Robust Scene Inference under Noise-Blur Dual Corruptions

Bhavya Goyal, Jean-François Lalonde, Yin Li, Mohit Gupta

Abstract—Scene inference under low-light is a challenging problem due to severe noise in the captured images. One way to reduce noise is to use longer exposure during the capture. However, in the presence of motion (scene or camera motion), longer exposures lead to motion blur, resulting in loss of image information. This creates a trade-off between these two kinds of image degradations: motion blur (due to long exposure) vs. noise (due to short exposure), also referred as a dual image corruption pair in this paper. With the rise of cameras capable of capturing multiple exposures of the same scene simultaneously, it is possible to overcome this trade-off. Our key observation is that although the amount and nature of degradation varies for these different image captures, the semantic content remains the same across all images. To this end, we propose a method to leverage these multi exposure captures for robust inference under low-light and motion. Our method builds on a feature consistency loss to encourage similar results from these individual captures, and uses the ensemble of their final predictions for robust visual recognition. We demonstrate the effectiveness of our approach on simulated images as well as real captures with multiple exposures, and across the tasks of object detection and image classification. Project: https://wisionlab.com/project/noiseblurdual

Index Terms—Low Light, Motion Blur, Scene Inference, Object Detection, Image Classification

1 Introduction

Imaging trade-offs are a fundamental characteristic of any computer vision system, and they are often exacerbated when the imaging conditions are challenging. One such challenging condition is low-light and motion (scene or camera). In such conditions, various types of corruptions are bound to be present in the image, and one can only compromise between them without ever removing them completely. For example, low light could cause the images captured by the camera to exhibit strong noise. While it is possible to mitigate noise by capturing longer exposures (or larger apertures), this often results in strong motion (or defocus) blur, leading to another kind of image quality degradation. Hence, noise and blur represent "dual corruptions"—reducing one (*e.g*, by adjusting the exposure) necessarily increases the other.

> *All happy families are alike;*
> *each unhappy family is unhappy in its own way.*
>
> Leo Tolstoy

Consider a set of images captured under low-light at varying exposures (Figure 1), thereby spanning the space of noise-blur "dual corruptions". Each image, being corrupted in its own way, offers a different "window" on the scene: moving objects will appear sharper when the exposure is lower, while static low-contrast regions will be more easily perceptible in longer exposures. In order words, while any single image from the set might never be optimal in challenging scenarios, the *set of images spanning the dual corruption space* contains much richer and complementary information that can be leveraged for performing robust scene inference even under challenging imaging scenarios.

In this paper, we propose the idea of performing scene inference in the space of noise-blur corruption. Our *key observation* is utilizing the "persistence of prediction" across differently degraded images of the same scene, significantly higher accuracy can be achieved as compared to performing inference on individual images. Figure 1 shows an example. Although differently degraded images have different low-level features, the semantic content remains the same across all images. We develop techniques that encourage similar predictions from individual captures, and aggregate the predictions across individual images for robust visual recognition.

We demonstrate the proposed approaches on two visual recognition tasks, namely image classification and object detection. We perform experiments on large scale datasets of real images with synthetic corruptions and show that performing inference on a set of dual corruption images outperforms conventional baselines in extreme low-light and motion conditions. Finally, we also show improved performance on real-world experiments using machine vision sensors.

Scope and Limitations: While implementing this idea requires capturing multiple exposures, most modern cameras already allow varying imaging parameters (*e.g*, exposure, aperture) in rapid succession. For example, modern cell phone cameras can take multiple snaps with a variety of exposures and fuse them to create an aesthetically pleasing image [1]. Increasingly, machine vision sensors [2] are also starting to perform exposure bracketing to capture high dynamic range (HDR) images for autonomous driver assist systems, while others go further and offer the capability of *simultaneously* capturing different exposure images via a spatially varying exposure sensor for HDR imaging [3] and

- *B. Goyal, Y. Li and M. Gupta are with University of Wisconsin-Madison, Madison, WI, 53715.*
  E-mail: bhavya@cs.wisc.edu
- *J. Lalonde is with Université Laval, Québec, QC, Canada*

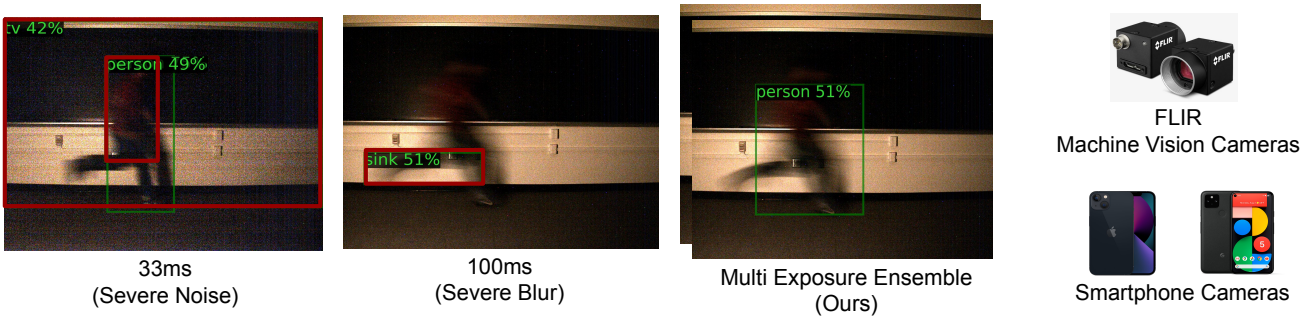


Fig. 1: **Multi Exposure Ensemble:** Figure shows a scene containing a fast moving object under low-light. Images with short exposure (a) and long exposure (b) suffer from dual corruption: noise and/or blur. Inference tasks like object detection on these images are severely affected: numerous false positives and wrong bounding boxes. (c) Our approach leverages multiple captures of varying exposures for robust inference: accurate and tight bounding boxes. (d) Such multi-exposure images are easy to capture with machine vision cameras or modern smart phone cameras (eg. Google pixel and iPhone) that uses burst photography for HDR imaging.

motion-deblurring [4]. These ongoing developments in camera technology, coupled with the proposed computational techniques can lead to the next generation of computer vision systems which will perform reliably even in non-ideal real-world scenarios (*e.g*, imagine an autonomous car driving on a dark night attempting to detect pedestrians) where it is extremely challenging for conventional algorithms to extract meaningful information reliably.

2 Related Work

Image Corruptions and Benchmarks. There has been some recent interest in simulating common image corruptions and benchmarking their adversarial effect on the performance of computer vision models, especially those relying on deep models [5], [6]. In parallel, developing robust visual inference methods has also received much attention. For example, a teacher-student framework was proposed [7] to improve image classification performance. Several noise and corruption models have been considered, including both physics-based [8] and learning-based [9]. Efforts in capturing real datasets of noisy images have also been pursued. A dataset of images captured in low light with annotations for object detection [10] has been collected. Another examples is the dataset containing low-light and corresponding well-lit cellphone images for denoising [11], which has recently been extended to videos in [12]. Most previous works simulate or collect real captures with image degradations like noise in low-light, but we consider a more challenging and practical setting where both low-light and motion are presented, and hence dual image degradation come into play.

Noise Removal and Deblurring. Due to its importance in image processing, denoising and/or deblurring degraded images has been a very popular topic for decades. Recently, numerous works have been proposed using neural networks for deblurring [13], [14], [15] and denoising [16]. For example, sparse denoising auto-encoder was considered for robust denoising [17]. A recent line of work proposes to perform joint denoising and inference on noisy images [18], [19], [20]. While existing image restoration methods can obtain high quality reconstructions, performing inference directly on the corrupted images does not require any pre-processing and is thus more efficient and as we demonstrate, can achieve increased robustness under severe image degradation. Alternatively, other methods aim to design cameras that produce better images directly, either by optimizing the hyperparameters of existing image signal processors (ISP) [21] or, by designing novel ISPs [22], [23], [24], [25]. These methods may, however, not entirely remove noise in challenging low-light situations, due to the fundamental limitation of the optics and sensors.

Inference on Corrupted Images. Many recent works tackle different inference tasks *directly* on images with common corruptions. Rozumnyi *et al.* [26] proposed a matting and deblurring network for faster inference for the detection of fast moving objects in videos. Cui *et al.* [27] designed a multitask auto-encoder for image enhancement, which leverages a physical noise model and ISP setting in a self-supervised manner to improve detection performance. Wang *et al.* [28] presented a framework for monocular depth estimation under low-light using self-supervised learning and demonstrate their results on nighttime datasets. Others have used knowledge distillation techniques for image classification under low-light [29], or for object detection by leveraging bursts of short exposure frames [30]. Goyal *et al.* [31] used a single photon camera and proposed to train on a wide spectrum of images at various SNR, with encouraging results on image classification and monocular depth estimation. Song *et al.* [32] introduced a technique for image matching using local descriptors and initial point-matching methods for extremely low-light images in RAW format. Wang *et al.* [33] proposed to learn the mapping relationship between representations of low and high quality images, and used it as a deep degradation prior (DDP) for image classification on degraded images. Adversarial Logit Pairing [34] also provides some robustness to the inference on noise and blur corruptions [5] by matching logits output of clean image with adversarial perturbed image.

Our goal is different from all previous approaches. We propose techniques that leverage the space of noise-blur dual corruptions rather than looking at a single image

corruption. We show that our approach is versatile for several downstream tasks, including image classification and object detection.

Leveraging Multiple Captures. Multiple exposures can be used to reconstruct high dynamic range (HDR) images [35], even in the presence of motion [36], [37]. Hasinoff *et al.* [38], [39] proposed ways to select settings for these multiple captures like ISOs and focus settings. The popularity of mobile photography has led to the further development of burst photography [1], which has been used for denoising [40], deblurring [41], [42], and super-resolution [43]. In sharp contrast, we exploit multiple exposures for high-level inference tasks such as classification and detection, rather than low-level image reconstruction.

3 Dual Corruption Space

Noise Blur Trade-off in Image Formation. We consider the relationship of noise and blur with the exposure time under low-light conditions and in the presence of scene (or camera) motion. The number of photons incident at a given pixel during a short exposure time is small under low-light conditions. Because of this, noise becomes dominant in the captured images and has to be properly modeled.

In the presence of scene/camera motion, let the photon flux (photons/second) at a pixel p on time t be $\phi_{p,t}$. The key is to consider that the incident flux at each pixel changes over time t, since the pixel may image different scene points due to scene/camera motion, resulting in an image x with motion blur. Assuming an exposure time Δt and a linear camera with quantum efficiency η, the raw reading at pixel p (without quantization) is given by

$$I_p = \int_0^{\Delta t} \phi_{p,t}\eta \; dt + z_p \tag{1}$$

where z_p is the noise at pixel p. Here we ignore the non-uniformity of photon response and noise [44], and consider three sources of noise.

- *Shot noise* z_p^s refers to the inherent natural variation of the incident photons due the Poisson process of photon arrival $\mathcal{P}$ and is modelled as the square root of the signal. Therefore, $z_p^s \sim \mathcal{P}\left(\int_0^{\Delta t} \phi_{p,t}\eta \, dt\right)$.
- *Readout noise* z_p^r comes from the process of quantizing the electronic signal as well as electrical circuit noise, which is modelled as a zero mean Gaussian with variance σ_r^2 at each readout. Namely, $z_p^r \sim \mathcal{N}(0, \sigma_r^2)$.
- *Dark current* z_p^d arises due to thermally generated electrons and also follows a square root relationship with signal with a variance of σ_d. We thus have $z_p^d \sim \mathcal{P}(\sigma_d \Delta t)$.

We further assume that z_p^s, z_p^r, and z_p^d are independent of each other, and follow an additive noise model [45], such that $z_p = z_p^s + z_p^r + z_p^d$ [44]. Thus, $\mathrm{Var}(z_p) = \mathrm{Var}(z_p^s) + \mathrm{Var}(z_p^r) + \mathrm{Var}(z_p^d)$. This leads to the derivation of the signal-to-noise ratio (SNR) for the captured images, given by

$$\mathrm{SNR} = \frac{\left(\int_0^{\Delta t} \phi_{p,t}\eta \; dt\right)^2}{\int_0^{\Delta t} \phi_{p,t}\eta \; dt + \sigma_r^2 + \sigma_d \Delta t}. \tag{2}$$

Under the presence of both low-light and motion, longer exposure time leads to improved SNR, as the noise increases slower than the signal. This, however, comes at a cost of increased motion blur in the captured images due to the integral of the incoming flux $\phi_{p,t}$. Hence, the exposure time allows us to trade off noise and blur in the image degradation space, which we term as *Dual Corruption Space*.

Dual Corruption. Our key idea is to leverage the spectrum of dual-corruption images by varying the camera parameters, resulting in a set $\mathcal{I} = \{x_1, ...x_N\}$ of images with different low-level characteristics (*e.g*, different amounts of blur and noise). For example, varying exposure time Δt creates a sequence of images where noise gradually decreases but the amount of blur increases. An example such sequence is shown in Figure 1. Since these images are captured simultaneously (or in rapid succession), we can assume that they have similar semantic content.

An Image without Noise and Blur. A special and theoretically interesting case in the dual corruption space is an *ideal clean image* x_{clean} captured using a very short exposure time ($\Delta t \to 0$) and without noise corruption ($z_p = 0$). Such an image is free of noise and blur. Despite physically implausible, this construct is sometimes convenient for our derivations.

4 Scene Inference under Noise-Blur Dual Corruptions

We consider scene inference tasks represented as an inference module $f(x) \equiv g \circ \varphi(x)$, where, without loss of generality, $\varphi(x)$ is a feature extractor, and g is a prediction module. Here, $\circ$ is the composition operator. $f(x)$, oftentimes represented by a neural network, maps an input image x into its semantic label y. This generic formulation covers several vision recognition tasks, including image classification where y is a categorical label, and object detection where y is a set of labeled bounding boxes. We further assume that this function $f(\cdot)$ is learned from data by minimizing a certain loss function.

Given a set of N noise-blur dual corruption images $\mathcal{X} = \{x_1, ...x_N\}$ capturing the same scene, our key intuition is that despite differences in low-level *image* features (*e.g*, pixel values), their *latent* features should remain similar. In what follows we formulate this intuition as a data prior, devise the training and inference schemes, and demonstrate interesting properties of the resulting method.

4.1 Robust Inference with Multiple Exposures

A simple prior is to assume that the latent features $\{\varphi(x_1), ...\varphi(x)_N\}$ follow a Gaussian distribution, centered at the ideal clean image x_{clean} and with a small variance ϵ^2. This prior ensures that with high probability the distance between any pair of latent features will stay in a small ℓ_2 radius controlled by ϵ^2. With such assumption, we arrive at the following conditional probability $p(y|x)$ for scene inference.

$$p(y|x) \propto p(y|\varphi(x))p(\varphi(x)|x), \tag{3}$$

where $p(y|\varphi(x))$ is given by the prediction module g, and $p(\varphi(x)|x) \sim \mathcal{N}(\varphi(x_{clean}), \epsilon^2)$ represents the data prior. We now describe the training and inference schemes based on this formulation, as illustrated in Figure 2.

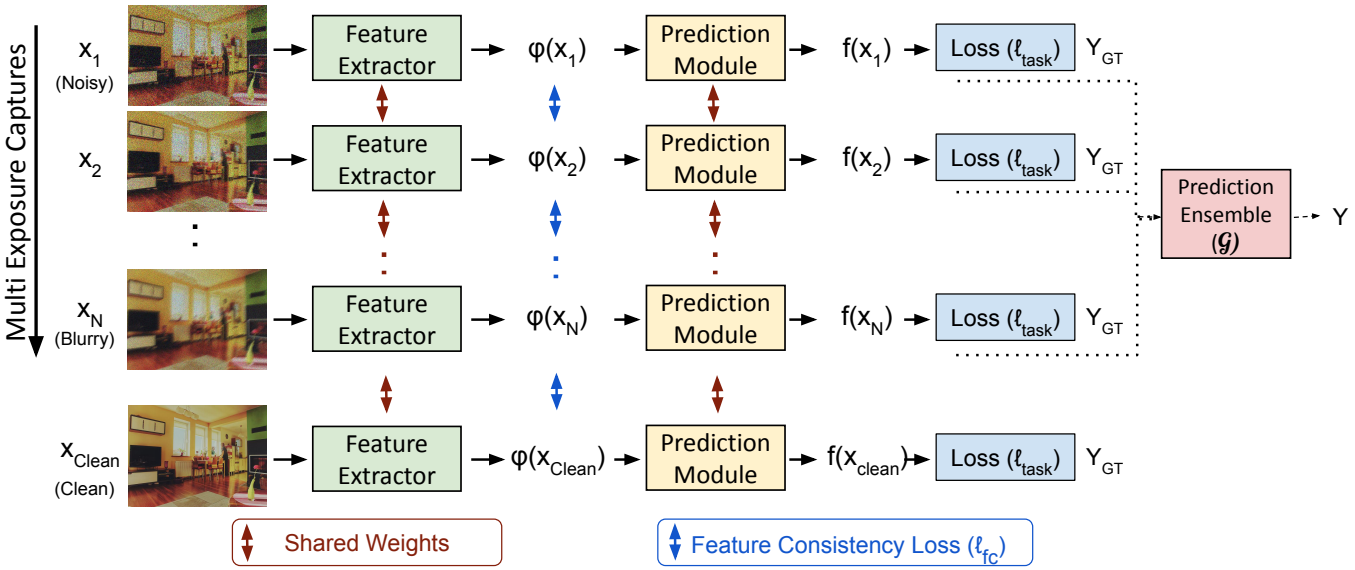


Fig. 2: **Architecture Overview:** Our approach trains an inference model using multiple captures of varying exposures, all containing the same semantic content but different amounts of noise-blur dual corruptions. We introduce feature consistency loss during training to enforce consistency of feature outputs from each individual captures. During testing (dashed lines), our model returns the ensemble prediction using each individual capture to produce final output for more robust prediction.

Training with Multiple Exposures. Given the ground-truth label y, minimizing the negative log likelihood of Equation 3 on a training sample (a set of images $\{x_i\}$ spanning the dual corruption space) leads to the following loss function

$$\ell = \sum_i^N \ell_{task}(p(y|\varphi(x_i)), y) + \frac{1}{\epsilon^2}\sum_i^N \|\varphi(x_i) - \varphi(x_{clean})\|_2^2. \quad (4)$$

Here, we slightly abuse the notation to replace the first term $-\log(p(y|\varphi(x)), y)$ with a more general task-specific loss $\ell_{task}(p(y|\varphi(x)), y)$. It is easier to consider the case of image classification, where the target y is a categorical variable. The term of $-\log(p(y|\varphi(x)), y)$ becomes the cross-entropy loss, commonly used for classification. When y moves beyond simple categorical or scalar outputs (*e.g*, for the object detection task), Equation 4 allows to plug in any loss function ℓ_{task} tailored for the task. On the other hand, the second term can be interpreted as a feature consistency loss, re-weighted by a coefficient as the reciprocal of the Gaussian variance $(1/\epsilon^2)$.

Our loss function in Equation 4 assumes that a reference clean image is available during training, as often the case in our experiments. When such a clean image is not presented, we simply replace the second term with its equivalent form that only involves the summation of pairwise distances between $\varphi(x_i)$ and $\varphi(x_j)$, *i.e*, $\frac{1}{2N\epsilon^2}\sum_{i,j}\|\varphi(x_i) - \varphi(x_j)\|_2^2$.

Inference with Model Aggregation. At inference time, the maximum likelihood estimation of Equation 3 is not viable without the clean image x_{clean}. Instead, we resort to using the ensemble of the predictions from individual multi-exposure images as the final output prediction. Our key intuition is that no individual capture in the dual corruption space captures all the necessary information that may be required for the robust inference, but the ensemble output is more effective as it uses the predictions from individual images that contribute with the relevant information individually. This is given by

$$f(\mathcal{X}) = \mathcal{G}(f(x_1), f(x_2)...f(x_N)), \quad (5)$$

where $\mathcal{G}$ is an aggregate function to get the ensemble prediction. $\mathcal{G}$ is highly flexible and often task-relevant. For example, for the image classification task, $\mathcal{G}$ could be a simple average operator over the probability outputs. For object detection, $\mathcal{G}$ might be a voting scheme of detected objects. By aggregating multiple model outputs, Equation 5 is conceptually similar to the well-known model ensemble [46].

Certified Robustness. When considering a classification problem with c categories (*e.g*, image classification), we notice an interesting link between our inference scheme and a well-known robust classifier [47]. Specifically, when $\mathcal{G}$ is an average operator and the decision is made by taking the category with the highest confidence from $f(\mathcal{X})$, our inference defined a "smoothed" classifier with certified robustness [47] under the Gaussian distribution

$$\begin{aligned} &\arg\max\ p(g(\hat{\varphi}(x)) = c), \\ &\text{where}\quad \hat{\varphi}(x) \sim \mathcal{N}(\varphi(x_{clean}), \epsilon^2). \end{aligned} \quad (6)$$

Cohen *et al.* [47] showed that such a classifier, if passes additional certification, is robust within a certain ℓ_2 radius around $\varphi(x_{clean})$. Intuitively, this indicates that our model will produce consistent results (the same as ones given by the clean image) for all corrupted images spanning the dual corruption space, should the Gaussian assumption is satisfied. We deem theoretic investigation into this direction as our future work.

5 Evaluation of Robust Scene Inference

We demonstrate the effectiveness of our method on two important scene inference tasks: object detection and image classification.

5.1 Object Detection

Instantiation. Figure 2 shows the overview of our approach using multi-exposure ensemble for the object detection task. We implement our approach using the single-stage FCOS architecture [48]. The output prediction of the FCOS model for image of size $H \times W$ consists of pixel-wise classification scores ($H \times W \times C$) for C object categories, centerness scores ($H \times W \times 1$) and bounding box coordinates regression outputs ($H \times W \times 4$). During inference, our ensemble predictor ($\mathcal{G}$), takes the pixel-wise classification scores, centerness scores and box coordinates, and returns their average at each FPN level. Loss function for the inference task (ℓ_{task}) is the same as defined in FCOS architecture (i.e. sum of focal loss, regression loss for bounding boxes and centerness loss. Refer [48] for details). Our feature consistency loss (ℓ_{fc}) is the L2 distance between feature outputs from the CNN network (final layer after global average pooling).

Datasets and Metrics. We evaluate our approach using three object detection datasets: Cityscapes [52], MS-COCO [53] and REDS [54]. Cityscapes consists of street scenes captured from a vehicle and consists of 8 categories related to autonomous driving with 2975 training and 500 test images. MS-COCO consists of 80 categories for general object detection with 118k training and 5k validation images. REDS consists of 120fps video sequences of 270 scenes captured by a high speed camera. Dataset represents images with common objects (like person, car, chair etc.).

The ground truth annotations provided in Cityscapes and MS-COCO are used for evaluation. We follow common conventions, train our models on their training sets, and report results on the validation sets. In contrast, REDS does not have object annotations. We thus use a pretrained Faster R-CNN object detector model [55] available in the Detectron2 platform [56] to obtain pseudo-ground truth annotations to create our evaluation benchmark containing 270 images with 2160 box annotations.

All results are reported using mean average precision (mAP) across multiple intersection-over-union (IoU) thresholds, following the COCO evaluation protocol [53].

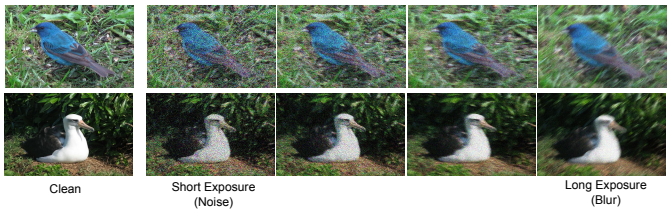
Low-light and Motion Blur Dataset Generation. All three datasets mentioned above contain images captured in sufficient light and no noticeable motion blur (scene or camera). Since there is no publicly available large-scale annotated dataset containing images captured in low-light and motion blur conditions, we simulate such conditions using various strategies, as described below.

- *REDS*: Since the REDS dataset contains video sequences captured by a 120fps camera, we first simulate low-light conditions for each individual frame of the sequence by adding Poisson noise (shot noise) and read noise. Multiple frames are then averaged together to generate images with motion blur that captures realistic motion conditions of camera or scene. In practice, we select a random frame from each video sequence, select a varying number of adjacent frames (from 0 to 3 on each side of the frame), and compute their average (after adding noise) to simulate blurry images with motion. This generates images with different exposures, examples of which are shown in Figure 3b.

(a) CityScapes

(b) REDS

(c) MS-COCO



(d) CUB-200-2011

Fig. 3: **Simulated Images:** Few examples of images with simulated noise and blur. CityScapes and REDS dataset images are generated by simulating low-light frames from high speed video sequence. MS-COCO and Birds dataset images are generated using a single frame by adding noise (shot noise and read noise) and blur (random motion blur kernel) of varying amounts.

- *CityScapes*: CityScapes provides low-fps video sequences around each annotated frame in the dataset (30-frame sequence captured at 17fps). We use a pretrained video interpolation network [57] to synthesize high-fps video sequence by increasing the frame rate by a factor of 4x. A motion-blurred image is then generated as with the REDS dataset, that is adding noise to each individual frame, and averaging a number of adjacent frames. Figure 3a shows examples of simulated low-light and motion-blur frames used for training and evaluation on the CityScapes dataset. The resulting images indeed represent realistic motion conditions under autonomous driving scenarios (like fast moving camera/car or moving pedestrians, other vehicles etc.).
- *MS-COCO*: As the MS-COCO dataset does not contain any video sequences, we simulate the blur and noise from a single image using the same procedure as the image corruptions benchmark in [5] by selecting varying severity of shot noise and motion blur. Specifically, the

Method	REDS				CityScapes				MS-COCO			
	mAP	APs	APm	APl	mAP	APs	APm	APl	mAP	APs	APm	APl
Clean Model	16.36	17.96	18.46	16.45	2.72	0.22	2.47	7.02	3.35	0.21	2.51	7.69
Stylized Training [6]	19.13	18.11	21.64	23.71	6.75	0.24	3.32	20.00	7.89	0.25	3.13	17.07
Single Exposure	30.17	20.27	25.75	36.88	18.07	3.96	15.77	35.54	21.25	6.58	22.39	33.88
Denoising (BM3D) [49]	30.25	20.28	25.90	37.08	18.01	3.82	15.53	35.97	21.78	6.76	22.78	34.43
Denoising (MPRNet) [50]	25.67	18.97	23.47	31.84	15.26	2.97	13.89	34.11	18.78	5.12	17.45	27.13
Deblurring [51]	30.68	18.82	26.36	36.02	17.67	3.63	15.90	34.67	12.42	2.52	11.77	21.11
Denoising [49] + Deblurring [51]	29.45	18.46	26.35	34.46	17.91	4.01	15.34	35.09	22.03	6.79	22.89	34.63
Short Exposures ($N = 4$)	30.81	18.41	26.53	36.02	18.46	4.33	15.97	35.86	22.17	6.87	23.91	35.11
Multi-Exposure Ensemble ($N = 2$)	33.76	14.67	27.64	40.81	19.36	5.11	17.23	37.66	23.11	8.01	25.87	36.09
Multi-Exposure Ensemble ($N = 4$)	**36.17**	14.15	**29.04**	**42.17**	**20.97**	**5.38**	**19.46**	**38.95**	**24.71**	**9.13**	**27.08**	**37.79**

TABLE 1: **Object Detection Results**: AP results on REDS, MSCOCO, and CityScapes datasets. Our approach of Multi-Exposure Ensemble (Ours) outperforms all baselines.

noisiest image has a shot noise level of 4 and a motion blur level 1. Subsequent levels in the dual corruptions are simulated by increasing the motion blur and decreasing the shot noise successively to generate 4 levels of dual corruptions. Figure 3c shows a few examples of simulated images. We note that, contrary to the other two datasets above, the blur simulated by this approach is not spatially varying.

Baselines. We compare our approach with the following set of baselines. All approaches use the same backbone for fair comparison. We evaluate all the methods using all four exposures and report the results for the best exposure settings.

- *Clean Model*: This baseline model is trained only on clean images, and evaluated on noisy images.
- *Stylized Training*: We follow the data augmentation approach of [6], who propose to augment training images with stylization for robustness.
- *Single Exposure*: We train a model on a dataset containing varying exposures and clean images, essentially considering distortions as a way to perform data augmentation [5]. For evaluation, we select the single exposure setting yielding the best performance and report those results. This baseline acts as an oracle for the selecting the best performing exposure time at inference time.
- *Denoising*: This baseline represents the conventional approach of denoising the noisy images under low-light conditions. We perform both training and inference on denoised images. Here, we experiment with the BM3D [49] and MPRNet [50] approaches for denoising the images.
- *Deblurring*: We also compare with the approach of debluring the images for scene inference, where we use a deblurring model [51]. We perform both training and evaluation of our model using deblurred images.
- *Denoising + Deblurring*: As the test images in low-light and motion blur have both noise and blur, we also compare with the approach of denoising (BM3D) followed by deblurring. Model is trained and evaluated using Denoised+Deblurred images.
- *Short Exposures*: This baseline compares with the approach of evaluating using multiple short exposures by using the ensemble prediction from N short exposure images. Model is trained with short exposure images.

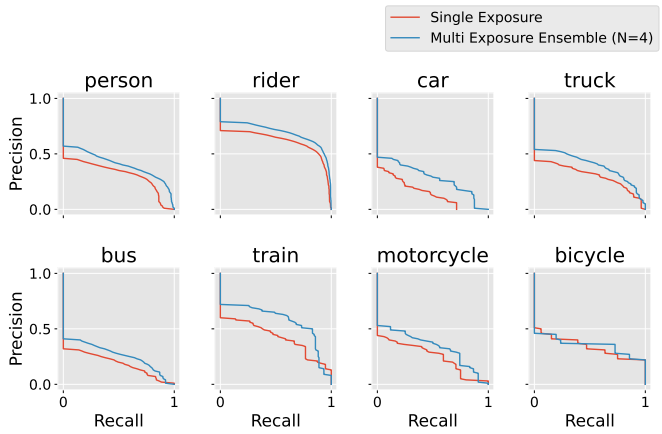
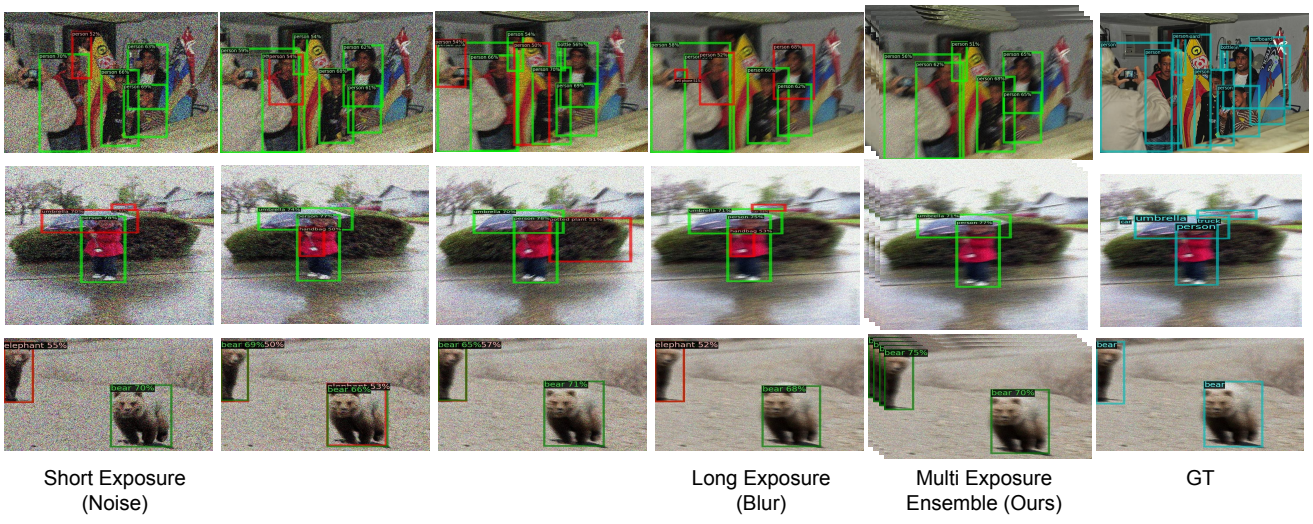


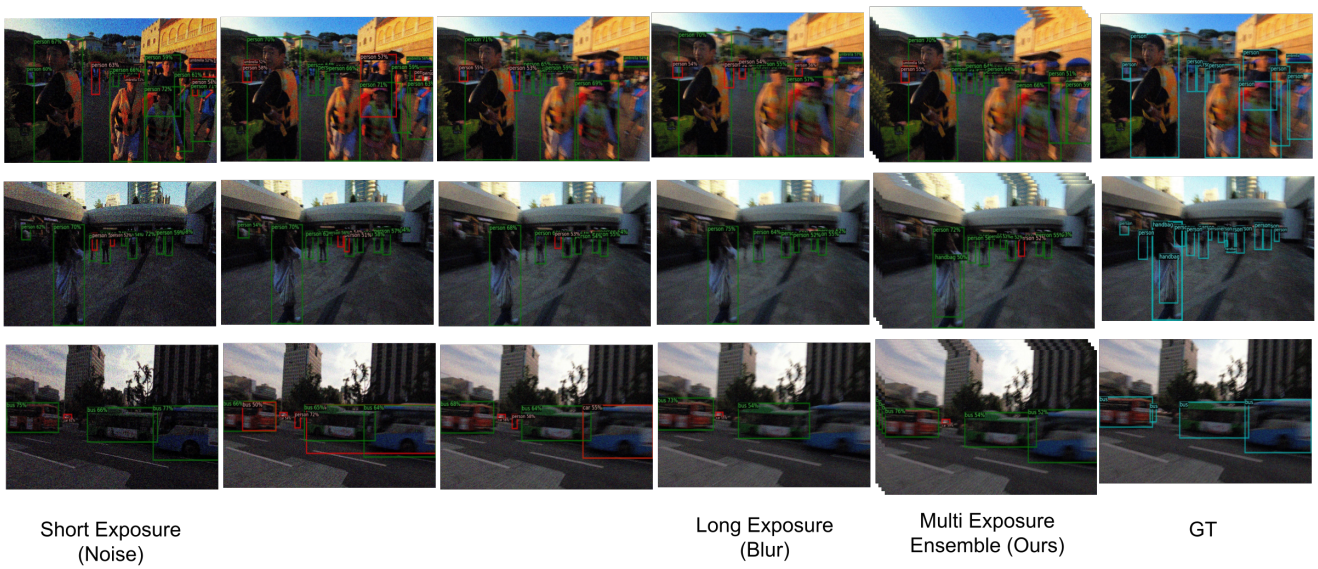
Fig. 4: **Precision Recall Curve** of our approach and baselines on CityScapes Dataset for all 8 categories with IOU threshold of 0.5. We see significant improvement for 'person' and 'car' categories, which are most common in the dataset.

Implementation Details. We used the official implementation of the FCOS architecture [58] for the object detection experiments, which is based on the Detectron2 framework [56]. ResNet-50 [59] with FPN was used as backbone for training and initialized with ImageNet pretraining weights for all our models. We followed the hyperparameters from Detectron2 to train our models. MS-COCO models were trained with a learning rate of 0.01, batch size of 16 for 90k iterations whereas CityScapes model were trained with a learning rate of 0.005, batch size of 8 for 24k iterations. REDS is used only for evaluation, in this case we use the model trained on MS-COCO.

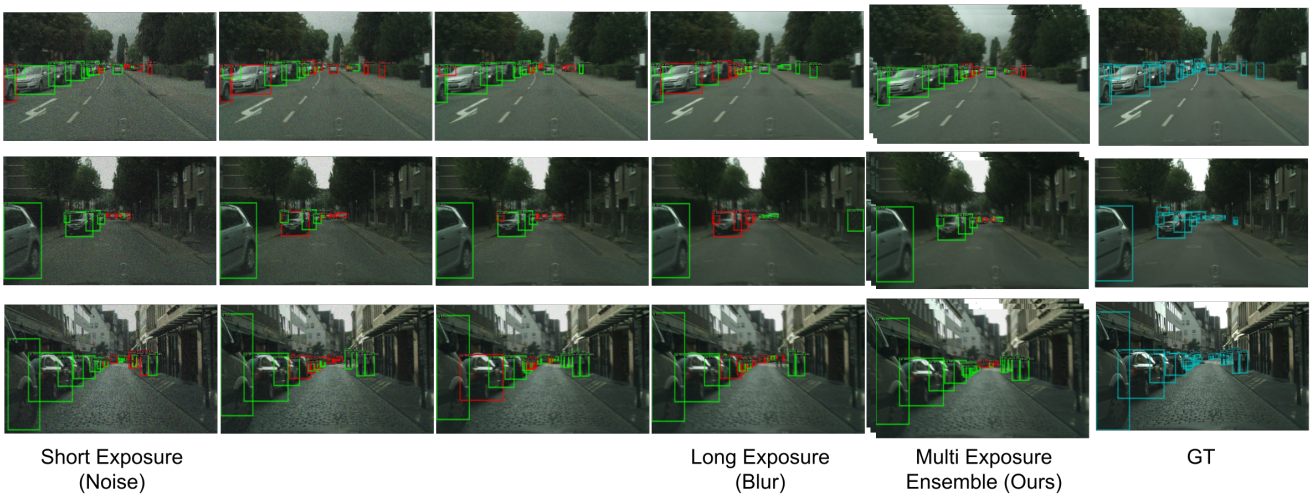
Results and Discussions. Table 1 shows the results (in mAP along with AP of small, medium and large objects) of our approach on all three datasets. Our method outperforms all baselines by a significant margin. Our approach beats Single Exposure baseline by 6% in REDS, 2.9% in CityScapes, and 3.5% in MS-COCO with four exposures. In other words, it is best to leverage all the dual-corruption images even if we knew the best possible single exposure ahead of time. Denoising provides improvements over Single Exposure baseline in some cases but is not as effective. Deblurring approaches does not show performance improvement over



(a) MS-COCO Dataset Results



(b) REDS Dataset Results



(c) CityScapes Dataset Results

Fig. 5: **Object Detection Results** for MS-COCO, REDS and CityScapes Dataset. Correct/Incorrect predictions are highlighted with green/red and ground truth boxes are highlighted with blue in the clean image. First 4 columns show results on single captures followed by a column with results from multi-exposure captures using our approach. Single Captures have a lot more false positives (red) while our approach effectively removes those cases (Better viewed on screen).

Method	Top-1	Top-5
Clean Model	6.13	13.45
Stylized Training [6]	9.51	17.83
Single Exposure	41.18	64.13
Denoising (BM3D) [49]	43.34	67.11
Deblurring [51]	39.13	60.45
Denoising [49] + Deblurring [51]	42.95	66.59
Short Exposures ($N = 4$)	45.16	69.84
Multi Exposure Ensemble ($N = 2$)	52.10	74.13
Multi Exposure Ensemble ($N = 4$)	**55.27**	**79.34**

TABLE 2: **Image Classification Results**: Top-1 and top-5 accuracy results on CUB-200-2011 dataset. Our approach of Multi-Exposure Ensemble outperforms all the baselines.

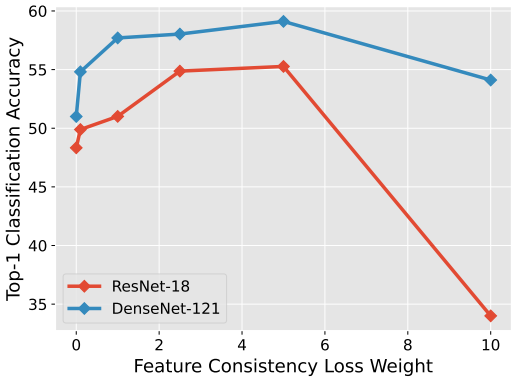


Fig. 6: **Ablation Studies:** Image classification results of our approach on CUB-200-2011 while varying feature consistency loss weight and backbone architecture.

Single Exposure baseline in most cases. This is because images contain both noise and blur and deblurring models are specialized to handle only blur. Deblurring+Denoising baseline also shows relatively minor performance gain. We see significant gain with Short Exposures (with 4 exposures) baseline, highlighting the benefit of ensemble prediction. However, since all the exposures are short, they all suffer from sever noise and have similar errors, and hence outperformed by our method. Our method provides large improvements even with two exposures, and increasing the number of exposures (from two to four) further increases the performance. This highlights that our approach benefits with more number of exposures as different exposures have a wide variety of dual corruption level.

Figure 5 shows representative qualitative examples of our approach for object detection and shows direct comparison with each individual exposure and its predictions. The correct/incorrect bounding boxes are highlighted in green/red and ground truth bounding boxes are highlighted in blue on the clean image (right). Our approach makes fewer false positive predictions (red) compared to the Single Exposure. Since individual single captures make different false positive predictions, the ensemble is able to remove those false positive boxes. Figure 4 shows the precision recall curve on CityScapes dataset for IOU threshold of 0.5 for all 8 categories in the dataset. We see a significant improvement in area under the curve for person and car category, which is the most common in the dataset.

5.2 Image Classification

Instantiation. Similar to object detection, our approach uses a shared CNN architecture as a feature extractor. In particular, we used a ResNet-18 [59] as the image classification architecture. The feature consistency loss ℓ_{fc} is defined as the L2 distance between the feature map output of the final layer (after global average pooling) to encourage consistent predictions. The model returns the average of the predictions (probability output) from multiple degraded images (as the ensemble operator $\mathcal{G}$) for the final output.

Datasets, Metrics, and Baselines. We use simulated images from the CUB-200-2011 image classification dataset [60]. CUB-200-2011 is commonly used for fine-grained image classification benchmarks and consists of 200 species of birds with 5,994 training images and 5,794 test images. All results are reported using top-1/5 accuracy on the test set, following the standard evaluation protocol for image classification. A set of baselines similar to the ones used in the experiments on object detection (Section 5.1) is considered here.

Simulating Noise and Blur. Since CUB only contains single images, we employ the same strategy to generate dual corruption images as for the MS-COCO dataset in the object detection experiments (Section 5.1). Figure 3d shows a few examples of simulated images.

Implementation Details. The model is trained with SGD with momentum of 0.9, base learning rate of 0.1 with cosine decay and batch size of 32 is used to train for 100 epochs.

Results and Discussions. Table 2 shows top-1 and top-5 accuracy of our approach on the simulated CUB dataset. We report results of our model using two and four exposure settings. Our method outperforms both baselines using a single exposure by a significant margin. Compared to choosing the single best exposure, our approach, with $N = 4$, attains an overall gain of 14.1% and 15.2% in top-1 and top-5 accuracy respectively. Our approach shows significant gains with only two exposures however having more number of exposures (from 2 to 4) further helps the overall performance.

Ablation Studies. We study the performance of our approach with varying weight for feature consistency loss. Figure 6 shows that our approach performs best for the weight factor of 5 image classification on CUB-200-2011 Dataset. We also evaluate the performance of our approach with another backbone architecture. Figure 6 shows similar performance gain using DenseNet-121 [61] which highlights the versatility of our approach as it can extend to different CNN feature extractors.

6 Experiments with Real Captures

Finally, we evaluate our approach on real images by capturing multiple simultaneous exposures of the same scene.

Camera Setup: Our setup includes four BlackflyS USB3 cameras [62] by Teledyne Flir. These are machine vision cameras that can capture colored images with a resolution of 1280×1024 with up to 175 frames per second. Same lenses (Tamron 8mm) are used for all cameras, which are stacked

Fig. 7: **Camera Setup** for capturing multiple exposure images

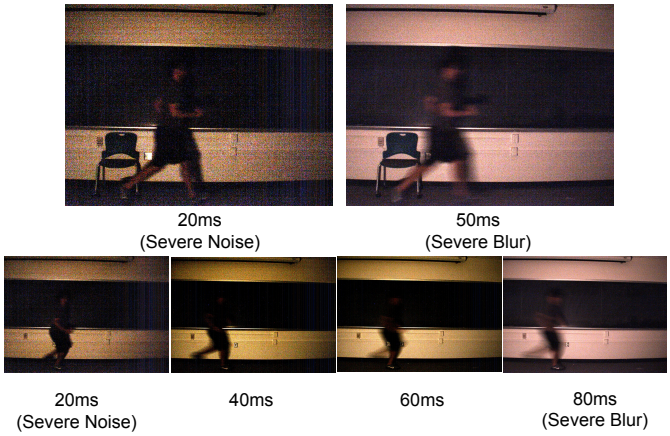


Fig. 8: **Examples of Real Captures:** Images captured with varying exposure settings with our multi camera setup. Images with shorter exposure have severe noise while images with longer exposure contain motion blur for the moving objects.

together to get similar (overlapping) fields-of-view. Aside from an approximate physical alignment of the cameras, no further alignment of the captured images is done as all cameras have similar fields-of-view, and the scene is sufficiently far away. Cameras are connected to a computer that triggers the simultaneous captures (software sync). Our complete setup is shown in Figure 7.

We use spinnaker SDK [63] provided by Teledyne to capture RAW images. Maximum available gain (18dB) for the camera is used and a gamma correction ($\gamma = 2.2$) is applied on the captures to get the final images. We set different exposure times for each camera and synchronously capture images using all the cameras.

Exposure Selection: We manually select the exposure times in order to span a wide range of exposures while ensuring that images are not too under- nor over-exposed. Our indoor scenes consist of fast moving objects in a very dark environment (∼0.25lux) lit by a single light source. We experiment with multiple settings depending on the lightning conditions including A) 20-30-40-50ms, B) 20-40-60-80ms, and C) 16-33-66-100ms. When evaluating our approach, we use two or four exposures, examples of which are shown in Figure 8.

Results and Discussions. We train our object detection models with the simulated images from MS-COCO dataset and evaluate the trained model on real captures. Figure 9 shows sample results with the real captures on two scenes. Both scenes consists of both fast moving and stationary objects under low-light. The prediction output from the individual exposure contain several false positives and inaccurate boxes. By leveraging the multiple exposures across the space of dual corruptions, our method is able to correctly detect all the objects with tight bounding boxes and remove false positive boxes.

Our approach shows performs better inference even with two exposures ($N = 2$). As we increase the number of exposures, the prediction improves as long as the exposures are not too noisy or blurry for inference (as that can deteriorate the performance of the ensemble prediction). One simple heuristic that performs well with our approach is to select exposure times around the *auto-exposure* value, as this ensures the frames are not too under- or over-exposed. We show more examples in the supplementary text with two and four exposures including failure cases.

7 Discussion and Future Outlook

Multi-Exposure Cameras: We demonstrated our approach of multi exposure captures by utilizing multiple cameras with similar (overlapping) fields-of-view. With cameras that are capable of capturing multiple images with varying exposures simultaneously [3], [4], multiple exposure images could be captured with a single camera, thus making it easier to perform spatio-temporal alignment. Our work can be considered as a preliminary proof-of-concept for an eventual implementation where a single camera can capture multiple exposure images. Demonstrating and evaluating our approach on such images is an important next step.

Exposure Selection for Multiple Captures: Most modern cameras have the functionality of *auto-exposure* that selects the exposure setting based on the lighting and motion conditions (light and motion metering) of the scene for the best image quality. The optimal exposure for inference is a function of the amount of light and motion (camera/scene) in the scene, and determining it automatically (for a single exposure) is an active area of research [64]. With the ability to capture multiple exposures, an important research problem is to develop generalized auto-exposure techniques for *multiple captures* that result in the best performance for the inference tasks under these challenging conditions.

Computational Considerations: Capturing, processing and performing inference on multiple exposures incurs a linear increase in computational cost. However, since many of these computations can be done in parallel, the increase in latency is small which is important for safety critical applications like autonomous driving. Our approach is agnostic to the number of exposures during inference, which allows inference systems to switch between multi-exposure settings (during challenging conditions of low-light and/or motion) and single exposure setting during less challenging conditions (day-time driving or slow/no motion). In practice, the inference system can operate at no computational overhead by using single exposure setting during most of the time (*e.g* daytime driving) and use multi-exposure setting during more challenging conditions (*e.g* night time driving).

Fig. 9: **Object Detection Results on Real Captures:** Scene in the first row contains an indoor scenario with two objects: a person (moving) and a chair (stationary). Single Exposures are severely affected by noise and/or blur: detects false positives or inaccurate bounding boxes. Scene in the second row contains a driving scenario with a car (moving) on the left and traffic light (stationary) in the front. Single Exposures fail to detect the moving car or the stationary traffic light. Multi-Exposure Ensemble approach (right) leverages multiple exposures and detects all objects with correct labels and tight bounding boxes in both scenes.

Dual Image Degradations: So far, we have considered the dual corruptions of noise and blur. In principle, a similar dual relationship exists between several other image degradation pairs, such as, rain and defocus blur [65], and snow and motion blur [66]. A promising research direction is to evaluate the proposed approach on other such dual pairs of image degradations, toward the goal of achieving 'all-weather' computer vision systems.

Acknowledgments

This research was supported in part by the National Science Foundation under the grants CAREER #1943149 and #2003129, Intel MLWiNS grant, and a SONY Faculty Innovation Award.

References

[1] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *TOG*, vol. 35, no. 6, pp. 1–12, 2016. 1, 3

[2] "Comma ai," https://comma.ai. 1

[3] S. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *CVPR*, vol. 1, 2000, pp. 472–479 vol.1. 1, 9

[4] C. M. Nguyen, J. N. Martel, and G. Wetzstein, "Learning spatially varying pixel exposures for motion deblurring," *arXiv preprint arXiv:2204.07267*, 2022. 2, 9

[5] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2018. 2, 5, 6

[6] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv preprint arXiv:1907.07484*, 2019. 2, 6, 8

[7] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *CVPR*, 2020. 2

[8] K. Wei, Y. Fu, J. Yang, and H. Huang, "A physics-based noise formation model for extreme low-light raw denoising," in *CVPR*, 2020. 2

[9] A. Abdelhamed, M. A. Brubaker, and M. S. Brown, "Noise flow: Noise modeling with conditional normalizing flows," in *ICCV*, 2019. 2

[10] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. and Image Under.*, vol. 178, pp. 30–42, 2019. 2

[11] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *CVPR*, 2018. 2

[12] R. Wang, X. Xu, C.-W. Fu, J. Lu, B. Yu, and J. Jia, "Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment," in *ICCV*, 2021, pp. 9700–9709. 2

[13] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," *NIPS*, vol. 27, 2014. 2

[14] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Learning to deblur," *IEEE TPAMI*, vol. 38, no. 7, pp. 1439–1451, 2015. 2

[15] J. Zhang, J. Pan, W.-S. Lai, R. W. Lau, and M.-H. Yang, "Learning fully convolutional networks for iterative non-blind deconvolution," in *CVPR*, 2017, pp. 3817–3825. 2

[16] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE TIP*, vol. 26, no. 7, pp. 3142–3155, 2017. 2

[17] F. Agostinelli, M. R. Anderson, and H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," in *NeurIPS*, 2013, pp. 1493–1501. 2

[18] Z. Liu, T. Zhou, H.-J. Wang, Z. Shen, B. Kang, E. Shelhamer, and T. Darrell, "Transferable recognition-aware image processing," *arXiv preprint arXiv:1910.09185*, 2019. 2

[19] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang, "Connecting image denoising and high-level vision tasks via deep learning," *IEEE TIP*, vol. 29, pp. 3695–3706, 2020. 2

[20] S. Diamond, V. Sitzmann, S. Boyd, G. Wetzstein, and F. Heide, "Dirty pixels: Optimizing image classification architectures for raw sensor data," *arXiv preprint arXiv:1701.06487*, 2017. 2

[21] E. Tseng, F. Yu, Y. Yang, F. Mannan, K. S. Arnaud, D. Nowrouzezahrai, J.-F. Lalonde, and F. Heide, "Hyperparameter optimization in black-box image processing using differentiable proxies." *ACM TOG*, vol. 38, no. 4, pp. 27–1, 2019. 2

[22] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajak, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. O. Egiazarian, J. Kautz, and K. Pulli, "Flexisp: a flexible camera image processing framework," *ACM TOG*, vol. 33, 2014. 2

[23] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM TOG*, vol. 35, pp. 1 – 12, 2016. 2

[24] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," *ICCV*, 2017. 2

[25] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018. 2

[26] D. Rozumnyi, J. Matas, F. Sroubek, M. Pollefeys, and M. R. Oswald,

"Fmodetect: Robust detection of fast moving objects," in *ICCV*, October 2021, pp. 3541–3549. 2
[27] Z. Cui, G.-J. Qi, L. Gu, S. You, Z. Zhang, and T. Harada, "Multitask aet with orthogonal tangent regularity for dark object detection," in *ICCV*, 2021, pp. 2553–2562. 2
[28] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in *ICCV*, 2021, pp. 16 055–16 064. 2
[29] A. Gnanasambandam and S. H. Chan, "Image classification in the dark using quanta image sensors," *ECCV*, 2020. 2
[30] C. Li, X. Qu, A. Gnanasambandam, O. A. Elgendy, J. Ma, and S. H. Chan, "Photon-limited object detection using non-local feature matching and knowledge distillation," in *ICCV Workshops*, October 2021, pp. 3976–3987. 2
[31] B. Goyal and M. Gupta, "Photon-starved scene inference using single photon cameras," in *ICCV*, 2021, pp. 2512–2521. 2
[32] W. Song, M. Suganuma, X. Liu, N. Shimobayashi, D. Maruta, and T. Okatani, "Matching in the dark: A dataset for matching image pairs of low-light scenes," in *ICCV*, 2021, pp. 6029–6038. 2
[33] Y. Wang, Y. Cao, Z.-J. Zha, J. Zhang, and Z. Xiong, "Deep degradation prior for low-quality image classification," in *CVPR*, 2020, pp. 11 049–11 058. 2
[34] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *arXiv preprint arXiv:1803.06373*, 2018. 2
[35] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *ACM SIGGRAPH*, 1997, pp. 369–378. 3
[36] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes." *ACM TOG*, vol. 31, no. 6, pp. 203–1, 2012. 3
[37] N. K. Kalantari and R. Ramamoorthi, "Deep hdr video from sequences with alternating exposures," in *Computer Graphics Forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 193–205. 3
[38] S. W. Hasinoff, K. N. Kutulakos, F. Durand, and W. T. Freeman, "Time-constrained photography," in *ICCV*, 2009. 3
[39] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *CVPR*, 2010. 3
[40] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *CVPR*, 2018. 3
[41] M. Delbracio and G. Sapiro, "Burst deblurring: Removing camera shake through fourier burst accumulation," in *CVPR*, 2015. 3
[42] M. Aittala and F. Durand, "Burst image deblurring using permutation invariant convolutional neural networks," in *ECCV*, 2018. 3
[43] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C.-K. Liang, M. Levoy, and P. Milanfar, "Handheld multi-frame super-resolution," *ACM TOG*, vol. 38, no. 4, 2019. 3
[44] M. Granados, B. Ajdin, M. Wand, C. Theobalt, H.-P. Seidel, and H. P. Lensch, "Optimal hdr reconstruction with linear digital cameras," in *CVPR*. IEEE, 2010, pp. 215–222. 3
[45] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *CVPR*. IEEE, 2010. 3
[46] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review*, vol. 33, no. 1, pp. 1–39, 2010. 4
[47] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*. PMLR, 2019, pp. 1310–1320. 4
[48] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *ICCV*, 2019, pp. 9627–9636. 5
[49] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007. 6, 8
[50] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *CVPR*, 2021. 6
[51] G. Carbajal, P. Vitoria, M. Delbracio, P. Musé, and J. Lezama, "Non-uniform blur kernel estimation via adaptive basis decomposition," *arXiv preprint arXiv:2102.01026*, 2021. 6, 8
[52] 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 *CVPR*, 2016. 5
[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014. 5
[54] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. M. Lee, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in *CVPR Workshops*, June 2019. 5
[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015. 5
[56] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019. 5, 6
[57] H. Sim, J. Oh, and M. Kim, "Xvfi: extreme video frame interpolation," in *ICCV*, 2021. 5
[58] Z. Tian, H. Chen, X. Wang, Y. Liu, and C. Shen, "AdelaiDet: A toolbox for instance-level recognition tasks," https://git.io/adelaidet, 2019. 6
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 6, 8
[60] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011. 8
[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708. 8
[62] "Teledyne flir blackfly s machine vision camera," https://www.flir.com/products/blackfly-s-usb3/?model=BFS-U3-13Y3C-C. 8
[63] "Spinnaker sdk," https://www.flir.com/products/spinnaker-sdk/. 9
[64] E. Onzon, F. Mannan, and F. Heide, "Neural auto-exposure for high-dynamic range object detection," in *CVPR*. IEEE, 2021. 9
[65] K. Garg and S. Nayar, "When Does a Camera See Rain?" in *ICCV*, vol. 2, Oct 2005, pp. 1067–1074. 10
[66] P. Barnum, S. G. Narasimhan, and T. Kanade, "Analysis of rain and snow in frequency space," in *IJCV*, 2008. 10

Bhavya Goyal is a PhD student in CS at University of Wisconsin-Madison, advised by Prof. Mohit Gupta. He received his bachelors in CS from Indian Institute of Technology, Delhi in 2016. His interests broadly include computer vision and computational imaging. He is particularly interested in learning based approaches for emerging sensing technologies.

Jean-François Lalonde is an Associate Professor in the Electrical and Computer Engineering Department at Université Laval. Previously, he was a Post-Doctoral Associate at Disney Research, Pittsburgh. He received a Ph.D. in Robotics from Carnegie Mellon University in 2011. His research interests lie at the intersection of computer vision, computer graphics, and machine learning. In particular, he is interested in exploring how physics-based models and data-driven machine learning techniques can be unified to better understand, model, interpret, and recreate the richness of our visual world. To this end, he has published 70 refereed papers which have been cited close to 4,000 times. He is actively involved in bringing research ideas to commercial products, as demonstrated by his 9 patents, several technology transfers with large companies such as Adobe and Meta, and involvement as scientific advisor for several high tech startups.

Yin Li is an Assistant Professor in the Department of Biostatistics and Medical Informatics and affiliate faculty in the Department of Computer Sciences at the University of Wisconsin-Madison. Previously, he obtained my PhD from the Georgia Institute of Technology and was a Postdoctoral Fellow at the Carnegie Mellon University. His primary research focus is computer vision. He is also interested in the applications of vision and learning for mobile health. Specifically, his group develops methods and systems to automatically analyze human activities for healthcare applications.

Mohit Gupta is an assistant professor of Computer Science at the University of Wisconsin-Madison. Before coming to Madison, he was a research scientist in Columbia University. He received his PhD from the Robotics Institute, Carnegie Mellon University. He directs the WISIONLab with research interests broadly in computer vision and computational imaging.

8 Supplementary Report: Robust Scene Inference under Noise-Blur Dual Corruptions

This document provides additional results that are not included in the main paper.

8.1 Results for Object Detection

Comparison to Baselines: We compare our approach with additional baselines. Table 3 shows performance of the model trained and evaluated on clean images. We also show the results of training and testing with a single corruption level. Results are included for four different noise-blur dual corruption levels (from 1 to 4) with increasing motion blur and decreasing the shot noise image. Comparing with clean images shows the impact of noise and blur degradation as the mAP drops significantly. Our approach utilizes clean images and corrupted images with feature consistency that helps the model to learn robust features. Our model outperforms these baselines by a significant margin using the same model capacity.

Method	REDS				CityScapes				MS-COCO			
	mAP	APs	APm	APl	mAP	APs	APm	APl	mAP	APs	APm	APl
Clean Training & Testing	78.21	52.94	73.91	84.33	33.36	10.40	32.26	54.70	38.59	22.9	42.28	49.56
Corruption Level 1 (Severe Noise)	23.46	16.14	24.08	26.27	14.06	1.82	11.80	30.98	20.26	6.18	21.18	32.73
Corruption Level 2	30.20	20.27	25.75	36.88	17.19	3.71	15.36	33.84	20.29	5.59	21.19	32.21
Corruption Level 3	27.85	19.75	23.80	35.09	17.07	3.26	15.45	32.89	20.21	6.35	20.94	32.70
Corruption Level 4 (Severe Blur)	26.78	15.51	20.13	33.53	15.94	4.21	14.73	30.39	20.47	6.45	20.94	32.32
Multi-Exposure Ensemble ($N = 4$)	**36.17**	14.15	**29.04**	**42.17**	**20.97**	**5.38**	**19.46**	**38.95**	**24.71**	**9.13**	**27.08**	**37.79**

TABLE 3: **Object Detection Results**: AP results on REDS, MSCOCO, and CityScapes datasets.

Results Visualization for Object Detection: Figure 10 shows examples where our approach outperforms the baselines. The first row of Figure 10 shows an example where one baseline predicts correct bounding boxes and our approach is as good as best single exposure baseline. Figure 11 shows a few result images with real captures using our approach and the baseline. Our method is more effective in predicting the correct bounding boxes with fewer false positive boxes.

Failure Cases for Object Detection: Figure 12 and 13 shows some failure cases where our approach performs worse than single exposure baseline. Since our approach relies on the average of output predictions, it fails to perform well when one of the exposure has too much degradation.

Fig. 10: **Object Detection Results on MS-COCO dataset:** Correct/Incorrect predictions are highlighted with green/red and ground truth boxes are highlighted with blue in the clean image. Single Exposures have a lot more false positives (red) while our approach effectively removes those cases. For the first scene, our approach produces tighter bounding boxes than individual predictions (Better viewed on screen).

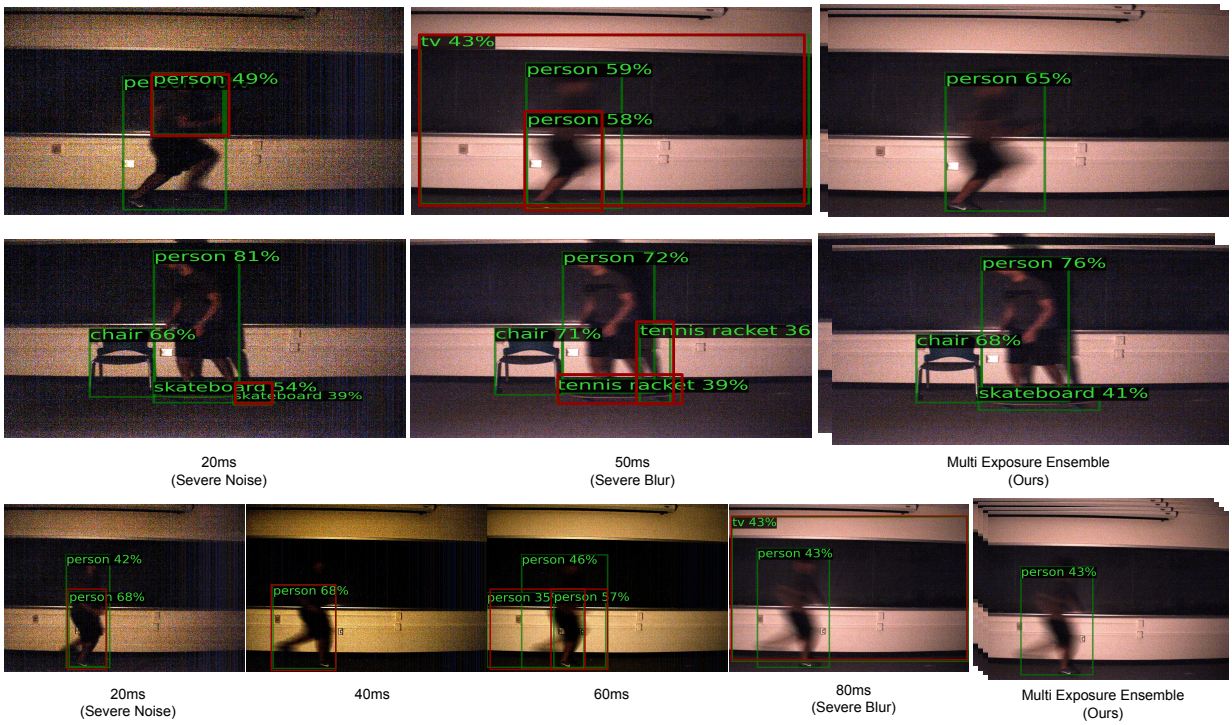


Fig. 11: **Object detection results with Real Captures:** Single Exposures are severely affected by noise and/or blur. The model detects false positives and inaccurate bounding boxes. Multi-Exposure Ensemble approach (right) leverages multiple exposures and detects all objects with correct labels and tight bounding boxes

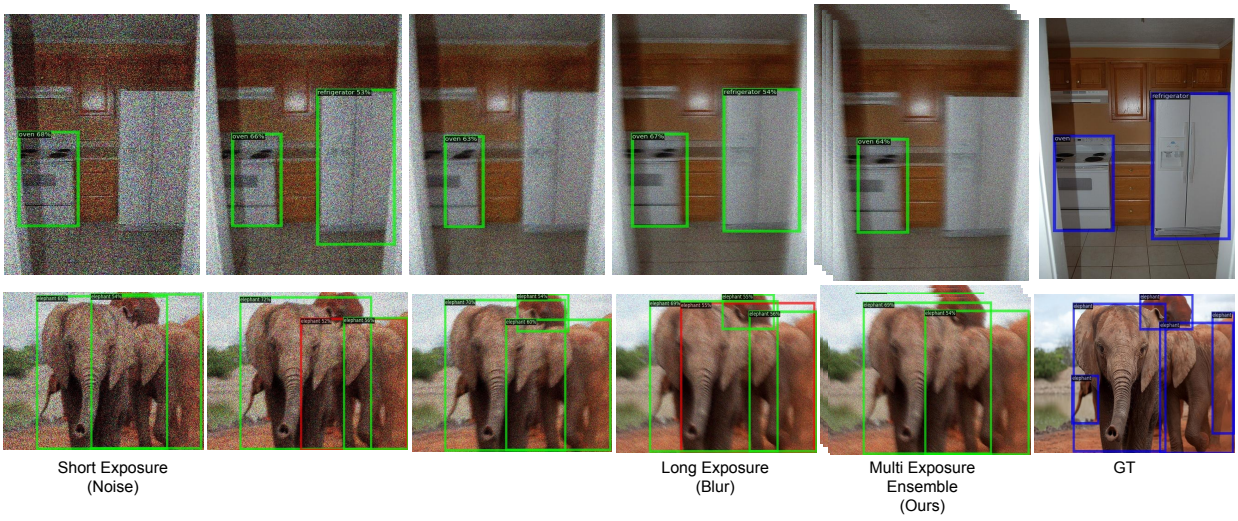


Fig. 12: **Object Detection Failure Cases on MS-COCO dataset:** Figure shows examples where single exposure performs better than our approach. First scene contains two objects and our approach fails to detect second object. Second scene contains a lot of overlapping ground truth bounding boxes and our approach fails to detects a few bounding boxes.

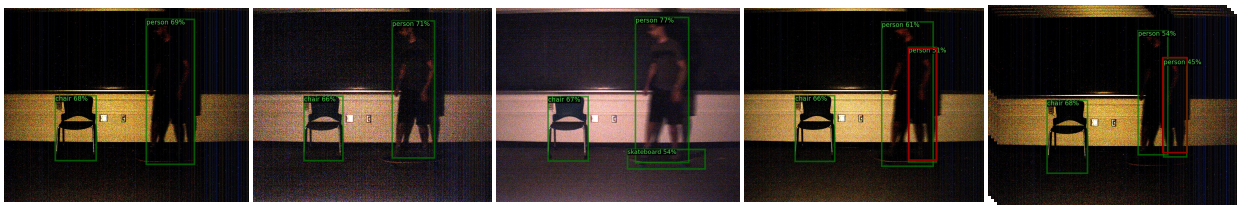


Fig. 13: **Object Detection Failure Cases on Real Captures:** Figure shows a failure case with the real captures where a single exposure (60ms) detects all three objects correctly whereas our model a detects false positive box and fails to detect skateboard object. Our model performs worse in cases when any single exposure has too much degradation.