extended period of time, since the markers lock onto their respective locations. The captured images and markers' coordinates are stored with their corresponding frame indices. Multiple recording sessions can be performed in the field and then downloaded for post processing on return. We support automatic translation to popular formats such as VOC PASCAL and KITTI.

Through this interface and workflow, WARHOL is able to achieve high quality labels similar to fully manual labeling, but with much less time and effort. Consider P scenes with K objects and N number of frames to be labeled. Manual labeling will take $O(N \cdot K \cdot P)$ human time, while WARHOL takes $O(K \cdot P)$ time, because only the initial object labeling phase requires human effort. This is a significant speed up especially when N is large (e.g. tens of thousands for a typical vision dataset). A quantitative comparison is provided in Section .

Experiments

We first evaluate the efficiency and accuracy of WARHOL . Next, we report the results of training our proposed detection network that uses the perspective-aligned bounding box coordinates provided by WARHOL .

WARHOL Dataset

We collected three datasets demonstrating the flexibility of WARHOL, and make these available to the community.

Data-Room: contains images of multiple objects in a room. For each image, object instances were labeled by different colored markers. Labeling was performed by an annotator walking around a room capturing images of each object at various viewpoints. Once markers were placed, ambient lighting was adjusted, and a variety of occluders were placed around the object such as paper notes, coffee cups, and boxes. A total of 9,255 images of 30 object classes were captured in 4 room types: kitchen, office room, conference room, and office common area.

Data-Part: contains images of labeled objects parts. This is crucial in applications where the user needs to interact with specific parts of an object (e.g. locating the jammed paper tray of a printer). This dataset comprises 10,872 images of 9 different articulated parts of 7 office printers in various offices and common areas.

Data-Shape: This dataset demonstrates WARHOL's capability to perform more complex annotations. It comprises 4 stuffed animals placed around an office common area: lizard, cheetah, giraffe, and zebra. For each animal, WARHOL was used to place virtual markers along its physical outline from one viewpoint. We additionally placed 4 markers as a reference box to bound the shape. A total of 5192 images were collected with an average of 48 points per animal.

The three datasets in total contain over 25,000 labeled images; typical examples are shown in Figure 3.

WARHOL Annotation Quality

In this experiment, we compare WARHOL quantitatively with image labeling tools that are widely used in the computer vision community.

We randomly selected 50 frames from the Data-Part dataset, containing on average 6 parts per frame for a total of approximately 300 parts. Each part is labeled by a polygon with 4 vertices. We compare WARHOL with representative manual labeling, semi-automatic video labeling, and crowd-sourced labeling methods. For manual labeling, we used an online tool (VIA) [2]. To obtain a measure of manual labeling consistency, two experts annotated a set of images and we computed IOU among the annotations. For semi-automatic video labeling, we use a popular video-based labeling method called VATIC [24], in which a user labels key frames and the software predicts the labels for the remaining frames by tracking visual features. One volunteer was hired to complete both tasks. For crowd-sourced labeling, we developed a web-based manual labeling interface that serves selected frames to Amazon Mechanical Turkers for annotation. The crowd workers were guided to label object parts using a few ref-

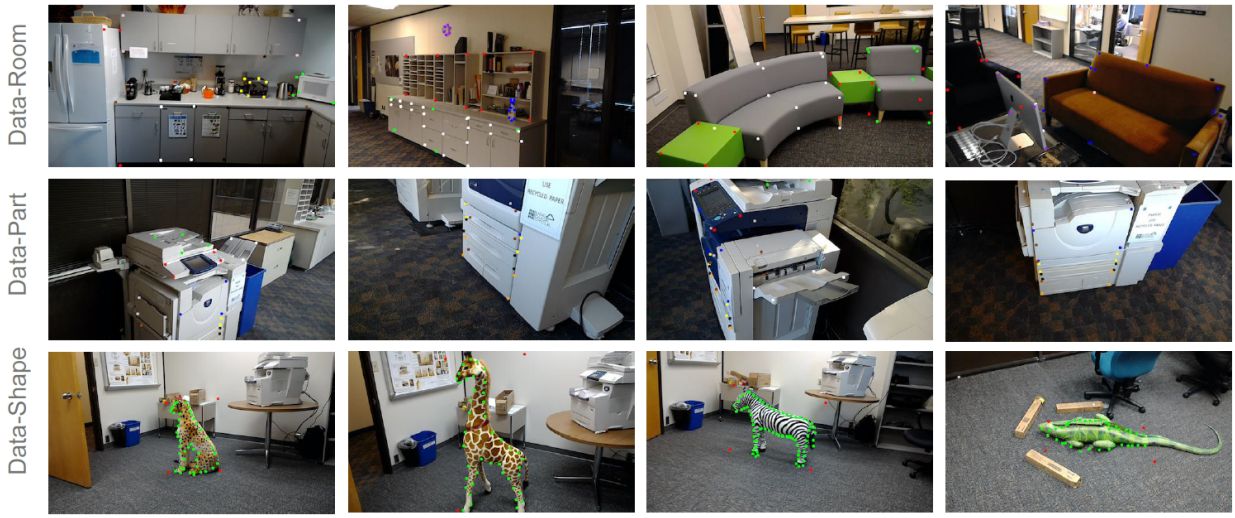


Figure 3. *Example images and marker labels from our datasets.*

erence images as exemplars.

Table 1 shows that the accuracy of WARHOL is comparable to that of manual annotation ($0.74/0.82 = 90\%$) with only a fraction (5%) of the time cost for labeling the selected frames. Moreover, the time required for manually labeling N frames is proportional to N; while WARHOL requires no additional time spent for labeling mutiple frames for the same objects in the same scene. For video labeling, VATIC required heavy user interventions due to the highly non-linear trajectories of objects and parts when captured using a Hololens. With heavy key frame adjustments (60% of total frames), it is able to achieve about 57% (0.47/0.82) accuracy in IOU. With even more key frame adjustments, we expect VATIC to further improve accuracy, but with human expended time approaching that of a fully manual process.

For crowd-sourced annotations, we assigned 150 HITs on Amazon MTurks and recruited 26 workers to complete our annotation task in two days. On average, there were three workers annotating each image. The final polygon was generated by majority votes of labeled pixels. Table 1 shows the comparison of the quality of the MTurk labels and WARHOL. Overall, MTurk labels have a comparable label accuracy but a much higher turnaround time.

Bounding Box Refinement Network (BRN)

In many real-time augmented reality applications, the bounding box from the object detection step is used to create virtual overlays that assist users in various tasks such as including machine repair [25], assembly or surgery [16]. In these domains, axis-aligned bounding boxes (AABB) are confusing since they "float" on top of the object rather than aligning with its 3D orientation and include large regions of the image that are not part of the object. Axis aligned bounding boxes are especially confusing in scenes with labels on many neighbouring parts, as they begin to overlap (see Figure 4a).

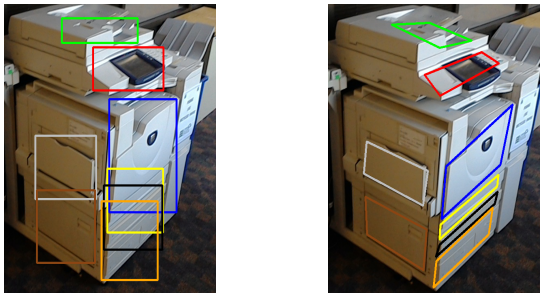
We conducted a user study to understand how bounding box geometry affects visual task guidance. We collected twelve images of large office printer and labeled 8 parts using both axis aligned bounding boxes (AABBs) and perspective-aligned bounding boxes (PABBs) at varying perspectives. We then asked five volunteers to identify which part was indicated by each bounding box in each image. Part identification accuracy was approximately 72% on AABB-labeled images, and as expected, 100% with the PABB-labeled images (images in the latter category were primarily used as a control). In addition, for the AABB-labeled images, 3 out of 4 instances had at least one user indicating the wrong part. It is also worth noting that more errors were made on images where the parts had large out-of-plane rotations (for example, see the blue, yellow, and orange bounding boxes in Figure 4). We believe this result indicates that tighter fits to object shapes (higher IOU) result in higher visual identification. Since WARHOL stores the true locations of markers and projects them into images using a camera model, it can readily generate bounding boxes that are correctly aligned to any perspective. To fully leverage the benefits of these object-aligned bounding boxes, we designed a Bounding Box Refinement network (BRN). It receives an image and a canonical AABB provided by a standard object detection network, and predicts the parameters of a spatial transformation that maps the AABB to a target perspective-aligned bounding box (PABB) that optimally fits the object's spatial extent. Figure 5 illustrates the network structure. We emphasize that BRN is a general purpose network that can be applied to arbitrary datasets beyond those collected through WARHOL. Our pipeline for predicting PABBs as generated by WARHOL is as follows:

1. Collect images and annotations using WARHOLby following the workflow described in Section .
2. For each annotation, calculate the AABBs for each detected object or part and use this data to train SSD [15].
3. The cropped regions for each detection output from SSD are collected and used to train BRN, which predicts PABBs.

The BRN, as a function of the input image I, minimizes $\|X_t - T(I)X_s\|_2$, where X_s are raw axis-aligned bounding box coordinates and X_t are the target coordinates that are aligned with the object's orientation. $T(I)$ is a 3×3 transformation matrix that can be flattened to a 9-dimensional vector.

		VATIC				
	Manual	Initialize	10% Keyframes	60% Keyframes	MTurk	WARHOL
IOU	0.82	0.19	0.27	0.47	0.71	0.74
Time (50 Frames)	133	1	2	10	48 hrs	6
Time (N Frames)	2.66N	1	0.04N	0.24N	N/A	6

Accuracy and efficiency comparison between manual labeling, VATIC, Amazon MTurks, and WARHOL. Manual labeling IOU is the inter-annotator agreement. We report VATIC performance after annotating and/or adjusting a varying number of keyframes. Times are in minutes unless specified otherwise.



(a) AABB are hard to associate with parts and overlap.

(b) PABB clearly align with object parts and are distinct.

Figure 4. *A comparison of axis-aligned bounding boxes (AABB) and perspective-aligned bounding boxes (PABB). Figure best viewed in color.*

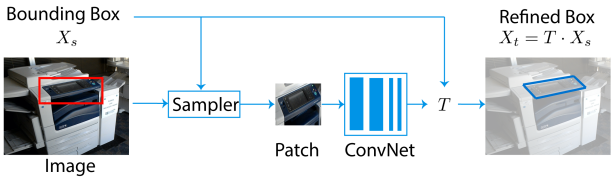


Figure 5. *Network structure of BRN.*

There are several transformation choices for $T(\cdot)$, including Rotation, Rotation+Scale, Affine, and Homography. It is relatively straightforward to modify the number of regression outputs in BRN to output any of the any of these transformations, of which homography is the most general. With simplicity and efficiency in mind, the network is designed with two convolutional filter layers, two batch normalization layers, and a fully connected regression layer.

We investigate BRN's ability to improve IOU over the standard axis-aligned bounding box. We first run the input image through an SSD object detector to obtain axis-aligned ROIs for each object, which are then supplied to BRN for refinement. We then compare different parametric BRN transformations with end-to-end coordinate regression and a rotated bounding box approach. The end-to-end method takes in an image region of interest and directly regresses coordinates of all bounding box points. The rotated bounding box approach predicts $[x1, y1, x2, y2, \theta]$ parameters [10] that represent a rectangular shape with arbitrary rotations.

Each dataset is randomly split into 90% training and 10% testing. On the Data-Part dataset, BRN with homography transformation outperforms all methods, with end-to-end training coming closest to a comparable performance. However on the Data-Shape dataset, BRN is significantly (70%) better than the end-to-end training. One critical reason is that learning four coordinates regression is much simpler than learning all boundary coordinates (Table 3). Figure 6 shows that the end-to-end training fails completely for complex shapes. On the other hand, BRN does not learn a direct coordinate mapping, but rather a simple transformation between image spaces.

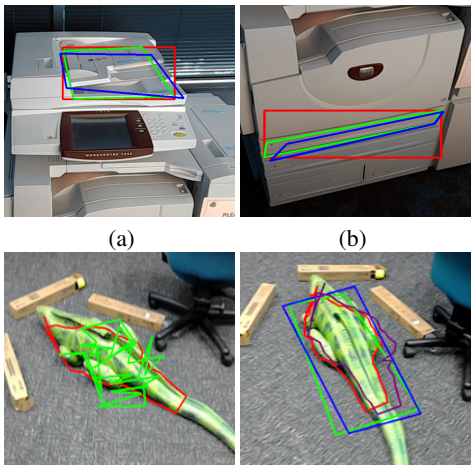


Figure 6. *(a) and (b): Example of Data-Part predictions, axis-aligned box (red), ground truth box (green), BRN estimated box (blue). (c) End-to-End predicted shape (green) compared with ground truth (red). (d) BRN estimated shape (purple) compared with ground truth (red)*

Method	AABB	End2End	xyxyθ	BRN(R+S)	BRN(H)
IOU	0.44	0.63	0.53	0.60	0.68
MSE	0.50	0.17	0.75	0.19	0.13

Comparison of different bounding box refinement methods on Data-Part dataset. BRN(R+S): BRN with rotation + scale. BRN(H): BRN with homography.

Method	AABB	End2End	BRN(H)
IOU (37 pts)	0.322	0.39	0.678
IOU (73 pts)	0.17	0.19	0.623

Comparison of different bounding box refinement methods on Data-Shape dataset, with different number of shape points. PABBs produced by BRN completely outperform AABBs.

Data Augmentation for Enhanced Object Detection

As mentioned earlier, standard object detection datasets including PASCAL VOC [9], DAVIS [17] and COCO [12] are often biased in terms of pose and lighting conditions, leading to mixed

results when used to train object detectors for real-world applications. In this experiment, we show that by using WARHOL, it is possible to quickly collect a large number of labeled images that can then augment these existing datasets to train a more robust detector with superior performance. We selected samples from the 'monitor' class in the PASCAL 2012 VOC datasets. We then used the WARHOL tool to collect 3000 additional images of six different types of monitors in the span of 2 hours. We combined this new dataset with PASCAL and trained an SSD detector with different combinations of training/testing splits. The result is summarized in Table 4. The first entry of this matrix [PASCAL (train) / PASCAL (test)] used a publicly available SSD model that was pre-trained on the PASCAL VOC dataset.

	PASCAL (test)	AR (test)
PASCAL (train)	0.52	0.39
PASCAL+AR (train)	0.66	0.49

Average precision with different training configurations of a SSD detector trained on monitors.

The PASCAL trained detector does not work well on our AR data because many monitor images in PASCAL are taken from a frontal view, and as a result, the detector performs poorly on non-frontal views in the AR data. When jointly trained with PASCAL and AR data, the SSD detector is able to improve performance significantly on both the AR dataset and the PASCAL dataset by about 20%. This suggests that the AR data provides valuable pose and environmental variations in training an object detector to complement the rich object type variation in the PASCAL data. This experiment illustrates the promise of WARHOL in efficiently filling the gap between the performance of detectors built on existing benchmark datasets and that of detectors applied in real world settings with enormous variations in pose, lighting and other confounding variables.

Conclusion

We propose an augmented reality application, WARHOL, that enables intuitive and efficient large scale data annotation with greatly reduced labeling time, and increased sample diversity when compared to standard data collection and labeling methods. WARHOL is particularly beneficial for annotating specialized objects with complex structure and/or articulated parts. It can be used to generate standard axis-aligned bounding boxes, perspective-aligned boxes, object key-points and detailed outlines quickly and accurately. It labels images 20 times faster than typical manual labeling, and is useful to rapidly create new datasets that can be used standalone to train accurate object detection systems, or combined with existing benchmark datasets to further improve state-of-the-art detectors. Additionally, we present a bounding box refinement network (BRN) that predicts perspective-aligned bounding boxes using an efficient neural network module with minimal computational overhead. It refines the axis-aligned bounding boxes extracted from a standard object detector to predict skewed boxes that more closely align with an object's visual extent. We demonstrate significantly improved IOU (50% improvement over axis-aligned bounding boxes) for general object detection applications. Finally we contribute three datasets collected using WARHOL that exhibit rich diversity in pose, lighting, and background clutter.

References

[1] Business wire, 2017. https://www.businesswire.com/news/home/20170619005183/en/Worldwide-Shipments-Augmented-Reality-Virtual-Reality-Headsets.

[2] VGG image annotator, 2017. http://www.robots.ox.ac.uk/ vgg/software/via/.

[3] How accurate is the hololens, 2019.

[4] Unity technologies, 2019. http://unity.com/.

[5] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. *What's the Point: Semantic Segmentation with Point Supervision*, pages 549–565. Springer International Publishing, Cham, 2016.

[6] B. Berlin and P. Kay. CSLI Publications, 1969.

[7] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler. Annotating object instances with a polygon-rnn. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4485–4493, 2017.

[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[9] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136, Jan. 2015.

[10] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo. R2cnn: Rotational region cnn for orientation robust scene text detection. *CoRR*, abs/1706.09579, 2017.

[11] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016.

[12] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[13] V. Mariano, J. Min, J.-H. Park, R. K. amd D. Mihalcik, D. Doermann, and T. Drayer. Performance evaluation of object detection algorithms. pages 965–969, 2002.

[14] D. T. Nguyen, B. Hua, L. Yu, and S. Yeung. A robust 3d-2d interactive tool for scene segmentation and annotation. *CoRR*, abs/1610.05883, 2016.

[15] C. Ning, H. Zhou, Y. Song, and J. Tang. Inception single shot multibox detector for object detection. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 549–554, July 2017.

[16] Y. Oyamada. A look into medical augmented reality. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 1–1, Sept 2014.

[17] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, June 2016.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. A Robust 3D-2D Interactive Tool for Scene Segmentation and Annotation. *IEEE Transactions on Visualization and Computer Graphics (2017)*, 2017.

[19] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, May 2008.

[20] M. A. Serrano, J. Gracía, M. A. Patricio, and J. M. Molina. *In-*

teractive Video Annotation Tool, pages 325–332. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[21] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *In IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[22] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[23] J. Valentin, V. Vineet, M.-M. Cheng, D. Kim, J. Shotton, P. Kohli, M. Nießner, A. Criminisi, S. Izadi, and P. Torr. Semanticpaint: Interactive 3d labeling and learning at your fingertips. *ACM Transactions on Graphics (TOG)*, 34(5):154, 2015.

[24] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21. 10.1007/s11263-012-0564-1.

[25] I. Wijesooriya, D. Wijewardana, T. D. Silva, and C. Gamage. Demo abstract: Enhanced real-time machine inspection with mobile augmented reality for maintenance and repair. In *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 287–288, April 2017.

[26] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and i will show you where it is. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3197, June 2014.

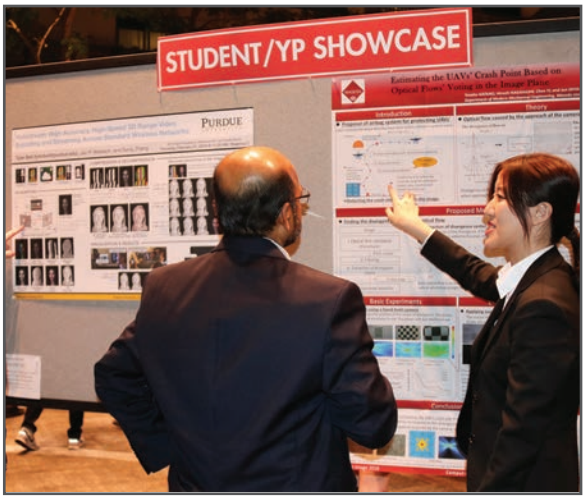
JOIN US AT THE NEXT EI!

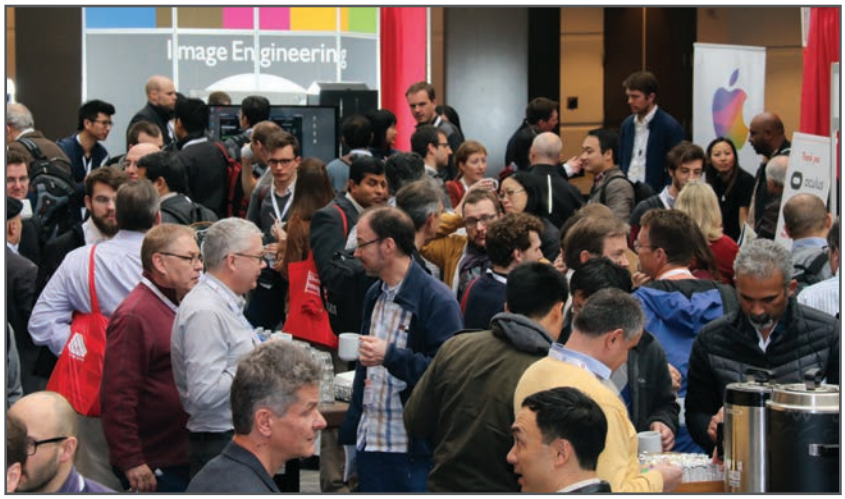
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org