Improving Random GUI Testing with Image-Based Widget Detection

Thomas D. White
tdwhite1@sheffield.ac.uk
Department of Computer Science,
The University of Sheffield
Sheffield, United Kingdom

Gordon Fraser
gordon.fraser@uni-passau.de
Chair of Software Engineering II,
University of Passau
Passau, Germany

Guy J. Brown
g.j.brown@sheffield.ac.uk
Department of Computer Science,
The University of Sheffield
Sheffield, United Kingdom

ABSTRACT

Graphical User Interfaces (GUIs) are amongst the most common user interfaces, enabling interactions with applications through mouse movements and key presses. Tools for automated testing of programs through their GUI exist, however they usually rely on operating system or framework specific knowledge to interact with an application. Due to frequent operating system updates, which can remove required information, and a large variety of different GUI frameworks using unique underlying data structures, such tools rapidly become obsolete, Consequently, for an automated GUI test generation tool, supporting many frameworks and operating systems is impractical. We propose a technique for improving GUI testing by automatically identifying GUI widgets in screen shots using machine learning techniques. As training data, we generate randomized GUIs to automatically extract widget information. The resulting model provides guidance to GUI testing tools in environments not currently supported by deriving GUI widget information from screen shots only. In our experiments, we found that identifying GUI widgets in screen shots and using this information to guide random testing achieved a significantly higher branch coverage in 18 of 20 applications, with an average increase of 42.5% when compared to conventional random testing.

CCS CONCEPTS

• **Software and its engineering → Software testing and debugging**; *Empirical software validation.*

KEYWORDS

GUI testing, object detection, random testing, black box testing, software engineering, data generation, neural networks

ACM Reference Format:
Thomas D. White, Gordon Fraser, and Guy J. Brown. 2019. Improving Random GUI Testing with Image-Based Widget Detection. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '19), July 15–19, 2019, Beijing, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3293882.3330551

ISSTA '19, July 15–19, 2019, Beijing, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6224-5/19/07…$15.00
https://doi.org/10.1145/3293882.3330551

1 INTRODUCTION

A Graphical User Interface (GUI) enables events to be triggered in an application through visual entities called widgets (e.g., buttons). Users interact using keyboard and mouse with the widgets on a GUI to fire events in the application. Automated GUI test generation tools (e.g., AutoBlackTest [15], Sapienz [14], or GUITAR [16]) simulate users by interacting with the widgets of a GUI, and they are increasingly applied to test mobile and desktop applications. The effectiveness of these GUI test generation tools depends on the information they have available. A naïve GUI test generator simply clicks on random screen positions. However, if a GUI test generator knows the locations and types of widgets on the current application screen, then it can make better informed choices about where to target interactions with the program under test.

GUI test generation tools tend to retrieve information about available GUI widgets through the APIs of the GUI library of the target application, or an accessibility API of the operating system. However, relying on these APIs has drawbacks: Applications can be written using many different GUI libraries and widget sets, each providing a different API to access widget information. Although this can be circumvented by accessibility APIs, these differ between operating systems, and updates to an operating system can remove or replace parts of the API. Furthermore, some applications may not even be supported by such APIs, such as those which draw directly to the screen, e.g., web canvasses [2]. These challenges make it difficult to produce and to maintain testing tools that rely on GUI information. Without knowledge of GUI widgets, test generation tools resort to blindly interacting with random screen locations.

To relieve GUI testing tools of the dependency on GUI and accessibility APIs, in this paper we explore the use of machine learning techniques in order to identify GUI widgets. A machine learning model is trained to detect the widget types and positions on the screen, and this information is fed to a test generator which can then make more informed choices about how to interact with a program under test. However, generating a widget prediction model is non-trivial: Different GUI libraries and operating systems use different visual appearance of widgets. Even worse, GUIs can often be customized with user-defined themes, or assistive techniques such as a high/low contrast graphical modes. In order to overcome this challenge, we randomly generate Java Swing GUIs, which can be annotated automatically, as training data. We explore the challenge of generating a balanced dataset that resembles GUIs in real applications. The final machine learning model uses only visual data and can identify widgets in a real application's GUI without needing additional information from an operating system or API.

In detail, the contributions of this paper are as follows:

- We describe a technique to automatically generate labeled GUIs in large quantities, in order to serve as training data for a GUI widget prediction model.
- We describe a technique based on deep learning that adapts machine learning object detection algorithms to the problem of GUI widget detection.
- We propose an improved random GUI testing approach that relies on no external GUI APIs, and instead selects GUI interactions based on a widget prediction model.
- We empirically investigate the effects of using GUI widget prediction on random GUI testing.

In our experiments, for 18 out of 20 Java open source applications tested, a random tester guided by predicted widget locations achieved a significantly higher branch coverage than a random tester without guidance, with an average coverage increase of 42.5%. Although our experiments demonstrate that the use of an API that provides the true widget details can lead to even higher coverage, such APIs are not always available. In contrast, our widget prediction library requires nothing but a screen shot of the application, and even works across different operating systems.

2 BACKGROUND

Interacting with applications through a GUI involves triggering events in the application with mouse clicks or key presses. Lo et al. [13] define three types of widgets in a GUI:

- Static widgets in a GUI are generally labels or tooltips.
- Action widgets fire internal events in an application when interacted with (e.g. buttons).
- Data widgets are used to store data (e.g., text fields).

In this paper, we focus on identifying action and data widgets. The simplest approach to generating GUI tests is through clicking on random places in the GUI window [9], hoping to hit widgets by chance. This form of testing ("monkey testing") is effective at finding crashes in applications and is cheap to run; no information is needed (although knowing the position and dimensions of the application on the screen is helpful). Monkey testing is now also commonly used to test mobile applications through tools like the Android Monkeyrunner [1].

GUI test generation tools can be made more efficient by providing them with information about the available widgets and events. This information can be retrieved using the GUI libraries underlying the widgets used in an application, or through the operating system's accessibility API. For example, Bauersfeld and Vos created GUITest [3] (now known as TESTAR), which uses the operating system's accessibility API to identify possible GUI widgets to interact with, and then randomly chooses from the available widgets during test generation. TESTAR has been applied to many industrial applications, including a web-based application from the rail sector [6]. The AutoBlackTest tool [15] relies on a commercial testing tool (IBM Rational Functional Tester) to retrieve widget information, and then uses Q-Learning to select the most promising widgets for interaction.

In contrast to these randomized approaches, GUI ripping [16] aims to identify *all* GUI widgets in an application to permit systematic test generation. However, a recent study by Nguyen et al. [17] found that, although GUI ripping enables effective testing and flexible support for automation, there are drawbacks, mainly related to the GUI ripping. For example, GUI trees have to be manually validated, and component identification issues can lead to inaccurate GUI trees being generated.

The problem of widget identification is not only relevant for test input generation, but also for asserting the test outcome. For example, the Sikuli [24] tool uses OpenCV, an image processing library, to match images of GUI widgets saved in a test against the current application's GUI. Matching an image of the button decreases the chance of tests failing if the application's GUI changes intentionally (e.g., by moving the button from the top of a GUI to the bottom). Sikuli also features assertions, in which tests can check that some part of the GUI exists after performing an interaction on the application under test. However, Sikuli tries to exactly match images of previously seen widgets, and therefore cannot be used to identify previously unseen widgets.

All the current approaches to testing an application through its GUI rely on an automated method of extracting widget information from a GUI. Applications and application scenarios exist where widget information cannot be automatically derived, and tools may fall back to random testing. However, object detection and image labelling may be able to help with this.

3 PREDICTING GUI WIDGETS FOR TEST GENERATION

In order to improve random GUI testing, we aim to identify widgets in screen shots using machine learning techniques. A challenge lies in retrieving a sufficiently large labeled training dataset to enable modern object recognition approaches to be applied. We produce this data by (1) generating random Java Swing GUIs, and (2) labelling screen shots of these applications with widget data retrieved through GUI ripping based on the Java Swing API. The trained network can then predict the location and dimensions of widgets from screen shots during test generation, and thus influence where and how the GUI tester interacts with the application.

3.1 Identifying GUI Widgets

Environmental factors such as the operating system, user-defined theme, or application designer choice effect the appearance of widgets. Each application can use a unique widget palette. When widget information cannot be extracted through use of external tools, e.g., an accessibility API, then this diversity of widgets presents a problem for GUI testing tools. For example, applications that render GUIs directly to an image buffer (e.g., web canvas applications) generally cannot have their GUI structure extracted automatically. Pixels are drawn directly to the screen and there is no underlying XML or HTML structure to extract widget locations. We propose a technique of identifying GUI widgets solely through visual information. This is an instance of object detection, i.e., the process of automatically extracting and "tagging" objects in an image.

Some methods such as Region-based Convolutional Neural Network (R-CNN) by Girshick et al. [12] work by selecting areas of the image to input through the neural network. Convolutional Neural Networks (CNN) identify patterns in images and can be expensive to compute, especially if a sliding window inputs subsets of

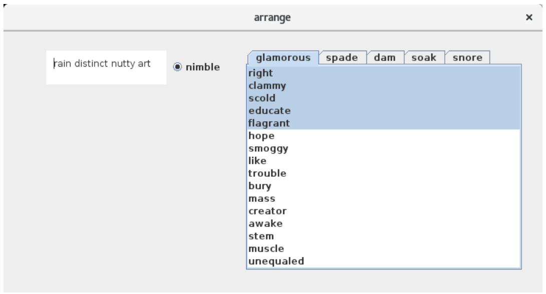


Figure 1: A generated Java Swing GUI. Each element has a random chance to appear.

Algorithm 3.1: RandomWidgetTree(*nodeCount*)

```
nodes = [Container]
while |nodes| < nodeCount
  do {nodes ← nodes ⋃ RandomWidgetType()
while (|nodes| > 1)
  do
    node ← sample(nodes, 1)
    parent ← sample(nodes, 1)
    if isContainer(parent) and node ≠ parent
      then
        parent.children ← parent.children ⋃ node
        nodes ← nodes \ node
return (nodes[0])
```

the image through a CNN multiple times. During GUI testing, it is beneficial to recognize widgets quickly so more actions can be performed in less time. Therefore, we use You Only Look Once (YOLO), proposed by Redmon et al. [18], which labels an image by seeding the whole image through a CNN once. YOLO is capable of predicting the positions and dimensions of objects in an image.

The input to YOLO is an image with width and height being a multiple of 32 pixels and equal in value. To predict labels, YOLO continuously downsamples the input image into $N \cdot N$ grid cells, where N is the width or height divided by 32 in the last layer. For example, if the input dimension is $(416, 416)$, YOLO will predict widgets in a $(13, 13)$ grid. YOLOv2 [19] is an extension to YOLO, predicting B boxes per grid cell. A cell is responsible for a prediction if the center of a box falls inside the respective cell.

Each box contains five predicted values: the location (x, y), the dimension $(width, height)$ and a confidence score for the prediction (c). Predicting multiple boxes per grid cell aids in training as it allows different aspect ratios to be used for each box in each cell. The aspect ratios are passed to the algorithm to modify the predicted width and height of each box. To calculate well-suited aspect ratios, the dimensions of all boxes in the training data are clustered into N clusters, where N is the number of boxes predicted per cell. The centroid of each cluster represents the aspect ratios to supply to YOLOv2.

A single class is predicted from C predefined classes for each grid cell. In total, this makes the network's output $N{\cdot}N{\cdot}(B{\cdot}5{+}C)$. We can now filter the predicted boxes using the confidence values. Boxes with a confidence value close to zero may not be worth investigating, and can be eliminated using a lower confidence threshold. This will be discussed further in section 3.3.

Algorithm 3.2: RandomJFrame(*width*, *height*, *nodeCount*)

```
procedure ApplyWidget(container,widget)
  swingComponent ← ComponentFromWidget(widget)
  i ← 0
  while i < |widget.children|
    do
      child ← widget.children[0]
      ApplyWidget(swingComponent, child)
      i ← i + 1
  container.add(swingComponent)

jframe ← JFrame<init>(width, height)
rootNode ← RandomWidgetTree(nodeCount)
ApplyWidget(jframe, rootNode)
return (jframe)
```

Using the YOLOv2 convolutional neural network, we can automatically identify GUI components in a screen shot. We chose the YOLOv2 algorithm for our network due to the speed it can process entire images, and the accuracy it achieves on predictions. Our implementation of YOLOv2 only uses greyscale images, so the first layer of the network only has a single input per pixel opposed to the three (r, g, b) values proposed in the original network [19].

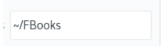
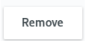
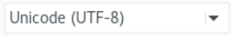
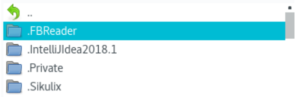
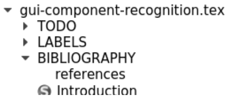
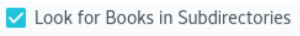
3.2 Generating Synthetic GUIs

One issue with using a neural network is that it requires large amounts of labeled data for training. To obtain labeled screen shots, we generate synthetic applications. A synthetic application is one with no event handlers, containing only a single screen with random placements of widgets. Generating GUIs allows precise control over the training data, such as ensuring that the generated dataset contains a balanced number of each widget. An example generated GUI can be seen in Figure 1. We use 11 standard types of widgets in generated GUIs, which are shown in Table 1.

To generate synthetic applications, we use the Java Swing GUI framework. Initial attempts at generating GUIs by entirely random selection and placement of widgets yielded GUIs that were not representative of ones encountered in real applications. Consequently, the resulting model performed poorly on real GUIs. To create more realistic GUIs as training data, our approach therefore generates an abstract tree beforehand, and uses this tree as a basis, assigning widgets to each node or leaf before generating the GUI.

First, we randomly choose a Swing layout manager and then generate a random tree where each node represents a GUI widget. Only widgets which can contain other widgets can be assigned child nodes in the tree, for example, a tab pane can have children representing other GUI widgets assigned to it, but a button cannot. Algorithm 3.1 shows how a random abstract tree of GUI widget types is generated. Here, the nodes list initially contains a "Container" widget type which is to eliminate an infinite loop later if the $RandomWidgetType()$ function returns no containers. The call to $RandomWidgetType$ randomly returns one of the 11 types of widgets. Each widget has the same probability of appearing but we found that some GUI widgets are constructed of others when using Java Swing, e.g., a Combo Box also contains a button, and a scroll bar contains two buttons, one at each end. When generating data, we have a unique advantage that we can balanced the number of

Table 1: Widgets that the model can identify in a Graphical User Interface

Widget	Description	Example
Text Field	Allows input from the keyboard to be stored in the GUI to be used later.	
Button	Allows an event to be triggered by clicking with the mouse.	
Combo Box	Allows selection of predefined values. Clicking either inside the box or the attached button opens predefined options for users to select.	
List	Similar to a Combo Box, allows selection of predefined values. Values are present at all times. Scrolling may be needed to reveal more.	 .. .FBReader .IntelliJIdea2018.1 .Private .Sikulix
Tree	Similar to a list but values are stored in a tree structure. Clicking a node may reveal more values if the node has hidden child elements.	 gui-component-recognition.tex TODO LABELS BIBLIOGRAPHY references Introduction
Scroll Bar	A horizontal or vertical bar used for scrolling with the mouse to reveal more of the screen.	
Menu	A set of usually textual buttons across the top of a GUI	
Menu Item	An individual button in a menu. Clicking usually expands the menu revealing more interactable widgets.	
Toggle Button	Buttons that have two states toggled by clicking on them.	
Tabs	Buttons changing the contents in all or part of the GUI when activated.	
Slider	A button that can be click-and-dragged in a certain axis, changing a value which is usually a numeric scale, e.g., volume of an application.	

widgets in GUIs, evenly distributing the widget types through the generated dataset. We found that *menu_items* has a dependency with a *menu*. Weighting *menu_items* with an equal probability to appear forced *menus* to also appear on nearly all generated GUIs. To balance the dataset, we lowered the probability for *menu_items*.

To generate a Swing GUI, Algorithm 3.2 walks through the generated tree. Each node is assigned a random position and dimension inside its parent. The position is randomly selected based on the layout manager. For example, with a GridLayout, we randomly assign the element in the current node to a random (x, y) coordinate in the grid. However, with a FlowLayout, the position does not matter as all widgets appear side by side in a single line. In algorithm 3.2 , the $container.add$ method call in the $ApplyWidget$ procedure is from the JComponent class of Java Swing, and the random position is seeded here depending on the current LayoutManager.

Once a Swing GUI has been generated, Java Swing allows us to automatically extract information for all widgets. This includes the position on the screen, dimension and widget type. This is similar to the approach current GUI testing tools use when interacting with an application during test execution.

3.3 A Random Bounding Box Tester

Once widgets can be identified, they may be used to influence a random GUI tester. We created a tester which randomly clicks inside a given bounding box. At the most basic level, a box containing the whole application GUI is provided, and the tester will randomly interact with this box. One of three actions is executed on the selected box: a) left click anywhere inside the given box; b) right click anywhere inside the given box; c) left click anywhere inside the given box and type either a random string (e.g., "Hello World!" in our implementation) or a random number (e.g., between -10000 and 10000 in our implementation). We use these two textual inputs to represent the most common use for text fields: storing a string of characters or storing a number. This random clicker is a version of Android Monkey [1] we implemented that uses conventional GUIs in place of Android ones. Algorithm 3.3 shows how the tester can interact with a provided box. In this algorithm, $rand(x, y)$ returns a random number between x and y inclusive. $LeftClick(x, y)$ and $RightClick(x, y)$ represent moving the mouse to position x, y on the screen and either left or right clicking respectively. $KeyboardType(string)$ represents pressing the keys present in $string$ in chronological order.

We can refine the box provided to this random tester using the trained YOLOv2 network. We randomly select a box with a confidence greater than some value C. When seeded to the tester, the tester will click on a random position inside one of the predicted widgets from the network.

Finally, we can provide the tester with a box directly from Java Swing. This implementation currently only supports Java Swing applications but will ensure that the GUI tester is always clicking inside the bounding box of a known widget currently on the screen. Our tool *GUIdance* is open-source and can be found and contributed to on GitHub[1].

[1] https://github.com/thomasdeanwhite/GUIdance

Algorithm 3.3: RandomInteraction(*box*)

```
interaction ← rand(0, 2)
x ← box.x + rand(0, box.width)
y ← box.y + rand(0, box.height)
if interaction == 0
  then LeftClick(x, y)
  else if interaction == 1
  then RightClick(x, y)
  else if interaction == 2
  then
    LeftClick(x, y)
    inputType ← rand(0, 1)
    inputString ← ""
    if inputType == 0
      then inputString ← "Hello World!"
      else
        inputNumber ← rand(-10000, 10000)
        inputString ← inputNumber.toString()
    KeyboardType(inputString)
```

4 EVALUATION

To evaluate the effectiveness of our approach when automatically testing GUIs, we investigate the following research questions:

RQ1 How accurate is a model trained on synthetic GUIs when identifying widgets in GUIs from real applications?

RQ2 How accurate is a model trained on synthetic GUIs when identifying widgets in GUIs from other operating system and widget palettes?

RQ3 What benefit does random testing receive when guided by predicted locations of GUI widgets from screen shots?

RQ4 How close can a random tester guided by predicted widget locations come to an automated tester guided by the exact positions of widgets in a GUI?

4.1 Model Training

In order to create the prediction model, we created synthetic GUIs on Ubuntu 18.04, and to capture different GUI styles, we used different operating system themes. We generated 10,000 GUI applications per theme and used six light themes: the default Java Swing theme, adapta, adwaita, arc, greybird; two dark themes: adwaita-dark, arc-dark, and two high-contrast themes which are default with Ubuntu 18.04. These are all popular themes for Ubuntu and were chosen so that the pixel histograms of generated GUI images were similar to that of real GUI images.

In total this resulted in 100,000 synthetic GUIs, which we split as follows: 80% of data was used as training data, 10% as validation data, and 10% as testing data. To train a model using this data, the screen shots are fed through the YOLOv2 network and the predicted boxes from the network are compared against the actual boxes retrieved from Java Swing. If there is a difference, the weights of the model are updated so next time the predictions are more accurate.

It is important to have a validation dataset to determine whether the model is over-fitting on the training data. This can be done by checking the training progress of the model against the training and validation dataset. During training, the model is only exposed to the training dataset, so if the model is improving when evaluated against the training dataset, but not improving on the validation dataset, the model is over-fitting.

With the isolated training data, we trained a model which uses the YOLOv2 network. It has been observed that artificial data inflation (augmentation) increases the performance of neural networks, exposing the network to more varied data during training [8, 22]. During training, we artificially increased the size of input data using two techniques: brightness and contrast adjustment. Before feeding an image into the network, there is a 20% chance to adjust the image. This involves a random shift of 10% in brightness/contrast and applies to only a single training epoch. For example, an image could be made up to 10% lighter/darker and have the pixel intensity values moved up to 10% closer/further from the median of the image's intensity values.

4.2 Experimental Setup

4.2.1 RQ1. To evaluate RQ1, we compare the performance when predicting GUI widgets in 250 screen shots of unique application states in real applications against performance when predicting widgets in synthetic applications. Screen shots were captured when a new, unseen window was encountered during real user interaction. 150 of the screen shots were taken from the top 20 Swing applications on SourceForge, and annotated automatically via the Swing API. The remaining 100 screen shots were taken from the top 15 applications on the Ubuntu software center and manually annotated. The model used to predict widget locations was trained on only synthetic GUIs, and in RQ1 we see if the model is able to make predictions for real applications.

YOLOv2 predicts many boxes, which could cause a low precision. To lower the number of boxes predicted, we pruned any predicted boxes below a certain confidence threshold. To tune this confidence threshold, we evaluated different confidence values against the synthetic validation dataset. As recall is more important to us than precision, we used the confidence value with the highest F2-measure to compare synthetic against real application screen shots. We found this value C to be 0.1 through parameter tuning on the synthetic *validation* dataset. However, the actual comparison of synthetic data against real GUI data was performed on the isolated *test* dataset, to avoid biases in this value of C being optimal for the validation dataset.

After eliminating predicted boxes with a confidence value less than C, we can compare the remaining boxes with the actual boxes of a GUI. In order to assess whether a predicted box correctly matches with an actual box, we match boxes based on the intersection over union metric.

Intersection-over-union (IoU) calculates the similarity of two boxes in two dimensional space. We calculate the IoU between the predicted boxes from the YOLOv2 network, and actual boxes in the labeled test dataset. The IoU of two boxes is the area that the boxes intersect, divided by the union of both areas. An IoU value of one indicates that the boxes are identical, and an IoU of 0 indicates the boxes have no area of overlap. See Figure 2 for an example of IoU values for overlapping boxes. The shaded area indicates overlap between both boxes. We consider a predicted box to be matched with an actual box when the IoU value is greater than 0.3.

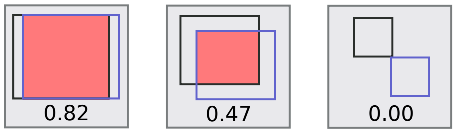


Figure 2: Intersection over Union (IoU) values for various overlapping boxes.

Table 2: The applications tested when comparing the three testing techniques.

Application	Description	LOC	Branches
Address Book	Contact recorder	363	83
bellmanzadeh	Fuzzy decision maker	1768	450
BibTex Manager	Reference Manager	804	309
BlackJack	Casino card game	771	178
Dietetics	BMI calculator	471	188
DirViewerDU	View directories and size	219	90
JabRef	Reference Manager	60620	23755
Java Fabled Lands	RPG game	16138	9263
Minesweeper	Puzzle game	388	155
Mobile Atlas Creator	Create offline atlases	20001	5818
Movie Catalog	Movie journal	702	183
ordrumbox	Create mp3 songs	31828	6064
portecle	Keystore manager	7878	2543
QRCode Generator	Create QR codes for links	679	100
Remember Password	Save account details	296	44
Scientific Calculator	Advanced maths calculator	264	62
Shopping List Manager	List creator	378	62
Simple Calculator	Basic maths calculator	305	110
SQuiz	Load and answer quizzes	415	146
UPM	Save account details	2302	530

4.2.2 RQ2. To evaluate RQ2, we use the same principle as in RQ1, however, the comparison datasets are the synthetic test data-set and a set of manually annotated screen shots taken from the applications in the Apple store. We gathered 50 screen shots of unique application states, five per application of the top 10 free applications on the store as of 19th January 2019. Again, each screen shot was taken when a new, previously unknown window appeared during real user interaction. The screen shots were manually annotated with the truth boxes for all widgets present.

4.2.3 RQ3. To evaluate RQ3, we compare the branch coverage of tests generated by a random tester to tests where interactions are guided by predicted bounding boxes. The subjects under test are 20 Java Swing applications, including the top six Java Swing applications from SourceForge and the remaining ones from the SF110 corpus by Fraser and Arcuri [10]. Table 2 shows more information about each application. We limited the random tester to 1000 actions, and conservatively performed a single action per second. Adding a delay before interactions is common in GUI testing tools, and using too little delay can produce flaky tests or tests with a high entropy [11]. Because of the delay, all techniques had a similar runtime. On a crash or application exit, the application under test was restarted. Each technique was applied 30 times on each of the applications, and a Mann-Whitney U-Test was used to test for significance. Although all the applications use Java Swing, this was to aid conducting experiments when measuring branch coverage and allow retrieval of the positions of widgets currently on the screen from the Java Swing API for RQ4. Our approach should work on many kinds of applications using any operating system.

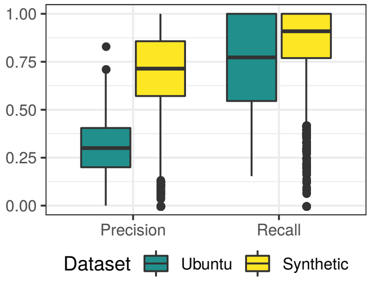


Figure 3: Precision and recall of synthetic data against real GUIs from Ubuntu/Java Swing applications.

4.2.4 RQ4. To answer RQ4, we compare the branch coverage of tests generated by a tester guided by predicted bounding boxes, to one guided by the known locations of widgets retrieved from the Java Swing API. The API approach is similar to current GUI testing tools, which exploit the known structure of a GUI to interact with widgets. We use the same applications as RQ3. We allowed each tester to execute 1000 actions over 1000 seconds. On a crash or application exit, the application under test is restarted. Each technique ran on each application for 30 iterations.

4.3 Threats to Validity

There is a chance that our model over-trains on the training and validation synthetic GUI dataset and therefore achieves an unrealistically high precision and recall on these datasets. To counteract this, we use the third test dataset when calculating precision and recall values for the synthetic dataset which has been completely isolated from the training procedure.

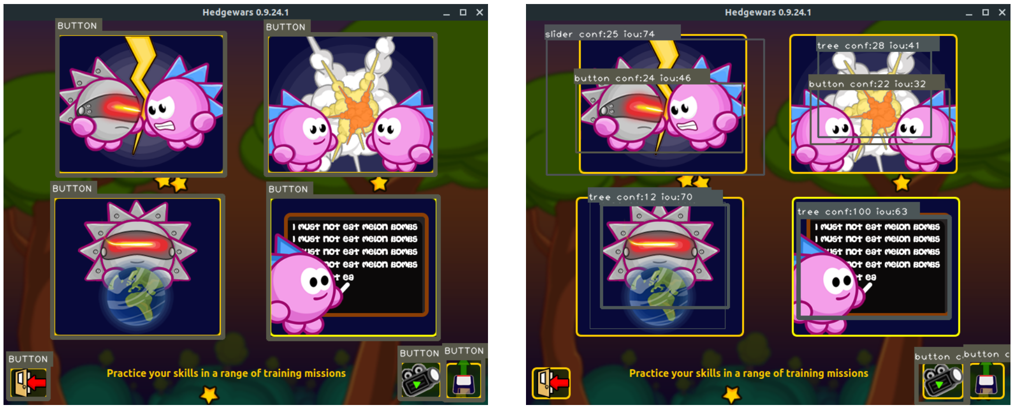
To ensure that our real GUI screen shot corpus represents general applications, the Swing screen shots were from the top applications on SourceForge, the top rated applications on the Ubuntu software center, and the top free applications from the Apple Store.

In object detection, usually an IoU value of 0.5 or more is used for a predicted box to be considered a true positive (a "match"). However, we use an IoU threshold of 0.3 as the predicted box does not have to exactly match the actual GUI widget box, but it needs enough overlap to enable interaction. Russakovsky et al. [21] found that training humans to differentiate between bounding boxes with an IoU value of 0.3 or 0.5 is challenging, so we chose the lower threshold of 0.3.

As the GUI tester uses randomized processes, we ran all configurations on all applications for 30 iterations. We used a two-tailed Mann-Whitney U-Test to compare each technique and a Vargha-Delaney A_{12} affect size to find the technique likely to perform best.

4.4 Results

4.4.1 RQ1: How accurate is a model trained on synthetic data when detecting widgets in real GUIs? Figure 3 shows the precision and recall achieved by a model trained on synthetic data. We can see that predicting widgets on screen shots of Ubuntu and Java Swing GUIs achieves a lower precision and recall than on synthetic GUIs. However, the bounding boxes of most widgets have a corresponding predicted box with an IoU > 0.3, as shown by a high recall value.



(a) Manually Annotated (b) Predicted Annotations

Figure 4: Manually annotated (a) and predicted (b) boxes on the Ubuntu application "Hedge Wars".

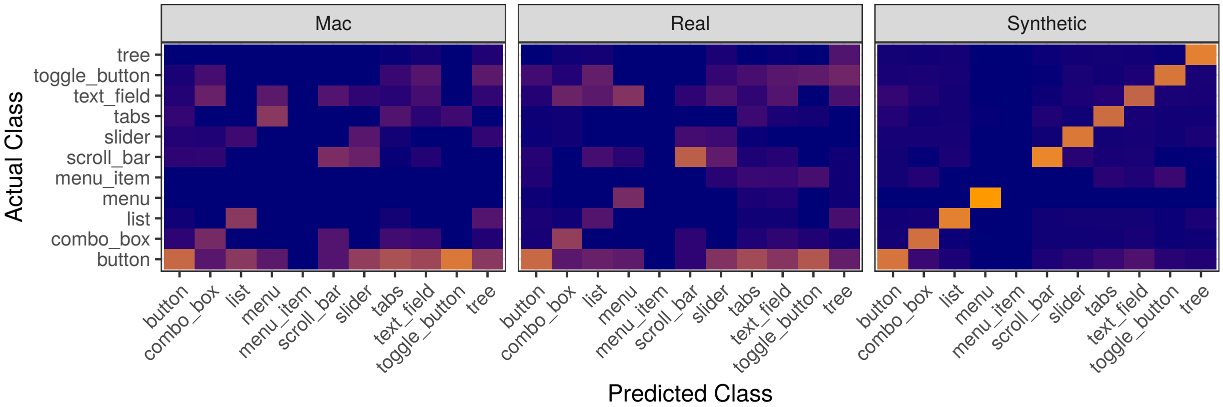


Figure 5: Confusion matrix for class predictions.

A low precision but high recall indicates that we are predicting too many widgets in each GUI screen shot. Figure 4a shows an example of a manually annotated image, and Figure 4b shows the same screen shot but with predicted widget boxes.

The precision and recall values only show if a predicted box aligns with an actual box. Figure 5 shows the confusion matrix for class predictions. An orange (light) square indicates a high proportion of predictions, and blue (dark) square a low proportion. We can see that for synthetic applications, most class predictions are correct. However, the model struggles to identify *menu_items* and this is most likely due to the lower probability of them appearing in synthesized GUIs. The network would rather classify them as a button which appears much more commonly through all synthesized GUIs.

From the confusion matrix, another problem for classification seems to be buttons. Buttons are varied in shape, size and foreground. For example, a button can be a single image, a hyper-link, or text surrounded by a border. Subtle modifications to a widget can change how a user perceives the widget's class, but are much harder to detect automatically.

While this shows that there is room for improvement of the prediction model, these improvements are not strictly necessary for the random tester as described in Section 3.3, since it interacts with all widgets in the same manner irrespective of the predicted type. Hence, predicting the correct class for a widget is not as important as identifying the actual location of a widget, which our approach achieves. However, future improvements of the test generation approach may rely more on the class prediction.

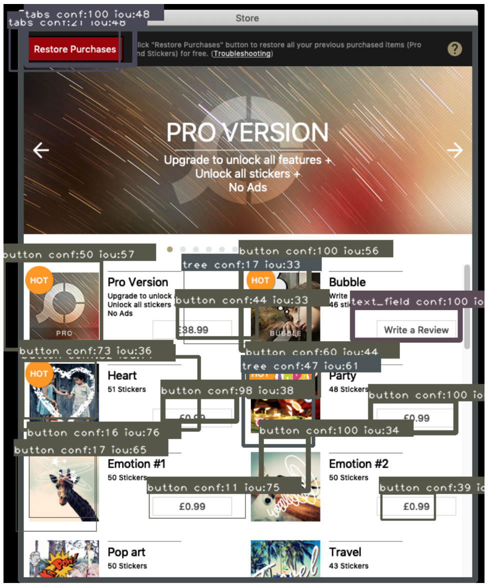
RQ1: In our experiments, widgets in real applications were detected with an average recall of 77%.

4.4.2 RQ2: How accurate is a model trained on synthetic data when detecting widgets on a different operating system?

To investigate whether widgets can be detected in other operating systems with a different widget palette, we apply a similar approach to RQ1 and use the same IoU metric, but evaluated on screen shots taken on a different operating system and from different applications.

Figure 7 shows the precision and recall achieved by the model trained on synthetic GUI screen shots. We again see a lower precision and recall on OSX (Mac) GUI screen shots compared to synthetic GUIs, but we still match over 50% of all boxes against predicted boxes with an IoU > 0.3.

A lower precision indicates many false positive predictions when using the OSX theme in applications. An observable difference between predictions on OSX and on Ubuntu is that our model has



lick "Restore Purchases" button to restore all your previous purchased items (Pro nd Stickers) for free. (Troubleshooting)

Figure 6: Predicted bounding boxes on the OSX application "Photoscape X".

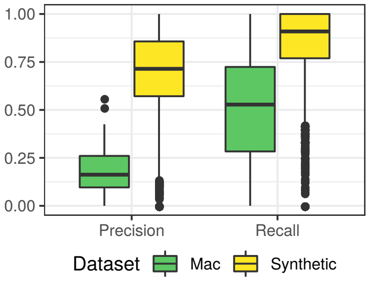


Figure 7: Precision and recall of synthetic data against real GUIs from Mac OSX applications.

greater difficulty in predicting correct dimensions for bounding boxes on OSX. See Figure 6 for correctly predicted boxes in the OSX application "Photoscape X".

One observation of applications using OSX is that none use a traditional menu (e.g. File, Edit, etc.). OSX applications instead opt for a toolbar of icons that function similar to tabs. Our model could be improved by including this data in the training stage.

For the purposes of testing, an exact match of bounding boxes is less relevant so long as the generated interaction happens somewhere within the bounding box of the actual widget. For example, if a predicted box is smaller than the actual bounding box of a widget, interacting with any point in the predicted box will trigger an event for the corresponding widget. IoU does not take this into account, and it is possible that a box will be flagged as a false positive if IoU < 0.3 but the predicted box is entirely within the bounds of the actual box. In this case, the predicted box would still be beneficial in guiding a test generator.

> *RQ2: GUI widgets can be identified in different operating systems using a model trained on widgets with a different theme, achieving an average recall of 52%.*

4.4.3 RQ3: What benefit does random testing receive when guided by predicted locations of GUI widgets from screen shots?

Figure 8 shows the branch coverage achieved by the random tester when guided by different techniques. Table 3 shows the mean branch coverage for each technique, where a bold value indicates significance. Here we can see that interacting with predicted GUI widgets achieves a significantly higher coverage for 18 of the 20 applications tested against a random testing technique. The A_{12} value indicates the probability of the tester guided by predicted widget locations performing worse than the comparison approach. If A_{12}=0.5, then both approaches perform similarly (i.e., the probability of one approach outperforming another is 50%); if A_{12}<0.5, the tester guided by predicted widget locations usually achieves a higher coverage. If A_{12}>0.5 then the tester guided by predicted widgets would usually achieve a lower coverage. For instance, take Address-book: $p_v(Pred, Rand) < 0.001$ and $A_{12}(Pred, Rand) = 0.032$. This indicates that the testing approach guided by predicted widgets would achieve a significantly higher code coverage than a random approach around 96.8% of the time when testing this application.

Overall, guiding the random tester with predicted widget locations increased coverage by an average of 42.5%. We can see that even on applications that use a custom widget set (e.g., ordrumbox), using predicted widget locations to guide the search achieves a higher coverage. The main coverage increases were in applications with sparse GUIs, like Address Book (24%⟶48%) and Dietetics (20%⟶54%). The predicted widgets also aided the random tester to achieve coverage where complex sequences of events are needed, such as the Bellmanzadeh application (22%⟶28%). Bellmanzadeh is a fuzzy logic application and requires many fields to be created of different types. Random is unlikely to create many variables of unique types but, when guided by predicted widget locations, is more likely to interact with the same widgets again to create more variables. The random tester is similar to Android Monkey and achieves a similar level of coverage to that Choudhary et al. [7] observed. The coverage levels achieved by the random tester show that it spends more time performing uninteresting actions, whereas it is far more likely to interact with an actual widget when guided by widget predictions

One notable example is the application JabRef, where unguided random achieved 6.6% branch coverage, significantly better than random guided by widget predictions which achieved 5.2%. JabRef is a bibtex reference manager, and by default it starts with no file open. The only buttons accessible are "New File" and "Open". The predicted boxes contain an accurate match for the "Open" button and a weak match for the "New File" button. If the "Open" button is pressed, a file browser opens, locking the main JabRef window.

As we randomly select a window to interact with from the available, visible windows, any input into the main JabRef window is ignored until the file browser closes. There are two ways to exit the file browser: clicking the "Cancel" button or locating a valid JabRef file and pressing "Open". There are, however, many widgets on this screen to interact with lowering the chance of hitting cancel, and it is near impossible to find a valid JabRef file to open for both the prediction technique and the API technique. Even if the "Cancel" button is pressed, there is a high chance of interacting with the "Open" button again in the main JabRef window.

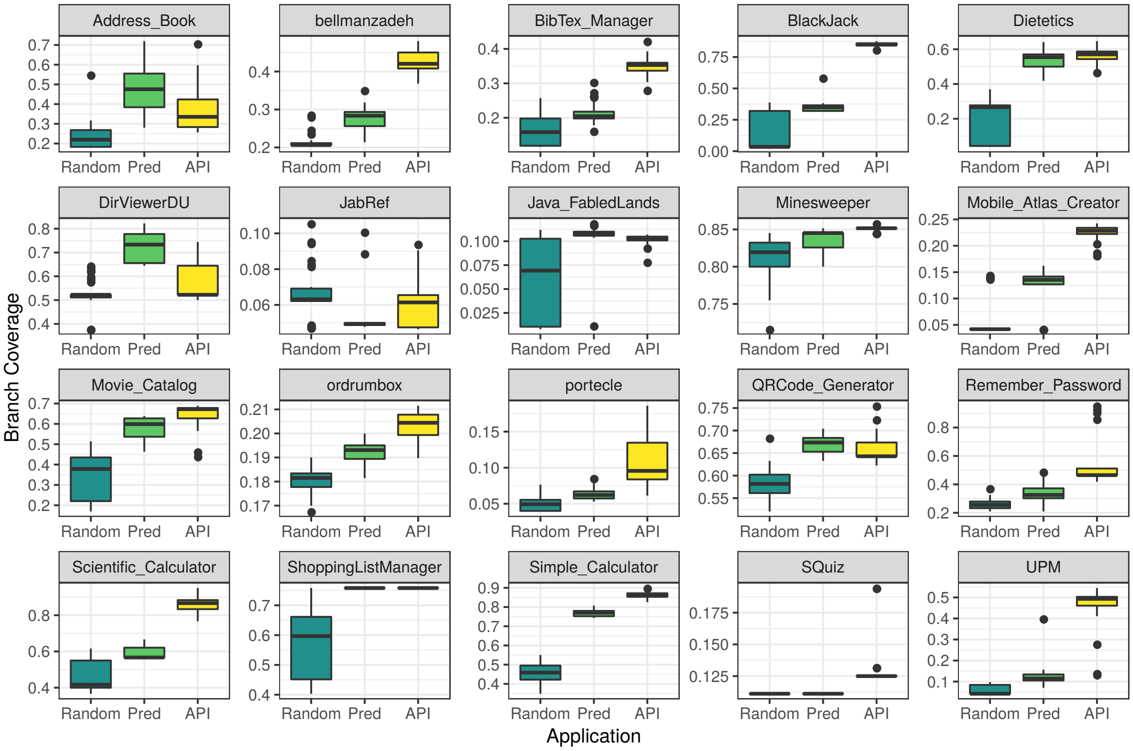


Figure 8: Branch Coverage achieved by a random clicker when clicking random coordinates, guided by predicted widgets positions and guided by the Swing API.

On the other hand, the random technique has a low chance of hitting the "Open" button. When JabRef starts, the "New" button is focused. We repeatedly observe the random technique click anywhere in the tool bar and type "Hello World!". As soon as it presses the space key, it would trigger the focused button and a new JabRef project would open. This then unlocks all the other buttons to interact with in the JabRef tool bar

RQ3: In our experiments, widget prediction significantly increased branch coverage by an average of 42.5% over random testing.

4.4.4 RQ4: How close can a random tester guided by predicted widget locations come to an automated tester guided by the exact positions of widgets in a GUI? Using GUI ripping to identify actual GUI widget locations serves as a "golden" model of how much random testing could be improved with a perfect prediction model. Therefore, Figure 8 also shows branch coverage for a tester guided by widget positions extracted from the Java Swing API. It is clear that whilst predicted widget locations aid the random tester in achieving a higher branch coverage, unsurprisingly, using the positions of widgets from an API is still superior. This suggests that there is still room for improving the prediction model further.

However, notably, there are cases where the widget prediction technique *improves* over using the API positions. One such case is DirViewerDU. This is an application consisting of only a single tree spanning the whole width and height of the GUI. If a node in the tree is right clicked, a pop-up menu appears containing a custom widget not supported or present in the API widget positions. However, the prediction approach correctly identifies this as an interactable widget and can generate actions targeting it.

Another example of this is in the Address Book application. Both guidance techniques lead the application into a state with two widgets: a text field and a button. To leave this GUI state, text needs to be typed in the text field and then the button needs to be clicked. If no button is clicked, an error message is shown and the GUI state remains the same. However, the information of the text field is not retrieved by the Swing API as it is a custom widget. The API guided approach then spends the rest of the testing budget clicking the button, producing the same error message. Predicted widget guidance identifies the text field, and can leave this state to explore more of the application.

RQ4: Exploiting the known locations of widgets through an API achieves a significantly higher branch coverage than predicted locations, however widget prediction can identify and interact with custom widgets not detected by the API.

5 RELATED WORK

Bajammal et al. undertook previous work in generating assertions for web canvasses [2]. Web pages have a Document Object Model (DOM), however web canvasses do not have this underlying exploitable structure so have limited testability for current testing tools. For web canvasses, applications draw directly to a buffer representing the screen contents seen by users. Bajammal et al. identified common shapes in web canvasses, generating assertions

Table 3: Branch coverage of random guided by no guidance, widget prediction and the Java Swing API. Bold is significance.

Application	**Pred**iction Cov.	**Rand**om Cov.	p_v (Pred, Rand)	$\hat{A}_{12}$(Pred, Rand)	**API** Cov.	p_v (Pred, API)	$\hat{A}_{12}$(Pred, API)
Address-Book	0.484	**0.235**	**<0.001**	**0.032**	**0.370**	**<0.001**	**0.237**
bellmanzadeh	0.276	**0.215**	**<0.001**	**0.048**	**0.425**	**<0.001**	**1.000**
BibTex-Manager	0.214	**0.160**	**<0.001**	**0.145**	**0.347**	**<0.001**	**0.998**
BlackJack	0.355	**0.167**	**<0.001**	**0.143**	**0.848**	**<0.001**	**1.000**
Dietetics	0.544	**0.197**	**<0.001**	**<0.001**	0.564	0.067	0.640
DirViewerDU	0.728	**0.522**	**<0.001**	**<0.001**	**0.576**	**<0.001**	**0.089**
JabRef	0.052	**0.066**	**<0.001**	**0.768**	0.060	0.608	0.540
Java-FabledLands	0.105	**0.056**	**<0.001**	**0.098**	**0.102**	**<0.001**	**0.122**
Minesweeper	0.837	**0.811**	**<0.001**	**0.170**	**0.850**	**<0.001**	**0.859**
Mobile-Atlas-Creator	0.120	**0.059**	**<0.001**	**0.199**	**0.224**	**<0.001**	**1.000**
Movie-Catalog	0.581	**0.328**	**<0.001**	**0.007**	**0.643**	**<0.001**	**0.826**
ordrumbox	0.192	**0.181**	**<0.001**	**0.056**	**0.203**	**<0.001**	**0.905**
portecle	0.063	**0.049**	**<0.001**	**0.121**	**0.106**	**<0.001**	**0.948**
QRCode-Generator	0.673	**0.582**	**<0.001**	**0.024**	**0.658**	**0.010**	**0.304**
Remember-Password	0.333	**0.255**	**<0.001**	**0.182**	**0.535**	**<0.001**	**0.968**
Scientific-Calculator	0.588	**0.469**	**<0.001**	**0.129**	**0.863**	**<0.001**	**1.000**
ShoppingListManager	0.758	**0.563**	**<0.001**	**0.032**	0.758	1.000	0.500
Simple-Calculator	0.769	**0.460**	**<0.001**	**<0.001**	**0.864**	**<0.001**	**1.000**
SQuiz	0.111	0.111	1.000	0.500	**0.130**	**<0.001**	**1.000**
UPM	0.125	**0.060**	**<0.001**	**0.040**	**0.460**	**<0.001**	**0.986**
Mean	0.395	0.277	0.050	0.135	0.479	0.084	0.746

from identified shapes through multiple image processing steps. Although shapes could be identified, identification and classification of GUI widgets is different. We may be able to expand on this work to generate assertions. We predict rectangles that contain widgets on a GUI. Hence, we can directly apply the assertion technique to generate assertions on the application under test.

Borges et al. link event handlers in the source code to the corresponding widget displayed on the screen [5]. However, this approach uses a tool that captures the current interface of an Android application, providing the known location of each widget. Our approach could possibly take the place of this tool in identifying GUI widgets for applications which the widget ripping tool does not currently support.

Previous work on applying context to GUI widgets involved searching for possible descriptive labels for each data widget. Becce et al. search above and to the left of data widgets for static widgets that can provide more information for testers about the type of data to input [4]. A coverage increase of 6% and 5% were found in the two applications tested when providing context about data widgets to the tool AutoBlackTest.

6 CONCLUSIONS

When applications have no known exploitable structure behind their GUI, monkey testing is a common fail-safe option. However, it is possible to identify widgets in a GUI from screen shots using machine learning, event if the model is trained on synthetic, generated GUIs. Applying this model during random GUI testing led to a significant coverage increase in 18 out of 20 applications in our evaluation. A particular advantage of this approach is that the prediction model is independent of a specific GUI library or operating system. Consequently, our prediction model can immediately support any GUI testing efforts.

Comparison to a gold standard with perfect information of GUI widgets shows that there is potential for future improvement:

- Firstly, we need to find a better method of classifying GUI widgets. A tab that changes the contents of all or part of a GUI's screen has the same function as a button, so they could be grouped together.
- We currently use YOLOv2 and this predicts classes exclusively: if a button is predicted, there is no chance that a tab could also be predicted. Newer methods of object detection (e.g. YOLOv3 [20]) focus on multiple classification, where a widget could be classified as a button and as a tab. This could improve the classification rate of widgets that inherit attributes and style.
- Whilst labor intensive, further improvements to the widget prediction model could be made by training a model on a labeled dataset of *real* GUIs. To lower effort costs, this dataset could be augmented with generated GUIs. The performance of the model is dependent on the quality of training data.
- Furthermore, in this paper we focused on a single operating system with various themes. However, it may be beneficial to train the model using themes from many operating systems and environments to improve performance when identifying widgets across different platforms.

Besides improvements on the prediction model itself, there is potential to make better use of the widget information during test generation. For example, if there are a limited number of boxes to interact with, it may be possible to increase the efficiency of the tester by weighting widgets differently depending on whether they have previously been interacted with (e.g., [23]). This could be further enhanced using a technique like Q-learning (cf. AutoBlackTest [15]).

ACKNOWLEDGMENTS

This work is supported by EPSRC project EP/N023978/2.

REFERENCES

[1] Android Developers. 2015. Monkeyrunner.

[2] M. Bajammal and A. Mesbah. 2018. Web Canvas Testing Through Visual Inference. In *2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)*. 193–203. https://doi.org/10.1109/ICST.2018.00028

[3] S. Bauersfeld and T. E. J. Vos. 2012. GUITest: A Java Library for Fully Automated GUI Robustness Testing. In *2012 Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*. 330–333. https://doi.org/10.1145/2351676.2351739

[4] Giovanni Becce, Leonardo Mariani, Oliviero Riganelli, and Mauro Santoro. 2012. Extracting Widget Descriptions from GUIs. In *Fundamental Approaches to Software Engineering*, Juan de Lara and Andrea Zisman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 347–361.

[5] Nataniel P. Borges, Jr., Maria Gómez, and Andreas Zeller. 2018. Guiding App Testing with Mined Interaction Models. In *Proceedings of the 5th International Conference on Mobile Software Engineering and Systems (MOBILESoft '18)*. ACM, New York, NY, USA, 133–143. https://doi.org/10.1145/3197231.3197243

[6] Hatim Chahim, Mehmet Duran, and Tanja EJ Vos. 2018. Challenging TESTAR in an industrial setting: the rail sector. (2018).

[7] Shauvik Roy Choudhary, Alessandra Gorla, and Alessandro Orso. 2015. Automated Test Input Generation for Android:Are We There Yet? (E). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 429–440.

[8] J. Ding, B. Chen, H. Liu, and M. Huang. 2016. Convolutional Neural Network With Data Augmentation for SAR Target Recognition. *IEEE Geoscience and Remote Sensing Letters* 13, 3 (March 2016), 364–368. https://doi.org/10.1109/LGRS.2015.2513754

[9] Justin E Forrester and Barton P Miller. 2000. An Empirical Study of the Robustness of Windows NT Applications Using Random Testing. In *Proceedings of the 4th USENIX Windows System Symposium*. Seattle, 59–68.

[10] Gordon Fraser and Andrea Arcuri. 2014. A Large-Scale Evaluation of Automated Unit Test Generation Using EvoSuite. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 24, 2, Article 8 (Dec. 2014), 42 pages. https://doi.org/10.1145/2685612

[11] Zebao Gao, Yalan Liang, Myra B Cohen, Atif M Memon, and Zhen Wang. 2015. Making system user interactive tests repeatable: When and what should we control?. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, 55–65.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE Computer Society, Washington, DC, USA, 580–587. https://doi.org/10.1109/CVPR.2014.81

[13] R. Lo, R. Webby, and R. Jeffery. 1996. Sizing and Estimating the Coding and Unit Testing Effort for GUI Systems. In *Proceedings of the 3rd International Software Metrics Symposium*. 166–173. https://doi.org/10.1109/METRIC.1996.492453

[14] Ke Mao, Mark Harman, and Yue Jia. 2016. Sapienz: Multi-objective Automated Testing for Android Applications. In *Proceedings of the 25th International Symposium on Software Testing and Analysis*. ACM, 94–105.

[15] L. Mariani, M. PezzÃÍ, O. Riganelli, and M. Santoro. 2011. AutoBlackTest: A Tool for Automatic Black-Box Testing. (May 2011), 1013–1015. https://doi.org/10.1145/1985793.1985979

[16] A. Memon, I. Banerjee, and A. Nagarajan. 2003. GUI Ripping: Reverse Engineering of Graphical User Interfaces for Testing. In *10th Working Conference on Reverse Engineering, 2003. WCRE 2003. Proceedings.* 260–269. https://doi.org/10.1109/WCRE.2003.1287256

[17] Bao Nguyen, Bryan Robbins, Ishan Banerjee, and Atif Memon. 2014. GUITAR: An innovative tool for automated testing of GUI-driven software. *Automated Software Engineering* 21 (03 2014). https://doi.org/10.1007/s10515-013-0128-9

[18] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. *CoRR* abs/1506.02640 (2015). arXiv:1506.02640 http://arxiv.org/abs/1506.02640

[19] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *CoRR* abs/1612.08242 (2016). arXiv:1612.08242 http://arxiv.org/abs/1612.08242

[20] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *CoRR* abs/1804.02767 (2018). arXiv:1804.02767 http://arxiv.org/abs/1804.02767

[21] O. Russakovsky, L. Li, and L. Fei-Fei. 2015. Best of both worlds: Human-machine collaboration for object annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2121–2131. https://doi.org/10.1109/CVPR.2015.7298824

[22] J. Salamon and J. P. Bello. 2017. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters* 24, 3 (March 2017), 279–283. https://doi.org/10.1109/LSP.2017.2657381

[23] Ting Su, Guozhu Meng, Yuting Chen, Ke Wu, Weiming Yang, Yao Yao, Geguang Pu, Yang Liu, and Zhendong Su. 2017. Guided, Stochastic Model-Based GUI Testing of Android Apps. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 245–256.

[24] Tom Yeh, Tsung-Hsiang Chang, and Robert C. Miller. 2009. Sikuli: Using GUI Screenshots for Search and Automation. In *Proceedings of the 22Nd Annual ACM Symposium on User Interface Software and Technology (UIST '09)*. ACM, New York, NY, USA, 183–192. https://doi.org/10.1145/1622176.1622213