# Improving Robustness of Semantic Segmentation to Motion-Blur using Class-Centric Augmentation

Aakanksha
Indian Institute of Technology Madras
aakankshajha30@gmail.com

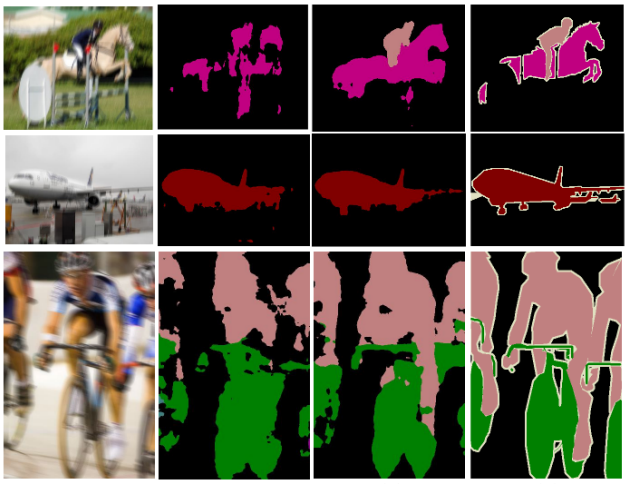
A. N. Rajagopalan
Indian Institute of Technology Madras
raju@ee.iitm.ac.in

## Abstract

*Semantic segmentation involves classifying each pixel into one of a pre-defined set of object/stuff classes. Such a fine-grained detection and localization of objects in the scene is challenging by itself. The complexity increases manifold in the presence of blur. With cameras becoming increasingly light-weight and compact, blur caused by motion during capture time has become unavoidable. Most research has focused on improving segmentation performance for sharp clean images and the few works that deal with degradations, consider motion-blur as one of many generic degradations. In this work, we focus exclusively on motion-blur and attempt to achieve robustness for semantic segmentation in its presence. Based on the observation that segmentation annotations can be used to generate synthetic space-variant blur, we propose a Class-Centric Motion-Blur Augmentation (CCMBA) strategy. Our approach involves randomly selecting a subset of semantic classes present in the image and using the segmentation map annotations to blur only the corresponding regions. This enables the network to simultaneously learn semantic segmentation for clean images, images with egomotion blur, as well as images with dynamic scene blur. We demonstrate the effectiveness of our approach for both CNN and Vision Transformer-based semantic segmentation networks on PASCAL VOC and Cityscapes datasets. We also illustrate the improved generalizability of our method to complex real-world blur by evaluating on the commonly used deblurring datasets GoPro and REDS .*

## 1. Introduction

Motion-blur has become ubiquitous in our lives driven largely by the compactness and affordability of light weight cameras. While camera quality has improved significantly, sensing technology cannot suppress blur completely yet. For handheld cameras and cameras mounted on moving vehicles, motion during capture is a major reason for the occurrence of blurred images. Most research in semantic segmentation especially those that are deep-learning based and state-of-the-art, focus on increasing accuracy [30], [38] and throughput [31], [23]. While these models have achieved significant gains, they are trained on clean data and consequently struggle to perform when presented with out-of-distribution data [13]. Such wrong predictions can have grave consequences for safety-critical applications like autonomous driving. Therefore, it becomes essential to focus on finding ways of making these models robust to unavoidable degradations like motion-blur.



(a) (b) (c) (d)

Figure 1. (a) Motion-blurred images, (b) Segmentation from a network trained on clean images (c) Segmentation after using our augmentation for training (d) Ground truth.

When trying to generalize to blurred images, an off-the-shelf approach would be to use a deblurring algorithm to deblur the image before continuing with the downstream task of segmentation. However, the generalization abilities of deblurring models are still subpar and they struggle to perform well for real blurred out-of-distribution images [37] . Additionally, many of these image restoration models process images at multiple-scales in a hierarchical fashion which improves the performance but also increases the latency and memory requirements [34,35]. Consequently, this two-stage approach of using deblurring as

pre-processing to obtain deblurred images before attempting segmentation is not viable for deployment and real-time applications. Hence, there is a strong need to devise single-stage methods that can bypass this step.

Some recent works have tried to focus on analysing and improving the robustness of existing models for different tasks to a spectrum of commonly encountered degradations. Unlike adversarial robustness, [11] defines robustness as the ability of a model trained on sharp images to retain competitive performance in images having degradations. [11,13,21] benchmark standard models for their robustness to multiple severity levels of sixteen commonly occurring degradations including motion-blur for the tasks of object recognition, object detection and semantic segmentation. Note that the motion-blur being considered here is spatially invariant and linear. [12] and [16] propose augmentation strategies to improve generic robustness of models across the sixteen degradations. While these augmentations have resulted in increased robustness, we believe that significant improvements can be made if degradation-specific and task-specific insights are exploited. [14] leverages the semantic segmentation annotation maps to increase the shape bias in the network which has been established to improve robustness [7]. However, this is only a task-specific augmentation and attempts to achieve improvements over all degradation types.

In this work, we attempt to make semantic segmentation robust to the presence of generic space-variant motion-blur. In particular, we develop a Class-Centric Motion-Blur Augmentation (CCMBA) strategy where we leverage the segmentation map annotations to introduce blur in specific regions of the image to enforce distinguishability and easier training. We randomly choose a subset of classes that we want to blur, blur the corresponding foreground image using a synthetic non-linear kernel, and then blend the blurred foreground image with the sharp background image. Since, motion-blur can be due camera ego-motion as well as dynamic scenes, the advantage of our augmentation strategy lies in better generalization to dynamic scenes due to its semantic class-centric nature. Fig. 1 shows the segmentation results obtained on motion-blurred images with and without our method.

Our contributions are the following :

- An effective data augmentation scheme for reliably segmenting out regions from motion blurred images without the need for deblurring.
- Our method is generic in nature and can be used with any supervised semantic segmentation network.
- While our model is trained on only synthetically generated data, the class-centric nature of our augmentation enables it to perform well on general dynamic blur datasets like GoPro and REDS, especially for common classes like humans.
- We report improved performance for DeepLabv3+ over baseline methods with 3.2% and 3% increase on PASCAL VOC and Cityscapes dataset, respectively, for the highest level of blur. We also achieve improvements on the Cityscapes-C dataset over previous works with a maximum 9% increase for highest levels of blur.

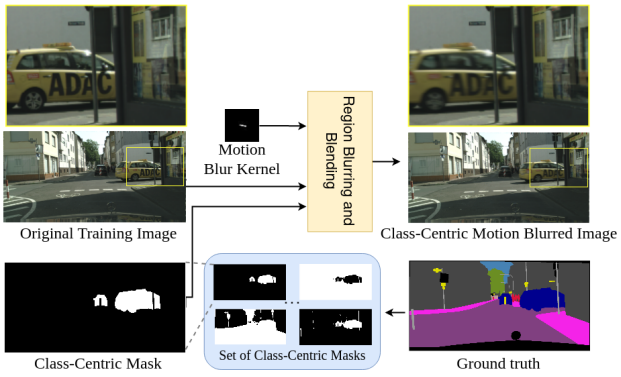



Figure 2. Class-Centric Motion-Blur Augmentation (CCMBA): Given a sharp image, its segmentation mask, and a motion-blur kernel, we synthetically blur the regions corresponding to a subset of classes present in the image to mimic dynamic scenes. When all classes are chosen, camera motion-blur is synthesized making our augmentation applicable for generic blur.

## 2. Related Works

### 2.1. Image Deblurring

In recent years, data-driven deep learning methods have achieved significant success in image deblurring and other low-level vision tasks using convolutional neural networks (CNNs). Earlier deep learning works [1, 29] estimated the blur kernel in order to obtain the deblurred image, drawing inspiration from traditional methods of deblurring. However, this method is not practical in real scenarios due to the complex nature of blur. DeepDeblur [22] proposed to directly map a blurry image to its sharp counterpart and this paradigm has been prevalent ever since. A series of state-of-the art models [33–35] have been encoder-decoder models with model sizes increasing with increasing performance. Some works also show that significant performance gains can be achieved by adapting transformer-based architectures for restoration tasks but these models are also bulky. These deblurring models while giving state-of-the-art performance on the training dataset, still struggle to generalize to real-world blurred images that are out-of-distribution [36]. Moreover, these methods are far from real-time. Some recent works [15, 18, 24] attempt to make deblurring more efficient and improve throughput but real-time deblurring with competitive performance remains elusive. Consequently, there is an imminent need for a single-stage approach that can retain competitive performance without the need for deblurring as a pre-processing step.

(a) (b) (c)

Figure 3. Space-invariant motion-blur for (a) Exposure level 1, (b) Exposure Level 2, (c) Exposure Level 3, with anxiety decreasing from left to right for each exposure level.

### 2.2. Robustness in Semantic Segmentation

Robustness of deep convolutional neural networks has been addressed in various benchmarks. Recent works, attempt to evaluate and increase the robustness of CNNs in various naturally-occurring degradations. [17] used defocus blur to reduce the impact of irrelevant background information for semantic segmentation. [28] examined the impact of blur on image classification and semantic segmentation using VGG-16 and found that using defocus blur augmentation leads to significant improvements in robustness for classification task but not for segmentation. Subsequently, [11] introduced the "ImageNet-C dataset" where the authors corrupt the ImageNet dataset by common image corruptions and benchmark various pre-trained models for robustness. [13] follows a similar methodology to benchmark segmentation models on Cityscapes-C and PASCAL-VOC-C. [5, 12, 32] improve the robustness of image recognition models to generic degradations using data augmentation. Based on the insight from [7] that deep neural networks trained on ImageNet seem to rely more on local texture instead of global object shape, [14] proposes a method to improve segmentation robustness to generic degradations by leveraging the semantic segmentation annotations. These methods cater to generic degradations and do not train on any specific degradation.

Motion-blur is a far more complex phenomenon than other degradations like Gaussian noise or shot noise due to its inherent dependency on scene complexity in terms of moving objects and occlusions, in addition to blur contributed by camera shake. This makes motion-blur a complex degradation to generalize to. Standard augmentation methods fail to take these intricacies into consideration.

We attempt to build an augmentation strategy that caters to all types of blur from space-invariant ego-motion blur to space-variant dynamic scene blur. Conventionally, soft segmentation in the presence of dynamic scene motion-blur has been modeled as an alpha-matting task. Given semantic segmentation masks, we synthetically model dynamic scene motion-blur by convolving segmentation maps of a set of randomly selected classes with a blur kernel to obtain an alpha-matte. We then blur the foreground image and blend it with the sharp background image to obtain a space-variant, class-centric motion-blurred image. When all the classes in the image are selected, we get space-invariant ego-motion blurred image. Since our method is tailored to handle generic blur it improves the robustness of segmentation models.

## 3. Methodology

In this section, we begin by describing the dataset synthesis process in Section 3.1 including a brief description of blur kernel generation which is followed by a detailed description of our Class-Centric Motion-Blur Augmentation (CCMBA) approach in Section 3.2.

### 3.1. Real or Synthetic Blur?

A major challenge to our task is the absence of appropriate datasets. To the best of our knowledge, no dataset exists that provides accurate annotations for semantic segmentation along with blurred and sharp pairs of real images. This restricts us to (i) capturing our own data and annotating it, or (ii) generating synthetic data for our experiments. The process of capturing and creating a new dataset with real, consistent blurred and sharp image pairs is non-trivial and requires special hardware setup [25]. An alternative approach is to capture high frame rate videos and average them to simulate synthetic motion-blur images akin to [22]. However, annotating sufficiently large number of sharp images for segmentation masks requires considerable time and effort. Hence, following existing works [8,11,13,20,26,36], we choose to generate synthetic data for our experiments.

Two possible approaches could be taken for synthesizing the dataset - (a) generate pseudo-ground truth segmentation maps for real blurred images by passing the corresponding sharp image through a pre-trained state-of-the-art semantic segmentation network, or (b) synthetically blur the images for which semantic segmentation map annotations are available. For the first approach, standard deblurring datasets like GoPro [22] can be used but the distributions of these datasets are significantly different from the distributions of segmentation datasets which renders them a poor choice. Additionally, if we use a pre-trained segmentation model, the quality of the resultant pseudo-ground truth segmentation maps is dictated by the generalizability of the pre-trained model across datasets. Training on such sub-optimal pseudo-ground truth segmentation maps would impose an upper limit on the performance that our model can achieve.

The second approach involves synthetically blurring standard datasets for segmentation like PASCAL-VOC, Cityscapes and MS-COCO [4, 6, 19] to obtain blur-sharp

pairs. The synthetic blur can be introduced in images in multiple ways. Some works advocate the use of Generative Adversarial Networks (GANs) because it enforces learning of more realistic blur [36]. But such methods do not provide any control or interpretability over the generated blur. Other works resort to a simpler approach and blur images by convolving them with blur kernels [26]. We too adopt a convolution-with-blur-kernels approach to synthesize different blurring situations including camera shake and dynamic scene blur.

To generate blur kernels, we employ the methodology described in [2] which is given in brief next.

#### 3.1.1 Blur Kernels Synthesis

The blur kernels generated in [2] are obtained by generating continuous camera motion trajectories and sampling them on a pixel grid. Two of the parameters that control the trajectories generated and are relevant to us are anxiety level (A) and exposure level (E). Anxiety level controls the amount of jerk and velocity changes in the trajectory. A higher anxiety level corresponds to more frequent changes in direction. Exposure level parameter models the exposure time of a camera. For our experiments, we consider a total of 3 exposure levels and 3 anxiety levels resulting in a total of 9 blur-severity levels. To synthesize a blur kernel corresponding to a particular exposure level '$e$' and anxiety level '$a$', we first generate a continuous trajectory with '$a$' as the parameter. To generate a trajectory, we consider starting velocity and position as $v_0$ and $x_0$ sampled from a uniform circle. Then, at every time step, the velocity is updated using the following acceleration vector,

$$\Delta v = a(\Delta v_g - Ix_t) + 2a|v|\Delta v_j, \tag{1}$$

where, '$a$' models the anxiety level, $\Delta v_g$ is random acceleration drawn from $N(0, \sigma^2)$, $Ix_t$ is a parameter modeling the inertia to retain position and $\Delta v_j$ models the jerk and is randomly sampled from a uniform circle. Subsequently, the generated trajectory is sliced off at time 'e' to simulate the varying exposure. The blur kernels are then computed by sampling the continuous trajectory on a regular pixel grid, using sub-pixel linear interpolation. Each kernel has a size of $32 \times 32$. The obtained motion-blur kernels are then centered by translating their barycenters to the center of the filters. We generate 12,000 blur kernels for every pair of anxiety and exposure values resulting in 108000 kernels for training. For evaluation, we generate another set of 108000 kernels and randomly sample from it. Anxiety levels are fixed at $A = [0.005, 0.001, 0.00005]$ and exposure times as $E = [1/25, 1/10, 1/5]$.

Note that [26] uses 5 levels of linear blur where the images blurred with kernels corresponding to higher exposure give rise to images that are not very commonly encountered in real-life. We argue that attempting to increase robustness to these less-frequently occurring and difficult-to-handle blur may not be necessary and can decrease the robustness of the model for the more relevant case of commonly occurring blurs. We restrict our experiments to exposure times corresponding to an approximate maximum of 15 pixels of blurring and use non-linear blur kernels to include the effect of handshake and jerks that can occur due to button press, etc. We define our blur levels as L1, L2 and L3 corresponding to different exposure times for our experiments with L3 corresponding to largest exposure time.


Figure 4. Examples of images generated by CCMBA. (a) Sharp image, (b) Image with aeroplane (top) and car (bottom) class blurred. (c) Image with only aeroplane (top) and only car (bottom) class sharp. (d) All classes blurred.

### 3.2. Class-Centric Blur Augmentation

Our goal is to improve the performance of semantic segmentation model in the presence of commonly encountered levels of blur while retaining comparable performance for sharp images. A simple approach to attempt would be to finetune pre-trained models on blurred images. Alternatively, the same model could be retrained using a combination of blurred and sharp images wherein the blurred images are sampled with a probability $p$. But, neither of these methods leverage the additional information that is available to us in the form of semantic segmentation annotations.

Conventionally, the task of alpha matting has been used to model transparent regions [10] as

$$I = \alpha.F + (1 - \alpha).B, \tag{2}$$

where $\alpha \in [0, 1]$, $F$ is the foreground object, and $B$ is the background. Motion-blur, especially, dynamic scene blur causes the pixels of the foreground object to smear along the boundaries and mix with the intensities of the background. Given a motion blurred image of a dynamic object, the task of segmenting out the motion blurred object can be achieved by computing the alpha matte [27].

Inspired by this, we attempt the inverse task of synthesizing class-centric motion blurred images by synthetically generating the blurred alpha matte for the selected classes using binary segmentation masks corresponding to those classes. The synthesized alpha matte is then used to mix the motion-blur of selected classes with the pixel intensities of the remaining classes. A brief overview of our approach is given in Fig. 2.

We refer to our approach as *Class-Centric Motion-Blur Augmentation* (CCMBA) strategy. The benefits of using such a strategy are two-fold. Firstly, blurring only the regions of an image belonging to a particular class simulates a dynamic scene being captured. As a result, the ability to generalize to complex blurs where space-variant dynamic scene blur is present in addition to camera shake is improved. Secondly, learning to segment out completely blurred images may be harder for the network as it may struggle to distinguish between regions having same colour but different texture because blurring leads to loss of texture. Blurring different image regions class-wise during training makes the network focus on improving performance specifically for the blurred region. As a result, over time, both sharp and blurred image features get learnt effectively. We now describe our CCMBA approach.

**Algorithm 1** Class-Centric Motion Blur Augmentation

**Input:** A tuple of sharp image $I_s$ and its semantic segmentation mask $M_s$ having $C$ semantic classes
**Output:** $I_{cblur}$ or $I_s$

Sample $p$ from $Unif(0, 1)$
**if** $p > 0.5$ **then**
  Randomly sample a blur kernel $K$
  $c \leftarrow$ Sample a number between 1 and $C$
  $c_{sub} \leftarrow$ Select a subset of $c$ classes of $C$
  $M_f$ = Sum($M_s[x]$ for $x$ in $c_{sub}$)
  $M_f[M_f > 0] = 1$
  $I_{cblur} = (M_f \cdot I_s) * K + (1 - (M_f * K)) \cdot I_s$
  return $I_{cblur}$
**else**
  return $I_s$
**end if**

Let $N$ be the total number of semantic segmentation classes in the dataset. Suppose we have an image $I_s$ of size $W \times H$ as input with $C$ semantic classes present in the image. We randomly select a $c < C$, as the number of classes to be blurred and sample $c$ of $C$ classes. Let this subset of selected classes be denoted as $C_{sub}$. Assuming a one-hot representation of the semantic segmentation annotations, the dimensions of the segmentation map is $W \times H \times N$. We combine their segmentation maps to get the foreground binary segmentation map $M_f$ by summing up along the channel dimension for classes in $C_{sub}$ followed by thresholding. $M_f$ is zero at all pixels which do not belong to the classes in $C_{sub}$ and is 1 otherwise. We blur this foreground map with a randomly selected (generated) blur kernel $K$ to obtain $M_{fb}$. Then we take the background masked version of the image $I_s \cdot M_f$ and blur it with the same kernel to obtain $I_{fb}$. To get the sharp background image $I_{bgs}$, we use $1 - M_{fb}$ to mask out the foreground regions from the sharp image using $I_s \cdot (1 - M_{fb})$. The final augmented image $I_{cblur}$ is then obtained by summing up the image with the blurred foreground $(I_{fb})$ and the image with the sharp background $(I_{bgs})$. These set of operations can be summarized as

$$I_{cblur} = (M_f \cdot I_s) * K + (1 - (M_f * K)) \cdot I_s \quad (3)$$

where '$\cdot$' denotes element-wise multiplication, '$*$' denotes convolution operation, $I_s$ is the sharp image, $K$ is the randomly selected blur kernel, $M_f$ is the binary mask corresponding to the set of $c$ classes to be blurred, and $I_{cblur}$ is the class-centric blurred image. Note that for each image, the probability of it undergoing CCMBA is given by $p$ while the sharp image is returned with $1-p$ probability in order to maintain performance on sharp images as well. We choose $p = 0.5$ for our experiments.

# 4. Experiments

In this section, we demonstrate the effectiveness of our CCMBA strategy for semantic segmentation in the presence of generic motion blur. We consider a standard CNN-based network, DeepLabv3+ [3], and a state-of-the-art vision-transformer-based network, Segformer [30], for our experiments and show results on two commonly used segmentation datasets – PASCAL-VOC [6] and Cityscapes [4]. We show that performance gains can be achieved using our augmentation scheme for any supervised segmentation method. We perform ablation studies to further support our claims.

In Section 4.1, we discuss the implementation details of our experiments followed by quantitative and qualitative results for space-invariant motion-blur and space-variant real motion-blur in Section 4.2 along with ablation studies.

## 4.1. Implementation Details

**Datasets:** We use two publicly available datasets in our experiments. PASCAL VOC [6] is a standard natural object segmentation dataset consisting of 21 classes with 10582, 1449, and 1456 images as training, validation and test splits. This includes additional training augmentation data according to [9]. Cityscapes [4] is an autonomous driving dataset having 19 classes with a total of 5000 high resolution images divided into 2975, 500, 1525 images for training, validation and testing. Cityscapes-C [13] dataset expands the Cityscapes validation set with 16 types of algorithmically generated corruptions. We generate the motion-blur subset of Cityscapes-C corresponding to severity levels S1, S2 and S3 for our experiments since it is not publicly available.
**Networks:** We select DeepLabV3+ [3] and Segformer [30] as networks for our experiments owing to their competitive performance and as representatives of their specific kinds of architectures. DeepLabV3+ [3] is a standard convolutional semantic segmentation model. It includes an atrous spatial pyramid pooling (ASPP) module which improved segmentation performance over previous works. Segformer [30] is a recent framework which unifies vision transformer

Table 1. Quantitative comparisons with baseline methods for DeepLabv3+ on PASCAL VOC and Cityscapes, and Segformer on Cityscapes for clean and blurred images with L1, L2, L3 blur levels. Best results are given in bold and second best are underlined.

| Method | VOC | | | | Cityscapes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DeepLabv3+ | | | | DeepLabv3+ | | | | Segformer | | | |
| | Clean | L1 | L2 | L3 | Clean | L1 | L2 | L3 | Clean | L1 | L2 | L3 |
| No-Retraining | **77.2** | 69.6 | 53.1 | 36.5 | 75.6 | 70.4 | 58.1 | 41.4 | 81.0 | 78.2 | 73.2 | 62.5 |
| Deblurring | - | 69.3 | 65.9 | 58.2 | - | 72.2 | 70.9 | 66.5 | - | 78.5 | 77.5 | 75.3 |
| Finetuning | 67.4 | 71.9 | 69.6 | 63.9 | 70.6 | 74.2 | 70.6 | 68.3 | 79.8 | **80.2** | **79.1** | **76.0** |
| MBA | 74.6 | 72.9 | 69.2 | 60.3 | 60.4 | 73.3 | 71.2 | 66.9 | 79.6 | 78.5 | 77.0 | 74.1 |
| CCMBA (Ours) | 76.5 | **74.6** | **72.1** | **66.0** | **76.2** | **75.6** | **73.6** | **70.4** | **81.1** | **80.2** | 78.7 | **76.0** |

backbones with lightweight multilayer perceptron (MLP) decoders to perform semantic segmentation and achieves state-of-the-art performance across datasets.

**Comparison Baselines:** Due to a dearth of previous works for comparison, we propose a set of comparison baselines to quantify the performance gains achieved by our method in terms of robustness to motion blur for semantic segmentation. Our first baseline is 'No-Retraining' where we consider a model trained on clean images and quantify its robustness to different levels of blur. The second baseline we compare with is 'Deblurring'. Using this baseline, we seek to benchmark the improvement that a deblurring preprocessing step can add for segmenting a blurred image. For experiments with this baseline, first, we blur our entire test set using blur kernels of one level. We then use a standard deblurring network, MPRnet [34], to deblur these images. The deblurred images are then passed through a semantic segmentation network trained on clean images to get performance corresponding to that particular level of blur. Our third baseline is 'Finetuning' where we want to make the network adapt to newer blurred images while retaining performance on previously seen sharp images. Towards this end, we first generate a set of blur kernels following Sec. 3.1.1. These are then used to blur each image in the training set. The same set of blurred images are used for finetuning the model pre-trained on clean images. Finally, our fourth baseline is 'Motion-Blur Augmentation' (MBA) where we seek to achieve robustness to motion blur by showing the network both clean and blurred images during training. Each sample is convolved with a randomly selected blur kernel with probability $p$ during training. Note that while such space-invariant motion-blur augmentation can model camera-motion blur, real generic blur is often composed of individual dynamic components which this augmentation scheme fails to take into account.

Table 2. Comparison with baselines and [13] on Cityscapes-C for DeepLabv3+ and Segformer. Best results are given in bold and second best are underlined.

| Method | DeepLabv3+ | | | | Segformer | | | |
|---|---|---|---|---|---|---|---|---|
| | Clean | S1 | S2 | S3 | Clean | S1 | S2 | S3 |
| No-Retraining | 75.6 | 71.2 | 65.7 | 56.9 | 81.0 | 77.8 | 74.6 | 68.8 |
| Finetuning | 70.6 | 72.4 | 70.4 | 68.1 | 79.8 | **78.0** | **76.0** | **73.1** |
| MBA | 60.4 | 57.9 | 54.5 | 50.8 | 79.6 | 77.4 | 75.3 | 71.9 |
| PbN [14] | 76.1 | 72.3 | 68.7 | 63.2 | - | - | - | - |
| CCMBA (Ours) | **76.2** | **74.0** | **72.3** | **68.9** | **81.1** | 77.7 | 75.9 | 72.3 |

We train DeepLabv3+ on PASCAL VOC and Cityscapes datasets and Segformer on Cityscapes dataset using our CCMBA strategy and compare their performances with each of the proposed baselines. To eliminate the need to repeatedly synthesize kernels for experiments, we take 12000 blur kernels corresponding to each of the 9 combinations of anxiety and exposure levels and save them before training. A separate set of such kernels is created for testing. Kernels are always randomly selected across blur levels unless specified otherwise. Images are blurred with reflection padding to ensure the naturalness of the image at the edges after blurring. For training the 'Finetuning' baseline, we reduce the base learning rate by a factor of 10. For training the MBA baseline and our CCMBA strategy, the image is blurred with a probability $p = 0.5$ with space-invariant and space-variant blurring, respectively. The evaluation metric used in our experiments is standard mean-Intersection-over-Union (mIoU). All models were written in Pytorch and were trained to convergence on NVIDIA GeForce RTX 3090 GPUs. A single GPU was used to train all models on PASCAL-VOC while two GPUs were used for Cityscapes.

For DeepLabv3+ on PASCAL VOC, we consider a batch size of 4 and crop size of $513 \times 513$ during training and a batch size of 16 with crop size of $768 \times 768$. For Segformer on Cityscapes, we consider a batch size of 16 with crop size of $1024 \times 1024$. In all our experiments, we use ResNet-50 as backbone for DeepLabv3+ and MiT-B2 as backbone for Segfomer unless specified otherwise. All the other hyperparameters and training setup are the same as those in the original papers.

## 4.2. Results

To establish the increased robustness of models trained with our augmentation, we compare quantitative results with a set of baseline methods and [14] for space-invariant blur in Table 1. We show qualitative comparisons for space-invariant blur in Fig. 5. Since no works exist that investigate the robustness of semantic segmentation for real dynamic scene motion-blur, we compare our qualitative results obtained for blurred and sharp images from GoPro and REDS dataset with the baseline methods in Fig. 6.

### Quantitative Results

We provide comprehensive comparisons with baseline methods in Table 1 for DeepLabV3+ on PASCAL VOC and Cityscapes dataset, and on Cityscapes dataset for Segformer. A general trend in performance is observed across all experiments. The 'No-Retraining' baseline performance drops severely as we move from blur levels L1 to L3. While 'Deblurring' baseline shows improved performance in the presence of blur, there is still a significant drop especially at blur levels L2 and L3. The 'Finetuning' baseline improves the performance across blur levels L1, L2 and L3 at the cost of performance drop for clean images which is not desirable. The 'MBA' baseline shows improved performance across all levels of blur while retaining performance for clean sharp images. Our CCMBA achieves performance gains of 2.3%, 3.5% and 3.2% over the best performing baseline for L1, L2, L3 blur levels for DeepLabv3+ on PASCAL VOC while retaining performance on clean images. For Cityscapes, performance gains of 1.9%, 3.4% and 3.1% are observed over the best performing baseline for L1, L2 and L3 blurs respectively while performance is retained for clean images. For Segformer, our CCMBA achieves comparable performance to the best case performance achieved by any baseline for each level of blur.

To establish the robustness of our approach, we show results for models trained using our strategy on the motion-blur subset of Cityscapes-C in Table 2. Compared to [14], our method achieves 2.3%, 5.2% and 9% gains on DeepLabv3+ over blur levels S1, S2 and S3. Note that this subset of blurring kernels has not been used for training our model. Our method also achieves 1.7%, 5.1 % performance gains over pre-trained Segformer for S2 and S3 and 2%, 0.5% gains over MBA baseline for Clean and S3 blur level while retaining comparable performance for the rest.

### Qualitative Results

Qualitative comparisons with baseline methods for PASCAL VOC dataset for space-invariant blur is given in Fig.5. Our results are consistent across clean and blurred images with minor deviations at highest blur level. Our approach is also able to segment out finer details better like the side-wing of the aeroplane in Fig.5 (a), the handle of the cycle in Fig.5 (b) and the leg of the rider in Fig.5 (c). Qualitative comparisons for space varying real blur is given in Fig.6 for one image from GoPro and two images from REDS. Our results are clearly more consistent between sharp and blur images. To appreciate the benefits of our method, for Fig.6 (a), compare the segmentation map regions corresponding the girl in foreground. Our approach gives a much more accurate segmentation map than others. For Fig.6 (b), zoom in on the couple in the center to appreciate our consistently better performance for blurred images. Note that the buses category is also getting predicted better across both blur and sharp images for our method while other works get confused. In Fig.6 (c), other methods are not able to segment out some humans in the blurred image, while our method is able to segment them out in both blurred and sharp images. The better generalization of our approach to real blur can be attributed to its class-centric nature where the network sees different subsets of class regions blurred at different instants of time during training, which effectively models dynamic scene blur along with egomotion blur.

### Ablations

Through our ablation studies, we investigate the effect of model parameter initialization, backbone size and the blur levels considered during training on our augmentation strategy. To investigate the effect of parameters initialization on our augmentation strategy, we use DeepLabv3+ with PASCAL VOC. We first train a the model from scratch using CCMBA with random initialization. We then repeat the same experiment with initialization from clean pre-trained network weights. Results in Table 3 establish that finetuning a pre-trained network using our augmentation scheme gives better results. We perform ablation studies for DeepLabV3+ on PASCAL VOC for MobileNetV2, ResNet-50 and ResNet-101 backbones. The consistent increase in performance with increasing backbone size in Table 4 shows the effectiveness of our approach. To demonstrate the necessity to train on all 3 levels of blur together in order to achieve best performance, we train our MobileNetV2 on PASCAL VOC for each blur level individually. The results in Table 5 clearly show that training with all 3 levels of blur is necessary.

Table 3. Ablation studies on network initialization.

| Initialization | mIoU | | | |
|---|---|---|---|---|
| | Clean | L1 | L2 | L3 |
| Random | 75.5 | 73.6 | 69.6 | 61.5 |
| Pre-trained | **76.5** | **74.6** | **72.1** | **66.0** |

Table 4. Ablation studies illustrating the applicability of our augmentation across network size.

| Backbone | Params | mIoU | | | |
|---|---|---|---|---|---|
| | | Clean | L1 | L2 | L3 |
| MobileNetV2 | 3.4M | 69.3 | 68.2 | 65.6 | 61.5 |
| ResNet-50 | 25.6M | 76.3 | 75.3 | 72.5 | 68.6 |
| ResNet-101 | 44.5M | **77.6** | **76.3** | **75.2** | **71.3** |

Table 5. Ablation studies for blur levels during training. Best results are given in bold and second best are underlined.

| Training Data | mIoU | | | |
|---|---|---|---|---|
| | Clean | L1 | L2 | L3 |
| Clean | **70.1** | 60.3 | 44.4 | 29.0 |
| L1 | 66.4 | 65.1 | 58.7 | 39.1 |
| L2 | 66.4 | <u>65.6</u> | <u>63.2</u> | 53.5 |
| L3 | 64.5 | 64.1 | 61.9 | <u>58.6</u> |
| L1, L2, L3 | <u>69.3</u> | **68.2** | **66.6** | **61.5** |

**Supplementary** We include in our supplementary more qualitative results for both PASCAL VOC and Cityscapes.

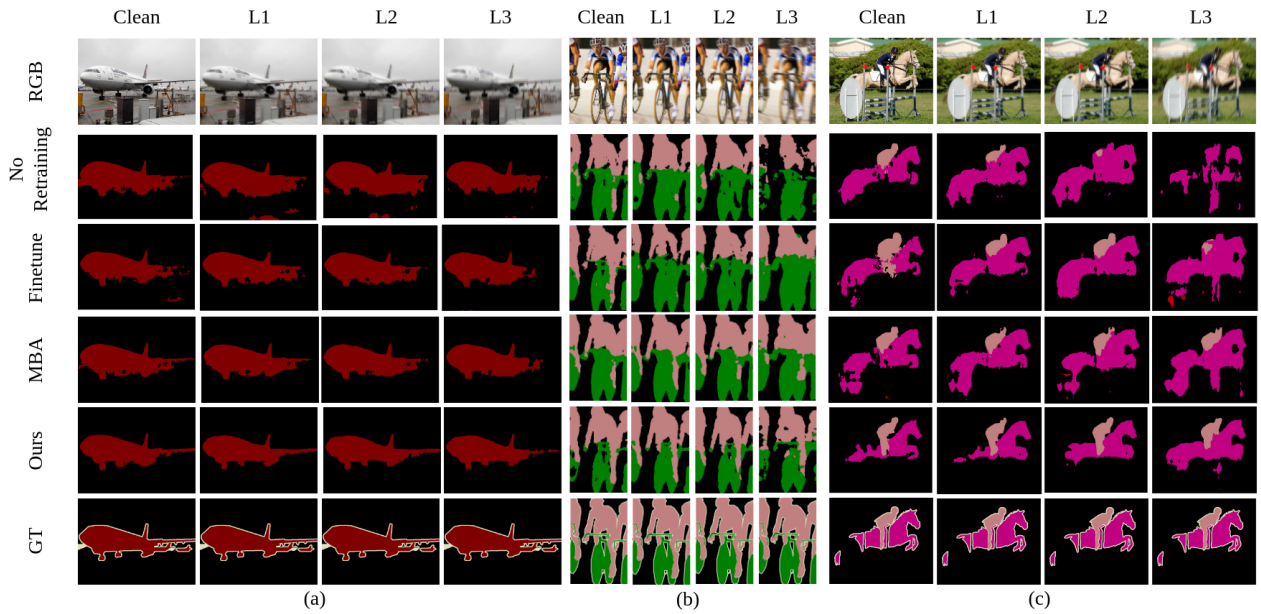



Figure 5. Qualitative results for space-invariant motion blur for DeepLabv3+ on PASCAL VOC. Note that our method CCMBA captures finer details better than all baselines.

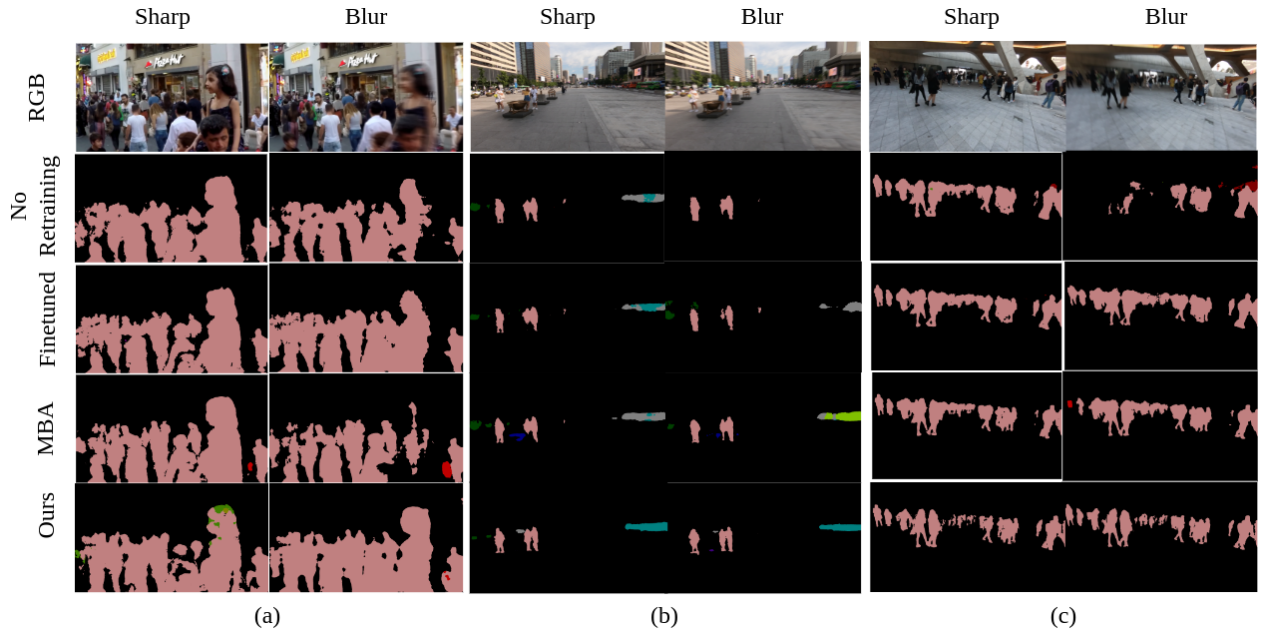



Figure 6. Qualitative comparisons with baselines for space-varying real blur on GoPro and REDS dataset.

We also include a detailed sensitivity analysis for each class and blur level alongwith an ablation study for the probability parameter '$p$'. We also include an ablation for the role of non-linear blur in our improved performance.

## 5. Conclusions

In this work, we attempted to improve the robustness of semantic segmentation performance in the presence of generic motion blur. We developed a class-centric motion-blur augmentation strategy inspired by the alpha-matting modeling of blur in literature. We selected a subset of classes and generate a synthetic alpha-matte by blurring the corresponding segmentation maps and subsequently blend it with the sharp background. Experiments demonstrate the effectiveness of our approach and its applicability to improving the robustness of any supervised semantic segmentation approach without any increase in model parameters and inference time.

## References


[1] Yuval Bahat, Netalee Efrat, and Michal Irani. Non-uniform blind deblurring by reblurring. In *Proceedings of the IEEE international conference on computer vision*, pages 3286–3294, 2017. 2

[2] Giacomo Boracchi and Alessandro Foi. Modeling the performance of image restoration from motion blur. *IEEE Transactions on Image Processing*, 21(8):3502–3517, 2012. 4

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 5

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 3

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007. 3, 5

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2, 3

[8] Dazhou Guo, Yanting Pei, Kang Zheng, Hongkai Yu, Yuhang Lu, and Song Wang. Degraded image semantic segmentation with dense-gram networks. *IEEE Transactions on Image Processing*, 29:782–795, 2019. 3

[9] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998, 2011. 5

[10] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR 2011*, pages 2049–2056. IEEE, 2011. 4

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 2, 3

[12] 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. 2, 3

[13] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8828–8838, 2020. 1, 2, 3, 5, 6

[14] Christoph Kamann and Carsten Rother. Increasing the robustness of semantic segmentation models with painting-by-numbers. In *European Conference on Computer Vision*, pages 369–387. Springer, 2020. 2, 3, 6, 7

[15] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[16] Jin-Ha Lee, Muhammad Zaigham Zaheer, Marcella Astrid, and Seung-Ik Lee. Smoothmix: a simple yet effective data augmentation to train robust classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3264–3274, 2020. 2

[17] Hao Li, Changjiang Liu, and Anup Basu. Semantic segmentation based on depth background blur. *Applied Sciences*, 12(3), 2022. 3

[18] Yuelong Li, Mohammad Tofighi, Junyi Geng, Vishal Monga, and Yonina C. Eldar. Efficient and interpretable deep blind image deblurring via algorithm unrolling. *IEEE Transactions on Computational Imaging*, 6:666–681, 2020. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[20] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Unsupervised domain-specific deblurring via disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10225–10234, 2019. 3

[21] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2

[22] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, July 2017. 2, 3

[23] Yuval Nirkin, Lior Wolf, and Tal Hassner. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4061–4070, 2021. 1

[24] Kuldeep Purohit and AN Rajagopalan. Region-adaptive dense network for efficient motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11882–11889, 2020. 2

[25] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020. 3

[26] Mohamed Sayed and Gabriel Brostow. Improved handling of motion blur in online object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1706–1716, June 2021. 3, 4

[27] Lin Hai Ting, Tai Yu-Wing, and Michael S Brown. Motion regularization for matting motion blurred objects. In *ACM SIGGRAPH 2010 Talks*, pages 1–1. 2010. 4

[28] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016. 3

[29] Subeesh Vasu, Venkatesh Reddy Maligireddy, and AN Rajagopalan. Non-blind deblurring: Handling kernel uncertainty with cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2018. 2

[30] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 5

[31] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 1

[32] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[33] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2

[34] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 1, 2, 6

[35] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[36] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 2, 3, 4

[37] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022. 1

[38] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, 2022. 1