Investigating CLIP Performance for Meta-data Generation in AD Datasets

Sujan Sai Gannamaneni[1], Arwin Sadaghiani[1], Rohil Prakash Rao[2], Michael Mock[1], Maram Akila[1]
[1]Fraunhofer IAIS, [2]University of Bonn
{sujan.sai.gannamaneni,arwin.sadaghiani,michael.mock,maram.akila}@iais.fraunhofer.de,
s6roraoo@uni-bonn.de

Abstract

Using Machine Learning (ML) models for safety-critical perception tasks in Autonomous Driving (AD) or other domains requires a thorough evaluation of the model performance and the data coverage w.r.t. the intended Operational Design Domain (ODD). However, obtaining the needed per-image semantic meta-data along the relevant dimensions of the ODD for real-world image datasets is non-trivial. Recent advances in self-supervised foundation models, specifically CLIP, suggest that such meta-data could be obtained for real-world images in an automated fashion using zero-shot classification. While CLIP was already reported to achieve promising performance on tasks such as the recognition of gender or age on facial images, we investigate to which extent less prominent and more fine-grained observables, e.g., presence of accessories such as spectacles or the shirt- or hair-color, can be determined. We provide an analysis of CLIP for generating fine-grained meta-data on three datasets from the AD domain, one of synthetic origin including ground truth, the others being Cityscapes and Railsem19. We also compare with a standard facial dataset where more elaborate attribute annotations are present. To improve the quality of generated meta-data, we additionally extend the ensemble approach of CLIP by a simple noise-suppressing technique.

1. Introduction

Rigorous evaluation of safety-critical autonomous systems is an important step towards building trust in their capabilities and limitations. Therefore, there is currently a strong focus on research into the identification of different failure modes of these safety-critical systems and methods to mitigate them [1, 7, 24, 29]. Identification of systematic weaknesses learnt by deep neural networks (DNNs) from training data is one such failure mode. Issues related to Fairness [58], where model bias w.r.t. age, gender, and ethnicity are extensively studied, can be considered as an expression of such systematic weaknesses. A typical example is the under-performance of models on dark-skinned people when used for person detection [9, 12]. However, models should also be evaluated for systematic weaknesses w.r.t. other human-understandable semantic attributes (*i.e.*, available meta-data).

Extending the Fairness example, person detection models might also display weaknesses w.r.t. accessories on the person like hats or sunglasses, angle or distance of the person to the camera, etc, which influence their visual appearance and likely their effective features. To identify such weaknesses, one would, however, require image-level semantic meta-data that contains the required granular information. To account for the lack of readily available semantic meta-data, several recent works have proposed different testing methods such as (i) identifying semantic clusters in the penultimate layers of DNNs being tested [14], (ii) identifying semantic clusters using embeddings from a cross-modal model representations [17], (iii) generating meta-data of objects using computer simulators and identifying systematic weaknesses along single semantic dimensions like distance, occlusion, etc using that meta-data [20,42,55]. While (i) and (ii) bypass the need for granular meta-data, they do have certain limitations, which we discuss in Sec. 2. Approaches based on synthetic data are useful for developing proofs-of-concept, but the safety augmentations built from these approaches might not be transferable to DNNs trained on real-world data due to domain gap to the used synthetic data. These issues highlight the importance of having granular meta-data for real-world image datasets to identify systematic weaknesses.

Also, from the AI Trustworthiness and certification perspective, recent specification standards [39] and expert groups [27] discuss the importance of considering data completeness or coverage. Research projects like KI-Absicherung[1] and several works [22, 33, 43] focusing on safety argumentations for DNNs used in Autonomous Driving (AD) have proposed defining operational design domains (ODDs). All these works additionally highlight the importance of having granular meta-data about objects in

[1]https://www.ki-absicherung-projekt.de/en/

images as extension of the currently available ground truth from a safety perspective. However, for most real-world datasets in AD, such fine grained meta-data is not available.

One way to potentially tackle this problem would be to formulate it as an image caption generation problem where the captions describe (in detail) the less prominent and fine-grained observables of the dominant object in an image, *e.g.*, accessories such as spectacles or shirt- or hair-color of a person. Instead of image captions of the form `"a photo of a person"`, we would like to generate captions of the form `"a photo of a` *`young black`* `man` *`standing in front of a shop`* `wearing a` *`hat`*`"`.[2] Through these captions, the necessary meta-data can be extracted to perform a systematic weakness analysis of a DNN used for, e.g., pedestrian detection. But, such granular caption generation is not trivial, as seen in the performance of image captioning approaches attempting dense captioning of images [31, 60].[3] However, with the latest advances in foundational models [6], where large DNNs containing billions of parameters trained on web-scale datasets containing millions of images show SOTA zero-shot performance on new domains and tasks, we have new tools to tackle this problem. In particular, CLIP [46], an important component of text-to-image models such as Stable Diffusion [50] and DALL-E [48] is a good candidate due to its rich latent space and impressive performance at zero-shot classification on benchmark datasets like Imagenet [13]. While the authors of CLIP have already reported its performance on classification of *gender* and *age* on facial images, we extend the evaluations to other datasets with focus on pedestrians and include captioning less prominent and fine-grained observables. Concretely, in this work, we evaluate if CLIP can indeed be used for this task by evaluating against standard datasets, some with existing metadata that we use as ground truth, and some datasets where we manually annotate a subset of the data.

2. Related work

The task of image captioning is the generation of a sequence of words that describe the content of an image meaningfully and in syntactically correct sentences [53]. It is an active field of research at the convergence of language description and image understanding, and several survey papers [28, 36, 53] have attempted to provide some structure to the large quantity of work. Broadly, methods prior to deep learning are based on description retrieval [45, 54] or template filling [18, 34] where captions are written by humans and then assigned to target images. These captions are, therefore, predefined and rigid. However, more recent deep learning-based approaches can generate novel captions. Typically, in these approaches, image content is first analyzed by a DNN, and subsequently, captions are generated by language models based on the image embeddings. With a focus on the more recent DNN-based approaches, Stefanini *et al.* [53] provides a taxonomy where the visual encoding models are split into two categories: (i) attention-based [35, 40, 56, 65, 67], and (ii) non attention-based [19, 49, 63], and the text encoding models are split into four categories: (i) LSTM-based [61, 62], (ii) CNN-based [2], (iii) Transformed-based [41, 59], and (iv) BERT-like [35, 67]. Based on the evaluation of different metrics for image captioning by Stefanini *et al.* [53], the top performing approaches on benchmark datasets are transformer-based methods like Unified VLP [67] and VinVL [65]. As mentioned, all these approaches generate one caption to describe an image, mostly related to the most prominent object in the scene. Dense captioning approaches [31, 60] which generate multiple captions per image are closer to our problem statement as they sometimes capture less prominent and more fine-grained observables. However, all these approaches do not use web-scale datasets, and their zero-shot capabilities are limited.

CLIP [46], on the other hand, has been trained on web-scale data and has remarkable performance on unseen datasets. We go into further detail about CLIP itself in Sec. 3. Several new CLIP extensions [4, 44, 52] have adapted CLIP to improve the generated captions and showed SOTA performance on benchmark datasets for image captioning [10] and Visual Question Answering (VQA) [21]. While these extensions adapt the CLIP architecture for VQA and plug in alternative text encoders for image captioning, we work with the established CLIP vision and text encoders that allows greater control over the semantic dimensions.

Approaches for finding systematic weaknesses in DNNs can be classified into two categories based on the type of data they are applied on, either structured or unstructured. For the former, *i.e.*, tabular data, approaches like SliceFinder [11], Sliceline [51], and sub-group discovery [3, 25] enumerate over various subset combinations and identify the top-k weakest subsets. By identifying these weaknesses, an actionable step is to collect more training data from the identified weak subsets and retrain the DNN. The main requirement for these approaches is the availability of semantic meta-data, which is easy to obtain for structured data and non-trivial for unstructured data (*e.g.*, real-world images). In this work, we show that some dimensions of semantic meta-data can be generated for the class *person* (or *pedestrian*) so that the above-mentioned approaches can

[2] This is only a representational text to highlight some of the interesting attributes. In our experiments, the prompts are defined so that one single attribute/meta-data is evaluated at one instance as discussed in Sec. 3.

[3] While not the same, dense captioning is closely related to our task. The main difference is that dense captioning focuses on captioning multiple objects and actions in an image while we focus on a single object and its attributes.

be applied.

For unstructured data like images, Domino [17] identifies subsets of data with weak performance by finding semantic clusters in the embedding space of the images generated using CLIP [46] while taking into account the images' classification performance. The identified clusters (or subsets) are then labeled in human-understandable form using a combination of a language model, *e.g.*, BERT [15], and CLIP. Spotlight [14], on the other hand, looks for semantic clusters using the activations of the penultimate layer of the DNN-under-test. The obtained clusters are manually evaluated and labeled by human experts. There are two problems with these approaches. One, both approaches perform the clustering on the representation space, while the methods working on structured data perform clustering on the high-level semantic space, making the latter approach more interpretable. Second, the final clusters, which are labeled manually or with a DNN, are assigned to a single dominant semantic description. For example, if all the images belonging to one cluster contain red shirts, an expert looking at the cluster might conclude that the systematic weakness is the presence of the red shirt. However, unlike the methods applied to structured data, a combination or impact of other factors is not considered as cause of weakness. Therefore, unlike the approaches for structured data, the contribution of methods like Domino and Spotlight to the safety argumentations is less strong. Approaches [20, 42, 55] used computer simulators like Carla [16] to generate meta-data of pedestrians in addition to the raw images and the default ground-truth. The DNN performance is then evaluated along individual semantic dimensions of the meta-data to identify weak spots. While the use of synthetic data is useful for developing proofs-of-concept, the safety argumentation for the DNNs used in safety-critical applications most probably needs to be made on real-world data. The results from these approaches might not be easily transferable due to the domain gap.

3. Probing ODDs with CLIP

In this section, we discuss, in further detail, about the use of ODDs in AD and their relation to our problem statement. Then we present the CLIP approach and explain the experiments already conducted in their paper and the difference to our experiments.

As motivated earlier, there is a lot of interest in the AD community to build safety augmentations by using operational design domains. Koopman and Fratrik [33] provide a list of dimensions along which the operational design domain can be structured and in which the AD vehicle should be validated. Zwicky boxes [5] was proposed as a way to develop operational design domains in the KI-Absicherung project [43], and Herrmann *et al*. [26] have used Zwicky boxes to develop ontologies for the perception function of AD. In Tab. 1, we provide a simplified ontology of the pedestrian class for the perception function which we intend to generate as meta-data with the help of CLIP. This subset ranges from dominant properties, such as clothing color, to highly fine-grained attributes, such as beard or eyeglasses. As it is, a priori, not clear how detection capabilities of a given DNN depend on such attributes,[4] we evaluate the captioning abilities of CLIP across this broad variety of attributes to open the possibility for future research.

Radford *et al*. [46] present CLIP as a pre-trained vision model capable of SOTA performance for zero-shot tasks on benchmark vision datasets similar to the capabilities of GPT-3 [8] in the NLP domain. It is trained on a web-scale dataset containing 400 million (image, caption) pairs collected from the internet. The training process consists of jointly training an image encoder (*e.g.*, Resnet-50 [23]) and a text encoder (*e.g.*, a standard transformer [57] with modifications described in Radford *et al*. [47]) such that the cosine similarity is maximized for all the correct pairings and minimized for all the incorrect pairings using a symmetric cross-entropy loss as used in contrastive learning. A detailed evaluation of 30 different datasets is provided, and it is shown that CLIP outperforms baselines trained on the benchmark datasets. In addition, the bias of CLIP models is evaluated over the FairFace benchmark dataset [32] by analyzing CLIP performance on dimensions of *gender*, *race*, and *age*. Here, the zero-shot CLIP model has, for the most part, a competitive performance to Fairface's own model. These encouraging results suggest that CLIP could be used as a meta-data caption generator for less prominent and fine-grained observables, which we evaluate in Sec. 5. For our experiments, we make use of the publicly available pre-trained CLIP ViT-B/32 model because of its inference time and wide adoption. We feed our input images containing persons to the image encoder to obtain the image embeddings and prompts of the form "`a photo of a young person`", "`a photo of an old person`" as text prompts to the text encoder. By calculating the cosine similarity of the embeddings and applying the softmax function, we can obtain the most likely caption that describes the image. The captions are designed to reflect the values of the semantic meta-data, compare Tab. 1.

3.1. Prompt ensembling as noise suppression

The reason for using prompt templates of the form "A photo of a {label}" instead of using the class names is the structure of the training data where a collection of (image, caption) pairs is used. Using only class names would lead to distributional shift [46]. In the CLIP paper, prompt engineering and prompt ensembling were shown to have a positive effect on the performance. Prompt engineering has been discussed [8] as a way to improve performance

[4] For a dependence on the dominant properties, see *e.g*. [20]

Semantic dimension	Attributes						
Gender	Male	Female					
Skin color	White	Dark					
Age	Young	Old					
Hair color	Black	Blond	Gray	Brown			
Clothing color	Yellow	Brown	Gray	Blue	Green	Red	White
Misc.	Beard	Eyeglasses	Goatee	Bald	Hat		

Table 1. A sample ontology inspired from the ontologies provided by Herrmann *et al*. [26]. The first lines represent dimensions and their possible attributes, while the last line, for brevity, provides a collection of binary attributes.

in GPT-3 type models and CLIP also shows improvement when prompts of the form "A photo of a {label}, a type of pet." are used, where more context about the object is provided. For our problem formulation, the main class is always a human and a caption would be some variation of `"a photo of a {} person"`. Replacing person with other words like "man", "woman", "girl", "boy" could add more context. However, to aggregate these results, one would require prompt ensembling. An example of prompt ensembling given by Radford *et al*. [46] is using multiple prompts of the form "A photo of a big {label}" and "A photo of a small {label}" where the adjectives "big" and "small" do not modify the main class but only provide more context. In their implementation, the ensembling is done by taking multiple text prompts, obtaining their text embeddings, and averaging them per-class. The cosine similarity with image embeddings is calculated with the ensemble average,*i.e*. with a single representation, and a softmax function is applied as the final step. Due to this reduction to single representations, the ensemble effectively functions as a linear classifier. In our experiments, we apply the softmax function prior to the class-wise averaging. This way, representations in the ensemble, which fit the image more closely, are emphasized. This serves as a noise-suppression technique, and we also obtain, effectively, a non-linear classifier.

4. Datasets

In this section, we discuss the four datasets that we use in our experiments.

CelebA dataset [37]: The CelebA dataset is a collection of 202599 images containing celebrity faces with 40 binary facial attributes (see Tab. 2 for all used dimensions). We make use of the aligned PNG images of resolution 178×218 provided by the authors.

AD datasets: In addition to the frontal face images, we use three AD datasets for our evaluations. First, we make use of synthetic data generated from the Carla simulator [16]. Inspired by Gannamaneni *et al*. [20], we generate a dataset of 10k images of resolution 1920×1280 and corresponding pedestrian meta-data using the provided modifications to the source code. Similar to their work, as pedestrian meta-data, we extract *Gender, Age, Skin-color, Shirt-color, Pant-color*. Second, we use the Cityscapes dataset, which is a collection of 5k images of urban street scenes obtained from 50 German cities. The images are of resolution 2048×1024 and taken from the ego-perspective of a vehicle. It is primarily used for semantic segmentation tasks and contains 30 different classes. Zhang *et al*. [66] created a subset of the Cityscapes dataset, which contains bounding boxes for the pedestrian class. Third, we use the RailSem19 dataset [64], which is a collection of 8500 images taken from the perspective of trains and trams with a focus on railway crossings in 1000 images. The primary labels are semantic segmentation maps with 19 classes, which we use to extract bounding boxes via connected components. The images are of resolution 1920×1080.

In all three datasets, as the primary focus is on pedestrian attributes, we crop the pedestrian images with the help of existing ground-truth bounding boxes. To maintain a constant aspect ratio, we do not use the bounding boxes directly but, based on their longer side, determine a square area around the pedestrian.[5] From the Carla dataset, we obtained 19090 individual cropped images of pedestrians by filtering out pedestrians with bounding boxes smaller than 1000 pixels to reduce noisiness in the data. Similarly, we use a filter size of 25k pixels in Cityscapes and RailSem19 and ensure that only single pedestrians are in the bounding boxes. We correspondingly obtained 60 and 63 individual cropped images. We deliberately chose a reduced selection of images for those datasets as manual evaluation by two human observers had to be performed.

5. Results

In the following section, we present the results of evaluating the performance of CLIP in generating meta-data captions of less prominent and more fine-grained observables of people by using images of celebrity faces in the CelebA dataset and cropped images of pedestrians in three AD datasets, see Sec. 4. The fine-grained observables we are interested in for our experiments are the semantic meta-data presented in Tab. 1.

[5] CLIP, by default, resizes images to a square format.

5.1. Datasets with fine-grained meta-data

In the first two datasets, CelebA and Carla, we have ground truth meta-data available for calculating performance metrics. Here, we perform two evaluations, one with a naive classifier and the other with an ensemble-based classifier. In the naive experiment setup, we use single prompts per-class for the binary cases, such that one prompt is for the presence of an attribute and a second for its absence. For example, `"a photo of a person wearing eyeglasses"` and `"a photo of a person not wearing eyeglasses"`. While antonyms exist for some attributes, *e.g.*, `young` and `old`, we use such a pattern for all attributes to maintain comparability, *e.g.*, by using prompts such as `"a photo of a not young person"`. For multi-class classifications, *e.g.*, when evaluating *shirt-color*, we use single prompts for each of the possible class values.

In the ensemble case, we, inspired by the original work [46], use multiple prompts per-class. In contrast to them, we, however, do not generate these prompts automatically based on high-frequency word lists. Instead, we build a smaller hand-crafted collection. For example, for the *age* dimension in CelebA, we make use of templates of the form `['a photo of a {} person', 'a photo of a {} man', 'a photo of a {} woman', 'a photo of a {} guy', 'a photo of a {} lady']`, where the {} are replaced by either elements from `['young', 'younger']` or `['old', 'older']` to indicate low or high age respectively. To maintain balance among the classes, we only use ensembles with equal numbers and semantically comparable prompts for every class. We aggregate these prompts as discussed in Sec. 3.1. We provide a comparison of our ensembling approach to CLIP's in the *supplementary material*. A detailed list of the prompts for the experiments along with the meta-data for the Cityscapes- and RailSem19-subsets are provided.[6]

In Tab. 2 showing the results for the CelebA dataset, we can see that our ensembling approach, right column, clearly outperforms the naive approach, left column. Furthermore, for dimensions *gender* and *age*, the results are comparable or better to the ones reported by CLIP [46] for the FairFace dataset.[7] For other less prominent dimensions like *hair-color*, wearing *hat*, . . . , we see the ensemble approach leads to improvement in performance in almost all cases. It must be noted that some of the dimensions are extremely unbalanced w.r.t. class distribution (*e.g.*, eyeglasses), and performance metrics like accuracy in such cases are not good enough for evaluation. We, therefore, look at per-class precision, recall, and F1 score to show the improvement for these dimensions. Performance on certain dimensions, see *smiling*, which, while not directly relevant for AD safety, shows the rich representation power of the CLIP model. By comparison to the results from existing works [30, 38], we can conclude that CLIP again achieves comparable performance for almost all of the attributes we evaluate. Note that CLIP is evaluated in a zero-shot fashion, while the other methods were explicitly trained on the CelebA dataset. One anomaly in performance is the dimension *skin-color*, where there is a drop in performance from naive to the ensemble-based approach. In the CelebA dataset, this maps to the binary attribute *pale*. As shown by the examples in Fig. 1, differentiating between these classes is non-trivial and labeling preference varies even among humans. We evaluate this dimension in further detail to understand the CLIP representation space. Looking at the performance metrics, we see a drop in recall for the *not-pale* class and a gain in recall for *pale*. For the ensemble-based approach, we expand the definition of what is considered as *pale* by using adjectives such as "sickly" or "bleached skin-color". Although this approach is balanced by comparable adjectives for the other class, such as "tanned" or "blushed", this expanded definition could be the reason for the improvement in recall values for the *pale* class. As *not-pale* is the dominant class, its decreased recall also leads to a reduction in overall accuracy. In Fig. 2, we can further highlight the challenge of separating this dimension by embedding the representations of the images in a 2D space using the cosine similarity distance between image and the mean representations of the respective classes (as derived from the ground truth). On the left, we see the visualization for *skin-color* where there is no clear separation, while in the middle the *gender* classes are easily separable. To investigate the separability of the representation further, we train a logistic regression on the full ground truth data, which, irrespective of any prompts, provides the "ideal" linear classifier for the given data.[8]

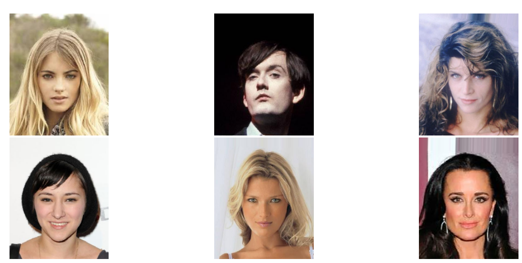

Figure 1. Example of challenging semantic dimensions in the CelebA dataset: Top row contains images of celebrities labeled as having pale skin. Bottom row contains images of celebrities labelled as having not-pale skin.

[6]https://github.com/sujan-sai-g/clip_evaluations_for_metadata

[7]Reported accuracy of class *gender* is 0.95 and *age* is 0.57 for category 'White', see [46]

[8]To achieve comparability to the prompt-based approach, we omit any intercept or regularization in the classifier. As can be seen from the precision-recall curve, the data is not separable but shows a strong gradient in space with *pale* images favoring one side.

Semantics	Attribute	Counts	Naive				Non-linear Ensemble			
			Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Age	Young	156734	0.78	0.80	0.95	0.87	0.86	0.91	0.91	0.91
	Not-young	45865		0.53	0.21	0.30		0.70	0.70	0.70
Gender	Male	84434	0.95	0.95	0.91	0.93	0.99	0.99	0.98	0.99
	Not-male	118165		0.94	0.97	0.95		0.99	0.99	0.99
Skin-color	Pale	8701	0.84	0.11	0.41	0.18	0.56	0.07	0.81	0.13
	Not-Pale	193898		0.97	0.86	0.91		0.98	0.54	0.70
Hair-color	Black	47323	0.77	0.93	0.64	0.76	0.78	0.94	0.65	0.77
	Blond	28252		0.81	0.93	0.87		0.83	0.93	0.87
	Gray	7928		0.76	0.69	0.72		0.81	0.65	0.72
	Brown	39167		0.65	0.83	0.73		0.64	0.86	0.73
Misc.	Eyeglasses	13193	0.97	0.86	0.55	0.67	0.97	0.74	0.86	0.80
	No eyeglasses	189406		0.97	0.99	0.98		0.99	0.98	0.98
	Hat	9818	0.92	0.35	0.73	0.47	0.96	0.56	0.74	0.64
	No Hat	192781		0.99	0.93	0.96		0.99	0.97	0.98
	Bald	4547	0.87	0.07	0.39	0.11	0.93	0.19	0.60	0.29
	Not Bald	198052		0.98	0.88	0.93		0.99	0.94	0.96
	Goatee	12716	0.53	0.05	0.37	0.09	0.90	0.26	0.30	0.28
	No Goatee	189883		0.93	0.54	0.68		0.95	0.94	0.95
	Beard	33441	0.81	0.23	0.06	0.10	0.84	0.69	0.10	0.18
	No Beard	169158		0.84	0.96	0.89		0.85	0.99	0.91
	Smiling	97669	0.81	0.74	0.94	0.83	0.87	0.88	0.86	0.87
	Not-smiling	104930		0.92	0.69	0.79		0.87	0.89	0.88

Table 2. The performance of CLIP in predicting different attributes on the celebrity images in the CelebA dataset.

Extending this to the AD domain, we first look at the performance of CLIP on the Carla dataset shown in Tab. 3. Similar to the earlier experiment, we see an improvement in performance from the naive to the ensemble prompts case. However, the overall performance is lower for dimensions like *age*, *gender*, and *skin-color* in comparison to FairFace, CelebA, and later experiments with Cityscapes and RailSem19. There could be two underlying reasons for this: First, there is a domain gap from real-world to computer-simulated data leading to generalization problems. Second, FairFace and CelebA have high-quality frontal images of peoples' faces. However, in AD datasets, we (mostly) use smaller images showing the person as a whole in more diverse contexts, for instance, w.r.t. occlusion, pose, brightness, etc. Such different contexts might play a role when interpreting the low performance on *shirt-* and *pant-color*, as color perception and also its rendering are strongly affected by the illumination and other factors, such as occlusion, which would reduce the effective number of visible colored pixels. Lastly, CLIP has significantly lower performance on *pant-color* than other dimensions. Through visual inspection and from calculating Pearson correlation of *shirt-* with *pant-color* predictions, we believe that CLIP focuses mostly on the dominant color in the image, and this dominates over concepts of shirt and pant. The correlation value for both (*i.e*. for overlapping colors in both dimensions) is 0.90.

5.2. Datasets without fine-grained meta-data

In Tab. 4, we have results of both the Cityscapes-subset and RailSem19-subset datasets. As mentioned, these datasets do not contain any ground truth regarding the fine-grained observables we are interested in, and the performance here is evaluated manually by looking at the images by two independent human observers. This experiment is conducted as a proof-of-concept to show that it is actually possible to transfer our learning from previous datasets to real-world data and annotate less prominent and more fine-grained observables. As these datasets do not have a significant variation in *skin-color*, we skip this dimension. The experiments here are conducted with ensemble-based approach only as it outperforms the naive approach in other experiments. Unlike in the Carla experiment, these are real-world datasets implying no domain gap due to synthetic images. However, these datasets also contain pedestrians in different poses, occlusions, and brightness. Therefore, *gender* and *age* still remain challenging dimensions in certain instances. The performance on *shirt-* and *pant-colors* is, however, slightly improved over the Carla dataset. Similar to the earlier experiment, the predictions of *pant-color*

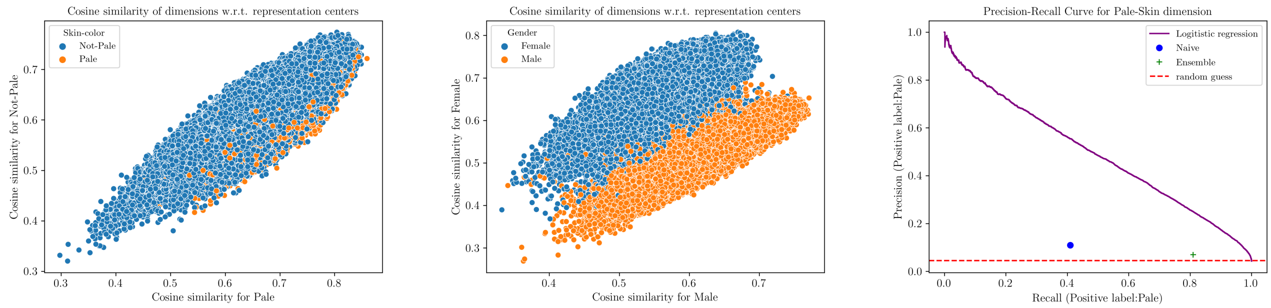

Figure 2. Left and middle: Cosine similarity of two dimensions *skin-color* and *gender* in the CelebA dataset calculated w.r.t. ground truth by taking a mean of image representations belonging to each group. Right: Precision-recall curve of a linear classifier along with performance values of the naive and ensemble approach for *skin-color* dimension.

Semantics	Attribute	Counts	Naive				Non-linear Ensemble			
			Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Age	Adult	14652	0.27	0.90	0.06	0.11	0.59	0.89	0.53	0.67
	Child	4438		0.24	0.98	0.39		0.34	0.79	0.47
Gender	Male	10009	0.67	0.63	0.77	0.69	0.71	0.63	0.93	0.75
	Female	9081		0.74	0.58	0.65		0.89	0.51	0.65
Skin-color	White	12119	0.70	0.64	0.81	0.72	0.73	0.68	0.83	0.74
	Dark	6971		0.78	0.60	0.68		0.81	0.65	0.72
Shirt-color	Yellow	3915	0.42	0.65	0.80	0.72	0.45	0.63	0.88	0.74
	Brown	5329		0.52	0.51	0.51		0.51	0.53	0.52
	Blue	1843		0.24	0.14	0.18		0.24	0.15	0.18
	Gray	4648		0.43	0.05	0.09		0.49	0.08	0.13
	Green	1233		0.34	0.48	0.40		0.51	0.42	0.46
	Red	1592		0.28	0.53	0.36		0.26	0.56	0.36
	White	530		0.13	0.62	0.21		0.13	0.45	0.20
Pant-color	Yellow	704	0.17	0.13	0.59	0.22	0.17	0.14	0.60	0.22
	Brown	3744		0.09	0.04	0.06		0.11	0.07	0.09
	Blue	7109		0.91	0.05	0.10		0.91	0.06	0.11
	Gray	1818		0.22	0.02	0.04		0.25	0.03	0.05
	Green	801		0.07	0.83	0.13		0.07	0.79	0.13
	Red	676		0.12	0.31	0.17		0.12	0.38	0.20
	Black	2572		0.55	0.13	0.22		0.45	0.10	0.16
	Orange	1071		0.78	0.80	0.79		0.74	0.82	0.78

Table 3. The performance of CLIP in predicting different attributes on the cropped images of pedestrians in Carla dataset.

are highly correlated with shirt-color with Pearson correlation of 0.71 and 0.79 on Cityscapes and RailSem19, respectively. While the reduced domain gap might explain the improvement, another major contributing factor could be the evaluation technique. As these datasets do not have any meta-data ground truth, and therefore two human observers were asked to validate whether the output of the CLIP model is plausible. For images with bad color reproduction, this might lead to a more lenient interpretation. While it is implausible to evaluate CLIP predictions on entire AD datasets, we believe the lenient approach to evaluation is a more realistic evaluation than what is possible with the fixed ground truth in Carla. This approach, however, also acts as labeling bias when comparing these results to those of the Carla experiment.

6. Conclusion

Coming from the direction of safety argumentation for safety-critical autonomous systems, specifications of operational design domains form an indispensable tool to analyze data coverage as well as to detect weaknesses of learned models. The latter is done, *e.g.*, by building data subsets

Semantics	Attribute	Cityscapes-subset					RailSem19-subset				
		Counts	Accuracy	Precision	Recall	F1 score	Counts	Accuracy	Precision	Recall	F1 score
Age	Young	48	0.65	1.00	0.56	0.72	40	0.68	0.86	0.60	0.71
	Old	12		0.36	1.00	0.53	22		0.53	0.82	0.64
Gender	Male	42	0.92	0.95	0.93	0.94	46	0.89	0.91	0.93	0.92
	Female	18		0.84	0.89	0.86	16		0.80	0.75	0.77
Shirt-color	Yellow	2	0.78	0.50	1.00	0.67	1	0.66	0.17	1.00	0.29
	Brown	5		0.67	0.80	0.73	10		0.73	0.80	0.76
	Grey	10		0.64	0.90	0.75	9		1.00	0.56	0.71
	Blue	8		0.80	0.50	0.62	8		1.00	0.62	0.77
	Green	6		0.75	1.00	0.86	7		0.42	0.71	0.53
	Red	2		1.00	1.00	1.00	7		0.57	0.57	0.57
	White	9		1.00	0.67	0.80	8		0.71	0.62	0.67
	Black	18		0.93	0.78	0.85	12		0.89	0.67	0.76
Pant-color	Yellow	0	0.48	-	-	-	0	0.50	-	-	-
	Brown	3		0.21	1.00	0.35	13		0.67	0.77	0.71
	Grey	15		0.71	0.67	0.69	10		0.86	0.60	0.71
	Blue	13		1.00	0.23	0.38	14		0.75	0.21	0.33
	Green	3		0.25	0.67	0.36	2		0.07	0.50	0.12
	Red	0		-	-	-	0		-	-	-
	White	2		0.17	0.50	0.25	2		0.50	1.00	0.67
	Black	24		1.00	0.42	0.59	20		0.90	0.45	0.60

Table 4. The performance of CLIP in predicting different attributes on the cropped images of pedestrians in Cityscapes-subset and RailSem19-subset datasets. Only our non-linear ensemble approach is used for this experiment.

based on the meta-data attributing each input to a part of the domain specification. However, such fine-grained meta-data is often not available in real-world datasets. Therefore, we investigated the zero-shot capabilities of CLIP to provide such information on a granular level of detail beyond previous tests of this model. For this, we introduce a simple softmax-based noise-suppressing technique to the CLIP prompt ensemble, which has proven robust in practice. The results, for many investigated aspects, are on-par with dedicatedly trained classifiers implying that CLIP may indeed be used to derive such annotations as well as for "weak" supervision of specialized tasks. This holds not only for commonly tested dimensions, such as *age*, *gender* or *hair-color*, but also for more fine-grained attributes, *e.g.*, wearing *eyeglasses* or *hats*. However, we also find dimensions, such as *pant-* or *shirt-color*, where this approach is challenged. This highlights the importance of human validation for its practical use. Specifically for the named dimensions, we observed that the performance of CLIP is better when evaluated on human-generated labels than on ground truth labels stemming from synthetic data. This raises the question of the granularity of the learned representations, *e.g.*, broad categories might work better than narrow ones. As a rule of thumb, the evaluation suggests that attributes more likely to appear in captions are currently resolved better. We investigated this granularity more closely on the highly challenging dimension *pale skin* of the CelebA dataset, which the current version of CLIP does not separate sufficiently even on the level of embeddings. The quality of such representations strongly impacts the performance of downstream tasks, as seen in our experiment. But, this likely transfers to other approaches, *e.g.*, Domino [17], that use (CLIP) representations, *e.g.*, for weakness detection, and likely will have short-comings w.r.t. such dimensions.

For future work, this leaves us with two directions: At first, given that generated fine-grained meta-data, or rather the underlying representations, are not always fully accurate, one needs to more closely investigate which degree of accuracy is needed for downstream tasks, *e.g.* to reliably detect weaknesses of DNNs. Second, given the broad implications on the performance of foundation models, it is necessary to better understand to which degree dimensions are separable, *i.e.*, resolvable. Ideally, one would like to substantiate the above rule of thumb and find ways to better detect or measure the quality of the representations w.r.t. their semantic content.

7. Acknowledgments

This work has been funded by the German Federal Ministry for Economic Affairs and Climate Action as part of the safe.trAIn project.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 1

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5561–5570, 2018. 2

[3] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015. 2

[4] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4662–4670, 2022. 2

[5] Matthias Bitzer, Martin Herrmann, and Eckart Mayer-John. System co-design (scode): Methodology for the analysis of hybrid systems. *at-Automatisierungstechnik*, 68(6):488–499, 2020. 3

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[7] Markus Borg, Cristofer Englund, Krzysztof Wnuk, Boris Duran, Christoffer Levandowski, Shenjian Gao, Yanwen Tan, Henrik Kaijser, Henrik Lönn, and Jonas Törnqvist. Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry. *arXiv preprint arXiv:1812.05389*, 2018. 1

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 1

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[11] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1550–1553. IEEE, 2019. 2

[12] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[14] Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022. 1, 3

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[16] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 3, 4

[17] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022. 1, 3, 8

[18] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer, 2010. 2

[19] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017. 2

[20] Sujan Gannamaneni, Sebastian Houben, and Maram Akila. Semantic concept testing in autonomous driving by extraction of object-level annotations from carla. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1006–1014, 10 2021. 1, 3, 4

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[22] Magnus Gyllenhammar, Rolf Johansson, Fredrik Warg, DeJiu Chen, Hans-Martin Heyn, Martin Sanfridson, Jan Söderberg, Anders Thorsén, and Stig Ursing. Towards an operational design domain that supports the safety argumentation of an automated driving system. In *10th European Congress on Embedded Real Time Systems (ERTS 2020)*, 2020. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[24] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A

simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 1

[25] Francisco Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29(3):495–525, 2011. 2

[26] Martin Herrmann, Christian Witt, Laureen Lake, Stefani Guneshka, Christian Heinzemann, Frank Bonarens, Patrick Feifel, and Simon Funke. Using ontologies for dataset engineering in automotive ai applications. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 526–531. IEEE, 2022. 3, 4

[27] High-Level Expert Group on AI (AI HLEG). Ethics guidelines for trustworthy AI. Technical report, European Commission, 2019. 1

[28] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 2

[29] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020. 1

[30] Sunhee Hwang, Sungho Park, Pilhyeon Lee, Seogkyu Jeon, Dohyung Kim, and Hyeran Byun. Exploiting transferable knowledge for fairness-aware image classification. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 5

[31] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 2

[32] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 3

[33] Philip Koopman and Frank Fratrik. How many operational design domains, objects, and events? In *SafeAI@AAAI*, 2019. 1, 3

[34] Siming Li, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, 2011. 2

[35] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2

[36] Xiaoxiao Liu, Qingyang Xu, and Ning Wang. A survey on deep neural network-based image captioning. *The Visual Computer*, 35(3):445–470, 2019. 2

[37] 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. 4

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[39] Wulf Loh, Andreas Hauschke, Michael Puntschuh, and Sebastian Hallensleben. VDE SPEC 90012 v1.0 - VCIO based description of systems for AI trustworthiness characterisation. Technical report, Verband der Elektrotechnik Elektronik Informationstechnik e.V. (VDE), 04 2022. 1

[40] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[41] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2286–2293, 2021. 2

[42] Maria Lyssenko, Christoph Gladisch, Christian Heinzemann, Matthias Woehrle, and Rudolph Triebel. From evaluation to verification: Towards task-oriented relevance metrics for pedestrian detection in safety-critical domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38–45, 2021. 1, 3

[43] Michael Mock, Stephan Scholz, Frédérik Blank, Fabian Hüger, Andreas Rohatschek, Loren Schwarz, and Thomas Stauner. An integrated approach to a safety argumentation for ai-based perception functions in automated driving. In *Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops: DECSoS, MAPSOD, DepDevOps, USDAI, and WAISE, York, UK, September 7, 2021, Proceedings 40*, pages 265–271. Springer, 2021. 1, 3

[44] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[45] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 5

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[48] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[49] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 2

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[51] Svetlana Sagadeeva and Matthias Boehm. Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2290–2299, 2021. 2

[52] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 2

[53] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: a survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022. 2

[54] Chen Sun, Chuang Gan, and Ram Nevatia. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE international conference on computer vision*, pages 2596–2604, 2015. 2

[55] Qutub Syed Sha, Oliver Grau, and Korbinian Hagn. Dnn analysis through synthetic data variation. In *Computer Science in Cars Symposium*, pages 1–10, 2020. 1, 3

[56] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[58] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 1

[59] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[60] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202, 2017. 2

[61] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 2

[62] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2621–2629, 2019. 2

[63] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision*, pages 4894–4902, 2017. 2

[64] Oliver Zendel, Markus Murschitz, Marcel Zeilinger, Daniel Steininger, Sara Abbasi, and Csaba Beleznai. Railsem19: A dataset for semantic rail scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 4

[65] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[66] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 4

[67] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020. 2