"Just Drive": Colour Bias Mitigation for Semantic Segmentation in the Context of Urban Driving

Jack Stelling
School of Computing
Newcastle University, UK
jackstelling2@gmail.com

Amir Atapour-Abarghouei
Department of Computer Science
Durham University, UK
amir.atapour-abarghouei@durham.ac.uk

***Abstract*—Biases can filter into AI technology without our knowledge. Oftentimes, seminal deep learning networks champion increased accuracy above all else. In this paper, we attempt to alleviate biases encountered by semantic segmentation models in urban driving scenes, via an iteratively trained *unlearning* algorithm. Convolutional neural networks have been shown to rely on colour and texture rather than geometry. This raises issues when safety-critical applications, such as self-driving cars, encounter images with covariate shift at test time - induced by variations such as lighting changes or seasonality. Conceptual proof of bias *unlearning* has been shown on simple datasets such as MNIST. However, the strategy has never been applied to the safety-critical domain of pixel-wise semantic segmentation of highly variable training data - such as urban scenes. Trained models for both the baseline and bias unlearning scheme have been tested for performance on colour-manipulated validation sets showing a disparity of up to 85.50% in mIoU from the original RGB images - confirming segmentation networks strongly depend on the colour information in the training data to make their classification. The bias unlearning scheme shows improvements of handling this covariate shift of up to 61% in the best observed case - and performs consistently better at classifying the "human" and "vehicle" classes compared to the baseline model.**

***Index Terms*—Fair AI, Bias Unlearning, Bias Removal, Semantic Segmentation, Convolutional Neural Networks**

I. Introduction

Recent years have seen a surge in the development of artificial intelligence (AI) systems; largely fuelled by the symbiosis of deep learning progress [1], [2], [3], [4], [5], [6], and advancements in micro/nanochip manufacturing — harnessing remarkable compute power. Ubiquitous deployment of AI throughout modern society means that practitioners have an ethical and moral responsibility in the dissemination of this cutting-edge technology.

Within the field of deep learning, convolutional neural networks (CNNs) have gained substantial traction in the last decade, and their performance in computer vision tasks is state of the art [7], [8], [9]. Due to the complex nature of these systems, a "black box" stigma is often attached to them; since deep learning networks are now achieving unprecedented levels of accuracy, external scrutiny and media spotlight demands a push towards transparency, fairness, and accountability [10], [11]. This has led to the more recent movement of "explainable artificial intelligence" (XAI).

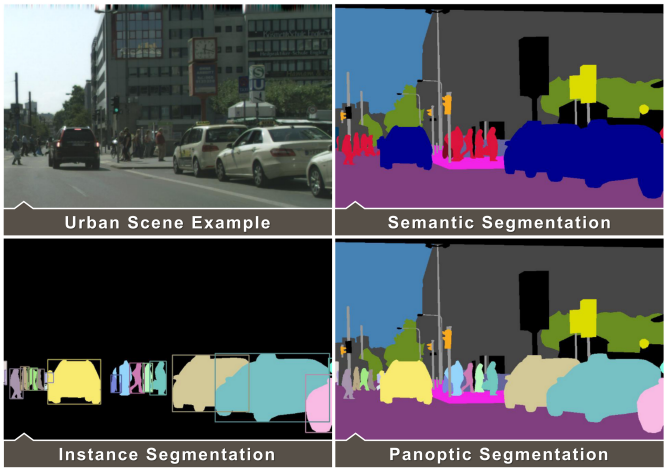

Fig. 1: Semantic, Instance and Panoptic segmentation [12].

The crux of developing fair AI is the elimination of bias – a long sought issue in any statistical modelling application. Bias can manifest itself in a multitude of guises which are generally quite context specific. For the purposes of this work, we will focus on algorithmic bias. Bias of this nature can creep into our models from training data, meaning our models exhibit the same systemic discrimination found in the wild. This bias can often be unknown and undetected – making for a particularly insidious force. Indeed, George Santayana warned us that *"those who do not remember the past are condemned to repeat it"*. If bias *is* allowed to creep into our models undetected, we risk propagating this bias throughout society; eventually distilling into a self-fulfilling prophecy — one which automates inequality.

This research focuses on colour bias which exists within urban scenes. Urban scene segmentation is at the heart of autonomous vehicle (AV) technology [13], and to date, many network architectures have been developed achieving impressive accuracy [5], [14]. This blinkered push towards optimal accuracy often neglects to consider biases within the training data; thus, any advancements in the field of bias mitigation in particular within the realm of image segmentation are significant.

Colour bias can manifest in many ways within urban scenes. Successful AV technology must be dependable in a multitude

of conditions, including densely populated urban streets, fog, rain, snow, glare, and seasonal changes. This highly variable distribution of data poses a great challenge. Clearly, the same tree that a segmentation model picks up on in summer might look very different in winter, or in New York would a CNN learn to categorise all yellow boxes as cars due to the number of taxis? We humans are extremely adaptable and can make decisions to correct our actions depending on external stimulus. Machines, however, struggle to perform well in edge cases or situations not encountered in the training phase. Consequently, a push towards robustness and generalisation is paramount for the evolution of safe AV technology.

The umbrella of image segmentation covers three main domains: instance segmentation [15], semantic segmentation [14], [5], and more recently a hybrid of both – panoptic segmentation [16] (Figure 1). In this work, we will consider semantic segmentation, which concerns categorising each pixel in an image into one of n predefined classes. This process splits the image into different regions based on what the pixels show and is the adopted methodology for AV systems. While this work focuses on semantic segmentation, the same principles can be applied to instance and panoptic segmentation or in fact many other computer vision tasks [17], [18], [19], [20].

The aim of this paper is to mitigate colour bias from semantic segmentation models trained on urban driving scenes. The primary contributions of this work are twofold:

- We provide empirical evidence that seminal semantic segmentation architectures do overfit to the colour information in highly variable urban scenes and, where possible, attempt to quantify this.
- We also demonstrate that a multi-headed network architecture can adversarially remove a *known* bias during training in a pixel-wise semantic segmentation model.

In the interest of reproducibility, transparency and progression, all code is publicly available in a project repository[1].

II. Related Work

We consider related work within the context of Semantic Segmentation (Section II-A) and Bias Removal (Section II-B).

A. Semantic Segmentation

For the purposes of semantic segmentation, fully convolutional networks (FCN) have gained the most traction in recent years after the work of Long et al. [21]. FCNs do not have any fully-connected layers, solely relying on convolutional layers passed to a classification layer. This preserves spatial information from the input and significantly reduces the number of parameters in the network, thus increasing computational efficiency.

Encoder-decoder style networks achieve impressive results with U-Net [22], SegNet [5] and DeConvNet [23] all adopting the same idea, whereby the first half of the network is a mirror image of the second half, albeit with different methodical nuances. SegNet [5] proposes a computationally fast and memory efficient network by saving the indices of the maximum values on the max-pooling operation – this is then used during upsampling in the decoder part of the network. U-Net utilises skip connections, allowing for information from the encoder part of the network to be used in the decoding procedure.

Later, the DeepLab family [24], [25], [14], [7] of architectures pushed the boundaries in semantic segmentation using atrous convolutions [25], effectively developing the Atrous Spatial Pooling Pyramid (ASPP) module to handle objects of different scales in the same image. ASPP is based on simultaneously performing dilated convolutions with different atrous rates and concatenating the resultant feature maps. This essentially combines multiple fields of view, tackling the issue of different scale objects. The DeepLab family have achieved state-of-the-art results on benchmark datasets with each iteration of the network.

More recently, the concept of attention has been applied to semantic segmentation tasks [26], creating a more efficient and higher performing model than using multiscale inference. Similarly to the problem that DeepLab attempted to solve with the ASPP module, Tao et al. [26] argue that fine detail (bollards, a person in the distance, etc.) is often better predicted with a scaled-up image size. Whereas large objects (roads, buildings, etc.) require more global context and downscaled images are generally more beneficial as the convolutional filters have a larger field of view and thus capture more context. Tao et al. develop a system whereby prediction for some pixels is performed using the scaled-up images and others use the scaled down images. Further, the ASPP module and other multiscale context methods e.g., PSPNet [27] are static and not learned, whereas relational methods build context based on image composition. This means that unlike [25], [14], [7], [27], the region of interest using an attention-based mechanism is not restricted to being square – this is advantageous in the context of urban scenes when the geometry is often a product of visual perspective, like a road sign in the foreground covering a long skinny rectangular patch. Tao et al. maintain state-of-the-art performance on the CityScapes [28] and Mallipary Vistas [29] datasets at the time of writing.

In this work, we adopt the DeepLabV3 [14] and SegNet [5] architectures to use as baselines for testing novel bias removal concepts in the context of semantic segmentation. DeepLab architectures have been shown [30] to be less susceptible to adversarial attacks, an increasing concern in safety-critical applications such as self-driving cars. Other segmentation models, however, can similarly be used given the overall pipeline of our approach.

B. Bias Removal

As mentioned in Section I, *"bias"*, in the context of urban scene segmentation, can manifest in many forms. Examples include adverse weather conditions experienced at test time when training data does not account for this, seasonality affecting physical colours (e.g., tree leaves, flora), seasonality affecting lighting/shadows/luminance, more obvious lighting

[1] https://github.com/JackStelling/BiasMitigation

fluctuations from night to day, reflection, shadow and different countries/localities using different colour systems for highway codes, among others. This is by no means an exhaustive list, as even edge cases such as sporting events, parades and accident blockades can cause out-of-sample differences that networks must tackle if we are to trust AV technology. These unknown perturbations cause a covariate shift from the input data that the models are trained on, which can cause adverse effects on performance due to the high intercorrelation between network weights.

Large-scale, finely annotated datasets for segmentation are expensive to obtain, requiring human annotation which can often take experienced workers up to 90 minutes an image to complete [28]. Due to this bottleneck, it is not feasible to create site specific training datasets for multiple locations where the cars will operate, thus robust generalisable models must be developed which perform well in a wide variety of situations. Models which can learn bias in the data and account for it are highly desirable. Not only does this problem exist within the sphere of AV technology, it also extends to many others, including the fields of augmented reality and virtual reality where indoor scene segmentation is paramount – another task which relies on a highly diverse input distribution.

Under the umbrella of bias removal, the taxonomy forms three natural groupings:

- Those seeking to increase generalisation and thus reduce the effects of bias via image augmentation.
- Those using the network architecture to attempt to remove or mitigate a known bias.
- Those attempting to learn the bias within a given dataset and mitigate it accordingly.

Image augmentation increases variability in the training data by adding variation to images. This synthetically increases the training set without affecting the information contained. Common perturbations include crops, sheers, flips, and colour jitter. This technique has been well researched in computer vision tasks, specifically image classification [31], [32].

Augmentation techniques have been adapted specifically to semantic segmentation where sheering and flipping images may not be the most appropriate approach. Kamann et al. [33] propose the use of a colour mask gained from alpha-blending the ground truth segmentation map with the input data during training, which they coin "Painting-by-Numbers". Building on the evidence of Geirhos et al. [34] who showed that CNNs are biased towards texture, Painting-by-numbers improves the robustness of semantic segmentation models to common image corruptions by making the texture of image classes less reliable and pushes the model to use geometry in the image to perform effective segmentation. This technique does not require more training data and is thus efficient during training. Also motivated by CNN textural reliance, Jackson et al. [35] explore style randomisation via altering colour and texture of the input image using style transfer whilst preserving semantic content, again showing an increase in accuracy.

Multi-head models have been explored [36], [37], [38] posing the ability of networks to *unlearn* a known bias. In fact, it has been shown in [38] that some models even exacerbate the biases that are known in the datasets after training is complete, thus, models are using the bias itself as a cue to make a certain categorisation. Kim et al. [36] demonstrate a proof-of-concept approach showing that after planting a synthetic colour bias in the MNIST dataset, the model uses the obvious colour cue for categorisation rather than geometry of the numbers. A second model head employs an iterative algorithm using reverse gradients to successfully "unlearn" the *known* colour bias, pushing the model to use the shape of the numbers as the cue for correct classification. To the best of our knowledge, such a technique has not been applied to highly variable domains such as semantic segmentation. As such, in this paper, we attempt to evaluate its success with semantic segmentation.

The contribution of this paper is to increase the robustness of CNNs via the mitigation of algorithmic bias - specifically colour bias found in highly variable urban road scenes. Building on the foundations of Kim et al. [36], we aim to implement an "unlearning" procedure within the network architecture itself rather than increase generalisability via augmentation of the input data. The unlearning procedure employs a multi-headed network to adversarially remove a target bias using reverse gradient loss. Semantic segmentation is used as a vehicle to assess the effectiveness of such a system, albeit the core principle should, in theory, be applicable to any deep learning architecture.

III. Removing a Known Bias

This work takes advantage of the work in [36] entitled *Learning Not to Learn* referred to, hereafter, as *LNTL*. This proof-of-concept paper synthesises colour bias into the MNIST [39] dataset and successfully uses a gradient reversal strategy to remove the colour information from the training data. In real data, we are not afforded the luxury of knowing exactly what the bias is and where it manifests – although it has implicitly been shown that CNNs can pick up on the wrong cues [34]. Thus, a comparison between standard training data, a greyscale baseline and an implementation of bias *unlearning* is tested. This strategy is a self-supervised task as we can extract the true colour labels from pixel values of the training data, which we already have.

A. Problem Statement

In theory, the set of all images that self-driving cars encounter at test time is drawn from one set. This set contains all possible situations that the car could ever encounter, whether it be across a desert track in midday sun or a snow storm in a mountain pass. Indeed, many landscapes could provide this challenge in a single journey - confirming that this rich theoretical variation is not hyperbole. Practicality dictates that the training dataset is a restricted subset, and due to this, it is not a true representation of situations we *could* encounter in the real world. In this sense, we assume the test set to be unbiased. We aim to train a network on biased data, attempt to systematically *unlearn* that bias during the training phase and then deploy the model to perform on unseen and unbiased test data.

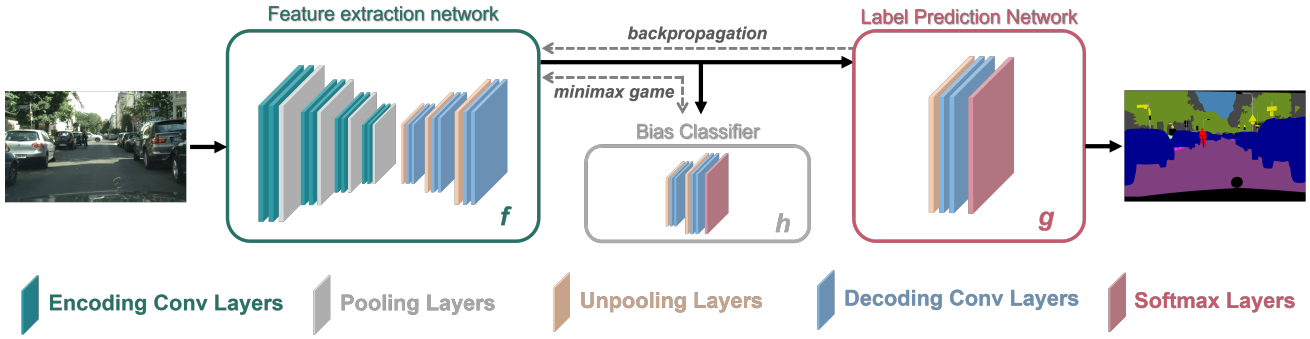

Fig. 2: Network architecture showing separate networks f, g and h and their roles in the system.

Training and test sets are normally split randomly, in an attempt to eliminate the domain gap between the two distributions X_{Train} and X_{Test}. Despite best efforts, a gap can still remain, and an even larger gap is the domain shift between X_{Test} and the actual distribution X. For this reason, the results of the bias mitigation strategy might appear slightly subdued. Analogously, results published for segmentation accuracy are inflated compared to performance in the real world. The Cityscapes dataset [28] - which consists of dash-cam footage from European cities - is partitioned on city for train, validation, and test sets, which means that model may lean on the nuances apparent in the training cities, making it an ideal candidate to use for the purposes of this work.

Our setup splits a standard semantic segmentation classification network (e.g. DeepLabV3), into a double-headed network – one with a pixel-wise classification head for the task of semantic segmentation and one with an auxiliary bias classification head. The networks are modular constructs consisting of a feature extractor network, $f : \mathcal{X} \mapsto \mathbb{R}^{\mathbb{N}}$, a pixel-wise classification network, $g : \mathbb{R}^{\mathbb{N}} \mapsto \mathcal{Y}$, and the bias classification head, $h : \mathbb{R}^{\mathbb{N}} \mapsto \mathcal{B}$, where N is the amount of feature maps produced by the embedding network f.

Figure 2 shows the implemented network architecture, with the sub-networks f, g and h, with the fork depicted at the last convolutional layer prior to classification. The precise architecture is left intentionally vague because the theory should apply to any system. Specific networks are discussed in more detail in Section IV-B.

B. A Caveat on Fork Placement

Interestingly, since f feeds its output feature maps into network g, the bias propagates through the network, and so the location of the fork is an arbitrary choice. Furthermore, $f \circ g$ will be void of any bias so long as h has done its job in correctly classifying the bias and the subsequent gradient reversal step successfully discourages f from using such cues. Prior work [40], [41] has shown that feature maps extracted at the start of a CNN generally contain low-level information, often containing blocks of colour and edges, whereas the final layer feature maps contain higher-level features with tightly integrated colour information.

Intuitively, we expect that the feature maps towards the end of the network would be the most suitable place to add the fork. The training regime requires a short burst of end-to-end training without including the bias classification network, h. This ensures that the network already has some classifying ability and weights are converging. If we do not allow this head start for the classifying network, when the auxiliary bias network is activated, a mode collapse situation could occur, where the network weights are unable to converge towards optimisation [36].

Thus, a fork located towards the end of the main bulk of the segmentation network would allow the weights upstream to be amended whilst the classification layer would be largely unaffected. We hypothesise that a different fork location further upstream would achieve the same result eventually but would take longer to reach convergence. Furthermore, the fork has been added leaving one convolutional layer before the SoftMax layer. This ensures learnable parameters remain in the classification head and allows for any reactive adjustments in network g from a change of its input, $f(x)$.

IV. Experiments

The following section details the experiments undertaken, and provides results and analysis. As such, in this section, the datasets are introduced (Section IV-A), an explanation of the specific networks used are provided (Section IV-B), evaluation metrics are explained (Section IV-C) and all experiments and results are discussed.

A. Datasets

Cityscapes [28] The Cityscapes dataset is a public and widely-used semantic segmentation benchmark. Cityscapes contains data from 50 cities and images are annotated with 30 semantic classes. The dataset contains over 20,000 coarsely annotated images and $\approx$ 5,000 finely annotated images providing pixel-level, instance-level and panoptic semantic ground truth labels. Raw images and segmentation masks are provided in portable network graphics format. Data contains both 16-bit High-Dynamic Range images and 8-bit Low-Dynamic Range format to use at an image size of 1024×2048. We evaluate our semantic segmentation performance on the official 500 annotated validation set.

SYNTHIA [42] The SYNTHIA dataset is also publicly available and comprised of nearly-photo-realistic images from a synthetically-rendered virtual city. The dataset used in this paper is the *synthia-rand-cityscapes* subset containing 9,400 images at a resolution of 1280×760, which contains labels compatible with Cityscapes, allowing for a fairer examination of the results. The dataset provides fine detail instance segmentation labels - thus data preprocessing is undertaken to create semantic labels from the data provided.

SYNTHIA images contain very realistic granular detail and complex scenes; some images contain very high numbers of pedestrians - all of which have pixel perfect annotation, leveraging the automatic ability to label images in a generated scene. Further, SYNTHIA has a rich variety of luminance, high scene diversity and vantage differences, making it a more variable sample space than Cityscapes. Input perturbations during rendering create similar images with slight nuances. When creating 70% / 30% training / validation sets, the data is split as though a temporal dependence exists within the data. This ensures that similar images do not occur in both the training and validation sets – providing a more representative interpretation of a real setting, where a test set is a wholly unseen set of images. This also ensures that we do not get over-optimistic metrics upon evaluation.

B. Network Architecture

Since the network f consists of a feature extraction sub-network, we have the flexibility to choose any architecture we wish. Indeed, we have positioned this paper for the task of mitigating colour bias in urban scenes, but we could equally apply the technique to other fields of computer vision; say, facial recognition de-biasing.

We have chosen to test two seminal semantic segmentation architectures, namely, DeeplabV3 and SegNet as discussed in Section II. DeeplabV3 is trained with multiple ResNet backbones with ImageNet [43] pretrained weights. Adopting this pretraining procedure could add noise to the results as it adds uncertainty about the origins of the CNN bias. Nevertheless, it is an efficiency trade off, and as mentioned in Section III-B, anything upstream from the fork can be unlearned. The fork is located directly after the concatenation of the ASPP module within DeepLabV3 and $f(X)$ outputs 1280 feature maps to both the auxiliary bias head, h, and the primary semantic classifier, g.

SegNet follows a simple symmetric encoder-decoder architecture. The encoder is topologically identical to the convolutional layers of the VGG16 [44] network, whilst the decoder upsamples hierarchically by using the indices of the corresponding max pooling operation from the encoding operation. Again, the fork for network h is placed before the last convolutional layer and $f(X)$ outputs 64 feature maps to both h and g.

C. Metrics

In semantic segmentation dealing with multi-class problems, we often encounter the issue of class imbalance. This occurs when background parts of an image, say buildings or sky, dominate the total pixel count compared to say that of a pedestrian or a red traffic light. Due to this, accuracy becomes an ambiguous metric; inaccuracy of minority classes gets overshadowed by the accuracy of majority classes. Furthermore, in the context of driving, the more infrequent classes are often the most important for human safety. As a result, it is common to use the intersection over union metric (IoU). IoU measures the ratio between the amount of overlap between the predicted and ground truth pixels and the total number of pixels taken up by the prediction and the ground truth. See Figure 3 for a visual interpretation.

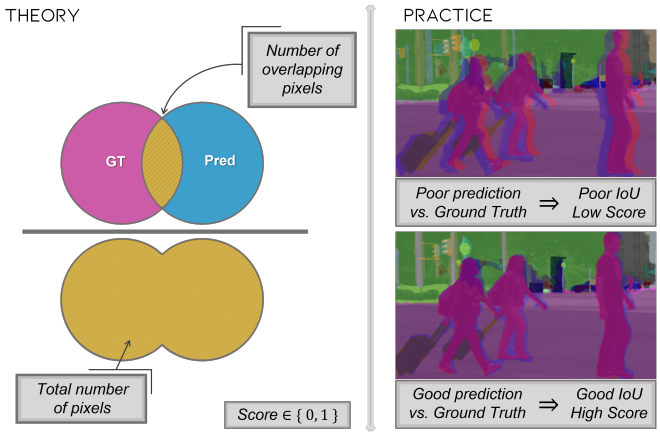

Fig. 3: Left pane: Visual demonstration of intersection over union calculation. Right Pane: Predicted segmentation masks from our trained DeepLabV3 model are overlaid on ground truth image masks to demonstrate realistic IOU scores.

It is customary to use the mean intersection over union (mIoU) which is quoted in this paper. However, during evaluation, we also compute and inspect the individual IoU per category to give a more granular understanding of the model performance. Scores theoretically fall between 0 and 1, although percentages are often quoted.

D. Comparing the Baseline and LNTL Schemes

All models are trained for 100 epochs until convergence. A learning rate of 0.001 is used with the Adam optimiser [45]. Learning rate decay is enforced with the scheduler reducing the learning rate by a factor of 0.1 every 40 epochs. Class weights are computed for all the training data so the cross-entropy loss function could allow for class imbalance, a common occurrence in urban driving scenarios.

Baseline models are trained for both SegNet and DeepLab, with inputs of colour training images and greyscale training images. Since the LNTL scheme penalises the classifier by using the gradient reversal module to adjust the weights, it was expected that the accuracy may suffer somewhat. We did, however, expect that the LNTL scheme would perform better than the networks making predictions using limited colour information, as is supplied in the greyscale training set.

This intuition is confirmed given the loss curves displayed in Figure 4; the LNTL scheme has penalised the loss compared to

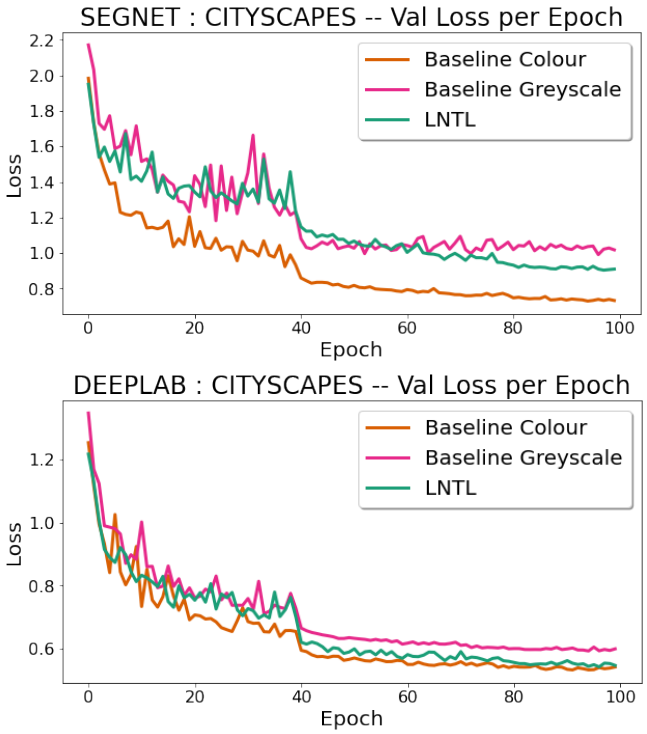

Fig. 4: Learning Not To Learn scheme vs. colour and greyscale baseline for SegNet and DeepLab segmentation architectures.

the converged colour input data for both SegNet and DeepLab architectures. Note that the SegNet minimum loss is 0.730, which DeepLab achieves after just 20 epochs of training. Of course, further image augmentation, extra training data and hyperparameter tuning could be used to drive down the loss even further and improve accuracy; however, to satisfy the project hypothesis, we only need a comparable canvas to test the concept of applying colour unlearning within the domain of urban scenes. In subsequent experiments, the Deeplab network is favoured for its slightly more stable learning, and higher accuracy.

E. Synthesising a Covariate Shift in Validation Data

In order to properly test our hypothesis that the LNTL scheme can unlearn colour information in highly variable scenes such as urban images, it is necessary to test both the baseline DeepLab architecture and the LNTL scheme on unseen data which contains a covariate shift from images it was trained on. To this end, three synthesised validation sets are created:

- converting the validation set to greyscale,
- applying random colour jitter to validation images,
- applying a colour invert transform to validation images.

Visual examples of these transformations are shown in Figure 5. Models are re-trained on the normal training set, both on SYNTHIA and Cityscapes datasets and only validated on the transformed images. All other parameters remain the same to allow for a fairer comparison. This enables us to monitor the network loss over the training cycle to see if the LNTL scheme slowly improves in its validation convergence. Monitoring loss in this way allows us to assess overall model health, since the network is technically focussed on *minimising* loss not *maximising* accuracy. Furthermore, during training, we can monitor the loss in the bias head to check for signs of divergence. This behaviour is welcomed and can be an indicator that colour information is being extracted out of the feature maps in network f. This makes categorisation in the bias head more difficult thus manifesting as a diverging loss in network h.

Intuitively, it is worth noting that although greyscale images remove a lot of the colour information, distances between pixel values can still be leveraged, synonymously for colour jitter, despite the stochasticity. Colour invert, however, maximises this difference and corrupts this relationship the most, disrupting spatial inter-correlations between different pixel values.

Figure 6 shows the disparity in loss between uncorrupted validation images and those with perturbed colour. Reference lines on the bottom of the graph highlight the extent to which the networks overfit to the colour in the training data. In the Cityscapes dataset, the LNTL scheme shows a clear improvement over the baseline method when colour invert is applied to the validation images. This suggests that the LNTL scheme has increased model robustness and has diminished its dependence on colour for categorisation. Although random colour jitter in Cityscapes validation has an increased loss, it appears to be converging on a downward trajectory.

Results are not as desirable for the SYNTHIA dataset. Firstly, we notice a significant reduction in disparity between uncorrupted and corrupted validation images. This may be due to the rich variety of luminance in SYNTHIA, as mentioned in Section IV-A. Colour invert seems to affect both schemes equally, whilst colour jitter is marginally favourable to the baseline scheme. As we may expect, jitter creates less of an impact on the loss than the invert in all cases.

Fig. 5: Input image manipulations.

F. Using Synthesised Weather Corruptions as a Proxy for Different Driving Conditions

From the Cityscapes data repository, researchers [46], [47] have created imitations of rain and fog over the normal Cityscapes training data. Ground truth labels are exactly the

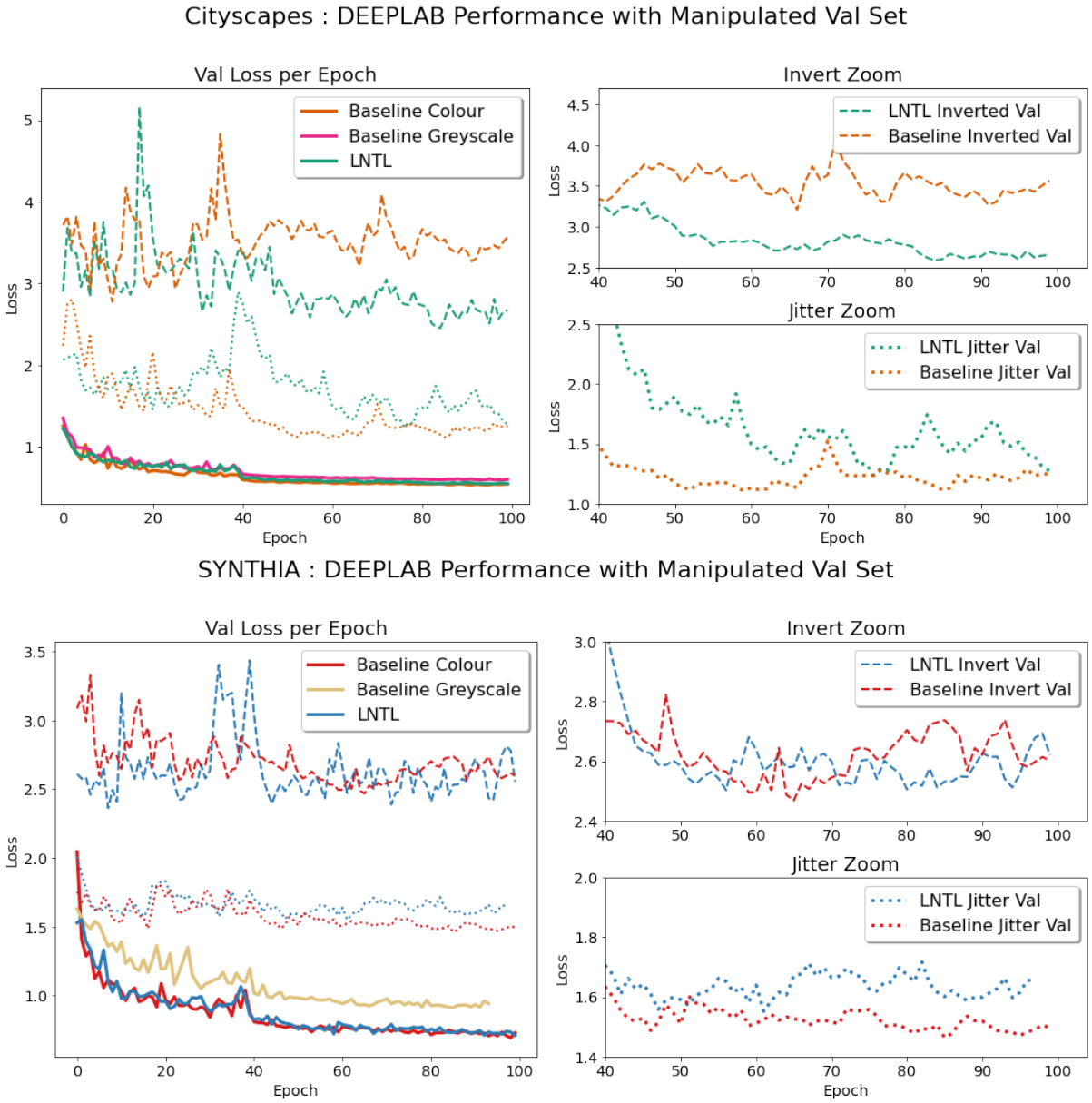

Fig. 6: Top Pane: Cityscapes dataset. Bottom Pane: SYNTHIA dataset. Large graph shows the validation results on the normal validation set and the results for manipulated validation sets with colour invert (dashed) and colour jitter (dotted). Adjacent are magnified and truncated plots to the last 40 epochs for the manipulations.

same as for the standard training data, so we can leverage this dataset to more closely resemble test images an autonomous vehicle may encounter in the wild. Different severities of image manipulation are provided. We randomly select one level of severity for each image, yielding a 295-image validation set for rain and 550-image validation set for fog.

Figure 7 uses the synthesised rain and fog validation images fed into the best trained model for each of the baseline and LNTL schemes. All prediction images contain more noise and demonstrate much poorer performance, as expected from the loss curves in Figure 6. Bounding boxes, denoting regions of interest, show that the LNTL scheme has managed to better segment the pedestrians in the first quadrant scene. The baseline model has produced nonsensical predictions, hallucinating pedestrians on the building in the scene. In the second quadrant "*Less Red More Tree*" we observe similar erroneous red patches - our debiasing method mitigates much of this and correctly identifies more tree through the synthesised fog. Similar pedestrian hallucinations can be seen in the tree in the fog in the third quadrant, and again the de-biasing scheme is more accurately able to segment the tree through the fog. Once more the fourth quadrant demonstrates the ability of the bias unlearning network to correctly classify a tree given a corrupted test image, where comparatively the baseline struggles. The fog

image corruptions perform better than the baseline on average (*see Table I*). Although both predictions contain noise and artefacts, results show a robustness to covariate shift at test time for the bias unlearning scheme. Table II shows congruence with qualitative observations with a clear improvement in the *Nature*, *Human* and *Vehicle* categories.

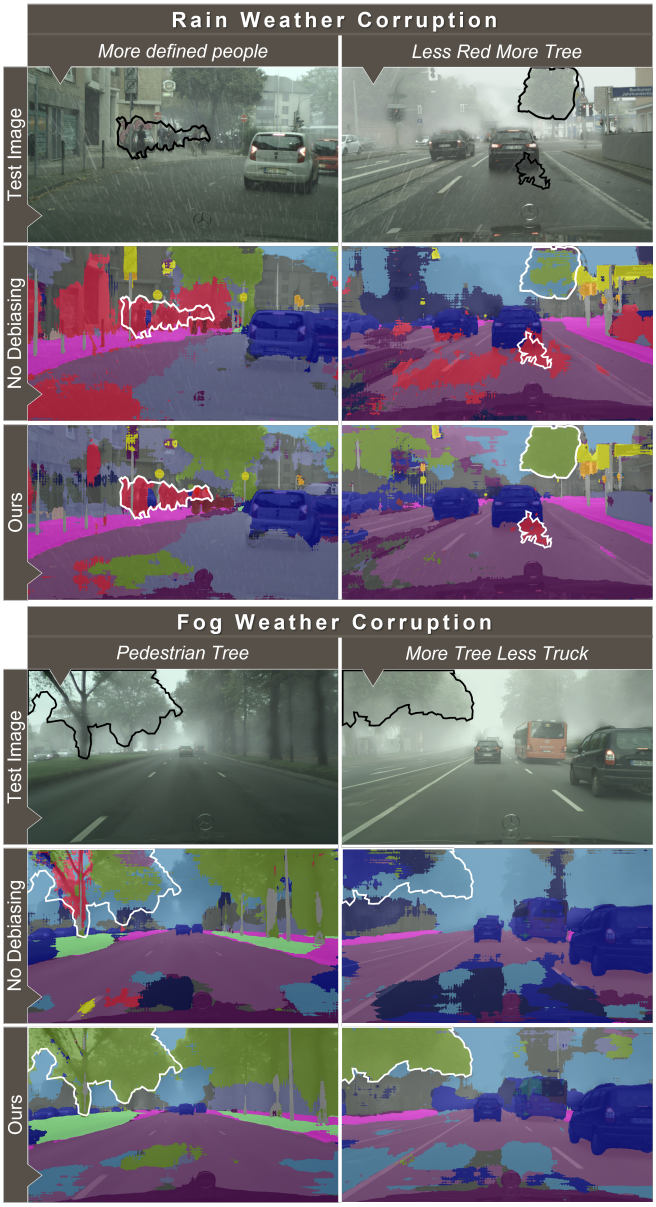

Fig. 7: Qualitative results from running the best models displayed in Figure 4 for the baseline and the LNTL models. Regions of interest are shown in boundaries.

G. Quantitative Results

A distinction between classes and categories is to be made; classes comprise the objects the model is seeking to predict used in the Softmax function of the network. In this case, we adopt the 20 Cityscape classes consisting of *Road*, *Sky*, *Car*, *Truck*, *Person*, *Rider*, etc. On the other hand, categories consist of groupings of classes – e.g. *Human*, *Vehicle*, *Nature* and so on. Table I provides average class-wise mIoU scores for both the DeepLabV3 baseline model and our bias unlearning model. Bold values highlight the best performers for each image transformation. The LNTL scheme not only performs marginally better on the original image, but it performs consistently better when validated on an out-of-distribution test image - only failing to beat the baseline in the *rain* validation set. This could be due to the relatively small size of the rain validation set compared to others tested, hindering model convergence.

TABLE I: Class-wise averages of minimum loss and mIoU validation results for a DeepLabV3 model trained on Cityscapes training set. Bold typeface indicates optimal results.

		mIoU (%)		**loss**	
		Baseline	*Ours*	Baseline	*Ours*
Original	RGB	58.50	**58.80**	**0.542**	0.546
Image Manipulation	Greyscale	36.20	**37.50**	1.173	**1.156**
	Invert	8.50	**13.70**	3.112	**2.423**
	Jitter	33.30	**34.10**	1.259	**1.180**
Weather Corruption	Rain	**39.40**	38.90	**1.190**	1.249
	Fog	52.80	**54.00**	0.873	**0.788**

The disparity between the manipulated validation sets and the RGB images empirically shows just how severe the lack of generalisability is - these results come from the same images only with colour corruptions added. This demonstrates that the networks *do* overfit to colour information available, highlighting the Occam's Razor nature of CNN learning, that is if the colour information is sufficient to drive down the loss, then why use another cue? Indeed, the results allow to quantify this degree of overfitting with an average reduction of 41.80% in mIoU score over all validation sets reported. The worst observed covariate shift is from RGB $\rightarrow$ Invert giving a mIoU reduction of 85.50%.

Table II shows the category-wise intersection over union scores providing a more granular understanding of the performance of the networks. The penultimate two rows display the average class scores and the performance increase, or decrease observed by the LNTL from the baseline. These values, and all others quoted in this paper, are displayed as a *percentage increase* from one percentage to the other, not simply a difference in percentages. Consistency is maintained in reporting value differences throughout our work.

Bold values in Table II highlight some interesting observations from these experiments. Firstly, the LNTL scheme performs consistently better in the "human", "nature" and "vehicle" classes than the baseline model. One subjective interpretation is that these categories in particular have a more unique, and largely unchanged geometry from scene to scene, i.e. the human form is distinctive and largely unchanged from individual to individual. Furthermore, the ASPP module of DeepLab handles objects at different scales. In contrast, the category-wise results show the model performing consistently worse at predicting the "flat" class - perhaps, in the same

TABLE II: Category-wise validation mIoU results for a DeepLabV3 model trained on Cityscapes training data. Observations of interest are highlighted in bold and discussed.

Categories	Normal		Invert		Jitter		Greyscale		Rain		Fog	
	Baseline	*Ours*	*Baseline*	*Ours*	*Baseline*	*Ours*	*Baseline*	*Ours*	*Baseline*	*Ours*	*Baseline*	*Ours*
Flat	93.6	93.1	32.8	55.8	80.3	74.6	88.7	85.7	71.9	55.1	94.3	92.3
Construction	85.0	85.4	20.5	40.5	56.6	64.1	61.9	64.4	55.6	47.5	72.6	73.7
Object	47.7	49.0	9.8	14.7	27.3	28.6	31.1	30.8	34.7	38.1	41.5	43.4
Nature	87.6	87.5	3.6	**34.6**	68.7	**72.0**	54.8	**56.2**	49.0	**58.5**	52.8	**58.1**
Sky	86.8	86.0	0.3	2.6	65.1	62.3	78.7	71.9	63.4	69.0	55.1	56.4
Human	66.5	**67.8**	13.9	10.1	19.3	**31.8**	29.8	**39.4**	48.5	**55.6**	63.3	**64.6**
Vehicle	86.3	**86.8**	17.0	**26.9**	38.6	**49.8**	49.5	**64.5**	46.9	**47.5**	80.8	**81.5**
Average	79.1	79.4	14.0	26.5	50.8	54.8	56.4	59.0	53.2	53.0	65.8	67.1
Ours ±%		+0.3		+89.3		+7.9		+4.6		-0.3		+2.0

vein, from buildings having no fixed geometry from scene to scene. This could mean that we have coerced the network to use more geometric-based cues to perform classification rather than colour. Quantitative results agree with qualitative observations. Furthermore, the debiasing scheme performs unanimously better in the "human" category than the baseline architecture. This is a promising result, given the safety-critical nature of the applications of segmentation technology.

V. Limitations and Future Work

Further work is needed to reinforce the reported findings. A deeper analysis of class-wise mIoU scores would provide more insight into precisely where the bias manifests within images. In particular, more granular understanding of the class-wise false-positives and false-negatives of predictions would be useful since in safety-critical applications such as autonomous vehicles, this is a vital requisite. This means that failing to correctly identify a pedestrian has attached with it more gravity than incorrectly classifying a pole, moreover - classifying a pedestrian as "road" has more consequence than classifying a pedestrian as a "rider".

Another consideration is analysing the extent to which augmentation techniques, e.g. [33], interact with the proposed bias unlearning scheme. Does the robustness offered by adequate augmentation reduce the performance observed in this work - or does it compliment it? In addition, urban scene data has high temporal dependence in the wild - a domain of active research and one which would be interesting to incorporate within our scheme.

Additionally, this project only focused on the mitigation of a known bias - colour. In the burgeoning world of big data, it is often unsurmountable to assess data for such bias. Furthermore, we are also at the mercy of our own biased representations. Amini et al. [48] tackle this issue with the use of variational autoencoders (VAEs). The proposed model actually learns, in an unsupervised manner, the latent structure of the input data and adaptively uses this learned latent distribution to selectively upsample underrepresented data points. This allows the model itself to determine the bias inherent within the data during training. Although Amini et al. demonstrate this technique through racial and gender bias in facial recognition systems, the idea itself is generalisable to multiple domains. We have shown that while colour bias does exist, the inter-correlation between variables in the input distribution may be more sophisticated than simply penalising a colour value.

VI. Conclusion

In this paper, we applied a colour bias unlearning scheme to highly variable images of urban road scenes as an iterative learning process during training. Our contribution empirically shows that semantic segmentation architectures do overfit to the colour within training data, and they struggle to generalise to unseen test data – even from a very similar input distribution, as seen in raw $\rightarrow$ weather manipulation experiments. In the worst case, when validating on a set with a colour invert transformation, reductions of 85.50% were observed. We demonstrate that the *unlearning* technique itself is viable, showing a qualitative improvement to both *stuff* and *things* classes in pixel-wise semantic segmentation, from a benchmark seminal architecture - mIoU metrics confirm this improvement. We observed a 62% increase in mIoU score for colour invert; when neglecting the result for colour invert, we still observe an average increase of 1.5% over all validation set manipulations tested. Furthermore, an average increase of 14.5% is observed for the "human" class, enhancing pragmatic performance in a safety-critical application such as autonomous driving. We position this paper to push towards robust, trustworthy technology - aiming for a transparent and explainable future in artificial intelligence, alleviating algorithmic bias.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2013.

[2] A. Atapour-Abarghouei and T. P. Breckon, "Real-Time Monocular Depth Estimation Using Synthetic Data with Domain Adaptation via Image Style Transfer," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018, pp. 2800–2810.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. Atapour-Abarghouei, S. Bonner, and A. S. McGough, "A King's Ransom for Encryption: Ransomware Classification using Augmented One-Shot Learning and Bayesian approximation," in *IEEE Int. Conf. Big Data*. IEEE, 2019, pp. 1601–1606.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, 2017, pp. 2481–2495.

[6] A. Atapour-Abarghouei, S. Bonner, and A. S. McGough, "Rank over Class: The Untapped Potential of Ranking in Natural Language Processing," *arXiv preprint arXiv:2009.05160*, 2020.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Euro. Conf. Computer Vision*, 2018, pp. 833–851.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Int. Conf. Machine Learning*, 2020, pp. 6105–6114.

[9] R. P. C. Kumar B. and Mohana, "YOLOv3 and YOLOv4: Multiple Object Detection for Surveillance Applications," in *Int. Conf. Smart Systems and Inventive Technology*, 2020, pp. 1316–1321.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017.

[11] R. C. Fong and A. Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation," in *IEEE Int. Conf. Computer Vision*, 2017, pp. 3449–3457.

[12] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic Segmentation," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.

[13] J. Hawke, V. Badrinarayanan, A. Kendall *et al.*, "Reimagining an Autonomous Vehicle," *arXiv preprint arXiv:2108.05805*, 2021.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Computer Vision*, 2017, pp. 2980–2988.

[16] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 12 472–12 482.

[17] A. Atapour-Abarghouei and T. P. Breckon, "Extended Patch Prioritization for Depth Filling within Constrained Exemplar-based RGB-D Image Completion," in *Int. Conf. Image Analysis and Recognition*. Springer, 2018, pp. 306–314.

[18] C. P. Papageorgiou, M. Oren, and T. Poggio, "A General Framework for Object Detection," in *Int. Conf. Computer Vision*. IEEE, 1998, pp. 555–562.

[19] A. Torralba and A. Oliva, "Depth Estimation from Image Structure," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1226–1238, 2002.

[20] A. Atapour-Abarghouei and T. P. Breckon, "A Comparative Review of Plausible Hole Filling Strategies in the Context of Scene Depth Image Completion," *Computers & Graphics*, vol. 72, pp. 39–58, 2018.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2015, p. 10.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[23] H. Noh, S. Hong, and B. Han, "Learning Deconvolution Network for Semantic Segmentation," in *IEEE Int. Conf. Computer Vision*, 2015.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2016.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2017, pp. 834–848.

[26] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical Multi-Scale Attention for Semantic Segmentation," *arXiv preprint arXiv:2005.10821*, 2020.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2017.

[28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[29] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *IEEE Int. Conf. Computer Vision*, 2017.

[30] A. Arnab, O. Miksik, and P. H. S. Torr, "On the Robustness of Semantic Segmentation Models to Adversarial Attacks," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018, pp. 888–897.

[31] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty," in *Int. Conf. Learning Representations*, 2020.

[32] J. Zhang, Y. Zhang, and X. Xu, "ObjectAug: Object-level Data Augmentation for Semantic Image Segmentation," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021.

[33] C. Kamann, B. Güssefeld, R. Hutmacher, J. H. Metzen, and C. Rother, "Increasing the Robustness of Semantic Segmentation Models with Painting-by-Numbers," in *Euro. Conf. Computer Vision*, 2020, pp. 369–387.

[34] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. Wichmann, and W. Brendel, "ImageNet-trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness," in *Int. Conf. Learning Representations*, 2019.

[35] P. T. Jackson, A. Atapour-Abarghouei, S. Bonner, T. P. Breckon, and B. Obara, "Style Augmentation: Data Augmentation via Style Randomization," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, p. 10.

[36] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning Not to Learn: Training Deep Neural Networks With Biased Data," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 9004–9012.

[37] M. Alvi, A. Zisserman, and C. Nellaaker, "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings," in *European Conf. Computer Vision Workshops*, 2018.

[38] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations," in *IEEE/CVF Int. Conf. Computer Vision*. IEEE, 2019, pp. 5309–5318.

[39] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324.

[40] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in *Int. Conf. Learning Representations Workshop*, 2014.

[41] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding Neural Networks Through Deep Visualization," in *Int. Conf. Machine Learning - Deep Learning Workshop*, 2015.

[42] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[44] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Int. Conf. Learning Representations*, 2015.

[45] D. P. Kingma and J. Ba, "Adam: A method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019.

[47] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic Foggy Scene Understanding with Synthetic Data," in *Euro. Conf. Computer Vision*, vol. 126, no. 9, 2018, pp. 973–992.

[48] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure," in *AAAI/ACM Conf. AI, Ethics, and Society*, 2019, pp. 289–295.