User Manual of DVM-CAR: A Large-Scale Automotive Dataset for Visual Marketing Research and Applications

Jingmin Huang*, Bowei Chen†, Lan Luo‡, Shigang Yue♯ and Iadh Ounis*

∗ School of Computing Science, University of Glasgow, UK
† Adam Smith Business School, University of Glasgow, UK
‡ Marshall School of Business, University of Southern California, USA
♯ School of Computer Science, University of Lincoln, UK

This document includes the following sections:

CONTENTS

I. Basics of DVM-CAR

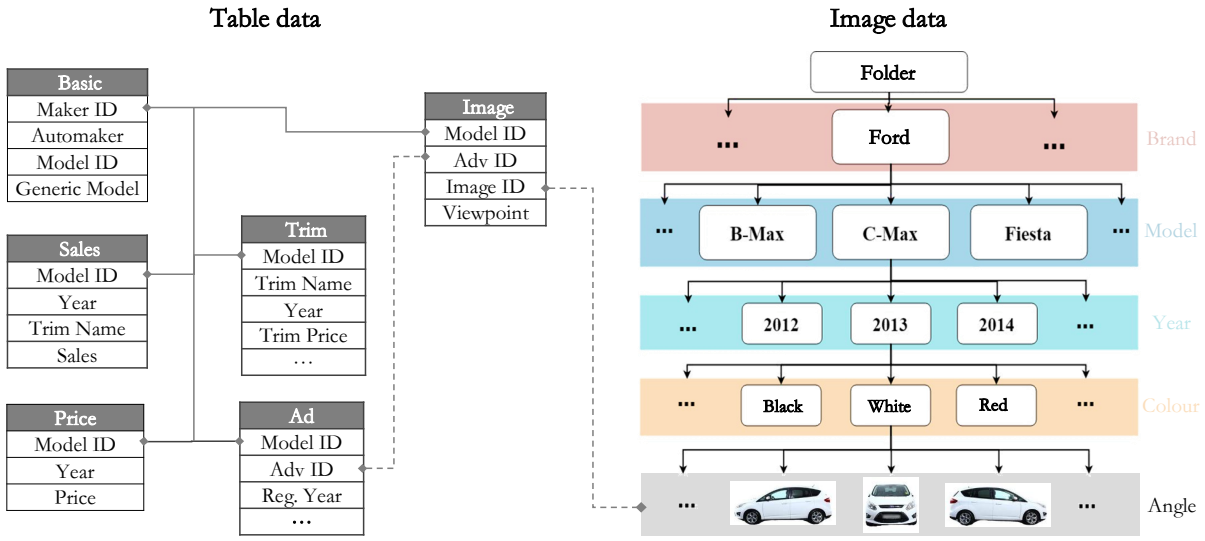


Fig. 1. **Structure of DVM-CAR dataset.**

This is the user manual of the DVM-CAR 2.0, a large-scale automotive dataset for visual marketing research and applications. Our initiative of creating this publicly available data is motivated by the recent growing trend of research interest in automotive aesthetics but there is no large-scale dataset available that can cover a wide range of needed variables and information. The targeted users of the DVM-CAR are multidisciplinary researchers who are interested in automotive aesthetics analytics and design, which are at the intersection of marketing, design studies, operations management, data science, and machine learning. Therefore, the DVM-CAR 2.0 can be used (but not limited to) the research and applications like understanding automotive aesthetics and consumer choice, AI-powered automotive exterior design, visual-based used car pricing.

Specifically, the DVM-CAR 2.0 contains car images, model specification and sales information about 899 car models that have been sold in the UK market over the last 20 years. As illustrated in Fig. 1, it comprises two data parts: the **image data** and the **table data**. The former contains 1,451,784 car images that have been deliberately cleaned and organized. While the latter includes six CSV tables that cover the non-visual attributes such as brand, price, sales etc. A general introduction to these tables is available in Table I. For easy access, the whole DVM-CAR is designed as a relational database. All the CSV tables could be joined by two primary keys, 'Genmodel ID' and 'Adv ID'. It means you can easily connect any two provided attributes by simply operations, which offers large flexibility for diverse data usage needs. On the other hand, the provided images are suitable for machine learning/deep learning applications. Pictures are labeled according to their viewpoints, unified to the same size, and strictly follow the same storage structure.

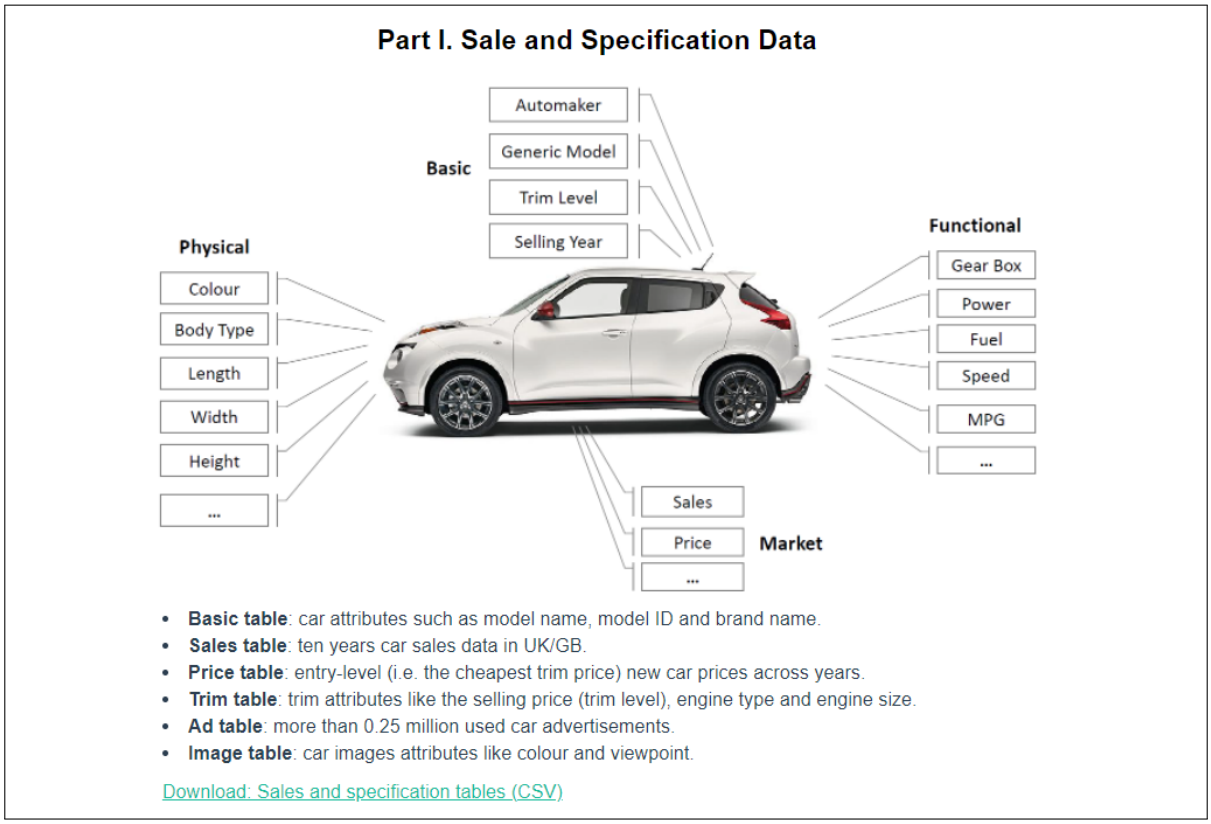
Our dataset resource paper "DVM-CAR: A Large-Scale Automotive Dataset for Visual Marketing Research and Applications" provides more details about the DVM-CAR 2.0 from a scientific perspective, including our motivation, how the dataset is designed, how the car images are processed, and where the dataset can be applied. Three simple application examples are also provided in the resource paper. It should be noted that users should only apply the DVM-CAR 2.0 for **non-commercial purposes only** as this dataset is under the CC BY-NC license. Next, in this user manual, we provide a step-by-step guidance on how to download and load the DVM-CAR 2.0.

TABLE I
OVERVIEW OF THE DVM-CAR

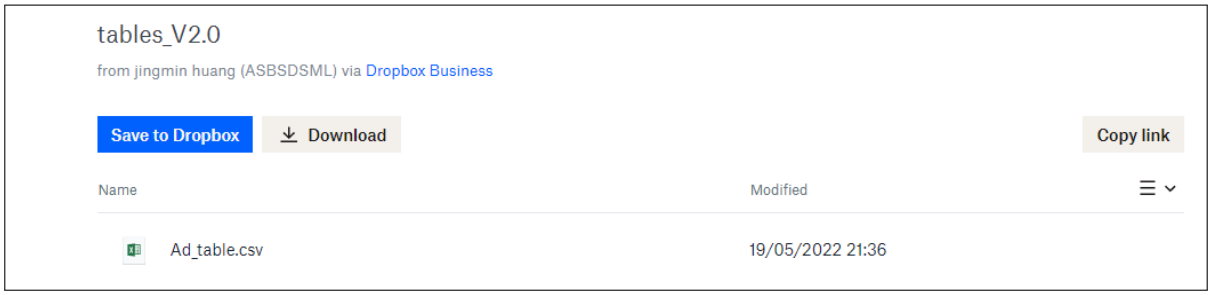
Image (Single zip file 13.6 GB size)	• Images are stored under the categorization '*brand-model-year-color*'. This structure allows researchers to easily locate images. • Images are provided in JPG format with resolution 300×300, and with background removed.
Tables (Six CSV files)	• *Basic Table*. This table is mainly for indexing other tables. It includes 1,011 generic models from 101 automakers. • *Sales Table*. It contains over 20 years car sales data of the UK market (based on the government released statics of new car registrations). In sum, it covers the sales of 773 car models from 2001 to 2020. • *Price Table*. It is designed for users who only need the basic price data of historical models. This table contains the entry-level (i.e., the cheapest trim price) new car prices of 647 models since 1998. • *Trim Table*. It includes 0.33 million trim-level information such as selling price, fuel type and engine size. • *Ad Table*. It shows more than 0.27 million used car advertisements information collected from online sources. It consists of variables like the advertisement's creation time and the car's registration year, cumulate mileage and selling price. • *Image Table*. It contains image information like predicted viewpoint, quality check results. Currently, it includes information of 1,451,784 car images, while expansions of this table are applicable in the future.

II. DATASET DOWNLOAD

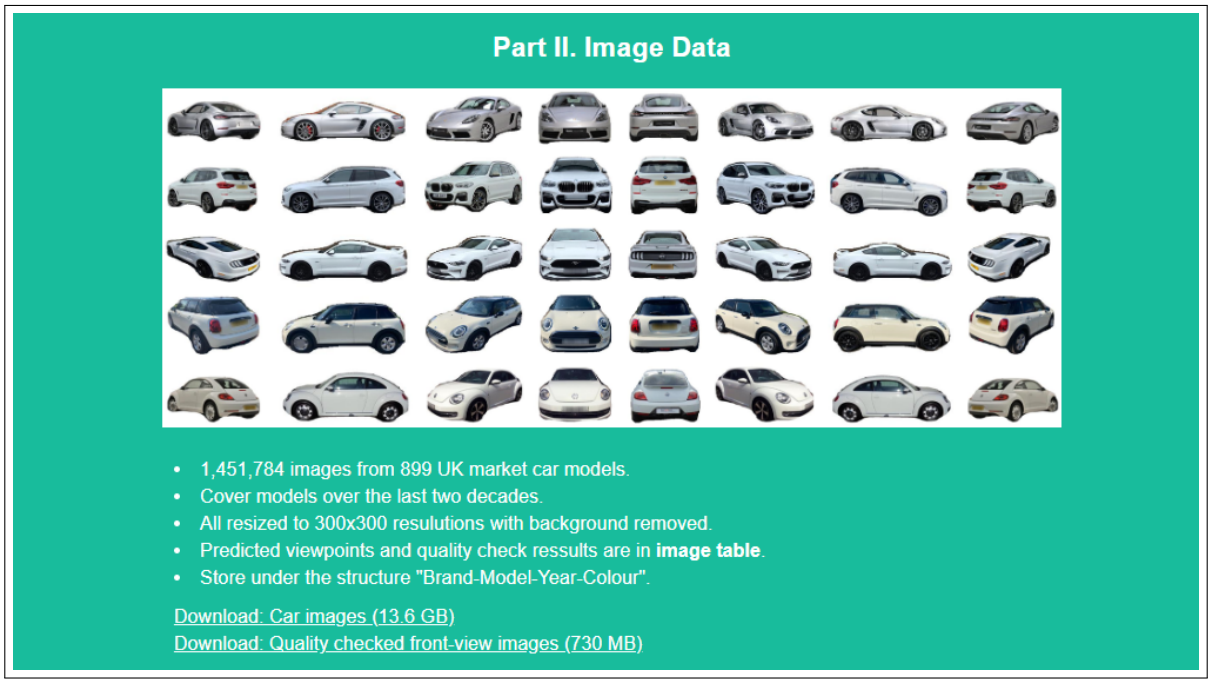
1. Visit the DVM-CAR webpage: https://deepvisualmarketing.github.io.
2. Click the link "Download: Sales and specification tables (CSV)" in section "Download Part I. Sale and Specification Data" to access the shared Dropbox folder of the table data.



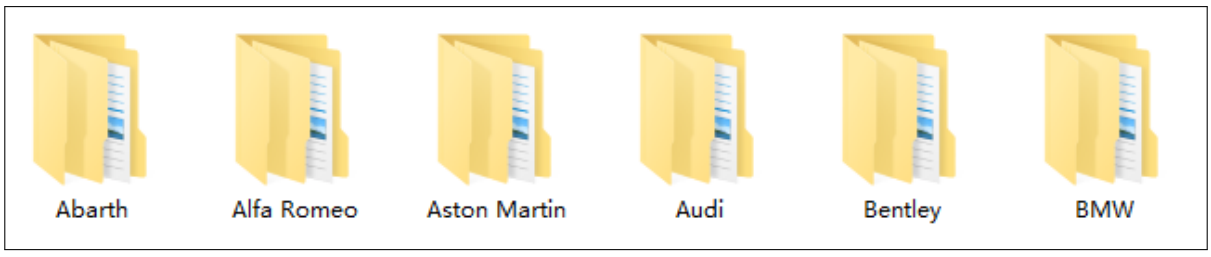
3. The shared Dropbox folder contains six tables of the DVM-CAR 2.0. The user can import the table data into their Dropbox accounts or download them directly (no need to log in).



4. One can follow the similar steps to download the image data. There are two links in Section "Part II. Image Data": "Download: Car images (13.6 GB)" and "Download: Quality checked front-view images (730 MB)". The former contains the whole set for 1,451,784 images, while the latter provides a smaller set with 61,827 front-view images that have been manually checked (which could be applied directly for some deep learning tasks). The user can clink the link to access the corresponding shared Dropbox folder of the image data.



5. The image data file is too big to preview in the Dropbox downloading page. The user can download the file first and unzip it. The root directory of the unzipped files contains 84 folders for various car brands.



6. The sub-folders of the unzipped files follow the '*brand-model-year-color*' structure, and the end-level folders contain car images in the JEPG format. The following are the the examples of the image data under directory ".../Abarth/595/2018/White". To understand a image's name, the string means: '{Maker name}$${Model name}$${Registration year}$${Color}$${Genmodel ID}$${Adv ID}$${Image index}'.

III. To Start

Considering our targeted users include social science researchers (e.g., marketing or management researchers), here we provide a simple illustration to extract information from the DVM-CAR 2.0. Let us assume that a user is interested in sales, prices and images of 2018 saloon models. He/she can extract those information following the steps as below. We use Python in the illustration but the steps can be followed using other programming or data analytics languages.

1. Load *Ad Table*, *Price Table* and *Sales Table* as dataframes. 'table_dir' should be changed to the user's data path.

```
import os
import pandas as pd

table_dir = r'C:\Users\DVM_CAR\Downloads\tables_V2.0'
ad_table = pd.read_csv(os.path.join(table_dir, 'Ad_table.csv'))
price_table = pd.read_csv(os.path.join(table_dir, 'Price_table.csv'))
sales_table = pd.read_csv(os.path.join(table_dir, 'Sales_table.csv'))
```

2. Use *Ad Table* to find 'Genmodel ID's of 2018 saloon models.

```
##Find the 2018 Ads
y2018_records = ad_table.loc[ad_table['Reg_year']==2018]

##Then get the Saloon Ads
saloon_records = y2018_records.loc[y2018_records['Bodytype']=='Saloon']

##Get the `Genmodel_ID' of these 2018 Saloon models
gid_saloon_2018 = set(saloon_records['Genmodel_ID'])
```

3. Find the prices of 2018 saloon models, and join *Price Table* with *Sales Table* on 'Genmodel ID'. It should be noted that all tables can be joined via 'Genmodel ID' or 'Adv ID'.

```
##Get the price records of Saloon models
saloon_prices = price_table.loc[price_table['Genmodel_ID'].isin(
                                                    gid_saloon_2018)]

##Only keep 2018 price records
y2018_prices = saloon_prices.loc[saloon_prices['Year']==2018]

##Join the price and sales dataframes to get the merged data
```

```
saloon_2018_price_sales = pd.merge(y2018_prices, sales_table[['Genmodel_ID'
   ,'2018']], how='left',left_on=['Genmodel_ID'], right_on=['Genmodel_ID'])
```

4. Use *Image Table* to list front-view images belonging to adverts of 2018 saloon models.

```
##Read the image table
img_table = pd.read_csv(os.path.join(table_dir, 'Image_table.csv'))

##Create a ID set for the found Saloon Ads
tar_ad_ids = set(saloon_records['Adv_ID'])

##Initial a empty list to store the found image IDs
tar_img_ids = []

##Check all image IDs and add the 2018 Saloon IDs into the tar_img_ids
for img_id in img_table['Image_ID']:
    if '$$'.join(img_id.split('$$')[:2]) in tar_ad_ids:
        tar_img_ids.append(img_id)

##Slice the datadaframe for target images
tar_imgs_recs = img_table.loc[img_table['Image_ID'].isin(tar_img_ids)]

##Remove images that not taken from front viewpoint
saloon_recs = tar_imgs_recs.loc[tar_imgs_recs['Predicted_viewpoint']==0]

##Extract the names of target images
saloon_2018_front_imgs = saloon_recs['Image_name']
```

5. Copy targeted images into a new folder.

```
from shutil import copyfile

##Define a new function that can copy images into a new folder
def copy_DVM_imgs(source_dir, tar_dir, imgs_names):

    ##Create the new folder if it is not exist
    os.makedirs(tar_dir, exist_ok=True)

    for fna in imgs_names:
        ##Extract the maker, model, year, and color from the image name
        relative_path = fna.split('$$')[:4]
        ##Reconstruct the absolute path to copy
        source_fpa = os.path.join(source_dir, *relative_path, fna)
        ##Generate the absolute path to paste
        tar_fpa = os.path.join(tar_dir, fna)
        ##Copy the image
        copyfile(source_fpa, tar_fpa)

##Declare the source folder (modify according to own path)
source_dir = r'C:\Users\DVM_CAR\Downloads\resized_DVM_v2

##Declare the target folder for pasting (modify according to own path)
tar_dir = r'C:\Users\DVM_CAR\Downloads\saloon_2018_front_designs'
copy_DVM_imgs(source_dir, tar_dir, saloon_2018_front_imgs)
```

IV. URL, HOSTING, LICENSE AND MAINTENANCE PLAN

The DVM-CAR 2.0 is publicly available at: https://deepvisualmarketing.github.io. This GitHub webpage provides data overview, usage statement, download links, timeline, and contact information of dataset developers. Users

can directly download the dataset from the provided Dropbox links. Besides this, for long-term maintenance considerations, this dataset is also hosted on Figshare with a persistent DOI https://doi.org/10.6084/m9.figshare.19586296.v1. This dataset is under the **CC BY-NC** license, in other words, it should be used for non-commercial purposes only. It should be pointed out that the DVM-CAR project has been carried out for over four years. It will be well maintained by the development team for a considerable long time.