GasBotty: Multi-Metric Extraction in the Wild

KEVIN DICK\textsuperscript{1,2,*}, (Graduate Student Member, IEEE), JOSHUA B. TANNER\textsuperscript{1,2,*}, FRANÇOIS CHARIH\textsuperscript{1,2}, (Graduate Student Member, IEEE), and JAMES R. GREEN\textsuperscript{1,2}, (Senior Member, IEEE)

\textsuperscript{1}Systems & Computer Engineering, Carleton University, Ottawa, ON, Canada
\textsuperscript{2}Institute of Data Science, Carleton University, Ottawa, ON, Canada
*Equal Contribution

Corresponding author: Kevin Dick (e-mail: kevin.dick@carleton.ca).

ABSTRACT The lived environment, particularly when proximal to roadways, is filled with multi-digit and multi-numbered values corresponding to advertised commodities. The reliable detection of single multi-digit values from natural imagery has been widely studied through the last decade (e.g. Street View House Numbers [SVHN]); however, extraction and assignment of the contextual meaning of those values are far more difficult given the diversity and unstructured nature of advertisements. To operationalize information extracted from detected values in the wild, the contextual meaning that those values represent is critical. To our knowledge, no large-scale visual dataset comprising multi-digit, multi-number values with associated context labels exists; we denote this class of problems as “multi-metric extraction in the wild”. In this work, we focus on the accurate detection and reading of gas prices, and their contextual association to gas grade and payment type. We provide complete annotations for the Gas Prices of America (GPA) dataset, comprising 2,048 training and 512 test images sampled across the United States including over 2,600 signs, 6,000 prices, 27,000 digits, and 7,800 gas grade and payment type labels. With these data, we develop the GasBotty predictor, a composite neural network model, and evaluate it over eight benchmark tasks of increasing difficulty. Finally, we define a new highly stringent, binary-type metric, denoted "All-or-Nothing Accuracy" (ANA), requiring that a predictor perfectly extract and correctly associate all information in a gas sign. Our proposed model achieves 72.9% ANA over the independent test set of 512 images, where conventional state-of-the-art models and both a uniform random and biased random predictor would all tend towards 0% ANA. GasBotty and the GPA dataset will serve as a valuable benchmark for the development of future multi-metric extraction in the wild systems.

INDEX TERMS Multi-Metric Extraction, Machine Vision, Optical Character Recognition

I. INTRODUCTION

Imagery and other sensor data sourced from increasingly instrumented circulating vehicles along (sub)urban, rural, and remote roadways offer diverse opportunity to detect and extract information from the lived and natural environments. Numerous studies have leveraged Google Street View images to this end\cite{1,2,3}. As visual creatures, humans navigating these roadways are presented with a plethora of advertisements for various commodities in a bid for attention and product interest. For certain commodities, such as gasoline, there additionally exist online repositories, namely GasBuddy, that aggregate pricing information from manually-entered crowd-sourced data\cite{1}.

As first proposed by Dick et al., the automated extraction of gas prices and correct association-mapping to contextual labels would herald an age of machine-augmented crowd-sourcing\cite{4}. A camera-laden vehicle, such as an autonomous vehicle (AV), with a model capable of detecting and reading values and associating them to context labels could, in theory, automatically update the GasBuddy repository.

While the automated detection and reading of text, digits, and multi-digit numbers from imagery of the natural environment has been extensively studied throughout the last decades\cite{5,6,7,8,9,10,11}, the development of models capable
Finally, the use of conventional performance metrics such as accuracy used in typical multi-digit detection tasks are woefully inadequate for evaluating the performance of models applied to multi-metric extraction, given the need to detect, assemble, and associate numerous visual elements. In cases with a small number of labels, a uniform-random (random_u) or biased-random (random_b) predictor achieves a non-zero low-level accuracy and be of little practical use. Therefore, we introduce a new highly-stringent and binary-type metric, denoted "All-or-Nothing Accuracy" (ANA), strictly requiring that all elements in the image are correctly detected and associated without tolerance for a single error.

Our main contributions, in four-fold summary:

- We generate a large-scale dataset of metric-specific annotations for gas prices and gas signs throughout the 49 mainland United States of America, greatly expanding the utility of the original GPA dataset for multi-digit, multi-number price detection and context mapping. This involved the manual segmentation of over 22,167 digits relating to over 4,766 prices and over 5,403 labels contained within 2,132 gas signs. These are made available to the research community in the following Dataverse repository: [GPA4MME Dataset](https://doi.org/10.1109/ACCESS.2022.3156578).

- We develop a highly accurate end-to-end model, denoted GasBotty, for the automated extraction of multi-number, multi-digit prices in the wild capable of accurately assigning contextual meaning to those values. All source code and model weights used in this work are available in the following GitHub repository: [github.com/GreenCUBIC/GasBotty](https://github.com/GreenCUBIC/GasBotty).

- We define the highly conservative "All-or-Nothing Accuracy" (ANA) metric to demonstrate our proposed model’s ability to accurately extract prices and associate them to contextual sign information.

- We experimentally validate the GasBotty performance over eight benchmark tasks and demonstrate the model’s capacity to effectively extract multiple prices in the wild.

II. RELATED WORK

Automated Extraction of Phenomena from Google Street View Imagery. The last decade has seen widespread usage of GSV to study phenomena from the streetscapes of the lived environment. Gebur et al. effectively estimated the demographic and socioeconomic characteristics of 200 US cities by identifying vehicle number and type from 50 million images [1]. Li et al. leveraged GSV to assess street-level greenery and vegetation within the East Village of Manhattan [2]. Indeed, the ability to systematically sample GSV imagery [18] enables the creation of large-scale spatial datasets of the lived environment.

Multi-Character Text & Multi-Digit Number Datasets in the Wild. Current benchmark datasets for the evaluation of text and number detectors include the Street View House Numbers (SVHN) dataset [8] comprising ~73K:26K digits for training:testing, the Natural Environment Optical...
Character Recognition dataset (NEOCR) comprising 659 images with 5,238 text fields [13], the Street View Text dataset (SVT) comprising 100/211:250/514 train-images/-words:test-images/-words [14], the COCO-text dataset comprising 63,686 images with 173,589 text regions [15], the ICDAR 2015 Incidental Scene Text dataset (ICDAR1500) comprising 1000:500 train:test images [19], and the French Street Name Signs dataset (FSNS) comprising over one million images [16]. However, none of these datasets provide associated context-labels or context-specific semantic meaning to the text and numbers they represent. Nonetheless, numerous effective methods have been developed from these sources.

Detecting Multi-Character Words & Multi-Digit Numbers from Natural Imagery. With the availability of several large-scale text and number datasets has come numerous methods to effectively extract these values from natural imagery. Early examples include Matan’s ZIP code reader using space displacement neural networks that preceded the advent of LeCun’s convolutional neural networks (CNNs) [5], [20]. In 2015, Goodfellow et al. developed a multi-digit number recognition model that achieved over 96% accuracy in recognizing complete street numbers from the SVHN dataset and reported 97.84% accuracy on a per-digit recognition task that rivaled the performance achieved by human operators at the time [9]. More recently, the EAST model significantly outperformed state-of-the-art methods on several datasets, including COCO-text, in terms of both accuracy and efficiency [10].

Context Association of Text/Numbers in Natural Imagery. The concept of context for multi-character text and multi-digit recognition is nebulous. In one instance, Liu et al. proposed structure inference nets that formulated object detection as a problem of graph inference where detected objects are treated as nodes and relationships between objects are modeled as edges [21]. In another instance, context is determined from the frequency of term occurrences in images based on a limited vocabulary contrasting Strongly Contextualized vocabularies from Weakly Contextualized and Generic vocabularies [12].

While the semantic segmentation of imagery also seeks to contextualise detected objects within images by assigning pixels to specific object classes, it has not yet been formulated as a label-association task or the prediction of relationships between detected objects (beyond linking disjoint regions of the same object). Consequently, in this work, we propose the first end-to-end evaluation of a composite model for multi-metric extraction in the wild as the inceptum of a field of research seeking to accurately extract information from the spatio-temporally sourced imagery of camera-laden vehicles.

III. APPROACH

In this section, we first describe the annotation process and composition of the GPA dataset, then explain how these data were leveraged to develop each component of our GasBotty composite neural network. For clarity, we distinguish be-
between (sub)images using the term "level": the scene-level image is the original 640 × 640 pixel image from GSV, the sign-level image contains only the advertised gas sign, the price- and label-level images only contain a single price or label, and digit-level images would only contain a single digit.

**TABLE 1.** Mapping Annotated Gas Label Instances to a Class.

| Ground Truth Grade Label | Example Mapping of Image Variants |
|--------------------------|-----------------------------------|
| Regular (Octane 87)      | ‘Regular’, ‘Reg’, ‘Unl’, ‘Unleaded’ |
| Mid-Grade (Octane 89-90) | ‘Plus’, ‘Mid-Grade’, ‘Mid’, ‘Special’, ‘Extra’, ‘Silver’ |
| Premium (Octane 91-94)   | ‘Premium’, ‘Prem’, ‘Supreme’, ‘Super’, ‘Super+’, ‘Ultra 93’, ‘V-Power’ |
| Diesel                   | ‘Diesel’, ‘Premium Diesel’, ‘DSL’ |

**A. DATASET: “GAS PRICES OF AMERICA” ACQUISITION & ANNOTATION**

As per [4], over 300,000 "seed" locations proximal to gas stations uniformly distributed across America were collected; of these, a GSV Image Collection App was used to collect 2,048 images containing advertised gas signs. Examples of relatively "easy" and "hard" instances are illustrated in Figs. 3 and 4, respectively. In Fig. 5, the State-wise geolocation of the images are summarized with notable outliers indicated. Only 1,024 sign-level masks and 2,048 cash prices for regular unleaded fuel (without the fractional component) were originally annotated [4]. We fully annotated the 2,048 image dataset by producing price-level, label-level (both gas grade and payment method), and digit-level masks. Grade labels were mapped to an octane rating (a measure of fuel stability) from the set: $g = \{\text{‘Regular’}, \text{‘Mid-Grade’}, \text{‘Premium’}, \text{‘Diesel’}\}$ (Table I); payment labels were mapped to classes from the set: $p = \{\text{‘Cash’}, \text{‘Credit’}, \text{‘Both’}\}$; digit-level masks were mapped to digit values with the addition of a special digit class for instances of "9/10" representing USD 0.009: $d = \{0, 1, \ldots, 9, 9/10\}$. Adding to the complexity and diversity of the dataset are the numerous fuel-type variants in advertisement; we map these instances to one of four fuel grades as seen in Table I.

As part of the annotation process, we relied on a single expert annotator to ensure consistency throughout the annotation process. Periodic co-author review of the annotations was performed for quality assurance and adherence to pre-determined definitions of the labelled elements. We visualize a number of exemplar annotated images with their contained bounding boxes to depict the annotation quality (Fig. 6).

**B. MODEL ARCHITECTURE: “GASBOTTY” COMPOSITE NEURAL NETWORK**

The end-to-end GasBotty architecture of component models and transforms is depicted in the overview Fig. 2. The overall processing pipeline can be broken down into the following subtasks:

1) From the input image (scene-level), segment the gas sign (sign-level mask) using a fine-tuned ResNet model (trained using the original 2,048 image,mask pairs of the GPA).

2) From the sign-level mask, we identify the four mask corners by applying Canny edge detection [22] and the Hough line transform [23], [24] to estimate the mask-edges which produces four clusters of points at the intersection of each Hough line to which we apply $k$-means ($k = 4$) [25] to obtain the centroid of each cluster.

3) With the estimated positions of the four sign corners, we extract from the scene-level image the perspective-distorted quadrangle corresponding to the advertised sign and apply keystone correction producing a sign-level image.
4) From the sign-level image, we independently detect both prices and labels using fine-tuned ResNets.
5) The detected labels are classified by fuel grade, and along with the detected locations of each price, a label-price association algorithm determines the optimal label-price assignment by minimizing the collective distance between price-label permutations measured using a weighted one-to-one Manhattan distance (described by Algorithm 1 in Section IV-G).
6) The detected price-level images are then provided to a fine-tuned ResNet digit detection model, and the detected digits are assembled into a final price.

Each of the learning models was selected following the comparison of both pre-trained and fine-tuned ResNet50, ResNet101, and ResNet152 models using 5-fold cross-validation over 75 epochs. While any class of model architecture could have been considered for the basis of these composite models, we opted for the three original ResNet models of varying complexity as first introduced in [26] to function as the basis of our proposed baseline for multi-metric extraction tasks. Each model’s hyperparameters were consistent with those optimized in [26]; we froze the model backbone and performed fine-tuning of the final layer of each model based on each given task. Specific details for each of the component ResNet models are described later.

The use of different model architectures (e.g. Faster R-CNN [27], YOLO variants [28], Single Shot Multibox Detectors [29], etc.) could be leveraged to further improve the performance and speed of similar multi-metric extractors. However, the determination of optimal model architectures and their hyperparameters are left to future work and those optimized models can be benchmarked against the performance of our selected ResNet models as part of this first version of GasBotty. For the purposes of establishing an initial multi-metric extraction benchmark, we chose to restrict our experimental framework to a single class of neural architecture.

C. RESNET MODEL ARCHITECTURE

The ResNet-based component models leveraged in this work are adapted from the three original ResNet models of varying complexity that were initially proposed by He et al. in [25]. While the internal model architecture was consistent with the ResNet50, ResNet101, and ResNet152 parameterization, we varied the input images and output layers to match the specific tasks of each network. The specifics of each network are illustrated in Fig. 7 along with the froze internal backbone.

In developing the GasBotty model, we leveraged the Keras RetinaNet environment that applies a preprocessing rescale to the input images of each of our ResNet models. This allows for variable sized inputs that are rescaled to reside within a consistent image size acceptable to the model. For simplicity, across all four of our ResNet component models, we leveraged the default image side ranges. The default min_side=800 parameter indicates that the image’s minimum side length in pixels will be equal to or greater than 800 pixels. The default max_side=1333 parameter indicates that, if following resizing, the image’s maximum side length exceeds 1,333 pixels, a subsequent resizing is performed until the maximum side length is equal to or less than 1,333 pixels. For example, our original Google Streetview image with size 640px×640px would be resized to 800px×800px.

More specifically for each model, the sign segmentation ResNet50 model takes as input the original 640px×640px three-channel RGB image and outputs a one-dimensional 640px×640px binary mask (where 1 indicates the detected sign and 0 otherwise). With Keras RetinaNet rescaling, the input image is rescaled to 800px×800px and the predicted 800px×800px output binary mask rescaled down to 640px×640px. The sign-segmentation model was trained on an 80/20 train:test split using annotator-provided ground-truth masks provided in the GPA dataset. The price detection ResNet50 model takes as input the keystone corrected sign image with variable-sized dimensions that are rescaled through Keras RetinaNet to a minimum side length of 800px and generates predicted bounding boxes around each of the detected prices within the image. The digit detection ResNet101 model takes as input the variable sized bounded box from the price detection model, rescales, and outputs bounding boxes around each of the detected digits and/or fractional component of the gas price (treated as a unique “digit”). Finally, the label detection ResNet152 model takes as input the same keystone corrected sign image with variable-sized dimensions that are rescaled through Keras RetinaNet to a minimum side length of 800px and generates predicted bounding boxes around each of the detected grade and/or payment type labels within the image. The details specific to each model are illustrated in Fig. 7 and additional architecture specifications are available in [26].

D. EVALUATION METRIC: “ALL-OR-NOTHING ACCURACY”

Given the number of grade, payment, and digit classes (|g| = 4, |p| = 3, and |d| = 11, respectively), a uniform or biased random predictor might achieve a low-range level of accuracy on a single subtask and yet be woefully impractical; given the complexity in both detecting and associating gas signage information, accuracy on its own cannot capture the complexity of the task. To address this for the task of multi-metric extraction in the wild, we propose "All-or-Nothing Accuracy" (ANA), defined for a given image i as:

\[
ANAt = \begin{cases} 
1 & \text{if } (\hat{g}_i = g_i) \wedge (\hat{p}_i = p_i) \wedge (\hat{d}_i = d_i) \\
0 & \text{otherwise}
\end{cases}
\]

where \(\hat{g}_i, \hat{p}_i, \hat{d}_i\) respectively represent the lists of predicted and associated grades, payment methods, and digits (including the fractional component and each assembled into multi-digit prices) and the \(g_i, p_i, d_i\) respectively represent the ground truth values for image i. The ANA metric strictly requires that all elements in the image be perfectly detected and associated without tolerance for a single error.
IV. EXPERIMENTS

We first provide an overview of the dataset and the available annotations and then describe how each component model or algorithm leveraged these data. Finally, we evaluate the performance of the GasBotty model over eight benchmark tasks of increasing difficulty, and evaluate it against two random models. Comparative results illustrate the utility of the fully annotated GPA dataset, the need for the ANA metric when evaluating multi-metric extractors, and the effectiveness of the GasBotty model as a first baseline against which future models might be compared.

A. DATASET OVERVIEW

The original release of the GPA dataset only contained 1,024 sign-level masks and the three digit cash price of regular, unleaded fuel. In Table 2 we list the number of annotations (both masks and values where available) for gas signs, prices, and labels. Prices were further annotated at the digit-level and the $\%$ was treated as an independent digit with a corresponding value of USD 0.009 (no other fraction appears in the datasets). The number and proportion of unique digits and labels are also reported; these frequencies are used as part of a biased random (random$^3$) predictor in Section IV-H.

This dataset contains 27,090 digits relating to over 6,001 prices and over 7,869 labels contained within 2,644 gas signs. Interestingly, the relative per-digit frequency of the training set roughly distribute according to Benford’s law [30], [31] with an average digit-frequency deviation of 1.27% as contrasted by the average digit-frequency deviation from a uniform distribution of 5.86%.

B. IMAGE-LEVEL SIGN SEGMENTATION

The first component model is a sign-segmentation network to localize the advertised gas sign within the scene-level image. We trained and compared two fine-tuned ResNet variants [26], ResNet50 and ResNet101, using the 2,048 sign masks and image pairs for 75 epochs using 5-fold cross-validation. Training was performed on Compute Canada infrastructure using an NVIDIA Tesla T4 GPU and 50GB of RAM over ~48 hours. We report the F1-score across epochs in Fig. [S].

Notably, the performance rapidly converges to a maximal performance between the ResNet50 and ResNet101 model and ultimately, the ResNet101 model performance was maximized at epoch 75. While both ResNets performing comparatively well, we selected the top-scoring ResNet101 as the better of the two model architectures (Fig. [S]).

C. ALGORITHMIC SIGN EXTRACTION

From the sign-level mask, we wish to identify the four mask corners from which we can extract the perspective-distorted quadrangle corresponding to the advertised sign and ultimately apply keystone correction to produce a perspective-corrected sign-level image as recommended in [32]. This extraction simplifies the detection task, facilitates annotation, and leads to improved label association. To this end, we first 0-pad the image with one pixel to account for out-of-scene signs and then apply a $3 \times 3$ Gaussian blur and Canny edge detection to delineate the bounds of the mask [22]. We then apply a Hough line transform [23], [24] to approximate the mask-edges using $\rho = 1$, $\theta = 1^\circ$, and a threshold of 37. The intersections of the Hough lines then form four clusters of points (if their angle are between $30^\circ$ and $150^\circ$) to which we apply the k-means ($k = 4$) algorithm [25] to obtain the centroid of each cluster, representing an estimate of the four sign corners enabling the extraction of the sign-segmented distorted quadrangle from the scene-level image. We then leverage these four points to produce a perspective-corrected variant of the advertised sign using Keystone correction (creating the sign-level image). The integration of these algorithmic components within the end-to-end pipeline is demonstrated in Fig. [S]. In later sections we describe how the enforcement of strict angular bounds may have limited the overall performance of our proposed model, particularly in the face of extreme perspectives of certain advertised gas signs. In future work, a model-driven approach, where we regress the four corners of a bounding polygon directly in lieu of a cascade of algorithmic image manipulations.

D. SIGN-LEVEL PRICE DETECTION

Leveraging sign-level imagery and price-level masks, we developed a price detection model that compares six ResNet variants: both pre-trained and fine-tuned ResNets 50, 101,

| Table 2. Summary of Annotations for the Training & Test Datasets |
|---------------------------------------------------------------|
| **Num.** | **Train** | **Test** |
| Images    | 2,048     | 512      |
| Signs     | 2,132     | 512      |
| Prices    | 4,766     | 1,235    |
| Digits: n (%) | 22,167 (100%) | 4,923 (100%) |
| 9         | 6,526 (29.4%) | 1,478 (30.0%) |
| 9/10      | 3,091 (13.9%) | — N/A — |
| 2         | 2,822 (12.7%) | 800 (16.2%) |
| 3         | 2,768 (12.5%) | 670 (13.6%) |
| 1         | 1,564 (7.06%) | 216 (4.38%) |
| 0         | 1,235 (5.57%) | 137 (2.78%) |
| 5         | 1,233 (5.56%) | 345 (7.00%) |
| 7         | 863 (3.89%)  | 226 (4.59%) |
| 4         | 838 (3.78%)  | 260 (5.28%) |
| 8         | 668 (3.01%)  | 639 (13.0%) |
| 6         | 559 (2.52%)  | 157 (3.19%) |
| Labels: n (%) | 5,403 (100%) | 2,466 (100%) |
| Regular   | 2,090 (38.7%) | 606 (24.6%) |
| Mid-Grade | 658 (12.2%)  | 174 (7.06%) |
| Premium   | 740 (13.7%)  | 198 (8.03%) |
| Diesel    | 988 (18.3%)  | 255 (10.3%) |
| Cash, Ca. | 534 (9.88%)  | 222 (9.00%) |
| Credit, Cc.| 393 (7.27%)  | 172 (6.97%) |
| Both, Ca./Cc.| — N/A — | 839 (34.0%) |

1 Tabulation of all instances available for training learning algorithms; digits sorted in rank-order by prevalence.

2 Tabulation of all target instances based on context-matching.

3 Incorporates the fractional $\%$ instances; otherwise ‘9’ : 727 and ‘9/10’ : 751.
and 152 (Fig. 8B). Models were trained under the same conditions as the sign-segmentation model: 75 epochs of 5-fold cross-validation using 1 × T4 GPU, 50GB RAM, ~48 hours. However, we report the mean average precision (mAP) as our performance metric.

The 95% confidence intervals reveal the majority of models achieve maximal performance within ~25 epochs. Even the pre-trained ResNet152 model demonstrates a comparable level of performance to the fine-tuned ResNet50 and ResNet101 models. Following 23 epochs, the ResNet50 achieves maximal performance, in agreement with the law of parsimony. From these results, the task of price-detection in natural imagery may effectively be considered a solved problem.

E. PRICE-LEVEL DIGIT DETECTION

Leveraging price-level imagery and digit-level masks, we develop a digit detection model that compared the same six ResNet variants under similar training conditions with a applied 5-fold cross-validation during training (Fig. 8C). We note that the performance across models is much more varied with the fine-tuned ResNet101 model rapidly reaching maximal performance at epoch 16.

F. SIGN-LEVEL LABEL DETECTION

Among the learning tasks, the development of a sign-level label detector using label-level masks was the most challenging. Using the same training parameters and configuration as the price- and digit-detectors, the fine-tuned ResNet152 model achieved maximum mAP at epoch 64. Based on the relative performance of models for each learning task, it appears label detection and classification is the most challenging. We posit that this may be due to the diverse representations possible for each label class and increasing the available training samples will result in more robust models. Promisingly, the four learning models each achieve a modest-to-high performance level useful to the end-to-end evaluation of the composite neural network.

G. ALGORITHMIC LABEL ASSOCIATION

With the extracted multi-digit price values and their relative positions within the image as well as the classified labels (grade & payment) with relative positions, we implemented a label-to-price association algorithm based on an element-to-element distance minimization (Algorithm 1). As an exhaustive search, the algorithm guarantees returning the optimal label-to-price mapping which for our purposes is considered optimal. For concision, we only describe the grade-to-price association algorithm. Briefly, a list of predicted prices and their coordinates, \( \hat{p} \), and list of predicted grades and their coordinates, \( \hat{g} \), are size-matched (Algorithm 1 lines 1-6). All possible permutations of \( \hat{g} \) are then generated, \( \hat{G} \), and a weighted one-to-one Manhattan distance is measured between the centroids of elements in the candidate permutation and the predicted prices (Algorithm 1 lines 7-14). The permutation of \( \hat{g} \) resulting in a minimized distance is the most likely candidate assignment. We weight the Manhattan distance penalty by type: vertical distance penalization is weighted by a factor \( 3 \times \) greater than horizontal distance and if the label is below the price, an additional \( 1 \times \) vertical distance is applied.

**Algorithm 1** Grade-to-Price Association

**Input:** list of detected prices & coordinates, \( \hat{p} \)

**list of detected grades & coordinates, \( \hat{g} \)

**Output:** grade-to-price assignment, \( \alpha \)

1: if \( |\hat{p}| > |\hat{g}| \) then
2: \( \hat{g} \leftarrow \hat{g} \times 2 \)
3: end if
4: while \( |\hat{p}| \neq |\hat{g}| \) do
5: add \( \emptyset \) to smaller list
6: end while
7: \( \hat{G} \leftarrow \) generate all possible permutations of \( \hat{g} \)
8: \( d_m, \alpha \leftarrow \infty, \emptyset \)
9: for each \( \hat{g} \) in \( \hat{G} \) do
10: \( d \leftarrow \) weighted 1-to-1 Manhattan distance of \( \hat{g} \) & \( \hat{p} \)
11: if \( d < d_m \) then
12: \( d_m, \alpha \leftarrow d, \hat{g} \)
13: end if
14: end for

| Task Id. | Visual Element | Fraction Inclusion | Metric | Predictor | Performance (\( \mu \pm \sigma \)) |
|----------|----------------|-------------------|--------|-----------|----------------------------------|
| T1 | Sign | N/A | F1 | GasBotty [ResNet101] | 0.8773 \( \pm \) 0.1150 |
| T2 | Price | w/o | Recall | EAST/Tesseract GasBotty | 0.1229 \( \pm \) 0.2567 0.9878 |
| T3 | Digit | N/A | Acc. | EAST/Tesseract random\_u random\_g GasBotty | 0.0493 \( \pm \) 0.1202 0.3276 \( \pm \) 0.0134 0.3635 \( \pm \) 0.0911 0.9721 |
| T4 | Label | N/A | Acc. | EAST/Tesseract random\_u random\_g GasBotty | 0.0039 \( \pm \) 0.1104 0.3152 \( \pm \) 0.0962 0.4534 \( \pm \) 0.0100 0.9570 |
| T5 | Price | w/o | ANA | EAST/Tesseract random\_u random\_g GasBotty | 0.0314 \( \pm \) 0.1743 0.0215 \( \pm \) 0.0195 0.0543 \( \pm \) 0.0347 0.9570 |
| T6 | Price | w/o | ANA | EAST/Tesseract random\_u random\_g GasBotty | 0.000 \( \pm \) 0.000 0.0213 \( \pm \) 0.0056 0.0240 \( \pm \) 0.0091 0.8710 |
| T7 | Price | w/ | ANA | EAST/Tesseract random\_u random\_g GasBotty | 0.000 \( \pm \) 0.000 0.0213 \( \pm \) 0.0056 0.0240 \( \pm \) 0.0091 0.8710 |
| T8 | All | w/ | ANA | EAST/Tesseract random\_u random\_g GasBotty | 0.000 \( \pm \) 0.000 0.0191 \( \pm \) 0.0027 0.0207 \( \pm \) 0.0054 0.7285 |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
H. EVALUATION OF THE END-TO-END PIPELINE

To fully evaluate the performance of the proposed GasBotty model, we considered eight benchmark tasks. These are organized in a way to function akin to an ablation study where the contribution of component models is demonstrated independently of other model components. This experimental design enables the determination of the relative difficulty of multi-metric subtasks. We illustrate the end-to-end performance of the GasBotty model on an example image in Fig. 9 interleaved among the resultant imagery are the component models and transforms that produce a given result. The individual benchmark sub-tasks, numbered as "T#" are outlined here and ordered in increasing difficulty:

- **T1:** Sign-level segmentation to predict the region of the image corresponding to the advertised gas prices.
- **T2:** Price-level recall comprising the proportion of correctly detected gas prices.
- **T3:** Digit-level accuracy comprising the proportion of correctly detected and classified digits.
- **T4:** Label-level accuracy comprising the proportion of correctly detected and classified labels (both grade and payment type).
- **T5:** All-or-Nothing extraction of at least one of the listed prices.
- **T6:** All-or-Nothing extraction of all of the listed prices (excluding fractions).
- **T7:** All-or-Nothing extraction of all of the listed prices (including fractions).
- **T8:** All-or-Nothing extraction of all listed prices, including fractions when present, as well as the correct association of grade of fuel and payment type.

We tabulate the results of each subtask in Table 3, comparing the performance to two variants of a random predictor: a uniform random (random_u) model and a biased random (random_m). The random_m generates predictions for each element (digits & labels) by drawing from a uniform distribution whereas the random_u is drawn from an a priori distribution of element-frequency in the test dataset (Table 2). Furthermore, as each image contains a variable number of target predictions, the random models are further advantage by also knowing the number of targets a priori (they predict the matching number of targets) which is not available to the GasBotty model nor comparative model. We generated a bootstrap distribution of random model predictions following $k = 100$ iterations and report the mean performance and standard deviation. Finally, we additionally compared the GasBotty and random performance to a task-adapted state-of-the-art (SOTA) scene text detection model selecting the Efficient and Accurate Scene Text (EAST) Detector [10] for detecting price and label bounding regions that are then passed as input to Google’s Tesseract OCR Engine [33] (denoted as EAST/Tesseract for their combined use) to read the contents of those bounding boxes.

I. COMPARISON TO A TASK-ADAPTED SOTA MODEL

The EAST detector is a SOTA scene text detection pipeline that directly predicts the bounding coordinate region of words, text lines, and number from natural scenes [10]. The output coordinate quadrangle for each text detection can then be read using a SOTA OCR method; in this work we leveraged the Tesseract OCR Engine developed by Google [33]. In combination, this SOTA pipeline could be effectively applied and the performance fairly compared to our proposed model for each of the sub-tasks considered in this work. To that end, we evaluated the performance of the SOTA EAST/Tesseract model on each of the sub-tasks considered in this work. Importantly, we note that this EAST/Tesseract model was not provided the same a priori information that were afforded to both the random models. However, given that the EAST/Tesseract model was not explicitly designed for multi-metric extraction in the wild (and this gasp-price specific use case) we considered much more lenient success criteria for the EAST/Tesseract model on these sub-tasks. A complete description of the methodological considerations are available in the Supplementary Materials. All model results are tabulated in Table 3.

Notably, for the first four subtasks T1–T4 the GasBotty model achieves a high-level of performance with T1: 0.8773 ± 0.1150% recall, T2: 98.78% recall, T3: 97.21% accuracy, and T4: 95.70% accuracy. For the latter two tasks (T3: digit- & T4: label-level prediction) both of the random models achieve a relatively low (yet non-zero) performance. However, the last four subtasks T5–T8 each leverage the ANA metric and we note that the performance of the random models rapidly drop to zero.

In fact, the only reason the random predictor have a non-zero performance is due to a single instance in the test set where a sign contains no price or label information, enabling the random predictors, as they have been defined, to achieve an ANA of 1 for that unique instance.

With a T5: 95.70% ANA, T6: 87.10% ANA, and T7: 87.10% ANA, the GasBotty model can effectively extract price-level information from imagery regardless of the presence of fractional digits or not (within the test dataset, GasBotty effectively extracts the fractional digit in every instance). Most impressive is the T8: 72.85% ANA for effectively extracting and correctly associating all sign information. Our model generalized well on the set of independent test images and therefore did not overfit the training data. This is further supported by our experimental design: each component model was trained using 5-fold cross-validation where the selected model for the composite GasBotty model was chosen when the training performance of each model began to plateau in performance. Certainly, the GasBotty model may be improved and serves as an initial benchmark. As the first multi-metric extractor in the wild, these are promising results.

In comparing these results to the SOTA EAST/Tesseract model outcomes, a number of notable (and often surprising) findings result. Foremost, in T2 the EAST/Tesseract model
experienced difficulty in identifying prices given that the Tesseract OCR Engine was often unable to identify at least one digit within the EAST detected bounding region. This resulted in a very large variation as seen in the reported standard deviation value. Based on the success criteria outlines for EAST/Tesseract in the Supplementary Material, as long as a single digit was detected in the predicted price bounding region, this was considered a "successful price detection". However, in the majority of cases, even a single detection was not detected. For multiple prices per test image, this resulted in a low price-detection recall and high standard deviation for those instances where a digit was in fact detected.

When applying the EAST/Tesseract model to digit-level detection in T3, we see a sharp drop in performance even as compared to the two random models considered. Given that the two random models benefit from a priori knowledge of the total number of expected digits in a given image, these models have a significant advantage. The EAST/Tesseract model, beyond misclassifying digits of an existing price bounding region, in mislocalizing even a single price (and therefore missing all potential digits therein), cannot achieve a comparable performance as the random models with a priori information of the number of expected digits per scene. Hence, the EAST/Tesseract model performs considerably worst than the two random models (5% EAST/Tesseract vs. 30% for random models). Ultimately, the GasBotty demonstrates that it has successfully learned this same task with a reported 97.21%.

For the label-detection task of T4, the EAST/Tesseract results are comparable. As described in the Supplementary Material, the EAST/Tesseract model is evaluated using a lenient criterion for success and yet achieves 3% of successful label detection as compared to the randomized models in the range of 31-45%. The GasBotty model achieved 95.7% for this task and is clearly the superior model.

Then for tasks T5-T8, both the EAST/Tesseract and proposed random models fail to fully detect the necessary information and each tend towards 0% ANA. The stringency of the ANA makes these tasks particularly challenging given that even a single error is intolerable (including the detection of fractions). Most impressively, the GasBotty achieves 73% ANA over the dataset indicating that it can perfectly extract and associate all gas sign for 358/512 test images.

Finally, to investigate whether or not model errors were geographically biased, we plotted the T8 ANA error rate by state (Fig. 10). Interestingly, several southern states (e.g., Arizona, Tennessee, and North Carolina) have relatively few errors while certain north-eastern states (e.g., Vermont & Iowa) have relatively high error rates. Given the relatively small number of samples distributed among states, wide variation in performance is expected. The proportion of errors across the US otherwise appear uniformly distributed and suggests generalized model performance across the USA should the GasBotty model be integrated as part of camera-laden vehicles, such as AVs.

J. LIMITATIONS

For certain benchmark sub-tasks listed in Table 3, a comparison to vanilla SOTA scene text-detection models is challenging given certain tasks require label association. While most SOTA models are not designed for the task of simultaneous object detection and label association, by the definition of these tasks and the stringency imposed by the ANA metric, some of these SOTA models can be compared (as demonstrated with EAST/Tesseract) however we demonstrated they may fare no better and in some cases worse than the random predictors we compared ourselves to (even with certain accommodated leniencies in sub-task criteria; see Supplementary Material). Consequently, it is important to highlight where these existing SOTA models could be improved and adapted to these (sub)tasks.

From our extensive experiments, we note that while the EAST model was capable of bounding a region of interest to a price and/or context label, the Tesseract OCR Engine was unable to read the contents. Further investigation into these failure cases suggests that the low-resolution representation of digits/labels for a number of reasons (small signage, blurring, obfuscation, truncation, etc.) results in missed detection(s). To improve upon these models, some fine-tuning will be required with the data and annotation presented in this work.

K. FUTURE DIRECTIONS

As previously described, no single model architecture has yet been leveraged for the end-to-end task of multi-metric extraction and future work will seek to optimize a unified model architecture in lieu of a composition of component models following a cascade of single-task convolutional neural networks. This work, therefore, serves as a key multi-metric extraction benchmark against which those models may be measured.

Furthermore, the optimization of individual component models and of end-to-end learned multi-metric extractors is left to future work which may compare themselves to the benchmark established in this work. Moreover, as tabulated in Table 3, the GasBotty selected component ResNet models leveraged in tasks T1-T4 already achieve a high-level of performance across the individual sub-tasks suggesting that other SOTA models could attain a similar or improved level of performance if specifically adapted to this task.

Our experiments were organized in a way to function akin to an ablation study where the contribution of component models is demonstrated independently of other model components. This enables the determination of the relative difficulty of multi-metric subtasks and our findings indicated that the most challenging facet of multi-metric extraction lies in the complete detection of all relevant elements and their final label detection and association stages.

Finally, beyond these limitations imposed by the task of multi-metric extraction, we would like to acknowledge that the dataset contributed in this work, in being built from Google Street View imagery, is predominantly limited to day-
time imagery (with a few rare exceptions as demonstrated in Fig. 4E), fair-weather, and summer-like environments. This potentially limits the utility of the GasBotty model, however, future work can focus on the generation of increasingly comprehensive datasets, potentially sourced from circulating AVs. Moreover, given our reported performance, our proposed GasBotty model could potentially be leveraged as part of the annotation process for such datasets.

V. CONCLUSION
The lived environment proximal to roadways are filled with multi-digit and multi-numbered values corresponding to advertised commodities that, if effectively extracted and associated to contextual meaning, would provide considerable value. To our knowledge, no large-scale visual dataset comprising multi-digit, multi-number values with associated context labels currently exists and in this work, we denote this class of problems as "multi-metric extraction in the wild". We focus on the accurate detection and reading of gas prices and contextual association to gas grade and payment type by: 1) fully annotating the Gas Prices of America dataset, 2) leveraging these data to develop the GasBotty composite neural network, and 3) evaluating it over eight benchmark tasks from an end-to-end pipeline of increasing difficulty using a new highly-stringent, binary-type metric, denoted "All-or-Nothing Accuracy". Our proposed model achieved 72.9% ANA and is as a benchmark against which future multi-metric extraction models might be evaluated.

REFERENCES
[1] Tinnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. Proceedings of the National Academy of Sciences, 114(50):13108–13113, 2017.
[2] Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. Assessing street-level urban greenerney using google street view and a modified green view index. Urban Forestry & Urban Greening, 14(3):675–685, 2015.
[3] Andrew G Rundle, Michael DM Bader, Catherine A Richards, Kathryn M Neckerman, and Julien O Teitter. Using google street view to audit neighborhood environments. American journal of preventive medicine, 40(1):94–100, 2011.
[4] Kevin Dick, François Charif, Jimmy Woo, and James R Green. Gas prices of america: The machine-augmented crowd-sourcing era. In 2020 17th Conference on Computer and Robot Vision (CRV), pages 158–165, IEEE, 2020.
[5] Ofer Matan, Christopher JC Burges, Yann LeCun, and John S Denker. Multi-digit recognition using a space displacement neural network. In Advances in neural information processing systems, pages 488–495, 1992.
[6] Ray Smith. An overview of the tesseract ocr engine. In Ninth international conference on document analysis and recognition (ICDAR 2007), volume 2, pages 629–633. IEEE, 2007.
[7] Parul Shah, Sunil Karamchandani, Taskeen Nadkar, Nikita Gulechha, Kaushik Koli, and Ketan Lad. Ocr-based chassis-number recognition using artificial neural networks. In 2009 IEEE International Conference on Vehicular Electronics and Safety (ICVES), pages 31–34. IEEE, 2009.
[8] Arival Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011.
[9] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082, 2013.
[10] Xinya Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 5551–5560, 2017.
[11] Xiaobing Wang, Yingying Jiang, Zhenbo Luo, Cheng-Lin Liu, Huyson Choi, and Sangjin Kim. Arbitrary shape scene text detection with adaptive text region representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[12] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2013 robust reading competitions: entries, results, and future directions. International Journal of Document Analysis and Recognition (IJDAR), 7(2-3):105–122, 2005.
[13] Robert Nagy, Anders Dicker, and Klaas van der-Wegen. Neocr: A configurable dataset for natural image text recognition. In International Workshop on Camera-Based Document Analysis and Recognition, pages 150–163. Springer, 2011.
[14] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In 2011 International Conference on Computer Vision, pages 1457–1464. IEEE, 2011.
[15] Andreas Veit, Tomas Matuska, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140, 2016.
[16] Raymond Smith, Chunhui Gu, Dar-Shyong Lee, Huiyi Hu, Ranjith Unnikrishnan, Julian Ibarz, Sacha Arnoud, and Sophia Lin. End-to-end interpretation of the french street name signs dataset. In European Conference on Computer Vision, pages 411–426. Springer, 2016.
[17] Kevin Dick, Joshua B. Tanner, François Charif, and James R. Green. Fully Annotated Gas Prices of America Dataset for Multi-Metric Extraction in the Wild, 2022.
[18] Kevin Dick, François Charif, Yasmina Souley Dosso, Luke Russell, and James R Green. Systematic street view sampling: High quality annotation of power infrastructure in rural ontario. In 2018 15th Conference on Computer and Robot Vision (CRV), pages 134–141. IEEE, 2018.
[19] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pages 1156–1160. IEEE, 2015.
[20] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 336(10):1995, 1995.
[21] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6985–6994, 2018.
[22] John Canny. A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, 8(6):679–698, 1986.
[23] Paul VC Hough. Method and means for recognizing complex patterns, December 18 1962. US Patent 3069654.
[24] John Illingworth and Josef Kittler. A survey of the hough transform. Computer vision, graphics, and image processing, 44(1):87–116, 1988.
[25] Anil K Jain and Richard C Dubes. Algorithms for clustering data. Prentice-Hall, Inc., 1988.
[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015.
[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.
[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.

[30] Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. American Journal of mathematics, 4(1):39–40, 1881.

[31] Frank Benford. The law of anomalous numbers. Proceedings of the American philosophical society, pages 551–572, 1938.

[32] Kevin Dick, Joshua B Tanner, and James R Green. To keystone or not to keystone, that is the correction. In 2021 18th Conference on Robots and Vision (CRV), pages 142–150. IEEE, 2021.

[33] Ray Smith. An overview of the tesseract ocr engine. In Ninth international conference on document analysis and recognition (ICDAR 2007), volume 2, pages 629–633. IEEE, 2007.

KEVIN DICK (M’15) received the B.Sc. degree in biology and computer science from McGill University, Montréal, QC, Canada in 2014 and is currently pursuing a Ph.D. degree in biomedical engineering (specialized in data science and bioinformatics) at Carleton University, Ottawa, ON, Canada.

He is currently a Research Assistant at the Carleton University Biomedical Informatics Collaboratory (cuBIC) and a Research Scientist for the Centre of Access to Information and Justice (CAIJ) at the University of Winnipeg. His current research includes machine learning and the use of high-performance computing infrastructure for biomedical informatic applications, computer vision to study phenomena of the natural environment, and the application of computational methods to solve digital access challenges.

JOSHUA B. TANNER received a B.Eng degree in mechanical engineering from Carleton University of Ottawa, ON, Canada in 2020. Joshua Tanner is currently pursuing a M.Eng degree in Electrical and Computer Engineering at Carleton University in Ottawa, ON, Canada.

He is a Research Assistant for Carleton University, Ottawa, ON, Canada. In the past he was a Mechanical Engineering Student for GasTOPS, Ottawa, ON, Canada and a Draftsman for Lee Valley and Veritas Tools, Ottawa, ON, Canada.

FRANÇOIS CHARIH earned a B.Sc. in biochemistry and a B.A.Sc. in chemical engineering from the University of Ottawa in 2016, before completing a M.A.Sc. in Electrical and Computer Engineering at Carleton University in 2019. He is currently a Ph.D. candidate at Carleton University where he researches the applications of artificial intelligence in peptide drug design.

JAMES R. GREEN (S’02–M’05–SM’10) received the B.A.Sc. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1998, and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from Queen’s University, Kingston, ON, Canada, in 2000 and 2005, respectively. He is currently a Professor with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada. His current research interests include machine learning challenges within biomedical informatics, patient monitoring, image analysis, computational acceleration of scientific computing, and the design of novel assistive devices.
FIGURE 5. Summary of gas station geolocation by state. The ordering of states is preserved in decreasing order of number of gas stations in each panel.

FIGURE 6. Example Annotations for Four Dataset Images. Each annotated element is bounded in a coloured box. Overlapped elements may be difficult to distinguish such as the individual digits bounding box over the containing price bounding box. Note that the colouring is arbitrarily applied by the annotation program and the same bounding box colouring does not necessarily imply the same label.
FIGURE 7. Visualization of the four ResNet Architectures, their Input Images, and Generated Outputs. For additional information on the architectural layers of each of our ResNet models, refer to the work of He et al. in [26]. It must be noted that the input images for each ResNet model, if below the minimum pixel amount of one axis, they are then upsampled to match that minimum criteria before being passed through the model.

FIGURE 8. Results of each of the learned component models for specific detection subtasks using the training images. (A) reports performance for the 5-fold cross-validation sign segmentation model results with the best ResNet101 model occurring at epoch 75. (B) reports the performance of six price detection models with the best ResNet50 model occurring at epoch 23. (C) reports the performance of the six digit detection models with the best ResNet101 model occurring at epoch 16. (D) reports the performance of six label detection models with the best ResNet152 occurring at epoch 64.
FIGURE 9. Example end-to-end pipeline applied to an image.

FIGURE 10. ANA error rate by state. Dashed states represent those for which evaluation data are unavailable.