PVD4RCV: A Photo-realistic Multi-Distortion Video Dataset for Benchmarking and Developing Robust Computer Vision Models.

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

Mohib Ullah
NTNU University
Gjovik, Norway
mohib.ullah@ntnu.no

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

Borhen-eddine Dakkar
L2TI, University Sorbonne Paris Nord
Villetaneuse, France
dakkarborheneddine@gmail.com

Zohaib Amjad Khan
L2TI, University Sorbonne Paris Nord
Villetaneuse, France
zohaib.amjad.khan@gmail.com

Faouzi Alaya Cheikh
NTNU University
Gjovik, Norway
faouzi.cheikh@ntnu.no

***Abstract*—This work addresses a significant gap in existing image and video databases commonly used in computer vision applications by introducing a unique and comprehensive database named Photo-realistic Multi-Distortion Video Dataset for Benchmarking and Developing Robust Computer Vision Models (PVD4RCV). A key innovation of PVD4RCV lies in its incorporation of some relevant physical factors (e.g. depth information, interaction of light with scene contents) inherent to video signal acquisition in constrained and complex real-world environments, which are used to generate realistic distortions in video sequences (e.g. local motion blur, local defocus blur). PVD4RCV includes a diverse collection of videos featuring common distortions, real-world scenarios, and contextual variations. It includes both original and degraded video versions, along with detailed annotations to support the development of advanced learning models, particularly for tasks such as distortion classification and object detection. This resource aims to advance research and applications in computer vision by providing a robust foundation for model training and evaluation.**

***Index Terms*—Dataset, Distortion, Object detection, Object tracking, Scene analysis.**

I. Introduction

Artificial intelligence (AI) has rapidly transformed scientific research by promoting data-driven approaches. While this shift has led to significant progress, it has also raised concerns regarding the lack of theoretical foundations aligned with traditional mathematical logic and reasoning [1], [2], [3]. Central to this paradigm is the role of high-quality datasets, which are essential for developing robust learning models [4].

In computer vision, robotics, and related fields, numerous datasets have emerged. However, the distortions are often applied without accounting for real-world physical factors such as scene depth, light interaction with scene contents and motion dynamics [5], [6], [7], [8]. This limits the realism of training data and can hinder generalization to real-world scenarios.

To address this gap, we introduce **PVD4RCV**, a novel dataset designed to simulate distortions using physically grounded models that integrates scene depth and other environmental parameters. For example, in PVD4RCV, the haze effect is applied according to the depth of the scene's contents, and the motion blur is generated according to the movement of each object. The blur caused by camera instability is also generated using a realistic physical model that takes account of the unpredictable nature of such movements. This guarantees a high level of photorealism in the data generated and thus contributes to the evaluation and construction of robust computer vision models. The key contributions of this paper are summarized below.

- **A Unique Database for benchmarking and development of Robust Computer Vision Models:** We provide the scientific community with a unique dataset dedicated to build robust models for solving computer vision problems such as object detection, visual object tracking, and distortion classification.
- **Photo-Realistic Video Dataset:** We introduce photo-realistic distortions taking into account relevant physical factors, enabling the development of efficient deep-learning architectures that enhance the robustness of computer vision models in complex and uncontrolled real-world environments.
- **New Research Directions:** Our work opens new avenues for dataset construction enabling research into deep-learning-based architectures for identifying and classifying distortions in video sequences.

II. Construction and presentation of the database

PVD4RCV contains 24 original videos, each 10 seconds long, and 672 associated distorted versions as well as the associated depth maps. These videos correspond to different

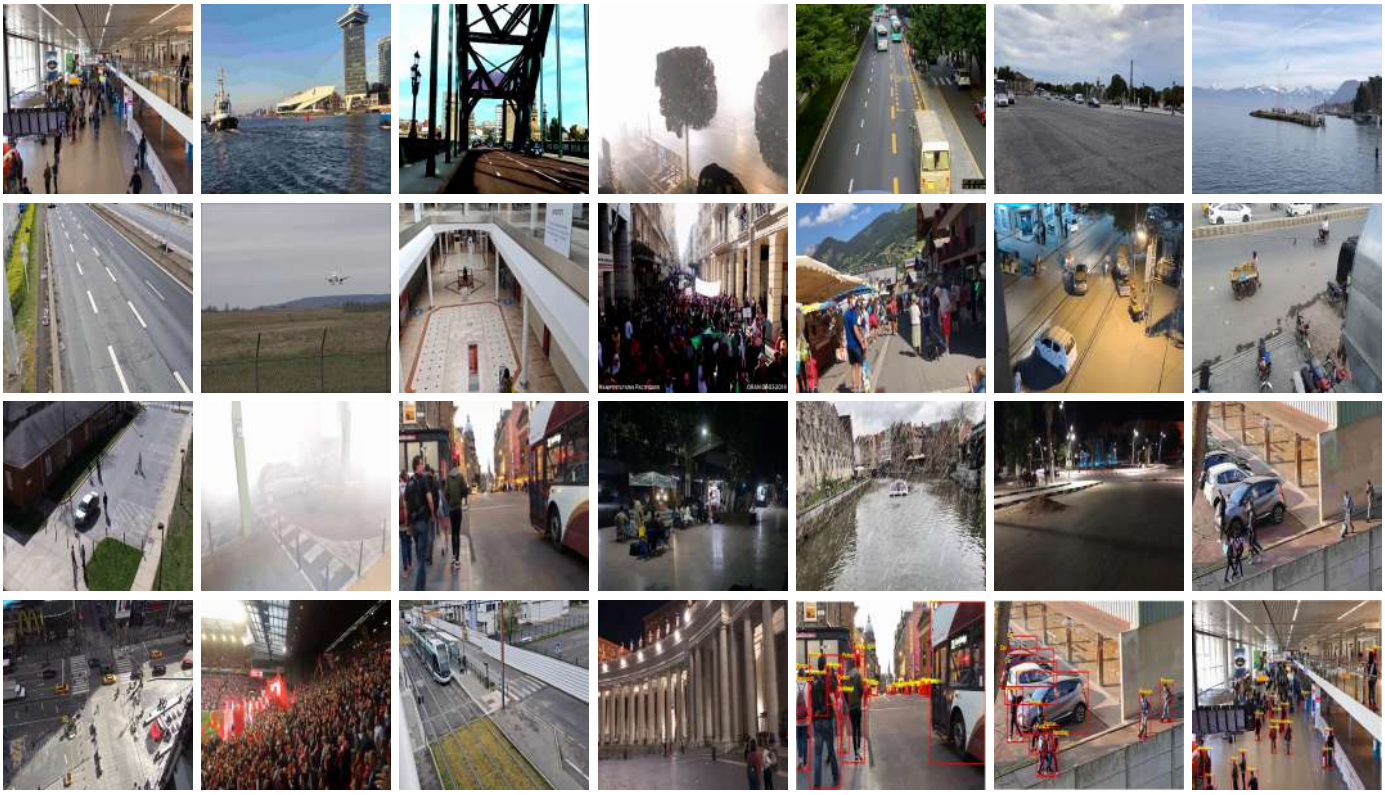

Fig. 1: Some samples from PVD4RCV dataset.

visual contents that are fairly representative of various real-world scenarios and are highly suitable for benchmarking and developing deep learning-based models to solve various problems encountered in computer vision such as object detection, visual tracking, distortion detection, and classification. Table I summarises the content of the dataset. The distortions are applied with 4 levels of severity to cover different levels of complexity and various plausible scenarios. PVD4RCV includes

TABLE I: Brief description of PVD4RCV

Parameter	**Value**
Number of Original Videos	24
Number of distorted videos	672
Resolution	1080 × 1920
Video duration	10 sec
Video frame rates	29.93-30 fps
videos type	MP4
Scenarios	traffic, parking, stadium, crowd, airport, street, train station, mall, sea navigation, city centre.
Distortion Types	compression, contrast, global blur motion, local/global defocus blur, noise, haze, rain, smoke
Distortion levels	4
Dataset size	21.2 GB

common distortions arising during capture, post-processing, video encoding and transmission. In-capture distortions are the most frequent and perceptually salient. Post-processing may introduce some side effects often due to uncontrolled contrast enhancement, denoising process or compression artifact removal. Other common artefacts and distortions may result from lossy video coding and transmission due to various factors such as motion estimation and compensation and packet loss. Table I summarises all distortion types considered in the built dataset. The in-capture distortions are generated by taking into account the depth information of the scene content using monocular depth estimation via the MiDaS model [9], enabling photorealistic and context-aware synthesis, as done in our previous work [10].

A. Global blur motion due to shakiness

Most databases consider two types of blur: global defocus and motion blur resulting from translational movement of the camera or objects in the captured scene [11]. However, real-world motion often involves pseudo-periodic, oscillatory and unpredictable movements, such as those from handheld or body-mounted cameras. Our database introduces a realistic model to generate this kind of blur using a pseudo-periodic oscillatory model described through the frequency modulated signal $s(t)$ given below.

$$s(t) = A(t)\sin\left(2\pi f(t)t + \phi\right), \tag{1}$$

where $A(t)$ and $f(t)$ denote time-varying amplitude and the instantaneous frequency, respectively, given by:

$$f(t) = f_0 + f_1\sin(2\pi\cdot 0.1\cdot t), \quad A(t) = A_0 + A_1\sin(2\pi\cdot 0.05\cdot t). \tag{2}$$

where ϕ is an arbitrary phase and f_0 and f_1 are two frequencies to be tuned according to the severity of the desired oscillations. In our experiment f_0 and f_1 are set to 25 and 5, respectively. The time varying amplitude $A(t)$ is also an oscillatory signal to count for the pseudo-periodic nature of the signal. Here the two key parameters A_0 and A_1 are set to 0 and 1. To generate the blur, the image is cropped using a trajectory derived from $s(t)$ and a spatio-temporal sliding window. An offset $\Delta_y(t)$, depending on the normalised value and sign of $s(t)$ defined below, is then used.

$$\Delta_y(t) = \Delta_y(t-1) + \lambda \frac{(s(t) - s(t-1))}{max(s(t))} \tag{3}$$

This offset is used to calculate the new position P' of the sliding window compared to its initial position $P_0(\frac{w}{2}, \frac{h}{2})$ as follow:

$$P' = P_0 + \Delta_y \tag{4}$$

Where λ is the maximum motion magnitude (here set to 40) of the sliding window, and $(\frac{w}{2}, \frac{h}{2})$ are the image centre coordinates. Finally, the resulting motion blur magnitude, noted ρ, is defined as:

$$\rho = \rho_0 |s(t) - s(t-1)|. \tag{5}$$

Where ρ_0 is the magnitude value related to the severity level of the blur motion distortion and set in range $[45, 90]$. This model allows to generate photo-realistic motion-induced blur as illustrated in Figure 2.

Fig. 2: Cropping process for generating shakiness

B. Global Defocus Blur

The blurring effect due to defocus can be simulated using low-pass filtering of the signal by an isotropic Gaussian represented by an impulse response h_σ. However, in the case of video captured in real conditions, this simulated de-focus must incorporate the variations in the blurring effect over time to account for the dynamic aspect of this distortion. One way of doing this is to make the parameter $\sigma(t)$ of the impulse response h_σ variable over time. The standard deviation of the Gaussian kernel filter is then expressed as follows:

$$\sigma(t) = \sigma_0(\beta(t) + 1) + 0.01 \tag{6}$$

with σ_0 denoting blur severity sets in range $[0.8, 2.3]$ and the signal $\beta(t)$ is given by:

$$\beta(t) = \sum_{i=1}^{n} A_i \sin(2\pi f_i t + \phi_i) \tag{7}$$

The signal $\beta(t)$ is normalised to $[-1, 1]$ and used as a ratio for the $\sigma(t)$ blur value to maintain consistent visual realism. While such temporal blur may not naturally occur in controlled scenes, its inclusion improves model robustness by exposing them to complex distortions.

C. Local Defocus Blur

Local defocus simulates focal plane adjustments using scene depth. Depth map D is segmented into foreground f, middle ground m, and background b via histogram-based thresholding, as shown in Figure 4.

Fig. 3: Illustration of the local defocus blur distortion

Using the 75th percentile of the depth distribution as the threshold, the focal depth μ is computed. Depth thresholds are defined as:

$$\delta_f = \mu \cdot th_f, \quad \delta_m = \mu \cdot th_m, \tag{8}$$

with $th_f = 0.8176$ and $th_m = 0.5$. Gaussian blur intensities are then computed:

$$\sigma_f = 0.5 + \frac{|\delta_f - threshold|}{threshold} \cdot 2.0 \cdot G_\sigma \tag{9}$$

$$\sigma_m = \sigma_f + \frac{|\delta_m - threshold|}{threshold} \cdot 1.5 \tag{10}$$

$$\sigma_b = \sigma_m + \frac{|\delta_m - threshold|}{threshold} \cdot 1.5 \tag{11}$$

The detailed implementation is given in Algorithm 1.

D. Rain Distortion

Rain simulation considers scene depth to account for visibility variations due to raindrop size and density. The scene is segmented into foreground (I_f), middle ground (I_m), and background (I_b), each blended with rain masks H whose size and density decrease with depth. The compositing is defined as:

$$I_f = 1 - (1 - I) \cdot (1 - \alpha H_f) \tag{12}$$

$$I_m = 1 - (1 - I_f) \cdot (1 - \alpha H_m) \tag{13}$$

$$I_d = 1 - (1 - I_m) \cdot (1 - \alpha H_b) \tag{14}$$

Algorithm 1 Local Defocus Blur

Input: Image I, Depth Map D, Gaussian Sigma G_σ
Output: Blurred Image I_{out}
$th_f \leftarrow 0.8176$, $th_m \leftarrow 0.5$
Compute histogram and cumulative histogram of D
Determine 75th percentile depth $threshold$
Compute mean focal depth μ
$\delta_f \leftarrow \mu \cdot th_f$, $\delta_m \leftarrow \mu \cdot th_m$
$\sigma_f \leftarrow 0.5 + \left(\frac{|\delta_f - threshold|}{threshold}\right) \cdot 2.0 \cdot G_\sigma$
$\sigma_m \leftarrow \sigma_f + \left(\frac{|\delta_m - threshold|}{threshold}\right) \cdot 1.5$
$\sigma_b \leftarrow \sigma_m + \left(\frac{|\delta_m - threshold|}{threshold}\right) \cdot 1.5$
for each pixel (i, j) in I **do**
  if $threshold > 100$ **then**
    if $D(i, j) > \mu$ **then**
      $I_{out}(i, j, :) \leftarrow I(i, j, :)$
    else if $\mu \geq D(i, j) > \delta_f$ **then**
      Apply Gaussian blur with σ_f
    else if $\delta_f \geq D(i, j) > \delta_m$ **then**
      Apply Gaussian blur with σ_m
    else
      Apply Gaussian blur with σ_b
    end if
  end if
end for

where α controls blending severity and is set in range $[1.5, 5.5]$. The higher the α value, the stronger and more intense the rain appears to be. This nested blending ensures realistic depth-dependent rain effects.

E. Haze and Smoke

Photorealistic hazy video frames are generated via a unified approach using haze masks extracted from real hazy scenes and blended with distortion-free sequences using a weighting process based on scene depth. The same procedure is used to generate video frames affected by smoke. Unlike complex deep learning methods [12], [13] or simple blending [14], this approach relies on physical depth to vary haze and smoke density, ensuring photorealism (Fig. 4).

Fig. 4: Illustration of haze distortion and the depth map

The haze/smoke mask H is then modulated pixel-wise by a factor $\kappa(i, j)$ proportional to the normalised depth $Depth_n(i, j)$ and a constant α_h, as described in Algorithm 2. The same approach is used to apply other atmospheric distortions to the video frames.

Algorithm 2 Haze Generation Algorithm

Input: Image I, haze mask H
Output: Distorted Image I_d

$\alpha_h \leftarrow 0.95$
for each pixel (i, j) **do**
  $Depth_n(i, j) \leftarrow \frac{Depth(I(i,j))}{Depth_{max}}$
  $\kappa(i, j) \leftarrow \alpha_h \cdot Depth_n(i, j)$
  $I_d(i, j) \leftarrow 1 - (1 - I(i, j)) \cdot (1 - \kappa(i, j) \cdot H(i, j))$
end for

III. Database Applications

The proposed database is designed for evaluating and training object detection, object tracking, and distortion classification models. Numerous studies [15], [16], [17], [18] have highlighted the impact of distortions on the performance of object detection and tracking models. This underscores the necessity of having access to distorted databases at various levels of severity with the corresponding annotations for the training of these models. The proposed PVD4RCV dataset furnishes ground truth, i.e. object labels and bounding boxes, to assess the robustness of the models against distortions.

Furthermore, PVD4RCV also can be used for evaluating and constructing efficient models for distortion classification [19], [20], [21]. It encompasses a range of scene contexts, including urban areas, roads, indoor environments, and natural environments, along with various distortion types at four levels of severity. These sequences are also useful for assessing methods for understanding scenes [22], [13] in complex environments for a substantial quantity of scenarios. Finally, scene depth ground truth can be used to test monocular depth estimation models [9] with a large range of distortions.

IV. Conclusion

In this paper we have presented a unique video database that is very useful for the evaluation and development of robust solutions for various computer vision tasks. What makes the proposed PVD4RCV database unique is that it takes into account physical parameters, in particular depth information and the interaction of light on objects as a function of their position and other physical aspects, often neglected in the existing datasets, in the process of generating photo-realistic distortions. By including realistic distortions like local blur and other depth-dependent distortions, the database generated in this way is extremely useful for developing robust models for solving computer vision problems such as object detection, visual tracking and distortion classification in real-world scenarios. Featuring both pristine and degraded sequences with detailed annotations, enables the benchmarking and improving various computer vision models. One of the avenues for future work is to extend the distortion generation algorithms developed in this database, dedicated to natural scenes, to other types of imagery and applications.

REFERENCES

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[2] B. Koçak, A. Ponsiglione, A. Stanzione, C. Bluethgen, J. Santinha, L. Ugga, M. Huisman, M. E. Klontzas, R. Cannella, and R. Cuocolo, "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects," *Diagnostic and Interventional Radiology*, pp. Epub–ahead, 2024.

[3] M. Fahad, N. E. Mobeen, A. S. Imran, S. M. Daudpota, Z. Kastrati, F. A. Cheikh, and M. Ullah, "Deep insights into gastrointestinal health: A comprehensive analysis of gastrovision dataset using convolutional neural networks and explainable ai," *Biomedical Signal Processing and Control*, vol. 102, p. 107260, 2025.

[4] E. S. Ortigossa, T. Gonçalves, and L. G. Nonato, "Explainable artificial intelligence (xai)–from theory to methods and applications," *IEEE Access*, 2024.

[5] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," *Domain adaptation in computer vision applications*, pp. 37–55, 2017.

[6] W. Sun, F. Zhou, and Q. Liao, "Mdid: A multiply distorted image database for image quality assessment," *Pattern Recognition*, vol. 61, pp. 153–168, 2017.

[7] S. Picard, C. Chapdelaine, C. Cappi, L. Gardes, E. Jenn, B. Lefèvre, and T. Soumarmon, "Ensuring dataset quality for machine learning certification," in *2020 IEEE international symposium on software reliability engineering workshops (ISSREW)*. IEEE, 2020, pp. 275–282.

[8] Y. Gao, Y. Cao, T. Kou, W. Sun, Y. Dong, X. Liu, X. Min, and G. Zhai, "Vdpve: Vqa dataset for perceptual video enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1474–1483.

[9] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[10] A. Beghdadi, A. Beghdadi, M. Mallem, L. Beji, and F. A. Cheikh, "Cd-coco: A versatile complex distorted coco database for scene-context-aware computer vision," in *2023 11th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2023, pp. 1–6.

[11] B.-E. Dakkar, A. Bourki, A. Beghdadi, and F. A. Cheikh, "Vstab-quad: A new video-stabilization quality assessment database," in *2023 11th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2023, pp. 1–6.

[12] Y. Zheng, A. Mi, Y. Qiao, and Y. Wang, "Realistic nighttime haze image generation with glow effect," in *Proceedings of the 2022 11th International Conference on Networks, Communication and Computing*, 2022, pp. 96–101.

[13] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[14] Z. A. Khan, A. Beghdadi, F. A. Cheikh, M. Kaaniche, E. Pelanis, R. Palomar, Å. A. Fretland, B. Edwin, and O. J. Elle, "Towards a video quality assessment based framework for enhancement of laparoscopic videos," in *Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, vol. 11316. International Society for Optics and Photonics, 2020, p. 113160P.

[15] 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.

[16] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Proceedings of the International Conference on Learning Representations*, 2019.

[17] R. G. Nieto, H. D. B. Restrepo, and I. 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)*. IEEE, 2019, pp. 2227–2231.

[18] A. Beghdadi, M. Mallem, and L. Beji, "Benchmarking performance of object detection under image distortions in an uncontrolled environment," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2071–2075.

[19] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[20] M. T. Hossain, S. W. Teng, D. Zhang, S. Lim, and G. Lu, "Distortion robust image classification using deep convolutional neural network with discrete cosine transform," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 659–663.

[21] Z. A. Khan, A. Beghdadi, M. Kaaniche, F. Alaya-Cheikh, and O. Gharbi, "A neural network based framework for effective laparoscopic video quality assessment," *Computerized Medical Imaging and Graphics*, vol. 101, p. 102121, 2022.

[22] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.