



HAL
open science

CD-COCO: A Versatile Complex Distorted COCO Database for Scene-Context-Aware Computer Vision

Ayman Beghdadi, Azeddine Beghdadi, Malik Mallem, Lotfi Beji, Faouzi Alaya Cheikh

► **To cite this version:**

Ayman Beghdadi, Azeddine Beghdadi, Malik Mallem, Lotfi Beji, Faouzi Alaya Cheikh. CD-COCO: A Versatile Complex Distorted COCO Database for Scene-Context-Aware Computer Vision. 11th European Workshop on Visual Information Processing (EUVIP 2023), Sep 2023, Gjovik, Norway. pp.1-6, 10.1109/EUVIP58404.2023.10323035 . hal-04366877

HAL Id: hal-04366877

<https://univ-evry.hal.science/hal-04366877v1>

Submitted on 3 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CD-COCO: A Versatile Complex Distorted COCO Database for Scene-Context-Aware Computer Vision

Ayman Beghdadi
IBISC Lab, Paris-Saclay University
Evry, France
aymanaymar.beghdadi@univ-evry.fr

Azeddine Beghdadi
L2TI Lab, Sorbonne Paris Nord University
Villetaneuse, France
azeddine.beghdadi@univ-paris13.fr

Malik Mallem
IBISC Lab, Paris-Saclay University
Evry, France
malik.mallem@univ-evry.fr

Lotfi Beji
IBISC Lab, Paris-Saclay University
Evry, France
lotfi.beji@univ-evry.fr

Fauzi Alaya Cheikh
Norwegian University of Science and Technology (NTNU)
Gjovik, Norway
faouzi.cheikh@ntnu.no

Abstract—The recent development of deep learning methods applied to vision has enabled their increasing integration into real-world applications to perform complex Computer Vision (CV) tasks. However, image acquisition conditions have a major impact on the performance of high-level image processing. A possible solution to overcome these limitations is to artificially augment the training databases or to design deep learning models that are robust to signal distortions. We opt here for the first solution by enriching the database with complex and realistic distortions which were ignored until now in the existing databases. To this end, we built a new versatile database derived from the well-known MS-COCO database to which we applied local and global photo-realistic distortions. These new local distortions are generated by considering the scene context of the images that guarantees a high level of photo-realism. Distortions are generated by exploiting the depth information of the objects in the scene as well as their semantics. This guarantees a high level of photo-realism and allows to explore real scenarios ignored in conventional databases dedicated to various CV applications. Our versatile database offers an efficient solution to improve the robustness of various CV tasks such as Object Detection (OD), scene segmentation, and distortion-type classification methods. The image database, scene classification index, and distortion generation codes are publicly available ¹.

Index Terms—Dataset, Deep learning, Depth, Distortion, Object detection, Scene analysis, Segmentation

I. INTRODUCTION

The interest in making databases available to the scientific community is becoming more and more important with the development of data-driven approaches, and in particular those based on deep neural network architectures. Few studies have been conducted to analyse the relevance and reliability of databases in the field of CV. However, we can point out some interesting studies where some attributes and descriptors have been introduced to measure the representativeness and the richness of the databases dedicated to the evaluation of image and video quality metrics [1], [2]. To the best of our knowledge, there have been no similar efforts to design realistic databases dedicated to improve methods developed for solving problems

in the field of CV. Here, we are interested in the detection or segmentation of objects in an uncontrolled environment and under various constraints related to the image acquisition conditions. OD is still a hot topic and many methods have been proposed during these last two decades [3], [4]. However, the impact of the distortions on the performance of the proposed OD solutions was often neglected apart a few studies limited to object recognition and image classification under specific distortions (noise and blur) [5] and OD under photometric and geometric distortions [6]. A previous study [7] highlighted the distortion impact on the OD performance through global and local distortions without any scene context consideration have been achieved, which proved the usefulness of data augmentation by using a distorted database to improve OD models robustness. Consequently, we propose a novel distorted image database with complex and photorealistic distortions. This database offers the diversity and quality of distortions necessary for designing robust deep-learning models, in particular OD models. For this, we introduced the local and realist atmospheric distortions in our database. Unlike the classic so-called global distortions applied to the entire image, local distortions apply to defined areas. Local distortions correspond to the local representation of distortions resulting from scene conditions due to object motion or position in the scene, such as motion blur from moving objects, defocus blur and backlight phenomena. The proposed atmospheric distortions attempt to better replicate the natural rain and fog phenomena by applying these distortions in a non-homogeneous manner. These new distortions consider scene context through scene depth and object annotation from MS-COCO’s ground truth for better photorealism. Furthermore, a manual annotation of the original COCO database was done to guide the choice of the distortion to be applied automatically to each image. In addition, a scene classification (indoor/outdoor) was performed to automatically manage the distortion intensity according to the type of scene. The main contributions of our study are summarized as follows:

¹<https://github.com/Aymanbegh/CD-COCO>

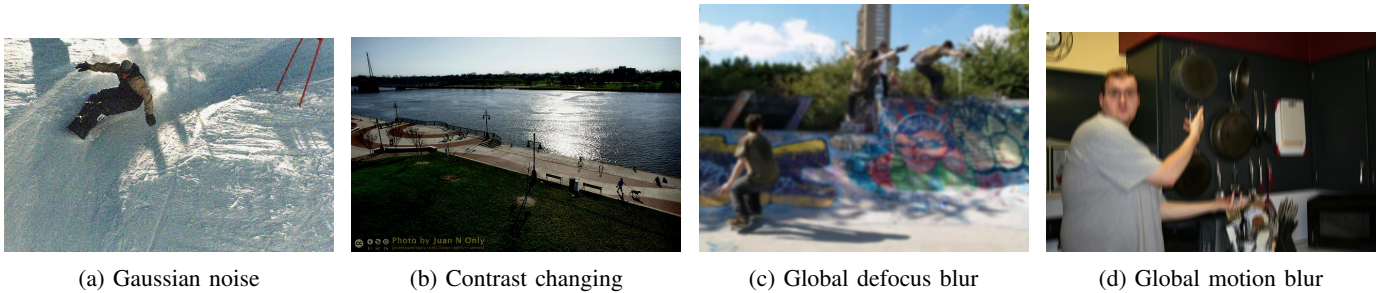


Fig. 1: Some examples of global distortions.

- New distortions with improved realism are introduced, describing common phenomena in computer vision through complex local and atmospheric distortions.
- This paper proposes efficient algorithms to generate local and global photorealistic distortions that are not included in any existing database.
- A novel dataset is built from the MS-COCO dataset, dedicated to the improvement of the robustness of the OD and object segmentation models against a broad type of distortion.
- The image database, the proposed database scene classification index, and distortion generation codes are publicly available.

The remainder of the paper is organized as follows. Section II summarizes previous related literature. Section III is devoted to detail the methods of generating complex distorted images. Then, section IV is dedicated to show dataset details. Finally, conclusions and perspectives are provided in section V.

II. RELATED WORK

Object detection in video sequences or still images is a research topic of great interest given the numerous applications in the computer vision field [8], [9] and especially in video surveillance [10]. With the development of deep learning methods and the availability of many databases dedicated to this problem, this field of research has seen a real progress. A comprehensive survey on deep learning based OD approaches is provided in [11]. However, most of the available databases do not consider real-world scenarios, especially images and videos captured in uncontrolled environments, which are affected by various types of distortions. In fact, many studies have shown that OD performance is strongly influenced by the quality of the images [5], [6], [12], [13], [7]. It is worth noticing that the number and types of distortions considered in these studies and the existing dedicated dataset are limited. Furthermore, the case of multiple distortions appearing simultaneously has not been taken into account in OD performance evaluation studies. Multiple distortion scenarios have been considered in a few studies on video quality assessment but in limited contexts [14], [15], [16]. Some interesting studies investigated the impact of various distortions on the performance of CNN-based OD architectures [17], [18]. However, all these studies are limited to a few distortions and do

not consider local distortions that really correspond to real scenarios. Indeed, if we take, for example, the blur caused by movement, it is usually simulated in a global way in the existing databases. While we know that in an observed scene there can be objects moving at different speeds and in different directions and therefore affected by blurs of different amplitudes and directions. The same applies to the defocusing blur, which depends on the depth of the objects in the filmed scene. In our database we have taken into account these aspects and others such as lighting effects that vary with the depth and geometry of objects. We have adopted the same approach concerning the distortions due to atmospheric phenomena such as rain and fog. Taking into account these aspects is not simple and it is one of the main originalities of our contribution.

III. COMPLEX DISTORTION GENERATION ALGORITHMS

Well-known global distortions have been applied to our database through classical distortions methods. In our case, global distortion refers to the classic distortions that apply more or less homogeneously to the entire image, regardless of the context of the scene. Thus, we applied global distortions for some images resulting from image acquisition (noise, compression, contrast changing) or camera (motion and defocus blur) conditions without considering the scene context (see fig.1). However, some images have specific scene contexts that require the application of local or atmospheric distortions using more sophisticated approaches. Our generated complex distortions use scene depth information, ground truth information from COCO annotations (object masks), and object and scene type to produce complex and photorealistic distortions. Scene depth information is obtained using the MiDaS depth estimation model [19].

A. Local motion blur

Local motion blur is a local application of motion blur phenomena to the annotated objects. It represents the cases of image acquisition where objects move rapidly in front of the camera. This local distortion requires the ground truth masks to define the pixel area where the blur motion is to be applied. Furthermore, the object mask is also used to determine the distortion orientation through a strategy specific to the nature of the object (object classes). Another dual strategy allows us to compute the motion magnitude applied to each object in the images. First, an interval of motion magnitude

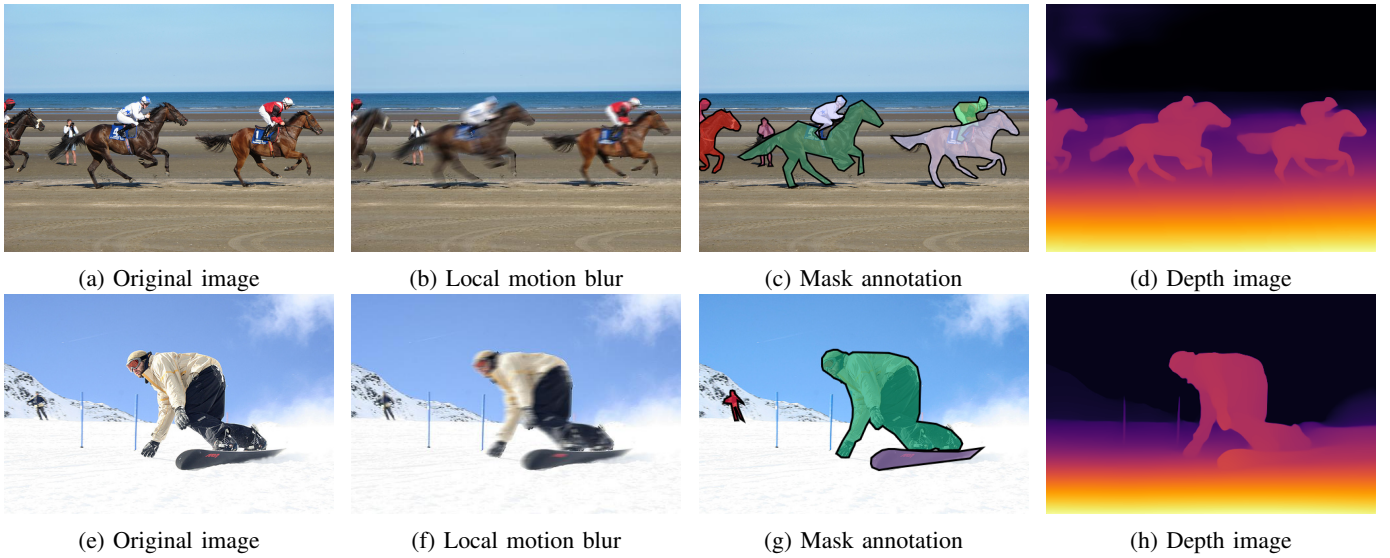


Fig. 2: Illustration of the local motion blur.

is derived from the nature of the object and prior knowledge about the speed of the object type. Then, a magnitude value is computed by considering the object's depth and the others. this value into the global scene context. Thus, each magnitude value is obtained by correlating the nature and depth of objects, ensuring the global consistency of each local blur motion distortions relative to each other. Object orientation is obtained by computing the angle between the X-axis and the ellipse's major axis containing the object. Then, a checking strategy of the orientation is adapted to apply a motion blur according to the object's nature. Furthermore, a checking of object interaction is achieved to prioritize the magnitude and orientation of higher-level objects on lower ones as shown in fig.3. This is done by correlating their depth proximity and

- 2) Object classification: create object superclasses by grouping objects together to think globally (vehicle, person, animal, food, etc...).
- 3) Compute the average depth of each annotated objects.
- 4) Calculate amplitude and orientation using depth and object type to distort each object individually.
- 5) Find interactions between objects by correlating their depth proximity and their overlapping bounding boxes to apply the same distortion to interacted objects.
- 6) Sort the objects according to their depth to adjust their motion amplitude for a global consistency of the scene and distortions.

B. Local Defocus blur

Local defocusing blur results from focusing only on only the background or foreground. To create a realistic defocus blur, we used successive smooth thresholding to create three distinct areas related to scene depth. This thresholding process is performed using a nonlinear smooth function Ω expressed as:

$$\Omega(x) = 1 - \frac{1}{1 + \exp -15(x - 0.5)} \quad (1)$$

Where x represents the keypoint depth normalized by the average depth of the closest object as follows:

$$x = \frac{\text{threshold} - p_i}{\text{threshold}} \quad (2)$$

Figure 5b illustrates the image splitting through the smooth thresholding of the scene depth to get the three different grounds. Foreground corresponds to depths with threshold coefficients higher than the high threshold, middle-ground to coefficients between the high and low thresholds, and the background for coefficients lower (see fig.5a). Then, the average depths δ , δ_m , and δ_b of the three grounds are computed to perform a proportional defocus blur related to the depth. We applied a cumulative defocus blur magnitudes λ , λ_m , and λ_b

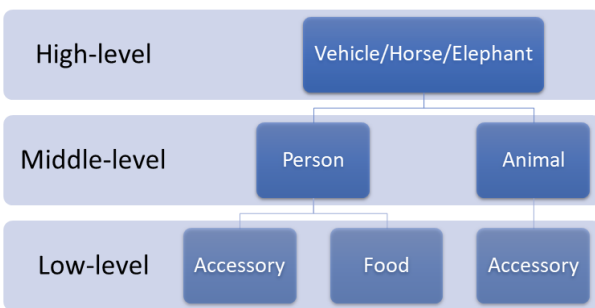


Fig. 3: Interaction hierarchy related to the object type

their bounding box overlap to ensure distortion consistency for linked objects. Thus, the magnitude and orientation of higher-level objects are applied to lower-level objects with which they are interacting. The complete algorithm follows the following steps:

- 1) Find the scene context: ski, riding, sport, skate or surf depending on present objects in the image.

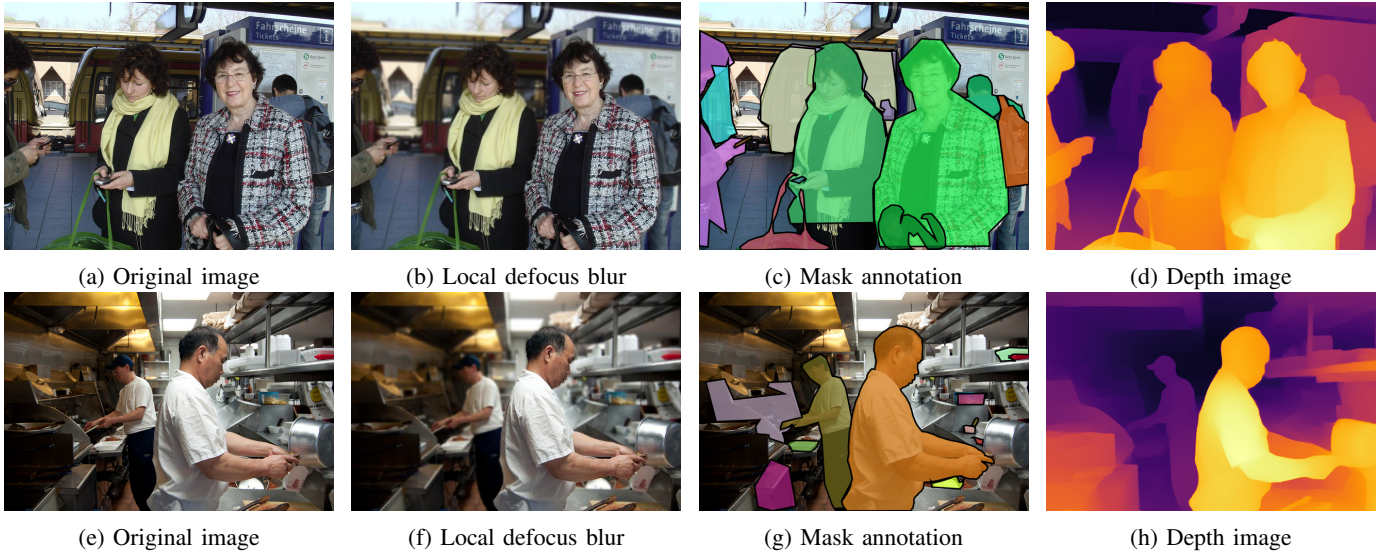


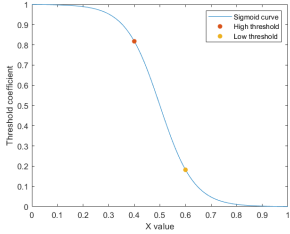
Fig. 4: Illustration of the local defocus blur.

going from foreground to background on each zone, expressed as follows:

$$\lambda = 0.5 + \frac{\delta_f - \text{threshold}}{\text{threshold}} \cdot 1.5 \quad (3)$$

$$\lambda_m = \lambda + \frac{\delta_m - \text{threshold}}{\text{threshold}} \cdot 1.2 \quad (4)$$

$$\lambda_b = \lambda_m + \frac{\delta_b - \text{threshold}}{\text{threshold}} \cdot 1.2 \quad (5)$$



(a) Sigmoid curve for smooth thresholding



(b) Threshold related to the depth

Fig. 5: Smooth thresholding related to the depth.

The aread bounded by the masks are distorted according to their corresponding defocus blur magnitude (λ , λ_m and λ_b), then fused to get the complete distorted image I_d . as shown in Fig. 4. The proposed local defocus blur algorithm is described in the algorithm 1.

C. Atmospheric distortion: the rain

Synthesizing the rain homogeneously, without any scene depth consideration, lacks realism. Indeed, the size and density of the rain depends on the distance from which it falls. Thereby, our rain generation algorithm used the method from algorithm 1 for performing a scene depth classification into foreground, middle-ground, and background. Each ground is assigned a rain intensity level that replicates the rain density. Note that the rain masks are obtained from images of flowing water like rain produced under experimental conditions.

Algorithm 1 Locale defocus blur algorithm

Input: Image I , Keypoints depth p_i

Output: Distorted Image I_d

Find the closest object depth: threshold

High threshold $th_f = 0.8176$

Low threshold $th_b = 0.182$

for each $p_i \in I$ **do**

$$\sigma = \frac{\text{threshold} - p_i}{\text{threshold}}$$

$$\Delta(p_i) = 1 - \frac{1}{1 + \exp(-15(\sigma - 0.5))}$$

if $\text{threshold} > p_i$ **then**

 Foreground $\leftarrow p_i$

else if $th_f \leq \Delta(p_i)$ **then**

 Foreground $\leftarrow p_i$

else if $th_b \leq \Delta(p_i)$ **then**

 Middleground $\leftarrow p_i$

else if $th_f > \Delta(p_i)$ **and** $th_b > \Delta(p_i)$ **then**

 Background $\leftarrow p_i$

end if

$\delta_f \leftarrow$ Foreground average depth

$\delta_m \leftarrow$ Middleground average depth

$\delta_b \leftarrow$ Background average depth

end for

We extract three rain densities from these rain masks by performing some erosion and dilation processes. These three rain sub-masks are applied for each ground according to a random constant α of blending, achieving image blending as shown in fig.6. It is worth noticing that the rain sub-masks are applied cumulatively from the foreground to the background as follows:

$$I_f = 1 - ((1 - I) \cdot (1 - (\alpha \cdot R_f))) \quad (6)$$

$$I_m = 1 - ((1 - I_f) \cdot (1 - (\alpha \cdot R_m))) \quad (7)$$

$$I_d = 1 - ((1 - I_m) \cdot (1 - (\alpha \cdot R_b))) \quad (8)$$

Where I , I_f , I_m and I_d are the original, foreground, middle-ground and final distorted images respectively. Likewise, R_f , R_m , and R_b are the three rain sub-masks. Thus, this approach

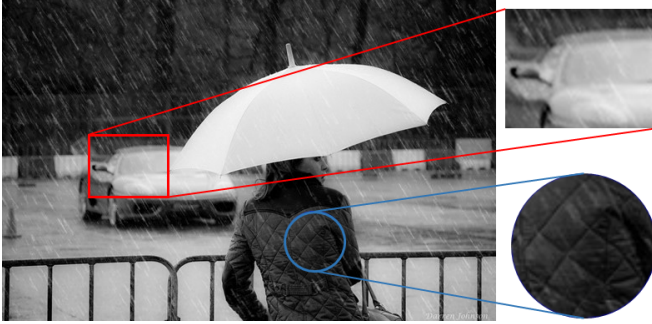


Fig. 6: Rain distortion example

applies only the fine rain corresponding to the distant rain stream to the distant parts of the scene for better realism, as shown in Fig.6. Our complex atmospheric rain improves the global consistency of distortion by considering the spatial relationship between scene depth and rain phenomena.

D. Atmospheric distortion: the fog

Generating synthetic fog is a complex task. Indeed, no mathematical fog model could have been used for generating synthetic fog. Thus, we opted to use fog masks extracted from images of experimental creations of fog with black background.

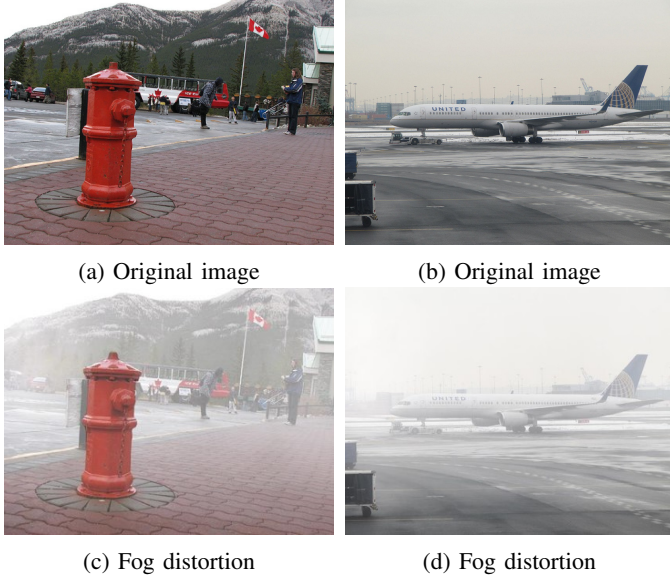


Fig. 7: Examples of atmospheric distortion: the fog.

Many masks have been extracted to provide a large fog sample with diverse densities and forms. These masks are applied to the original images, seamlessly blending through a mask adjustment. However, applying a mask homogeneously produce a non-realistic fog. Considering the scene depth for applying the mask to match the thick fog effect in real cases seems crucial. Thereby, we carry out this mask H with a

variable factor $\kappa(i, j)$ proportional to the normalized depth $Depth_n(i, j)$ of each image pixel $I(i, j)$ and a constant value α as summarized in algorithm 2.

Algorithm 2 Fog generation algorithm

Input: Image I , fog mask H

Output: Distorted Image I_d

$\alpha = 0.95$

for each pixel $i, j \in I$ **do**

$$Depth_n(i, j) = \frac{Depth(I(i, j))}{Depth_{max}}$$

$$\kappa(i, j) = \alpha \cdot Depth_n(i, j)$$

$$I_d(i, j) = (1 - (1 - I(i, j)) \cdot (1 - (\kappa(i, j) \cdot H(i, j)))) \cdot 255$$

end for

To give the images generated more photo-realism, the thickness of the fog is adapted to the depth of the observed scene, reproducing the effect of fog accumulation, as shown in figure 7.

E. Local Backlight

Local backlight distortion is generated by applying a local contrast enhancement process to the luminance component by using the object segmentation mask. This pixel-wise intensity transformation takes into account the position of the light source and that of the illuminated object on which the effect is to be brought out. This operation, which is nothing more than tone mapping, is applied to three preselected intensity intervals, semi-automatically and randomly. Figure 8 illustrates this type of photometric distortion.

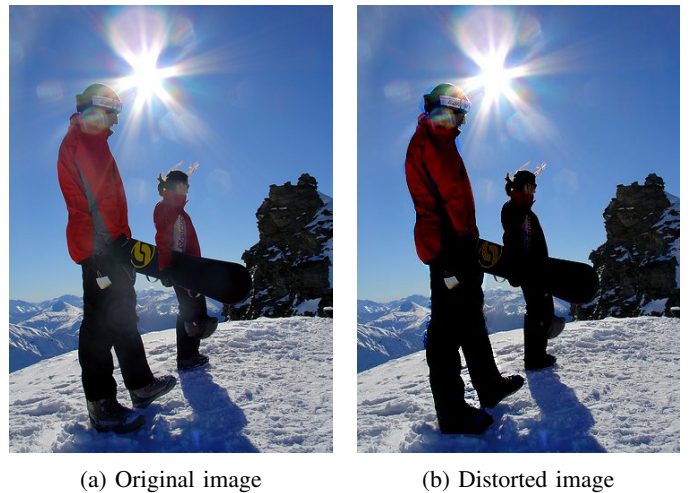


Fig. 8: Local backlight distortion

IV. DATASET

The generated dataset consists of more than 123K images with 80 object classes organized in three sets: 95K, 5K and 23K images for train, validation and test sets respectively. The ground truth annotation provides the objects' classes, bounding boxes, and masks for each image, which can be used for training object detection, and segmentation models.

A. Distortions

Our distorted dataset is composed of ten distortion types, 5 global distortions, 2 global atmospheric distortions and 3 local distortions. In order to generate the different distortions in a coherent and relevant way, a first scan of all the images is performed to prepare the distortion assignment protocol according to the semantic content of the scene and the context. The different distortions are automatically applied to the

TABLE I: Distribution of distortions.

Distortion type	Number of images	Ratio
Compression artefact	17989	15.3%
Contrast changing	18038	15.4%
Gaussian noise	18055	15.4%
Global motion blur	18018	15.3%
Global defocus blur	17792	15.1%
Fog	787	0.7%
Rain	845	0.7%
Local Backlight	296	0.3%
Local defocus blur	7061	6.0%
Local motion blur	18625	15.9%

images previously annotated during the first process. Images annotated as global distortions are then distorted by one of the global distortion types chosen randomly (see table I).

B. Scene classification

The observed scenes are classified into indoor and outdoor scenes based on the context of the images. Indoor scenes were attributed to scenes where most information is included in indoor environments (room, building, hall, vehicle interior, etc.). Conversely, outdoor scenes correspond to open environments. The table II summarises the scene classification of our dataset.

TABLE II: Scene classification.

Scene type	Number of images
Indoor scene	45884
Outdoor scene	72404
Skiing scene	4434
Surfing scene	3635
Skating scene	3603
Sport scene	11965

V. CONCLUSION

In this study, we presented novel local and global complex distortions generated by reliable algorithms considering the scene context to achieve a high level of photo-realism. The proposed database will improve not only OD algorithms but also many scene analysis, classification and image segmentation methods, providing a more complete and beneficial framework for deep learning-based methods. As a perspective, it would be interesting to enrich this database with other distortions and in particular those related to atmospheric perturbations such as the heat diffusion effect and pollution. Another aspect that could be considered in the future is to incorporate pose object estimation when applying distortion.

REFERENCES

- [1] Stefan Winkler. Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):616–625, 2012.
- [2] Muhammad Ali Qureshi, Azeddine Beghdadi, and Mohamed Deriche. Towards the design of a consistent image contrast enhancement evaluation measure. *Signal Processing: Image Communication*, 58:212–227, 2017.
- [3] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.
- [4] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.
- [5] Tejas S Borkar and Lina J Karam. Deepcorrect: Correcting dnn models against image distortions. *IEEE Transactions on Image Processing*, 28(12):6022–6034, 2019.
- [6] Zhongqi Lin, Zengwei Zheng, Jingdun Jia, Wanlin Gao, and Feng Huang. MI-capsnet meets vb-di-d: A novel distortion-tolerant baseline for perturbed object recognition. *Engineering Applications of Artificial Intelligence*, 120:105937, 2023.
- [7] Ayman Beghdadi, Malik Mallem, and Lotfi Beji. Benchmarking performance of object detection under image distortions in an uncontrolled environment. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2071–2075. IEEE, 2022.
- [8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.
- [9] Constantine P Papageorgiou and Tomaso Poggio. A trainable object detection system: Car detection in static images. 1999.
- [10] Azeddine Beghdadi, Muhammad Asim, Noor Almaadeed, and Muhammad Ali Qureshi. Towards the design of smart video-surveillance system. In *2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, pages 162–167. IEEE, 2018.
- [11] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE access*, 7:128837–128868, 2019.
- [12] Sebastian Cygert and Andrzej Czyżewski. Toward robust pedestrian detection with data augmentation. *IEEE Access*, 8:136674–136683, 2020.
- [13] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. Robust and accurate object detection via adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16622–16631, 2021.
- [14] Zohaib Amjad Khan, Azeddine Beghdadi, Mounir Kaaniche, and Faouzi Alaya Cheikh. Residual networks based distortion classification and ranking for laparoscopic image quality assessment. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 176–180. IEEE, 2020.
- [15] Roger Gomez-Nieto, José Francisco Ruiz-Muñoz, Juan Beron, César A Ardila Franco, Hernán Darío Benítez-Restrepo, and Alan C Bovik. Quality aware features for performance prediction and time reduction in video object tracking. *IEEE Access*, 10:13290–13310, 2022.
- [16] Roger Gomez Nieto, Hernan Dario Benitez Restrepo, and Ivan Cabezas. How video object tracking is affected by in-capture distortions? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE, 2019.
- [17] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.*, 20:184:1–184:25, 2019.
- [18] 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. *CoRR*, 2019.
- [19] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.