CarVer: Setting the Standard for Face Verification with Caricatures

Sara R. Davis, Bryson Lingenfelter, Kevin McElhinney, Shamik Sengupta and Emily M. Hand
CSE Dept.
University of Nevada, Reno
Reno, NV, USA
sarad@nevada.unr.edu

***Abstract*—Caricatures exaggerate the most prominent features of a face, highlighting an individual's unique features. Research in psychology and neuroscience point to the potential of caricatures for improved human face recognition in non-ideal settings and as a tool for understanding how humans perceive faces. Unfortunately, existing research in the area of caricature face verification is limited, due to lack of a well-constructed, large-scale dataset. In this work, we show that the largest existing dataset, WebCaricature, and its associated evaluation protocols do not meet acceptable standards of quality for face verification. We present a new caricature dataset, CarVer, and introduce a new face verification evaluation standard.**

This work builds on past work by expanding and improving the largest caricature face verification dataset, providing new evaluation standards, and introducing an end-to-end deep learning pipeline for the problem of caricature face verification.

***Index Terms*—Webscrape, Dataset Collection, Face Verification, Prominent Feature Recognition**

I. Introduction

Caricatures are artistic renderings of a human face that distort prominent features while still maintaining their resemblance to the original, veridical face; veridical is defined as the ground truth face [5]. Humans are capable of verifying if a caricature and veridical image belong to the same person implying that automated face verification using a combination of veridical images with caricatures is possible. The task is complicated by artists choosing to distort different features, or to distort them in different ways, as shown in Figure

Despite the wide variation in representation, a good caricature is still recognizable as the original subject. Knowing how faces are interpreted and understood by humans is particularly valuable in automated face recognition research. This has led to the study of caricatures across fields, especially in psychology and neuroscience [5], [9], [17], [22]. Currently, the study of caricatures in computer science is relatively limited. Most works try to generate caricatures from veridical images [10], [14], [24], [27] with varying levels of success. Others attempt to perform verification with caricatures, but are extremely limited in their implementation by either not using deep learning [12] or using a convoluted evaluation protocol [6], [8].

Progress in this area is further hindered by a lack of datasets of suitable size [1], [6], [12], [18], or a lack of images of acceptable quality [4], [8]. The largest of these datasets, WebCaricature, attempts to set the standard for caricature verification [8]. However, we show that the images in this dataset are often of unacceptable quality, either by not being a caricature, or being unrecognizable as the target identity. We also show that the evaluations used by WebCaricature [8] inflate their metrics, result in significant loss of image information, and in some cases, are overly complicated and could be replaced by the use of a traditional convolutional neural network (CNN) feature extractor, such as VGG-16 or ResNet-50.

Our main contributions are the following:

- A new set of end-to-end CNN baseline methods for caricature face verification.
- A new dataset, CarVer, comprised of 229 identities, 1638 caricatures, and 3148 veridical images, with *at least* 5 caricatured images per identity.
- Detailed analysis of the methods employed by the most

This material is based upon work supported by the National Science Foundation under Grants No. 1909707 and 2302187. Standard disclaimers apply.

Fig. 1. Examples of variation in caricature representation of the same person. The left most image is the photo (veridical face) and subsequent images are caricatures. All images are taken from our CarVer dataset. Note that while there is wide variation in representation of the veridical image, the identity of each of the caricatures is quite obvious.

well-known work on caricature verification, WebCaricature, as well as the dataset itself [8].

- A cleaned version of WebCaricature, WebCaricature-Clean, which we combine with CarVer to create CarVer-WebCaricature, which consists of 359 distinct identities, $5,436$ caricatures, and $9,025$ veridical images.

II. Related Works

This work focuses primarily on trying to improve the caricature/veridical face recognition introduced by [8]. We review other relevant research in caricature/veridical face verification.

The use of caricatures to perform face verification and identification has been an active area of research in psychology and neuroscience for decades [13], [16], [20]. [20] found that participants identified faces faster using simple line drawings of caricatured faces as compared to veridical faces. [16] found that caricatures were accurately identified more often and faster than veridical images, and caricatures of familiar faces were recognized with the best accuracy. [17] found that caricatures of unfamiliar faces also improved verification rates by approximately 30%. [5] found that using caricatures led to better recognition of unfamiliar faces across the entire human lifespan, that it improved low-resolution face verification in older adults, and that verification of faces of other races also improved. [9] found that faces are recognized better when they are first learned as a caricature and then shown as a veridical image, rather than vice versa. [17] found that above a certain rate of exaggeration, caricature verification is actually hindered– in other words, caricatures need to have a reasonable resemblance to the original face.

Though psychology research has shown that the use of caricatures improves human verification and recognition, work in computer science using caricatures is rather limited. Recent work generates a caricature from a photo [10], [14], [24], [27], but does not try to understand or utilize caricatures to improve verification or recognition. [12] attempts to exploit caricatures to improve verification. However, the dataset is small, and does not use modern deep learning methods. [1] introduces a method to match caricatures to veridical images by extracting facial attribute features from photos, but requires manual labeling of facial attribute features on caricatures, which is time consuming. Furthermore, they compute feature importance using genetic algorithms, which are extremely slow compared to deep learning.

The most comprehensive, recent work in automated caricature verification is WebCaricature [8], which provides an end-to-end framework for face verification and identification using caricatures. Their framework first detects the face and landmarks, then crops the image using several proposed cropping methods. Facial features are extracted using SIFT and VGG. PCA is applied on the features without reducing dimensionality. They show that SIFT extraction does not perform as well as using a VGG-16 CNN. Traditional face matching methods, such as Euclidean distance, are then employed. This ultimately creates a disjointed framework for face verification that requires multiple models. Furthermore, the authors only perform verification on caricature/veridical pairs, and do not introduce caricature/caricature or veridical/veridical pairs to the framework, which we show severely impacts the framework's efficacy. [8] also provides the largest publicly available dataset for caricature verification – WebCaricature. Through our analysis, we highlight several problems with both the dataset and approach provided in [8] as detailed in the following sections.

III. Dataset Collection

A. Collection and Labeling of CarVer

We introduce a new dataset for Caricature Verification, which we call CarVer. We compile a list of names of public figures. Dataset imbalance can lead to models learning unintentional bias [11], [23], [26], so we try to collect identities from a diverse sample of ethnicities and as close to a $50/50$ gender balance as possible. After a cursory Google image search, we discard any public figure from the list for which at least 5 *representative* caricature images cannot be found, since any fewer would lead to disproportionately learning identities with more images to learn from. A representative caricature image is one that is a) exaggerated in various ways while b) still recognizable as the target identity and c) not under exaggerated in such a way that it would be considered a drawing or cartoon. 1. After confirming that there are enough available caricature images of each public figure, we use a web scraper to pull as many veridical images and caricatures of the 229 identities as possible. We then manually review each image and confirm that if it is a caricature, it meets our definition of a representative caricature, and if it is a veridical image, it is not blurry and that it is of the target identity. We also confirm that there are no duplicate caricatures or veridical images in the dataset. This results in CarVer having $1,638$ caricatures and $3,148$ veridical images, across 229 public figure identities. The number of caricatures for each identity ranges from 5 to 23 and the number of veridical images for each identity ranges from 5 to 27

All images in CarVer are labeled with 68 landmarks using Face-Alignment [2], which performs well on most veridical images and a small selection of caricatures. We manually review the landmarks and adjust any as needed.

B. Evaluation Against WebCaricature

WebCaricature [8] is the most recent, comprehensive publicly available caricature dataset. We compare it directly to our CarVer in order to make our contributions clear. Though WebCaricature consists of $6,042$ caricatures and $5,974$ veridical images over 252 identities, we find that there are many quality issues with the dataset itself. The top row of Figure 2 provides examples of caricatures from WebCaricature whose identities are not immediately clear. This indicates that they are not representative caricatures and should not be included in the dataset. The bottom row of Figure 2 shows images from WebCaricature that are not a caricature, but rather a drawing, cartoon, or veridical image incorrectly labeled as a caricature. Additionally, we find that there are many instances where the

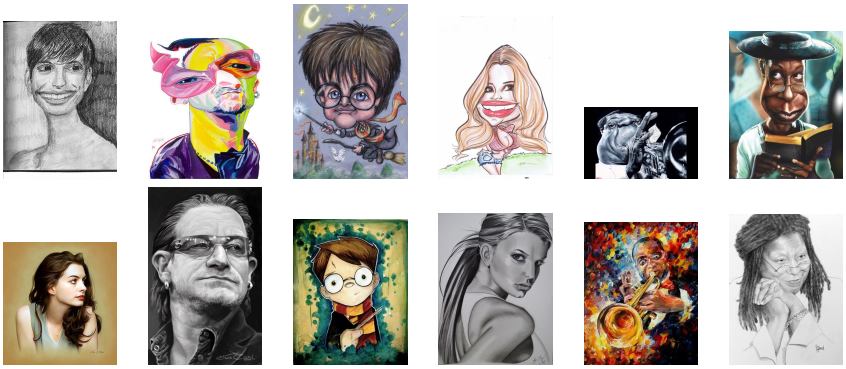

Fig. 2. Images from WebCaricature that are not of acceptable quality to be included in a computer vision dataset. The first row contains images where the identity is not immediately obvious without knowing who the person is. The second row contains images where the image is not a caricature, but rather a painting, drawing, or cartoon.

Data Set	Male %	Female %
WebCaricature	71.83	28.17
CarVer	57.64	42.36

TABLE I
PERCENT OF EACH GENDER IN CARVER AND WEBCARICATURE.

dataset contains duplicate images, or images that are not of the target identity.

We label the CarVer and WebCaricature datasets with gender and race using wikipedia reported race and gender. Imbalanced datasets can lead to models learning unintentional bias. The gender and race compositions of the WebCaricature and CarVer datasets can be found in Table I and Table II, respectively. Not only does CarVer improve gender diversity by increasing female representation from WebCaricature's 28.17% to 42.36%, but we also find that our dataset increases racial representation. Black, Hispanic, and South Asian representation are all drastically increased in our dataset, though Caucasian representation is still high.

Direct comparison of WebCaricature to CarVer reveals that there are 73 overlapping identities between the datasets. Within these identities, there are 223 overlapping images. This indicates that our dataset could extend WebCaricature by an additional 155 identities and $4,563$ images. Furthermore, WebCaricature only ensures that there is one caricature per identity, while we ensure that there are at least five representative caricatures for each identity. We also find that of all of the veridical images in WebCaricature, 13.11% are grayscale, while only 2.09% in CarVer are grayscale. [8] uses a VGG-16 network to extract features, after pretraining on VGG-Face. Importantly, we note that VGG-Face contains 25 overlapping identities with WebCaricature. **This means that the network has already seen approximately 10% of the dataset identities prior to testing, and that the segregation of train and test data is broken by their protocols; this led to higher reported performance than what would have been achieved if the train and test identities had been properly separated.**

Given the aforementioned problems with WebCaricature, we clean the dataset through manual review. We remove any caricature from the dataset that is not representative of the target identity or is not a caricature, and we remove any image that is a duplicate, or is not of the target identity. Overall, we remove $2,341$ images, 97 of which are veridical images, and $2,244$ of which are caricatures. This leaves WebCaricature with $3,798$ caricatures and $5,877$ veridical images. Over 95% of the veridical images removed are duplicates, while the remaining veridical images are removed because they are either not veridical images or not of the target identity. Of the caricatures removed, approximately 61% are removed for not being a caricature, and 37% are removed for not being representative of the target identity. Less than 2% are removed for being a duplicate or not of the target identity. Of the original 252 identities in WebCaricature, 226 have images removed. Our analysis revealed that 49 identities should not be included in WebCaricature due to them having fewer than 5 caricatures per identity. After cleaning, WebCaricature should consist of only 203 identities, not the reported 252. Our cleaning process revealed that 1% of all veridical images, 37% of all caricatures and 19% of all identities in WebCaricature were not up to acceptable standards. This modified version of WebCaricature is called "WebCaricature-Clean", which we combine with our dataset, CarVer, to create a composite dataset that is referred to as "CarVer-WebCaricature" in later sections.

Data Set	Caucasian %	Asian %	South Asian %	Black %	Pacific Islander %	Hispanic %
Web Caricature	75.40	11.90	0.50	9.92	0.00	2.38
CarVer	54.15	7.86	4.80	18.34	1.75	13.10

TABLE II
PERCENT OF EACH RACE IN CARVER AND WEBCARICATURE.

]

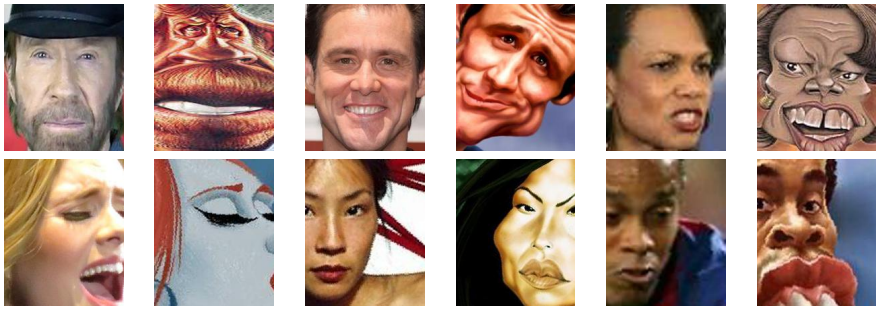

Fig. 3. Images where the alignment method cropped out useful information for caricature and veridical image alignment. The top row contains Bounding Box (BB) Based alignment images, while the bottom row contains Eye Location (EL) Based alignment images [8].

IV. Evaluation Protocols

A. Face Alignment

[8] implements three unique alignment techniques: Bounding Box (BB), Eye Location (EL), and Landmark (L) Based. Our work is primarily focused on trying to apply an end-to-end deep learning framework to caricature recognition. The authors of [8] state that one of their alignment methods, Landmark Based, is not suitable for feature extraction used by deep learning, and therefore it is not applicable to our research and is not implemented. BB alignment enlarges the face bounding box by 1.2 to crop the images. EL alignment identifies eye location, resizes the image to make eye distance 75 pixels apart, and then creates a crop box that is 80 pixels left and right of the center between the eyes and 70 pixels above and below the center between the eyes. Both BB and EL alignment calculate the angle between eyes, and use it to rotate the image until the eyes are horizontal. The authors of [8] find that BB alignment performs better than EL alignment. Unfortunately, we find that using either method results in many of the caricatures and veridical images being cropped in such a way that the hair, ears, and jawline are missing, examples of which can be seen in Figure 3.

While traditional face verification crops close to the chin, forehead, and sides of the face, we note that caricatures are exaggerated faces, where features such as ears, chin, lips, and hair are often enlarged. Cropping them out misses this essential information. For example, Condoleezza Rice's signature look includes earrings and her lips often extend past her chin in caricatures, so cropping her ears and lower face area out results in loss of that information. We also note that EL alignment results in some features that are necessary in traditional veridical image recognition being cropped out, which can also be seen in Figure 3. Side facing images usually result in a large amount of the farthest eye being cropped out, as well as the lips and chin, as seen in all of the example images in the bottom row of figure 3.

We implement two additional alignment methods, Enlarged Bounding Box (EBB) and Revised Bounding Box (RBB) Based, in an effort to find more effective cropping methods for caricatures. EBB aligns faces using the angle between the eyes, like the BB and EL alignment methods. EBB enlarges the bounding box by 1.5 rather than 1.2 in BB. RBB enlarges the face bounding box of veridical images by 1.3, and the entire caricature image is used in order to evaluate the contribution of environmental effects on caricature verification. A side by side comparison of the four alignment methods is shown in Figure 4.

Fig. 4. Examples of various crop types using CarVer dataset images. The top row contains veridical images, while the bottom row contains caricatures. From left to right: BB, EL, EBB and RBB.

B. CarVer Evaluation Protocol

The goal of face verification is to determine if two images are of the same person. The dataset is partitioned into 10 folds with unique identities in each fold. This is later used in the verification framework to perform 10-fold cross validation. [8] only compares veridical to caricatured images. We call these pairs "Mixed Match Only" (MMO). In an effort to make our evaluation protocols as relevant to traditional face verification as possible, we construct pairs that compare two veridical images, two caricatures, and a caricature to a veridical image. We call these pairs "All Pairings" (ALL). In later sections we compare the performance of MMO and ALL pairings, and show that ALL pairings increase performance without being unduly influenced by the caricature/caricature and veridical/veridical pairings.

Additionally, we implement the restricted and unrestricted settings introduced in [7], [8]. The restricted setting only labels pairings as match or not match and only pre-determined pairs can be considered, while the unrestricted setting labels only identity so that as many pairs as possible can be generated for training. To provide benchmark consistency, we utilize the same unrestricted pairings across folds, which we will release with the code for this work. For WebCaricature, we use the provided [8] restricted and unrestricted settings for the MMO pairings. For ALL pairings using WebCaricature, we utilize the same identities provided by [8] for restricted and unrestricted in each fold, and randomly generate additional veridical/veridical and caricature/caricature pairs using the available images for each identity. For the WebCaricature ALL pairings, we generate an equal number of veridical/veridical, caricature/caricature, and caricature/veridical pairings, with as many as possible using the smallest set of pairing types to govern how many matches are generated. We ensure a balance between matches and non-matches by selecting one non-matching pair for every matching pair in the WebCaricature ALL setting. The same process is repeated for the generation of the WebCaricature-Clean ALL and MMO pairings.

To generate the CarVer ALL pairings set, we use the same process as WebCaricature ALL pairings. The process to create CarVer MMO pairings is a combination of the WebCaricature MMO pair method and the CarVer ALL pair method. Like the provided WebCaricature MMO pairing sets, we select $\frac{1}{3}$ of all possible matches, and for each match, we generate a non-match. This, again, ensures a balanced dataset with an equal number of matches to non-matches, and with proportionally the same sampling of matches as WebCaricature MMO. For CarVer-Webcaricature, we generate the combined pairings by concatenating and shuffling the WebCaricature-Clean and CarVer pair lists.

V. Verification Framework

A. Caricature Verification Settings

The goal of CarVer's verification framework is to perform face verification on veridical/veridical, caricature/caricature, and veridical/caricature pairings using well documented, end-to-end deep learning methods. This is in direct contrast to WebCaricature [8], which uses several disjointed methods to construct their framework. In essence, our goal is to improve upon the caricature dataset and verification framework standards set forth by [8]. We perform verification using the proposed framework on WebCaricature so that it can be compared to performance on WebCaricature-Clean, CarVer, and CarVer-WebCaricature.

Evaluation of the proposed verification framework is performed on WebCaricature-Clean, CarVer, and CarVer-WebCaricature using all four alignment types – Bounding Box (BB), Eye Location (EL), Enlarged Bounding Box (EBB), and Revised Bounding Box (RBB). The original WebCaricature is only evaluated using BB and EL alignments, so that our results can be directly compared to those presented in [8]. We evaluate each dataset/alignment combination in restricted and unrestricted settings and with MMO and ALL pairings. With these various settings, 56 unique settings are evaluated, each using 10-fold cross-validation.

B. Caricature Verification Pipeline

We utilize the standard VGG-16 architecture [25] pre-trained on ImageNet as the starting point for our verification framework. The final fully-connected layer is adjusted to have 64 outputs instead of 4096 to reduce the number of parameters in an effort to avoid overfitting. We fine-tune this model on the CelebA dataset [15] using face verification rather than attribute prediction as the target task. All overlapping identities between CelebA, WebCaricature and CarVer are removed before pre-training. This is particularly important, because the results reported by [8] are artificially inflated by pre-training on VGG-Face which includes overlapping identities with WebCaricature. This pre-training utilizes a Cosine Embedding Loss [21], Adam optimizer, and a learning rate of 0.0001 over 10 epochs. We perform the pre-training five times, and the model with the median F1 performance on the CelebA data is used in the next step.

The final step involves fine-tuning the above model in the 56 unique settings detailed in the previous subsection. For each of the 56 settings, the model is fine-tuned for verification using 10-fold cross validation. WebCaricature provides an $0th$ fold, which is used for hyperparameter tuning. Cosine Embedding Loss and Adam optimizer are once again used in this fine-tuning and the learning rate is dropped to 0.000001.

VI. Results and Discussion

We evaluate our end-to-end verification framework using F1, F1 median, Area Under the Curve (AUC), and Verification Rate at .1 (VR.1) and at .01 (VR.01) so that results can be directly compared to [8]. We find that the trends in the F1 scores, AUC, and VR are similar. Thus we present only the F1 score in the paper, providing analysis where the F1 trend is not consistent with the AUC or VR. We supply a full F1, F1 median, AUC, and VR table in our GitHub repo.

Because we run multiple settings, alignments, and restrictions with each dataset, and use 10-fold cross validation, we provide the average of each statistic. Note that these metrics

are achieved with a VGG-16 CNN [25] after pre-training on CelebA for verification with all overlapping identities removed. There are $12,016$ images in WebCaricature, $9,675$ images in WebCaricature-Clean, $4,786$ images in CarVer, and $14,238$ images in CarVer-WebCaricature. We sample proportionally from CarVer and WebCaricature to form the combined set, so we anticipate that its metric performance should be better than WebCaricature-Clean, but worse than CarVer because there are proportionally more WebCaricature images. The best scores in each table are shown in **bold**. We present our results in the following subsections.

A. Dataset Variation

Dataset	F1
WebCaricature [8]	0.75± 0.05
WebCaricature-Clean	0.77± 0.04
CarVer	**0.82± 0.05**
CarVer-WebCaricature	0.78± 0.04

TABLE III
THE F1 FOR WEBCARICATURE, WEBCARICATURE-CLEAN, CARVER, AND CARVER-WEBCARICATURE. ALL VALUES ARE AVERAGED ACROSS FOLDS.

We utilize four datasets for our tests: WebCaricature, WebCaricature-Clean, CarVer, and CarVer-WebCaricature. The average F1 performance for each dataset are shown in Table III. The CarVer dataset outperforms the WebCaricature, WebCaricature-Clean, and CarVer-Webcaricature in all metrics. Furthermore, the F1 standard deviation is similar for all datasets, indicating that while the CarVer dataset has improved performance over others, the improvement is not unduly influenced by a particularly well-performing fold or that a small subset of images inflates the improved metric.

F1 on WebCaricature-Clean is higher than WebCaricature. However, cleaning WebCaricature results in an AUC decrease from 0.79 to 0.77, and verification rate performance decrease at .1 (0.44 to 0.40) and .01 (0.10 to 0.09). We note that the improvement in F1 and decrease in standard deviation of WebCaricature-Clean indicates that the dataset has had outliers removed from each fold that unduly influence the AUC and verification rate performance before cleaning. F1 decreases from CarVer to CarVer-WebCaricature from 0.82 to 0.78, but the standard deviation is decreased. This is likely due to the larger size of the dataset and increased influence of WebCaricature-Clean as it is larger than CarVer.

B. Alignment Variation

We utilize four alignment methods for our tests: Bounding Box (BB), Eye Location (EL), Enlarged Bounding Box (EBB), and Revised Bounding Box (RBB). The F1 for each alignment method averaged across datasets are shown in Table IV. The best performance across all metrics is achieved by our EBB alignment method, where the F1 is 0.82. The next best performance across all metrics is our RBB performance. The worst performance across all metrics is obtained using the Eye Location alignment from [8]. This indicates that not only is the eye location alignment not appropriate for caricature verification, but the size of the bounding box surrounding the faces in both veridical images and caricatures is important. Too much information, as in the RBB method, results in performance decrease. Too little information, due to poor early crops, as seen in Figure 3, results in even worse performance. Simply expanding the bounding box crop size so that more of the caricature face and veridical image face are included results in the best performance.

Alignment Method	F1
Bounding Box [8]	0.79± 0.04
Eye Location [8]	0.74± 0.04
Enlarged Bounding Box	**0.82± 0.04**
Revised Bounding Box	0.80± 0.04

TABLE IV
THE F1 FOR BOUNDING BOX, EYE LOCATION, ENLARGED BOUNDING BOX, AND REVISED BOUNDING BOX ALIGNMENT. ALL VALUES ARE AVERAGED ACROSS FOLDS AND DATASETS.

Dataset and Alignment	F1
WebCaricature [8] BB [8]	0.78± 0.05
WebCaricature-Clean BB [8]	0.78± 0.04
CarVer BB [8]	**0.82± 0.03**
CarVer-WebCaricature BB [8]	0.79± 0.03
WebCaricature [8] EL [8] [8]	**0.76± 0.04**
WebCaricature-Clean EL [8]	0.72± 0.02
CarVer EL [8]	**0.76± 0.04**
CarVer-WebCaricature EL [8]	0.74± 0.03
WebCaricature-Clean EBB	0.79± 0.03
CarVer EBB	**0.86± 0.03**
CarVer-WebCaricature EBB	0.81± 0.03
WebCaricature-Clean RBB	0.78± 0.03
CarVer RBB	**0.86± 0.01**
CarVer-WebCaricature RBB	0.79± 0.02

TABLE V
THE F1 FOR EACH POSSIBLE ALIGNMENT TYPE: BOUNDING BOX (BB), EYE LOCATION (EL), ENLARGED BOUNDING BOX (EBB), AND REVISED BOUNDING BOX (RBB)) BY DATASET (WEBCARICATURE, WEBCARICATURE-CLEAN, CARVER, AND CARVER-WEBCARICATURE). ALL VALUES ARE AVERAGED ACROSS FOLDS.

To further analyze the contribution of the new dataset to each type of alignment, we calculate the F1 for each alignment type by dataset, shown in Table V. We find that for BB, EBB, and RBB, CarVer outperforms the other three datasets, which is consistent with the overall results shown in Table III. For EL alignment, WebCaricature slightly outperforms CarVer by having better verification rates (0.40 to 0.39 and 0.10 to 0.06). However, as previously mentioned, EL alignment has the worst overall performance, and WebCaricature's slight improvement over CarVer for this type of alignment is likely due to wide variation in features shown in each crop.

C. Setting Variation

We utilize two settings for our verification experiments: Restricted [7], [8] and Unrestricted [7], [8]. As shown in Table VI, the best performance is achieved with the Restricted setting. Though the identities are not altered across folds for restricted and unrestricted settings, there are more images used in the unrestricted setting for each dataset because unrestricted maximizes the number of pairs possible while maintaining

Restriction Setting	F1
Restricted [7], [8]	**0.80± 0.05**
Unrestricted [7], [8]	0.76± 0.04

TABLE VI
THE F1 FOR RESTRICTED AND UNRESTRICTED SETTINGS. ALL VALUES ARE AVERAGED ACROSS FOLDS.

Setting Parameter Combination	F1
WebCaricature Restricted [7], [8]	0.78± 0.05
WebCaricature-Clean Restricted [7], [8]	0.78± 0.04
CarVer Restricted [7], [8]	**0.82± 0.06**
CarVer-WebCaricature Restricted [7], [8]	0.79± 0.04
WebCaricature Unrestricted [7], [8]	0.75± 0.03
WebCaricature-Clean Unrestricted [7], [8]	0.75± 0.04
CarVer Unrestricted [7], [8]	**0.80± 0.03**
CarVer-WebCaricature Unrestricted [7], [8]	0.77± 0.02

TABLE VII
THE F1 FOR RESTRICTED AND UNRESTRICTED SETTINGS PER DATA SET TYPE. ALL VALUES ARE AVERAGED ACROSS FOLDS.

balance, whereas the restricted setting takes a fixed percentage of all possible pairings while maintaining balance. The smaller training set is easier to learn from because there are fewer images per identity to learn from. The random Restricted sampling unintentionally resulted in sampling photos and caricatures that were more consistent in pose, lighting, etc than the Unrestricted sampling. This is why the Restricted outperforms the Unrestricted setting, but the standard deviation for the restricted setting is slightly higher.

To further analyze the contribution of CarVer to each type of alignment, we calculate the F1 of each alignment type by dataset, shown in Table VII. We find that for both Restricted and Unrestricted settings, CarVer outperforms the other three datasets, which is consistent with the overall results shown in Table III. There are higher levels of standard deviation throughout the restricted and unrestricted evaluation shown in Table VI and VII than compared to dataset variation, or alignment variation. This is to be expected given the small size of the dataset.

D. Pairing Variation

Pairing Type	F1
MMO [7], [8]	0.75± 0.04
ALL	**0.80± 0.04**

TABLE VIII
THE F1 FOR MMO AND ALL PAIRINGS. ALL VALUES ARE AVERAGED ACROSS FOLDS, RUNS, AND DATASETS.

We utilize two types of pairings: mixed match only (MMO) and all-type pairings (ALL). MMO pairings only use pairs that consist of a caricature and a veridical image. ALL pairings use pairings that are a balanced mixture of caricature and caricature, veridical and veridical, and veridical and caricature. Table VIII shows that ALL pairings outperforms MMO pairings. This is despite the fact that the ALL pairings datasets are much larger because they require three matches for every no match pair, while MMO only requires a single match pair for every no-match pair (veridical/caricature). This indicates that using

Pairing Parameter Combination	F1
WebCaricature MMO [7], [8]	0.73± 0.03
WebCaricature-Clean MMO [7], [8]	0.74± 0.04
CarVer MMO	**0.78± 0.06**
CarVer-WebCaricature MMO	0.75± 0.03
WebCaricature ALL	0.80± 0.03
WebCaricature-Clean ALL	0.78± 0.04
CarVer ALL	**0.84± 0.06**
Carver-WebCaricature ALL	0.79± 0.03

TABLE IX
THE F1 FOR MIXED MATCH ONLY (MMO) AND ALL-TYPE PAIRING PER DATA SET TYPE. ALL VALUES ARE AVERAGED ACROSS FOLDS.

all possible pair types improves performance, but also means that the influence of each pair type on metric performance must be analyzed to ensure that the addition of veridical/veridical or caricature/caricature pairs do not unduly increase performance.

We find that in all possible combinations of dataset, image alignment type, restriction setting, and pair type, the false positive rate, false negative rate, true positive rate, and true negative rate for caricature/veridical pairs is similar to those of veridical/veridical and caricature/caricature pairs. In other words, we do not find that caricature/veridical pairings have a higher rate of being mis-verified than the caricature/caricature and veridical/veridical pairs. Details are provided as a full table in our GitHub repo. These results indicate that the extension of pair types to include caricature/caricature and veridical/veridical pairs increases performance, despite the increase in dataset size.

To further analyze the contribution of the new dataset to each type of pairing, we calculate the F1 of each pairing type by dataset, shown in Table IX. We find that for both MMO and ALL pairing types, CarVer outperforms the other three datasets, which is consistent with the overall results shown in Table III. There are higher levels of standard deviation throughout the pairing results in Table VIII and IX than compared to dataset variation, or alignment variation. This aligns with our results from restricted/unrestricted settings again due to the small size of the dataset.

VII. CONCLUSION

In this work, we introduce a new standard for the problem of caricature face verification, providing a new dataset – CarVer – new evaluation protocols, new alignment methods and an end-to-end verification pipeline. We improve over the previous benchmark – WebCaricature [8] – the only available method and dataset for comparison. Performance on WebCaricature is increased by our cleaning process, as evidenced by metric improvement across the board. We also find that the WebCaricature verification pipeline does not appropriately pre-train the network, causing artificial metric inflation as reported in [8]. The best performing dataset is our CarVer with standard deviation similar to other datasets. This indicates that our dataset collection methods and cleaning methods for existing datasets are an improvement over [8].

We find that image alignment and dataset have a large impact on performance. Our Enlarged Bounding Box (EBB)

alignment has the best performance compared to the other three methods, and Eye Location (EL) alignment has the worst performance as it removes necessary facial information from images. If we consider alignment methods by data set type, CarVer has the best performance in each parameter combination. This indicates that there is a strong correlation between dataset type and performance. Finally, we find that F1 scores of alignment types are significantly varied between alignment types, which indicates that the alignment has a significant impact on performance.

We show that our work improves upon [8], we create a larger clean dataset, introduce an improved face verification pipeline that is cohesive and can be used to set face verification standards, and introduce improved alignment methods. Future work should consider improving the face identification standards introduced by [8], exploring additional alignment that does not use eye angle to rotate the image, and should use the CarVer-WebCaricature dataset to begin exploring prominent feature prediction.

Considerations, Limitations, and Societal Impact: When comparing our results to [8], pre-training must be discussed. As previously mentioned, our pre-training removed all overlapping identities, but [8] did not. This artificially inflates their metrics. Furthermore, we simplify the verification process and provide an end-to-end caricature verification pipeline that is capable of comparing caricatures to caricatures, veridical images to veridical images, and veridical images to caricatures. This makes the system more robust and capable of being applied to face verification in non-ideal settings. However, our work is limited by a cursory understanding of the way in which each altered parameter affects the overall performance of the system. Furthermore, we rely heavily on F1 to determine the best set of weights in each parameter pairing, and F1 may not always be the best metric. We provide the F1 Median, AUC, and verification rates in our GitHub repo. As with any face-verification system, there are ethical concerns for our research use. Other research has been used to unduly target minority groups [19]. Conversely, many face-verification systems do not take into account racial bias inherent to small, undiversified data sets [3]. We attempt to alleviate this problem by improving race and gender balance in our dataset.

References

[1] Bahri Abacı and Tayfun Akgül. Matching caricatures to photographs. *Signal Image and Video Processing*, 9:1–9, 12 2015. 1, 2

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 2

[3] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020. 8

[4] Elliot J. Crowley, O. Parkhi, and Andrew Zisserman. Face painting: querying art with photos. In *BMVC*, 2015. 1

[5] Amy Dawel, Tsz Ying Wong, Jodie McMorrow, Callin Ivanovici, Xuming He, Nick Barnes, Jessica Irons, Tamara Gradden, Rachel Robbins, Stephanie C Goodhew, Jo Lane, and Elinor McKone. Caricaturing as a general method to improve poor face recognition: Evidence from low-resolution images, other-race faces, and older adults. *Journal of experimental psychology. Applied*, 25(2):256–279, 2019. 1, 2

[6] Jatin Garg, Skand Vishwanath Peri, Himanshu Tolani, and Narayanan C. Krishnan. Deep cross modal learning for caricature verification and identification (cavinet). In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 1101–1109, New York, NY, USA, 2018. Association for Computing Machinery. 1

[7] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 5, 6, 7

[8] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. Webcaricature: a benchmark for caricature recognition. In *British Machine Vision Conference*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[9] Marlena L Itz, Stefan R Schweinberger, and Jürgen M Kaufmann. Caricature generalization benefits for faces learned with enhanced idiosyncratic shape or texture. *Cognitive, affective, & behavioral neuroscience*, 17(1):185–197, 2017. 1, 2

[10] Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. Stylecarigan: Caricature generation via stylegan feature map modulation. *arXiv preprint arXiv:2107.04331*, 2021. 1, 2

[11] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI*, pages 111–117, 2000. 2

[12] Brendan F. Klare, Serhat S. Bucak, Anil K. Jain, and Tayfun Akgul. Towards automated caricature recognition. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 139–146, 2012. 1, 2

[13] Michael B Lewis. Are caricatures special? evidence of peak shift in face recognition. *European Journal of Cognitive Psychology*, 11(1):105–117, 1999. 2

[14] Pei-Ying Chiang Wen-Hung Liao and Tsai-Yen Li. Automatic caricature generation by analyzing facial features. In *Proceeding of 2004 Asia Conference on Computer Vision (ACCV2004), Korea*, volume 2. Citeseer, 2004. 1, 2

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[16] Robert Mauro and Michael Kubovy. Caricature and face recognition. *Memory & Cognition*, 20(4):433–440, 1992. 2

[17] Alex H. McIntyre, Peter J. B. Hancock, Josef Kittler, and Stephen R. H. Langton. Improving discrimination and face matching with caricature. *Applied Cognitive Psychology*, 27(6):725 – 734, 2013. 1, 2

[18] Mishra A. Mishra A., Nandan Rai S. and Jawahar C. V. Iiit-cfw: A benchmark database of cartoon faces in the wild. In *"VASE ECCVW*, 2016. 1

[19] Paul Mozur. One month, 500,000 face scans: How china is using a.i. to profile a minority. *New York Times*, Apr 2019. 8

[20] Gillian Rhodes, Susan Brennan, and Susan Carey. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive psychology*, 19(4):473–497, 1987. 2

[21] Nivedita Rufus, Unni Krishnan R Nair, K. Madhava Krishna, and Vineet Gandhi. Cosine meets softmax: A tough-to-beat baseline for visual grounding, 2020. 5

[22] Claudia Schulz, Jürgen M. Kaufmann, Lydia Walther, and Stefan R. Schweinberger. Effects of anticaricaturing vs. caricaturing and their neural correlates elucidate a role of shape for face learning. *Neuropsychologia*, 50(10):2426–2434, 2012. 1

[23] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Mining data with rare events: A case study. In *19th IEEE*

International Conference on Tools with Artificial Intelligence(ICTAI 2007), volume 2, pages 132–139, 2007. 2

[24] Yichun Shi, Debayan Deb, and Anil K Jain. Warpgan: Automatic caricature generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10762–10771, 2019. 1, 2

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 5, 6

[26] Byron C Wallace and Issa J Dahabreh. Class probability estimates are unreliable for imbalanced data (and how to fix them). In *2012 IEEE 12th international conference on data mining*, pages 695–704. IEEE, 2012. 2

[27] Zipeng Ye, Ran Yi, Minjing Yu, Juyong Zhang, Yu-Kun Lai, and Yong-jin Liu. 3d-carigan: An end-to-end solution to 3d caricature generation from face photos. *arXiv preprint arXiv:2003.06841*, 2020. 1, 2