IAC–24–B2.2.10.x91024

Towards Domain Gap Bridging via Synthetic VIS Sensor Model

Michele Bechini[*], **Lucia Bianchi**[†], **Michèle Lavagna**[‡]

Autonomous spacecraft relative navigation has gained significant attention in recent years due to its potential for future space missions. A key component of this technology is relative state estimation, often performed using monocular images captured by cameras operating in the visible light spectrum. The ESA's Satellite Pose Estimation Challenge (SPEC) of 2019 highlighted that the highest accuracy in pose estimation is achieved using CNN-based feature regression combined with PnP solvers. However, SPEC2021 demonstrated accuracy drops when pose estimation models trained on synthetic images are tested on real-world images, due to inherent differences between synthetic and actual image data. To address this issue, this paper proposes a novel synthetic VIS (Visible Spectrum) sensor model. This model enhances the realism of synthetic images by accurately simulating the image formation process, from the reception of photons by the detector to the output image, including noise characteristics typically present in real sensors but not in standard rendering pipelines. The developed model relies on sensor datasheet parameters to approximate photon reception and applies disturbances characteristics of CMOS and CCD detectors, including fixed pattern noise, resulting in synthetic images that closely resemble real-world conditions. The developed model has been compared with frames acquired by a real camera, demonstrating its high-level fidelity and superior representativeness compared to commonly used additive white Gaussian noise. Furthermore, the VIS sensor noise has been applied as an image augmentation technique in the training phase of a pose estimation CNN, leading to significantly improved performance metrics on both synthetic and mockup frames, despite the model being trained only on synthetic images, effectively bridging the domain gap.

keywords: Domain Gap, Monocular Images, Sensor Noise, Relative Navigation, Pose Estimation,

1. Introduction

Autonomous spacecraft relative Guidance, Navigation, and Control (GNC) systems are critical technologies for future space missions that require high autonomy [1] to perform time-sensitive tasks across various scenarios such as spacecraft rendezvous and docking, active debris removal, and on-orbit servicing, where precise relative navigation is essential to ensure mission success [2–6]. One of the most challenging applications of autonomous relative navigation involves target-chaser systems [7], especially when the target is non-cooperative, meaning no communication link or light-emitting markers exist on the target. Onboard relative state estimation must be autonomous, due to the lack of accuracy and timeliness of ground-based solutions, relying solely on sensors and computing resources available on the chaser. Among the available sensors, monocular cameras operating in the Visible Spectrum (VIS) are commonly preferred for capturing relevant measurements for the GNC system, due to their heritage, reliability, and the ability to provide detailed images of the target, which can be processed to estimate the relative pose between the chaser and the target spacecraft [7, 8].

The advent of publicly available datasets like the Spacecraft Pose Estimation Dataset (SPEED) [9] and international competitions such as the ESA's Satellite Pose Estimation Competitions (SPEC2019 [10] and SPEC2021 [11]) has fueled the development of VIS camera-based navigation solutions. These resources provide synthetic images to train algorithms, pushing forward the state of the art in spacecraft relative navigation. SPEC2019, the first Satellite Pose Estimation Competition, played a pivotal role in benchmarking the performance of various pose estimation algorithms. Namely, the dataset developed for this competition, SPEED, features a synthetic training set, while the test set was split into two parts, a synthetic one and a mock-up-based one. The outcomes of SPEC2019 proved that the best option to deal with autonomous relative pose estimation is to use CNN-based feature regression coupled with classic PnP solvers [10]. However, the outcomes also highlighted a critical issue, i.e., most top-

[*] Polytechnic University of Milan, Italy, michele.bechini@polimi.it

[†] Polytechnic University of Milan, Italy, lucia.bianchi@polimi.it

[‡] Polytechnic University of Milan, Italy, michelle.lavagna@polimi.it

performing architectures showed a significant drop in accuracy when applied to mock-up images, which are more representative of real-world conditions than synthetic datasets. This degradation was primarily due to the substantial domain gap between synthetic training images and the real images seen only during testing [11]. Many algorithms excelled on synthetic frames but failed to generalize well to mock-up images, indicating that these models were overfitting to the synthetic data characteristics rather than learning robust features transferable to real-world scenarios. The SPEC2021 competition and its related dataset, SPEED+ [12], further underscored the domain gap problem. Namely, the dataset was developed to stress further the gap between training and testing images, featuring a synthetic-only huge training set and two mock-up-only testing splits acquired under different (and severe) lighting conditions [11]. Despite advancements in pose estimation architectures, the competition revealed that most algorithms continued to perform poorly on mock-up images unless using huge models or domain-discriminator and adaptation modules, both impractical for space applications. This gap was attributed to the differences in lighting, noise, and textures between the two types of images, which most architectures failed to account for during training [11].

The SPEC2021 results confirmed that the degradation in pose estimation accuracy seen in SPEC2019 was not an isolated issue but a systemic problem affecting most current architectures. The competition highlighted the need for improved training methodologies and dataset validation to bridge the gap between synthetic and real-world images toward reliable autonomous navigation systems deployable in actual space missions [11]. A valuable approach to bridge the domain gap was presented in [13], where the authors leveraged multitask learning and dedicated image augmentations. Namely, the augmentations add to the synthetic frames during training optical artifacts (e.g., sun flares and random erase) and texture randomization via AI-based neural style transfer methods. These techniques allowed a higher level of abstraction on the learned features, leading to an effective boost of the performance on mock-up images. Similarly, the work presented here focuses on image augmentations. Namely, by noticing that one of the main differences between a synthetic and an actual frame is the presence of the detector in the latter one, the work presented was dedicated to the development of a high-fidelity model for the detector and the sensor to obtain the noised version of the input noiseless images, achieving a higher degree of realism than the standard additive white Gaussian noise that is commonly applied to synthetic images (e.g., in SPEED and SPEED+ synthetic frames), constituting the main contribution of this paper. The noise level obtained from the model has been successfully compared with actual frames acquired with a CMOS sensor and with frames noised by using an additive white Gaussian noise, demonstrating the high fidelity of the proposed model and the enhancement of the photorealism with respect to the Gaussian noise. Further, to establish the efficacy in generalizing CNN trained on synthetic images also to mock-up ones, the pipeline to apply the high-fidelity noise from the introduced sensor model has been adopted as an augmentation during the training of a YOLOv8s-pose model trained on synthetic SPEED images, proving the enhancement of the performance in both target and keypoint detection in synthetic images and mock-up frames from SPEED+.

The paper provides a brief overview of already available detector and image noise models in Sec. 2, while the detailed description of the proposed model is reported in Sec. 3. Sec. 4 provides the comparison of the developed model with respect to an actual sensor, while the outcomes of the training of the YOLOv8 model using the proposed approach as image augmentation are discussed in Sec.5. Lastly, the main outcomes and possible future developement are summarized in Sec. 6.

2. Literature Review

VIS sensors (i.e., photon detectors operating in the visible spectrum of light) convert the incoming photons that hit the sensor into a digital representation as Analog-to-Digital Units (ADUs) or Digital Numbers (DNs) in grayscale images. The sensitive part of VIS sensors typically consists of photosensitive semiconductor materials, such as silicon. When incident photons strike this material, they interact with its atoms, transferring their energy to electrons within the semiconductor. This process moves electrons from the valence band to the conduction band, leaving positively charged holes and generating electron-hole pairs. These electron-hole pairs constitute electric charge carriers, with the number of pairs generated proportional to the incident photon's energy. The charge carriers are then collected and stored, initiating the conversion of photons into electrical charge within the sensor. Quantum Efficiency (QE) is a critical factor in this process, as it quantifies the efficiency of the photon-to-charge conversion.

QE is wavelength-dependent and varies with the material's properties, impacting the number of charge carriers produced per absorbed photon. Notably, the overall efficiency of the photon-to-charge conversion is also affected by sensor design factors (e.g., fill factor and optical throughput efficiency) [14] but, as per EMVA1288 standards [15,16], the QE value reported in datasheets refers to *total* QE, i.e., already including these sensor-dependent factors (when not differently specified). The charges collected are then converted into voltages by dedicated capacitors. The voltages are then amplified, digitalized, and quantized into discrete levels by a dedicated Analog Digital Converter (ADC) with characteristic ADC Gain (measured in $[e^-/ADU]$). This last step enables the sensor to represent the captured light intensity in ADU or DN.

Notably, the two most widely adopted sensors are CCD (Charge-Coupled Device) and CMOS (Complementary Metal-Oxide-Semiconductor) sensors [14] that differ in the quantification of the stored charge. CCD sensors follow a systematic charge transfer process across the chip with a centralized reading, where the analog-to-digital converter transforms the charge from each photosite (i.e., pixels) into a digital value. In contrast, CMOS sensors employ multiple transistors at each photosite to amplify and guide the charge through conventional wires, i.e., adopting an individual reading of each photosite. These differences in the charge reading modes lead to the following main differences between CCD and CMOS [17]:

- Image Quality: CCD sensors excel in producing high-quality, low-noise images, while CMOS sensors are more susceptible to noise.
- Light Sensitivity: CMOS sensors have lower light sensitivity due to transistor interference.
- Power Consumption: CCD sensors consume considerably more power than CMOS counterparts.
- Cost: CMOS sensors are cost-effective due to standard manufacturing processes.

CCD and CMOS can be described through the same mathematical model [15]. Hence, given the input photons per pixel μ_p received by the sensor, it is possible to compute the pixel response in DN μ_{DN} as:

$$\mu_{DN} = QE\,\frac{\mu_p}{G_{ADC}} \qquad [1]$$

where G_{ADC} is the ADC Gain in $[e^-/ADU]$. The mean charge generated can be retrieved as $\mu_{e^-} = QE\,\mu_p$. Notably, the input photons can be retrieved from a calibrated source during hardware calibration phases as per EMVA1288 standards, or the value for μ_p can be approximated as in [15,18] by knowing the irradiance on the sensor surface.

2.1 *VIS Sensor Noise*

The pixel response computed in Eq. 1 refers to the mean value since the true value given as output will be affected by noises that arise during the image generation phase. In a first approximation, the main noises affecting VIS images are the photon shot noise, the dark current shot noise, and the readout noise [14, 18, 19]. Photon shot noise arises from random fluctuations of photons hitting the sensor thus it can be modeled as a Poisson process as a function of the mean value of photons collected during the exposure time [19]:

$$N_{ph\_shot,\,p} \sim \mathcal{P}(\mu_p) \qquad [2]$$

The dark current shot noise is due to the random generation of charges within the sensor, even in the absence of incoming light [19], arising by temperature-dependent stochastic random processes. Hence, they can be modeled as a Poisson distribution. The mean value for the dark current shot noise is a function of the exposure time t_{exp} and the average dark current D_R (which is a function of the sensor temperature and is expressed as $[e^-/px/s]$). Hence, the dark current shot noise is modeled as [19]:

$$N_{dc\_shot,\,e^-} \sim \mathcal{P}(t_{exp} D_R) \qquad [3]$$

The readout noise is not signal-dependent. It occurs during the reading of the electrical signal and is commonly associated with the electronic components and circuitry used to amplify and digitize the analog signal generated by the sensor [20]. In first approximation, it can be modeled as an additive white Gaussian noise with zero mean and prescribed standard deviation that is a characteristic of each camera, i.e., $\mathcal{N}(0, \sigma^2_{read})$. This approximation is well established [15,16], but recent studies [20] proved that, in low-light conditions, a Tuckey-Lambda (TL) distribution [21] can better approximate the readout noise of real camera assemblies due to the higher weight associated to the distribution tails. Hence, by adopting the TL distribution, the readout noise can be modeled as [20]:

$$N_{readout,\,e^-} \sim \mathcal{TL}(\lambda, 0, \sigma_{read}) \qquad [4]$$

where the shape parameter, λ, shall be estimated during the camera noise characterization. Namely, if

$\lambda = 0.14$, the TL distribution approximates a Gaussian distribution while, if $\lambda < 0$, it corresponds to heavy-tail distributions, making it suitable for modeling data with outliers and extreme values.

Notice that the photon shot noise in Eq. 2 shall be multiplied by $QE\,/G_{ADC}$ to retrieve it in DN, while both the dark current shot noise in Eq. 3 and the readout noise in Eq. 4 shall be divided by G_{ADC} to retrieve their values in DN since they are already in electrons. The final noise value in DN can be retrieved by computing the root mean square of the noise sources in Eqs. 2 – 4 [15, 18, 20]:

$$N_{DN} = \sqrt{\sigma^2_{ph\_shot,\,DN} + \sigma^2_{dc\_shot,\,DN} + \sigma^2_{readout,\,DN}} \quad [5]$$

$$= \frac{1}{G_{ADC}} \sqrt{\mu_p\, QE^2 + t_{exp} D_R + \sigma^2_{read,\,e^-}} \quad [6]$$

The general equation provided above can be used to apply a cumulative noise to the image by using an additive white Gaussian noise with standard deviation as in 6. Notice that this procedure does not take into account fixed pattern noises, like Photo Response Non-Uniformity (PRNU) and Dark Signal Non-Uniformity (DSNU), that characterize each sensor and that can be relevant in low light or high gain conditions for CMOS sensors [19, 20], generating row and column noises that can affect the IP algorithms. Their contributions shall be estimated within the noise characterization phases [20] and included in a noise model with higher fidelity due to their relevance, especially for CMOS, as detailed in the next sections. Further, the additive Gaussian white noise does not retain the characteristics of each noise component (e.g., the dependency of the photon shot noise from the incoming light), resulting in a "uniform" noise map and losing all the modeling effort discussed above. Moreover, the noise can not be related to the parameters of the image acquisition system, e.g., the gain setting, the bit depth, the exposure time, and the operating temperature. Consequently, the work presented here builds a pipeline from the single noise contribution reported above and already available in the literature to apply a high-fidelity noise that preserves the characteristics of each contribution, also including the fixed pattern noises contribution and, differentiating from the model in [20], applicable even in the absence of the physical sensor, relying only on its datasheet.

It is acknowledged that images can also be affected by optical artifacts that may depend on both the optical characteristics of the imaging system and the sensor type. The most common optical artifacts include lens flares [22], stray lights [23], and blooming [24]. These artifacts originate from complex phenomena that can be evaluated only by sophisticated image rendering tools that leverage physically-based powerful and complex raytracers with the possibility of simulating full lens set [22, 25]. Consequently, they are usually not modeled or included in VIS sensor models. A simple yet effective approach to include these artifacts is approximating and applying them in postprocessing, disregarding the physical phenomena that originate them and focusing on the effects on images, as in [13, 26].

3. VIS Sensor Model

The photon detector model has been developed starting from models already available in the literature discussed in Sec. 2.1 and comprises the primary noise sources for CCD and CMOS sensors. The model implemented leverages the linear photon detector model prescribed by the EMVA1288 standard [15] and the models implemented in [19, 20], subdividing the image generation process into sequential steps belonging to the *photon space*, the *electron space*, and *digital number space*, as shown in Fig. 1.

3.1 *Photon Space*

The input for the pipeline in Fig. 1 is a matrix $I_{ph,\,init}$ shaped as the detector array size considered. $I_{ph,\,init}$ contains in each element the mean number of photons collected by the corresponding $(i,\,j)$-th element of the detector. This value can be retrieved by knowing the radiation received by the sensor as in [15, 19] or, when dealing with noiseless images, it can be approximated by knowing the ADC Gain G_{ADC} and the total QE as:

$$I_{ph,\,init} = I_{clean} \frac{G_{ADC}}{QE} \quad [7]$$

where I_{clean} is the noiseless image. Notice that the G_{ADC} given in datasheets compliant with EMVA1288 standards are in 16-bit quantization. Notice that most cameras allow setting the gain level using values in decibels (dB) while providing the base 16-bit gain in $[e^-/ADU]$. Hence, the correct gain value in $[e^-/ADU]$, corresponding to the selected level in dB, shall be retrieved before applying Eq. 7 and the pipeline in Fig. 1. Once that $I_{ph,\,init}$ is retrieved, the *Photon shot noise* is defined from a Poisson distribution and summed to the mean photon values as:

$$\begin{aligned} I_{ph\_shot,\,ph} &= \mathcal{P}(I_{ph,\,init}) \\ I_{ph} &= I_{ph,\,init} + I_{ph\_shot,\,ph} \end{aligned} \quad [8]$$

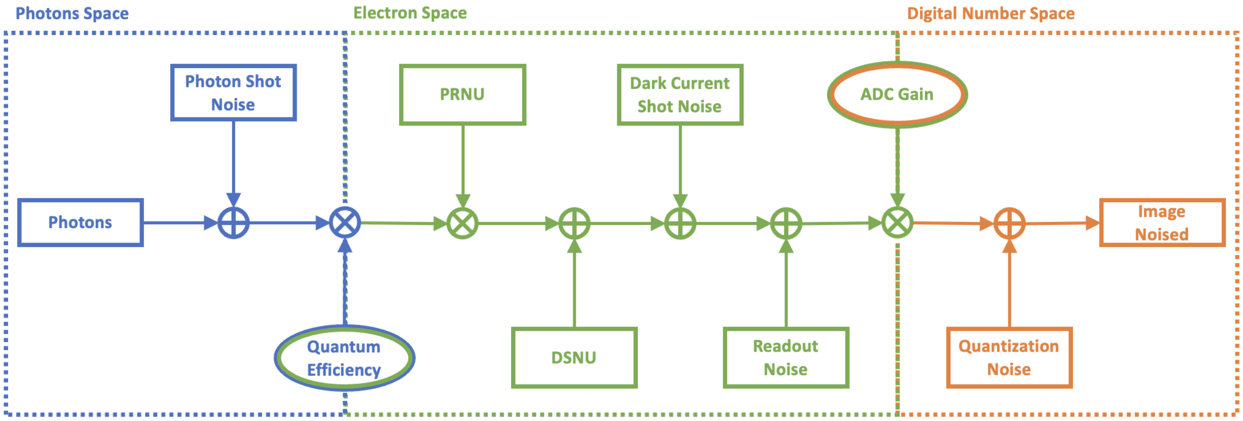


Fig. 1: Photon Detector Model.

Notice that most of the random number generators using the Poisson distribution directly output the I_{ph} matrix with the Poisson shot noise signal already added (i.e., $I_{ph} = \mathcal{P}(I_{ph,\,init})$. The matrix I_{ph} is then moved to the *electron space* by applying the QE:

$$I_{e^-,\,init} = I_{ph}\,QE \qquad [9]$$

3.2 *Electron Space*

In the *electron space*, the first noise source to be applied is the *PRNU*, a fixed pattern noise that arises from the non-uniform response of each pixel under uniform lighting conditions due to variations in the material composing the sensor [19]. The PRNU is expressed as a percentage of the electrons detected [15, 19]. In agreement with EMVA1288 standards, the PRNU can be characterized by column-wise, row-wise, and pixel-wise components. These components are reported in the datasheets that respect the latest EMVA1288 standards. In contrast, these components shall be estimated during the calibration phase using a uniform certified lighting source [15] for datasheets referring to previous releases of the standard. In case of unavailability of calibration parameters or datasheet values, the PRNU is modeled by assuming standard deviation values for each component in the range from 0.1% to 1% of the incoming light [15, 27]. The PRNU map can be defined for a general detector by sampling from a Gaussian distribution with zero mean and standard deviation $\sigma_{PRNU,\,col}$ a single row and repeating it for all the rows of the array size. The same approach is repeated for the row component by sampling a single column using $\sigma_{PRNU,\,row}$ and repeating for all the columns. Concerning the pixel-wise contribution, a map of equal size to the camera array is sampled from a Gaussian distribution with standard deviation $\sigma_{PRNU,\,pix}$. The sum of these three maps is the overall PRNU map K_{PRNU}. The PRNU map is generated only once for each sensor model, in agreement with [19]. Notice that the map shall be multiplied by the detected electrons since it contains scaling values. Hence, the overall electrons with applied the PRNU noise can be computed as [19]:

$$I_{PRNU,\,e^-} = I_{e^-,\,init}\,(1 + K_{PRNU}) \qquad [10]$$

The second noise applied in the electron space is the *DSNU*. This noise is not light-dependent and can also be detected in images acquired with capped lens [15]. The DSNU resembles the same characteristic of the PRNU from the modelization point of view. Still, its components (column-, row-, and pixel-wise) are usually defined in electrons in datasheets compliant with the EMVA1288 standards. Also, the DSNU can be estimated from bias frames (short exposure time and capped lens) acquired with the actual sensor during the noise characterization phase [15]. The DSNU noise map in electrons $I_{DSNU,\,e^-}$ can be retrieved from the values of $\sigma_{DSNU,\,col}$, $\sigma_{DSNU,\,row}$, and $\sigma_{DSNU,\,pix}$ available in datasheets by following the same procedure already outlined for the PRNU. Even in that case, the DSNU map is generated only once for each sensor model, being a fixed pattern noise. Please notice that sampling the PRNU and DSNU noise components from Gaussian distributions does not allow retrieving the actual spatial frequencies of these noises that characterize each sensor. Hence, these models shall be refined further to include the noise components obtained from the real sensor through a dedicated analysis.

The subsequent additive noise is the *dark current shot noise*. The dark current shot noise and its model have already been detailed in Sec. 2.1. The dark current shot noise map in electrons $I_{dc\_shot,\,e^-}$ is computed from a Poisson distribution by knowing the exposure time t_{exp} and the dark current $D_R(T)$ [19]. The value for D_R at a reference temperature can be retrieved from a datasheet compliant with EMVA1288 standard in $[e^-/s]$. Alternatively, if the datasheet does not provide this value, it can be retrieved from dark images (variable exposure time at fixed temperature and lens capped).

The last noise source in the electron space is the *readout noise*. Also, the readout noise has already been discussed in Sec. 2.1. The readout noise map in electrons $I_{read,\,e^-}$ is retrieved from a Tuckey-Lambda (TL) distribution where the shape parameter λ is assumed as equal to 0.14 (i.e., approximately Gaussian distribution) in a first approximation and the standard variation is retrieved from the sensor datasheet. The parameter λ and the standard deviation can also be retrieved from bias images captured with the actual sensor during its characterization [15, 20]. Notice that the readout noise will include all the noise sources not directly modeled that still affect the signal in a real scenario (e.g., thermal noise and source follower noise) [20].

The final electron matrix I_{e^-} can be retrieved as in Eq. 11 by adding the contribution of these last three additive noise sources to the electron counts retrieved after the introduction of the PRNU (see Eq. 10).

$$I_{e^-} = I_{PRNU,\,e^-} + I_{DSNU,\,e^-} + I_{dc\_shot,\,e^-} + I_{read,\,e^-} \qquad [11]$$

The final electron matrix I_{e^-} is rounded to integer values and then shifted into the *digital number space* by using the ADC Gain:

$$I_{DN,\,init} = \frac{round(I_{e^-})}{G_{ADC}} \qquad [12]$$

3.3 *Digital Number Space*

The only noise source in the *digital number space* is the quantization noise. The quantization noise arises when an analog continuous signal is quantized into a discrete digital signal. The quantization noise map $I_{q,\,DN}$ is sampled from a uniform distribution within the range $[-0.5,\ 0.5]$ and it is added to $I_{DN,\,init}$ from Eq. 12 to retrieve the final image array in DN:

$$I_{DN} = I_{DN,\,init} + I_{q,\,DN} \qquad [13]$$

Notice that the output value from Eq. 13 is rounded and clipped to the range $[0, 2^{16} - 1]$ in 16-bit to account for possible saturations before rescaling the image to the original bit depth.

It is emphasized that the ADC Gain will affect only the noises if the model described above is applied as VIS sensor noise generator since the gain is used in Eq. 7 to retrieve the initial photon values from the clean image and in Eq. 12 the same gain value is used to retrieve the DN values for the noised image. Hence, if the model schematized in Fig. 1 is adopted as a VIS sensor noise generator, the noise level increases if the ADC Gain in dB raises.

4. Comparison with an Actual Sensor

A VIS CMOS sensor has been characterized to extract the noise parameters needed as input to the model discussed above to assess its validity. Namely, the noisy images generated with the model are compared with those captured with the real sensor. Please notice that the light-dependent noise sources are not included in the actual sensor characterization phase due to limitations in the available facility. Hence, only the DSNU, the dark current shot noise, and the readout noise have been retrieved. Due to that, a complete validation of the model, including photon shot noise and PRNU, is still missing and is forecasted for the next development. Despite that, it is worth mentioning that the terms included in the conducted comparison give the highest contributions to the overall final noise [15]. Further, it is remarked that the characterization of the CMOS sensor was required since its datasheet is compliant with an old version of the EMVA1288. Despite that, it is worth underlying that all the parameters extracted from the noise characterization reported in Sec. 4.1 are available in the datasheet compliant with the latest EMVA1288 v.4, please refer to [15] for a sample datasheet.

4.1 *Real Detector Noise Characterization*

The camera adopted is the Teledyne FLIR CM3-U3-13Y3C (ex Point Grey Chameleon 3), a CMOS camera whose datasheet can be retrieved online from the producer's website. This camera has been selected since it is the model employed in the PoliMi-DAER facility dedicated to GNC algorithm tests [28]. The objectives are to estimate the shape parameter and the standard deviation for the readout noise, characterize the DSNU in terms of column-, row-, and pixel-wise components, and evaluate the dark current contribution. The bias and dark frames are acquired to perform the characterization following standard procedures adopted to evaluate these noise sources for astronomic photography. Namely, 200 bias frames have been acquired with a capped lens (ensuring that light does not reach the sensor), exposure time of 0.01 ms, and gain equal to 3 dB. The dark frames have been acquired using a gain of 3 dB, increasing values of exposure time (namely $t_{exp} = [1, 10, 20, 500, 999]$ ms), and increasing temperature of the sensor (namely $T = [15, 19, 23, 27, 30]$°C). Ten dark frames have been retained for each temperature and exposure time value, resulting in 250 images overall. All the images have been acquired using the Mono16 format (i.e., 16-bit grayscale format) and saved as *.raw* files to avoid compression and preserve the original image quality. Due to their nature, bias frames allow retrieving both the readout noise parameters and the DSNU (also named master bias frame) that can be further analyzed. Setting the exposure time to the minimum allowed by the camera entails lowering the dark current shot noise effects that can be assumed to be negligible. On the contrary, the dark frames are captured using long exposure times, making it possible to include the contribution of the dark current shot noise. Hence, once the readout noise and the DSNU are estimated, it is possible to remove their contribution from the dark frames and retrieve insights on the temperature and exposure-dependent noise.

4.1.1*DSNU Characterization*

The DSNU image (i.e., the master bias) is estimated in DN from the pixel-wise average of all the bias frames acquired. The averaging reduces the random noises while preserving the fixed pattern noise [15]. Due to a lack of DSNU information on the camera datasheet, all the parameters to characterize this fixed pattern noise are retrieved from the non-synthetic DSNU. The column and row noise components are recovered by computing the mean, per column and per row, respectively, of the DSNU image. The values retrieved in DN are converted into electrons using the known ADC Gain to provide a characterization independent from the gain adopted. The DSNU image (scaled for visibility), with reported row and column noise components, is shown in Fig. 2. The columns component is the main contribution to the DSNU for the CMOS sensor adopted simulated. The column noise shows a discretization into three bands at about 11, 6, and 1 electrons with a spatial frequency that leads to higher noise values in the column with an index between 0 and 250. The row noise component is more well-shaped and lower in electrons than the column noise, with an evident spatial frequency highlighted from the plot in Fig. 2. The pixel-wise component of the DSNU is characterized by retrieving its standard deviation in electrons as in Eq. 14 by assuming that all the noise components are independent.

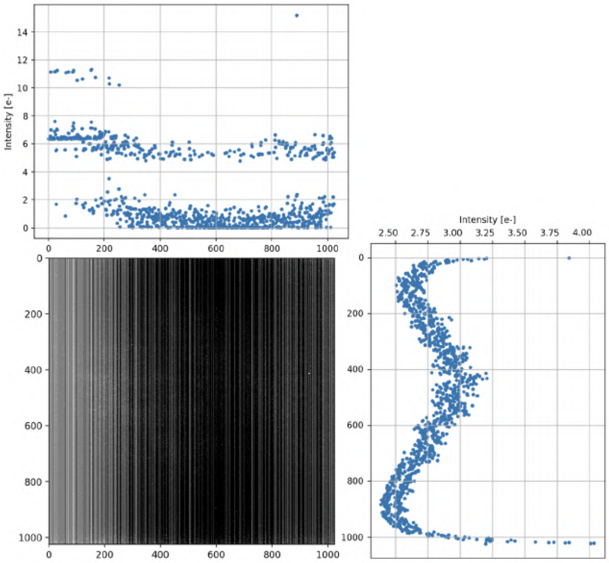


Fig. 2: DSNU image (magnified intensity) with column (top) and row (right) noise components in e^-.

$$\sigma_{DSNU,\,pix} = \sqrt{\sigma^2_{DSNU} - \sigma^2_{DSNU,\,col} - \sigma^2_{DSNU,\,row}} \qquad [14]$$

Notice that σ^2_{DSNU} is the variance in electron squared of the acquired DSNU, while $\sigma^2_{DSNU,\,col}$ and $\sigma^2_{DSNU,\,row}$ are the variances in electrons squared of the column and row components of DSNU. The estimated standard deviations for each DSNU noise component are $\sigma_{DSNU,\,col} = 2.82\ e^-$, $\sigma_{DSNU,\,row} = 0.21\ e^-$, $\sigma_{DSNU,\,pix} = 1.1\ e^-$.

A synthetic DSNU can be computed in first approximation as discussed in Sec. 2.1, i.e., without considering the spatial frequencies highlighted in Fig. 2 and relying only on the values for $\sigma_{DSNU,\,col}$, $\sigma_{DSNU,\,row}$, and $\sigma_{DSNU,\,pix}$ being the only DSNU-related values reported on EMVA1288-compliant datasheets. A visual and histogram comparison between the DSNU estimated from bias images and the synthetic DSNU generated (named *Synthetic Gaussian*) in DN is shown in Fig. 3, where both the DSNUs are scaled equally for visibility. Fig. 3 demonstrates that the synthetic Gaussian DSNU retains the main noise features of the actual DSNU, being primarily composed by the column-wise noise with slightly visible row-wise components. Despite that, the synthetic Gaussian DSNU has a mildly lower intensity in DN than the real one, also highlighted by the histogram

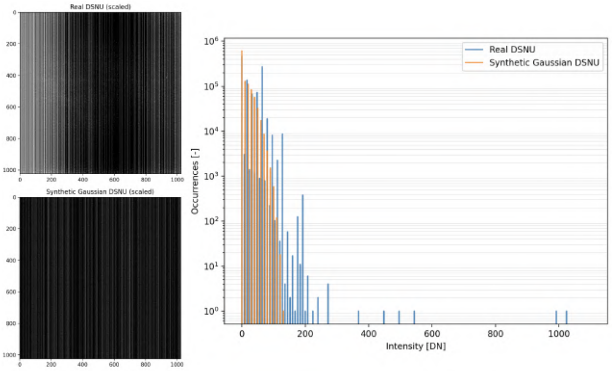


Fig. 3: Comparison of Real and Synthetic Gaussian DSNU.

comparison. Notice that the actual DSNU shows a few defected pixels, with associated intensity levels higher than 200 DN, not included in the model. As expected, the synthetic Gaussian DSNU does not resemble any spatial frequency highlighted in the actual DSNU due to the model adopted. Despite the limitations pointed out, the simple approach to approximate the DSNU is still valuable since the column-wise noise component of the actual CMOS is retrieved. Hence, it could be adopted to test IP algorithms in a scenario more representative of the actual camera behavior than those cases in which an additive Gaussian noise is applied to the entire image.

It is acknowledged that a high-fidelity approximation of the actual DSNU that retrieves even the spatial frequencies of the column- and row-wise component can be achieved by using Gaussian Multivariate Mixture (GMM) or polynomial fitting to the actual values, as in [26]. An example of the synthetic DSNU achieved by following the procedure detailed in [26] is reported in Fig. 4 As it can be noticed from Fig. 4, the synthetic DSNU better approximates the real DSNU both in terms of noise intensity (as shown by the histogram comparison) but also in terms of spatial frequency of the primary noise contributions.

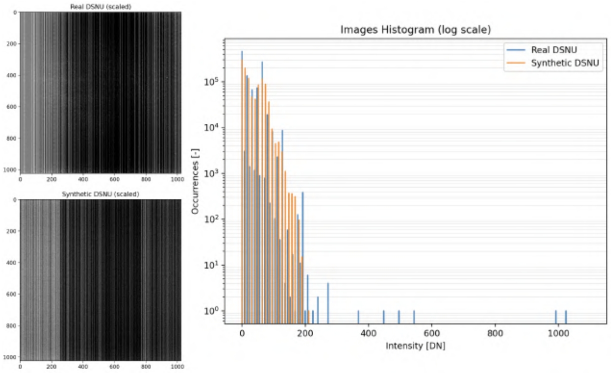


Fig. 4: Comparison of Real and Synthetic DSNU.

It is acknowledged that, by having the actual sensor available, it is possible to directly use the real DSNU inside the model outlined in Sec. 3 to replicate the camera behavior. In case the actual sensor is not available, the Gaussian approximation (leveraging the standard deviations of DSNU noise components from datasheets) allows retrieving a synthetic DSNU that resembles the main features of the actual DSNU (e.g., the relative magnitude of each noise component). The knowledge of the spatial frequencies and mean values of each noise component of the actual sensor is required to increase the representativeness of the synthetic DSNU without actually having the sensor due to the effect on the accuracy of the model, as pointed out from the results achieved with the accurate synthetic DSNU.

4.1.2 *Readout Noise Characterization*

The readout noise standard deviation is estimated using the bias frames and the master bias (i.e., the DSNU) computed by averaging all the bias frames. To correctly retrieve the readout noise standard deviation in DN, the master bias is removed from all the bias frames to remove the fixed pattern noise, making the residual noise independent from the light and the exposure time. Hence, all the noise sources contributing to the residual noise are collected in the readout noise. The standard deviation of all the residual images is computed in DN. Then, the readout noise standard deviation is estimated as the mean value of the standard deviations of the residual images. The readout noise standard deviation σ_{read} is converted into electrons by applying the ADC gain. The standard deviation of the readout noise estimated from the acquired bias frames and the extrapolated DSNU is about $\sigma_{read} = 2.37$ electrons. The model adopted for the readout noise samples the noise values from a TL distribution hence, the shape parameter $lambda$ shall be estimated. The shape parameter has been defined within this work by comparing the mean histogram of all the residual images converted to electrons against the histogram derived from a readout noise image defined using the TL distribution with fixed mean and scale equal to σ_{read}. The estimated shape parameters that best approximate the mean histogram is $lambda = -0.23$ since its decay toward zero is smoother (due to the higher weight given to the tails) than a Gaussian-like distribution, leading to a better approximation of the noise components with a higher content in electrons, as shown in Fig. 5. This behavior is in agreement with the histograms reported in [15]It is acknowledged that more accurate methods to retrieve the shape parameters exist (e.g., leveraging more refined statistical approaches) and can be exploited to fine-tune the TL distribution, as pointed out in [20].

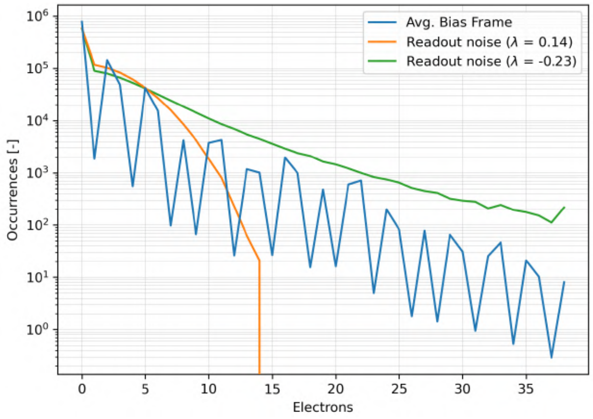


Fig. 5: Comparison of Gaussian-like and Heavy-Tail Readout noise distributions.

4.1.3 *Dark current Characterization*

The dark current D_R can be estimated from the dark images due to their long exposure time that makes relevant the contribution of the dark signal $S_{dark} = t_{exp}\, D_R$ [19]. Namely, the noise content in dark images is given by the DSNU, the readout noise and the dark current shot noise. It is possible to estimate the dark signal from the dark frames, captured with increasing exposure time and constant temperature, by measuring the linear increment of the mean value in DN, or electrons, as a function of the increment of the exposure time [15]. The dark current is then given by the slope of the linear increment detected. This procedure assumes that the exposure time does not affect the DSNU and the readout noises. A linear least square regression is performed on the dark signal data as a function of the exposure time to evaluate the slope accurately [15]. Namely, following the procedure outlined in [29], ten dark frames are captured for each operating temperature and exposure time. Then, the mean signal is computed by averaging the mean values for each dark frame at each operating condition (temperature and exposure time couples) while discarding the minimum and maximum mean values for each batch of 10 images [29]. The linear regression is then performed on the mean data. By performing the procedure outlined above for more operating temperatures, it is possible to retrieve a relation between the dark current D_R and the operating temperature. The relation between D_R and the temperature is fitted using an exponential function, as shown in Fig. 6, leading to the following relation:

$$D_R(T) = A \cdot \exp(B \cdot T) = 0.136 \cdot \exp(0.14 \cdot T) \quad [15]$$

Where the temperature of the sensor T is adopted in Celsius degrees and the coefficients $A = 0.136$ $[e^-/s]$ and $B = 0.14$ $[°C^{-1}]$ are retrieved from the exponential fitting shown in Fig. 6. Notably, the definition of a dark current temperature dependency is relevant for IP algorithms applied to long exposure images (i.e., when the dark current shot noise can give a high contribution to the overall noise), making it possible to simulate a VIS sensor noise coherent with the current operating temperature of the sensor itself. Namely, due to the exponential behavior of $D_R(T)$ and the long exposure time, the effect on the output noise intensity can be non-negligible.

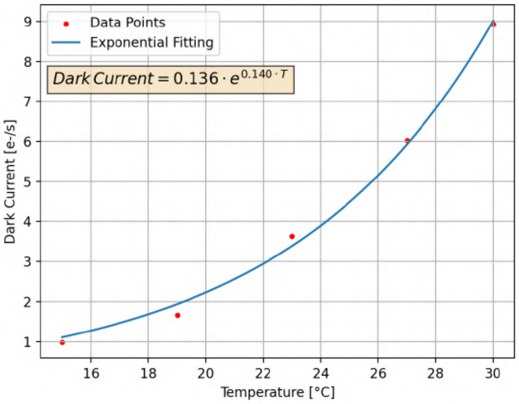
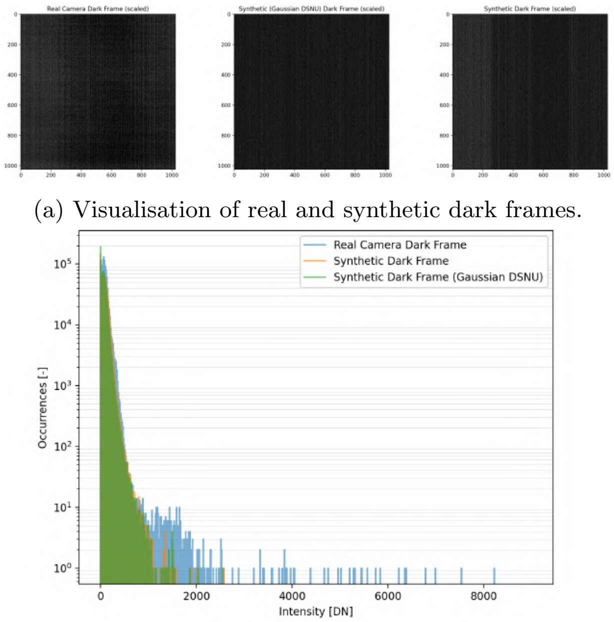


Fig. 6: Exponential fitting of dark current temperature dependency.

4.2 *Model Representativeness Verification*

The noise parameters estimated for the real sensor as in Sec. 4.1 have been adopted within the VIS sensor model outlined in Sec. 3 to generate a synthetic dark frame to be compared with a batch of dark frames acquired with the VIS camera adopting the same parameters. The dark frames are captured arbitrarily using an ADC Gain of 3dB and an exposure time of 0.5 seconds with a measured sensor temperature of 30°C. The input for the noise model is an array of zeros simulating that no photons within the VIS range hit the detector. As a consequence, both the photon shot and the PRNU noises will give a null contribution. To properly assess the modeling choices, both the accurate and the Gaussian synthetic DSNUs are adopted. The outcomes are reported in Fig. 7 by means of visual (Fig. 7a) and histogram comparison (Fig. 7b). Please notice that the dark frames shown in Fig. 7a are scaled uniformly to allow visual comparisons. The visual comparison of the dark frames in Fig. 7a reflects the outcomes of the analyses performed during the characterization of the DSNU. Namely, the synthetic Gaussian dark frame shows a lower intensity of the noise features associated with the DSNU contribution than both the real and the accurate synthetic dark frames. Despite that, both the models for the DSNU allow to generate synthetic dark frames that show a distribution of the noise into column-, row-, and pixel-wise components comparable to the real dark frame distribution, with the dark frames computed using the accurate synthetic DSNU capable of approximating also the spatial distribution of the fixed pattern noise. Despite that, the synthetic images show a lower intensity of the row-wise component given by the DSNU with respect to the real dark frame. This disparity can be attributed to an underestimation of the row-noise component at a high spatial frequency on the central portion of the frame not correctly retrieved by the GMM and polynomial models adopted. Despite that, both the synthetic images approximate the actual dark frame in terms of histograms (i.e., in terms of pixel-intensity count) as shown by Fig. 7b. The histograms computed for the two synthetic images almost overlap and follow the histogram computed for the dark frame after the initial peak at low pixel intensity. The disparity in the pixel count for intensity values higher than 1000 DN is due to outliers given by defected pixels [15] and the aforementioned row-noise component retrieved with lower intensity in the synthetic images. The mean and variance of the images involved in the comparison are reported in Tab. 1, further quantifying the accuracy of the model. Notice that the values for the non-synthetic images are averaged from the scores of 10 random dark frames. The values reported in the table confirm the analyses performed on the plots in Fig. 7, with a mean value of the Gaussian synthetic frame that is 16.6% lower than the reference mean value of actual images. The differences in the variance values are related to the unmodeled hot/cold pixels (i.e., the outliers), as already noticed in the discussion of the histogram comparison.



(a) Visualisation of real and synthetic dark frames.

(b) Histogram comparison for real and synthetic dark frames.

Fig. 7: Comparison between real and synthetic dark frames.

To further prove the validity of the Tuckey-Lambda distribution adopted for the readout noise, a histogram comparison between the real dark frame and two synthetic images generated by using both the reference ($\lambda = -0.23$) and a Gaussian-like ($\lambda = 0.14$) distribution to model the readout noise is shown in Fig. 8. As already noticed from Fig. 5, the Gaussian-like distribution is not well suited to model the readout noise since it does not approximate the noise component at high intensity but lower occurrences,

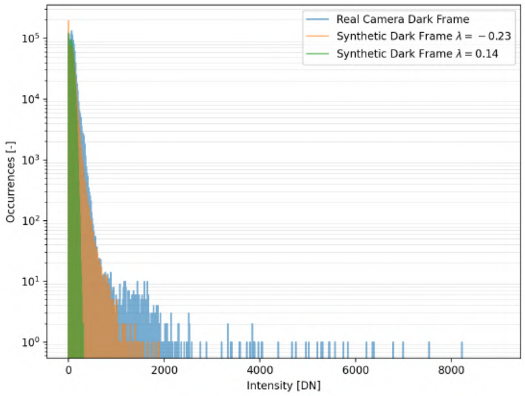


Fig. 8: Effect of shape parameter λ in readout noise model.

Table 1: Pixel intensity mean and variance for real and synthetic dark frames

Source	**Intensity Mean [DN]**	**Intensity Variance [DN^2]**
Real Dark Frames	78.4	5586.2
Synthetic Dark Frame	81.5	4531.5
Synthetic Dark Frame (Gaussian DSNU)	65.4	3736.9

i.e., the tails of the distribution, leading to the selection of a shape parameter for the Tuckey-Lambda distribution that best fit the higher weight of the tails. Please notice that, in agreement with Fig. 5 and Fig. 8, the deviation from the Gaussian behavior is substantial for the camera characterized. Hence, the widely adopted approach of applying only a Gaussian noise to synthetic images is a strong approximation of the actual behavior of a VIS sensor that well approximates only the initial noise peak, i.e., the low-intensity noise with the highest occurrences. It is remarked that even the shape parameter is a characteristic of each sensor, as proven in [20]. Hence, it shall be carefully evaluated for each camera, also because the readout noise is the main contribution to the overall noise, as highlighted by Fig. 8.

The outcomes of the comparison between the non-synthetic dark images acquired with a standard VIS camera and the model defined within this work point out that the model offers a better approximation of the actual VIS sensor noise than the simpler additive white Gaussian noise. Even if the spatial frequencies of the fixed pattern noise are not retrieved (i.e., using the Gaussian approximation of the DSNU), the model recovers the main noise features of the actual sensor, even if the overall mean intensity is slightly lower than non-synthetic images. Detailed knowledge of the DSNU with known spatial frequencies for both the column- and row-wise noise components is mandatory to further improve the model's representativeness. It is acknowledged that a proper validation of the model shall also include the light-dependent noise sources and a more detailed evaluation of all the parameters, including the actual ADC Gain. Despite that, evaluating these parameters requires a facility with dedicated instruments (e.g., a calibrated noise source [15]).

5. Application as Image Augmentation

The VIS sensor model pipeline detailed in Sec. 3 has been applied as image augmentation techniques during the training of a CNN to assess the improvements of the trained model in bridging the domain gap between synthetic and mock-up images by leveraging the more realistic noise applied to the synthetic training images. The scenario selected is the relative pose estimation from monocular images and the selected CNN is the YOLOv8s-pose, being a lightweight model capable of performing target detection and keypoints regression through a single inference. The model has been trained on synthetic SPEED images [9], being widely adopted as the benchmark dataset for monocular relative pose estimation task [30–32]. Namely, the training, validation, and test splits have been extracted from the original training set of SPEED in lack of available annotations for the actual SPEED synthetic test set, similarly to [32]. The mock-up frames from the *lightbox* and *sunlamp* subset from SPEED+ [12] have been adopted as additional test sets. Notably, the keypoint and Region of Interest (ROI) annotations not available for SPEED and SPEED+ images have been extracted by applying the procedure in [26, 32]. Overall, the dataset comprises 7680 synthetic SPEED training images, 1920 synthetic SPEED validation images, 2400 synthetic SPEED testing frames, 2791 mock-up testing images from SPEED+ sunlamp, and 6740 mock-up testing frames from SPEED+ lightbox. Please notice that the frames in the test sets for both SPEED and SPEED+ datasets are never used during the training phase of the YOLOv8-pose model.

The performance in target detection of the selected YOLO model on the test set is evaluated using the intersection-over-union (IoU) index and Average Precision (AP). IoU measures the overlap percentage between the predicted and ground truth bounding boxes, with a higher IoU indicating greater accuracy. AP represents the area under the precision-recall curve, where precision is the ratio of correct predictions (true positives) to all predicted boxes (true positives + false positives), and recall is the ratio of true positives to the total ground truth boxes (true positives + false negatives). Precision-recall curves are generated by varying the IoU threshold, and the mean AP (mAP) is computed by averaging AP values across different IoU thresholds, from 50% to 95%, in 5% increments. A higher AP_{50}^{95} value indicates more accurate ROI detection. The keypoint regression performances of the YOLOv8s-pose model have been evaluated using the Object Keypoint Similarity (OKS) and the keypoint Average Precision (AP). The OKS is a standard metric for keypoint regression that measures the proximity of predicted to ground truth keypoints, calculated using the equation:

$$OKS_i = \exp\left(-\frac{d_i^2}{2\sigma_i^2 k_i^2}\right) \qquad [16]$$

Here, d_i is the Euclidean distance between the estimated and ground truth i-th keypoint, σ_i is the falloff factor (assumed constant at $1/n_{kp}$, where n_{kp} is the total number of keypoints), and k_i^2 scales OKS based on the ROI area. OKS for keypoints is analogous to IoU for bounding boxes, and Average Precision (AP) for keypoints is calculated as the area under the precision-recall curve for different OKS thresholds. The AP_{50}^{95} is the mean AP for OKS thresholds from 50% to 95%, with a higher AP_{50}^{95} indicating more accurate keypoint detection.

The YOLOv8s-pose model has been trained on synthetic SPEED images, adopting different augmentation strategies added to the baseline model to assess the capabilities of bridging the domain gap between synthetic and mock-up images. By analyzing the mock-up images in the SPEED+ lightbox and sunlamp datasets and leveraging the VIS camera sensor model and its noise sources, a dedicated augmentation named *VIS sensor augmentation* has been developed. Namely, this augmentation is adopted to improve the model performances on mock-up images, leveraging a more representative noise. In doing so, all the hyperparameters of the model described in Sec. 4 have been randomized for each processed image to avoid overfitting the CNN on a single noise level. Namely, the values obtained from the characterization of the actual CMOS sensors have been taken as mean values for Gaussian distributions from which the actual values used for each frame of the training set have been sampled. An example of a SPEED synthetic frame and its augmented version using the VIS sensor augmentation is shown in Fig. 9.

The baseline model has been trained firstly by using the default augmentations of YOLOv8 (i.e., additive Gaussian noise, random rotations, blurring, etc.) and, in a second run, by adding the VIS sensor augmentation introduced in this work to all the images. The training hyperparameters have been maintained equal among the two runs to ensure consistency. These parameters are not reported here for brevity but they can be retrieved from [26]. The trained models have been evaluated in terms of mean IoU, mean OKS, and their respective AP_{50}^{95} on the test images from SPEED, SPEED+ lightbox, and SPEED+ sunlamp. The outcomes of this ablation study are reported in Tab. 2 and Tab. 3. The outcomes point out that the VIS sensor noise augmentation effectively increases all the evaluated metrics in synthetic and mock-up images, for both ROI and keypoint detection. It is worth pointing out a huge improvement in ROI detection for SPEED+ mock-up images, more intense in lightbox images than in sunlamp ones. Also, the keypoint regression metrics for muck-up images are strongly improved by the introduction of the VIS sensor noise augmentation, even if the relative increment is slightly lower than the metrics associated with the ROI detection. It is worth pointing out that the improvements of the metrics for both ROI and keypoint detection are higher for SPEED+ lightbox images than for those belonging to the SPEED+ sunlamp set due to the features of the images themselves. Namely, the images in the sunlamp set also contain optical artifacts such as flares, blooming, and saturations that make the dataset extremely complex, also because these artifacts are not present in the training images nor the augmentations considered in this work. Despite that and even if the achievement of a top-performing model is beyond the scope of this work, it is acknowledged that the artifacts can still be included via dedicated augmentations, as in [13, 26], leading to high-level performances also in sunlamp images. The outcomes for the synthetic test set from SPEED registered using the VIS sensor noise augmentation point out that training with a randomized noise higher and more complex than the actual noise already included in the images is beneficial in increasing the overall accuracy.

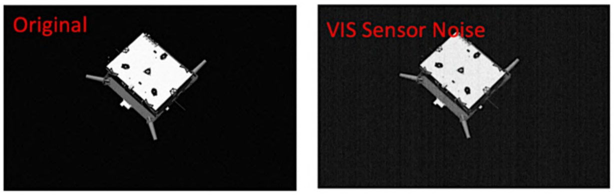


Fig. 9: Effect of VIS sensor augmentation on synthetic SPEED images.

Tab. 4 offers a comparison of the metrics for the SPEED images achieved by the YOLOv8s-pose model trained in this work (already provided in Tab. 2) with those scored by top-performing architectures available in the literature. It is worth pointing out that, as a side result, The YOLOv8s-pose

Table 2: Ablation study for YOLOv8s-pose augmentations, object detection results.

Model	Augmentation	SPEED		SPEED+ Lightbox		SPEED+ Sunlamp	
		ROI mAP_{50}^{95}	IoU mean	ROI mAP_{50}^{95}	IoU mean	ROI mAP_{50}^{95}	IoU mean
M1	Baseline	0.981	0.9653	0.687	0.8280	0.627	0.8349
M2	+ VIS Sensor Noise	**0.990**	**0.9758**	**0.843**	**0.8939**	**0.723**	**0.8682**

Table 3: Ablation study for YOLOv8s-pose augmentations, keypoint regression results.

Model	Augmentation	SPEED		SPEED+ Lightbox		SPEED+ Sunlamp	
		Keypoints mAP_{50}^{95}	OKS mean	Keypoints mAP_{50}^{95}	OKS mean	Keypoints mAP_{50}^{95}	OKS mean
M1	Baseline	0.987	0.9747	0.472	0.6267	0.185	0.3928
M2	+ VIS Sensor Noise	**0.994**	**0.9895**	**0.634**	**0.7226**	**0.262**	**0.4587**

model M2 (i.e., the one trained with the VIS sensor noise augmentation) is the best performing in target detection tasks, with the highest IoU score on SPEED images (+1.1% than the YOLOv5s adopted in [32], +2.2% than the YOLOv5s adopted in the SLN [31]) and mean AP_{50}^{95} (+3.3% with respect to the YOLOv5s adopted in [32], +0.5% than the YOLOv5s adopted in the SLN [31]). Moreover, by comparing the keypoint regression mean AP_{50}^{95} scored by the HR-Net32 adopted in [31] ($AP_{50}^{95} = 98.97\%$) with the ones achieved by model M2 ($AP_{50}^{95} = 99.40\%$), it arises that the YOLOv8s-pose is also the best-performing model on keypoint regression task for SPEED images, up to the author's knowledge. These side results highlight that coupling multi-task learning and the introduced VIS sensor noise augmentation effectively enhances the performances of the CNN, outperforming previous architectures.

Table 4: Performance comparison of target detection CNNs on SPEED images.

Method	Mean IoU	Mean AP_{50}^{95}
SLAB Baseline [33]	91.9%	N.A.
UniAdelaide [30]	95.34%	N.A.
SLN (YOLOv5s) [31]	95.38%	98.51%
YOLOv5n [32]	95.42%	95.7%
YOLOv5s [32]	96.46%	97.6%
YOLOv8s-pose M1 (our)	96.53%	98.1%
YOLOv8s-pose M2 (our)	**97.58%**	**99.0%**

6. Conclusions and Future Works

Autonomous spacecraft relative navigation via monocular images has been a highly active field of research in recent years. One of the most studied and effective solutions is to leverage CNNs to process the incoming images and then rely on classic PnP solvers to retrieve the relative pose with respect to a non-cooperative known target. The main drawback of this approach is that CNNs need to be trained on image datasets representative of the actual operative environment. Due to the lack of actual spaceborne images, this need is addressed by using synthetic images. Despite that, recent studies revealed that most architectures feature a strong drop in performance when trained on synthetic images and tested on mock-up frames, due to the huge domain gap between the sets. By noticing that one of the main differences between synthetic and mock-up frames is the presence of a detector and a sensor in the latter ones that are usually not accounted for in the image synthetic generation phase, the work in this paper aims at the development of a sensor model that can be applied to synthetic images to generate high-fidelity sensor noise, enhancing the representativeness of noise synthetic images and possibly improving the domain bridging capabilities of CNN trained with those images.

Comparing the actual images acquired from a CMOS sensor against the synthetic images noised using the widely adopted techniques of adding a white Gaussian noise with a standard deviation that includes the contribution of photon shot noise, dark current noise, and readout noise, it arose that the simple Gaussian model is ineffective in capturing both the actual noise characteristics and, most importantly, the fixed pattern noises that may be not negligible for CMOS sensors in low light and high gain conditions or for long exposure images. Notably, the fixed pattern noises can strongly affect the image processing algorithms leveraging edge detection due to their row-wise and column-wise patterns. Consequently, a VIS sensor model that describes the image formation steps from the reception of the photons on the detector to the final pixel intensity in DN has been developed by relying on detailed models for the principal noise sources (including fixed pattern noises) through all the image formation steps. This work provides an extensive analysis of all the noise sources and the models adopted, highlighting that the readout noise contribution to the overall noise is the highest in short exposure and that a Tuckey-Lambda distribution allows increasing the accuracy of this noise component. Remarkably, the developed model can be applied by using camera noise parameters from calibration and, remarkably, from datasheets, with a slight degradation of the fidelity mostly due to the strongest approximations in the fixed pattern noises, making it possible to approximate the camera without actually having the camera physically available and a dedicated facility.

To assess the representativeness of the model introduced within this work, its dark frames, i.e., images generated without incoming light, making light-dependent noise sources such as photon shot noise and PRNU negligible, have been compared against those acquired using an actual CMOS sensor. Despite the possibility of leveraging camera parameters from datasheets, a noise characterization of the actual CMOS adopted in this work has been performed due to the available datasheet non-compliance with the latest EMVA1288 prescriptions. It is worth pointing out that the procedure adopted in this work is general and can be applied to all camera models, resulting in a detailed guideline. The comparison between the introduced model and the actual CMOS confirms the high fidelity of the VIS sensor model developed. Namely, the VIS sensor model, either approximating the DSNU with Gaussian distribution or leveraging the more accurate GMMs and polynomial fittings, offers a better approximation of the actual CMOS dark frames than the widely adopted additive white Gaussian noise, with only small deviations from the intensity of the actual frames, more noticeable for the case of Gaussian approximation of the DSNU. Further, it has been proven that a Gaussian distribution for the readout noise leads to a good approximation of the first peak of the histogram, while the adopted Tuckey-Lambda distribution offers a better approximation due to the heavier weight applied to the tails of the distribution itself. It is acknowledged that the comparison between the introduced model and the actual frames lacks the evaluation of the light-dependent noise sources to be properly validated. This limitation arose from the unavailability of a calibrated light source in the facility. Consequently, the further development of the work presented here shall comprise a detailed validation of the developed model against actual frames acquired in a facility with a calibrated light source, in compliance with the EMVA1288.

Lastly, the application of the developed VIS sensor model during the training of a CNN suitable for pose estimation is discussed. Namely, the developed model has been introduced during the training phase as a custom augmentation applied to all the synthetic images in the training set. To avoid overfitting to a single camera noise model, the camera parameters used as input for the VIS sensor model have been randomized. The comparison with the outcomes achieved with and without the VIS sensor noise augmentation proved the enhancement of the performances on both synthetic and mock-up images in both ROI and keypoint regression offered by using a more realistic noise due to the introduced VIS sensor model. Namely, the performance increment in the synthetic images is such that the model trained with the VIS sensor noise augmentation is the best-performing model available in the literature in ROI and keypoint detection on SPEED synthetic frames, up to the author's knowledge. Concerning the mock-up images from SPEED+, the scores reveal increments of about 15% in ROI and keypoints AP_{50}^{95}, about 7% in ROI IoU, and about 10% in keypoint OKS for SPEED+ lightbox images. It is acknowledged that the enhancement of the performances in percentage for the SPEED+ sunlamp set is lower than the lightbox set due to the presence of optical artifacts (e.g., flares) in the former dataset that are not included in the training images and that strongly affect the performances of the adopted CNN. Despite that, all the outcomes achieved constitute a major outcome and confirm the effectiveness of the proposed VIS sensor model in enhancing the fidelity of the noise applied to synthetic images, strongly contributing to bridging the domain gap between synthetic and actual images and, consequently, allowing the training of CNN model more robust and better suited for real-case scenarios.

References

[1] Joseph A. Starek, Behçet Açıkmeşe, Issa A. Nesnas, and Marco Pavone. Spacecraft autonomy challenges for next-generation space missions. In Eric Feron, editor, *Advances in Control System Technology for Aerospace Applications*, pages 1–

48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[2] Louis Breger and Jonathan P How. Safe trajectories for autonomous rendezvous of spacecraft. *Journal of Guidance, Control, and Dynamics*, 31(5):1478–1489, 2008.

[3] Christophe Bonnal, Jean-Marc Ruault, and Marie-Christine Desjean. Active debris removal: Recent progress and current trends. *Acta Astronautica*, 85:51–60, 2013.

[4] Stefano Silvestrini, Jacopo Prinetto, Giovanni Zanotti, and Michèle Lavagna. Design of robust passively safe relative trajectories for uncooperative debris imaging in preparation to removal. In *2020 AAS/AIAA Astrodynamics Specialist Conference*, volume 175, pages 4205–4222. Univelt, 2020.

[5] Robin Biesbroek, Sarmad Aziz, Andrew Wolahan, Stefano Cipolla, Muriel Richard-Noca, and Luc Piguet. The clearspace-1 mission: Esa and clearspace team up to remove debris. In *Proc. 8th Eur. Conf. Sp. Debris*, pages 1–3, 2021.

[6] Paolo Lunghi, Marco Ciarambino, and Michèle Lavagna. A multilayer perceptron hazard detector for vision-based autonomous planetary landing. *Advances in Space Research*, 58(1):131–144, 2016.

[7] Roberto Opromolla, Giancarmine Fasano, Giancarlo Rufino, and Michele Grassi. A review of cooperative and uncooperative spacecraft pose determination techniques for close-proximity operations. *Progress in Aerospace Sciences*, 93:53–72, 2017.

[8] Lorenzo Pasqualetto Cassinis, Robert Fonod, and Eberhard Gill. Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft. *Progress in Aerospace Sciences*, 110:100548, 2019.

[9] Mate Kisantal, Sumant Sharma, Tae Ha Park, Dario Izzo, Marcus Märtens, and Simone D'Amico. Spacecraft pose estimation dataset (SPEED). *Zenodo*, February 2019.

[10] Mate Kisantal, Sumant Sharma, Tae Ha Park, Dario Izzo, Marcus Märtens, and Simone D'Amico. Satellite pose estimation challenge: Dataset, competition design, and results. *IEEE Transactions on Aerospace and Electronic Systems*, 56(5):4083–4098, 2020.

[11] Tae Ha Park, Marcus Märtens, Mohsi Jawaid, Zi Wang, Bo Chen, Tat-Jun Chin, Dario Izzo, and Simone D'Amico. Satellite pose estimation competition 2021: Results and analyses. *Acta Astronautica*, 2023.

[12] Tae Ha Park, Marcus Märtens, Gurvan Lecuyer, Dario Izzo, and Simone D'Amico. Next Generation Spacecraft Pose Estimation Dataset (SPEED+). *Zenodo*, October 2021.

[13] Tae Ha Park and Simone D'Amico. Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap. *Advances in Space Research*, 2023.

[14] Anup Bharat Katake. *Modeling, Image Processing and Attitude Estimation of High Speed Star Sensors.* PhD thesis, Texas A&M University, 2006.

[15] European Machine Vision Association. EMVA 1288 Standard: Standard for Characterization of Image Sensors and Cameras. Technical Report Release 4.0 Linear, EMVA, 2021.

[16] European Machine Vision Association. EMVA 1288 Standard: Standard for Characterization of Image Sensors and Cameras. Technical Report Release 4.0 General, EMVA, 2021.

[17] P. Jerram and K. Stefanov. 9 - cmos and ccd image sensors for space applications. In Daniel Durini, editor, *High Performance Silicon Imaging (Second Edition)*, Woodhead Publishing Series in Electronic and Optical Materials, pages 255–287. Woodhead Publishing, second edition edition, 2020.

[18] Alan M. Didion, Austin K. Nicholas, Joseph E. Riedel, Robert J. Haw, and Ryan C. Woolley. Methods for passive optical detection and relative navigation for rendezvous with a non-cooperative object at mars. In *AAS/AIAA Astrodynamics Specialist Conference*, Snowbird, Utah, August 19-23 2018.

[19] Mikhail Konnik and James Welsh. High-level numerical simulations of noise in ccd and cmos photosensors: review and tutorial. *arXiv preprint arXiv:1412.4031*, 2014.

[20] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8520–8537, 2022.

[21] Brian L Joiner and Joan R Rosenblatt. Some properties of the range in samples from tukey's symmetric lambda distributions. *Journal of the American Statistical Association*, 66(334):394–399, 1971.

[22] Sungkil Lee and Elmar Eisemann. Practical real-time lens-flare rendering. *Computer Graphics Forum*, 32(4):1–6, 2013.

[23] David J. Lamb and Lloyd W. Hillman. Computer modeling and analysis of veiling glare and stray light in Fresnel lens optical system. In Robert E. Fischer and Warren J. Smith, editors, *Current Developments in Optical Design and Optical Engineering VIII*, volume 3779, pages 344 – 352. International Society for Optics and Photonics, SPIE, 1999.

[24] Miguel A. Medina, Jason A. Mazzetta, and Stephen D. Scopatz. Image bloom testing and analysis. In Zia ur Rahman, Stephen E. Reichenbach, and Mark A. Neifeld, editors, *Visual Information Processing XIX*, volume 7701, page 77010M. International Society for Optics and Photonics, SPIE, 2010.

[25] Jean-Claude Perrin. Fast and accurate modeling of stray light in optical systems. In Georges Otrio, editor, *International Conference on Space Optics — ICSO 2000*, volume 10569, page 105691A. International Society for Optics and Photonics, SPIE, 2017.

[26] Michele Bechini. *Monocular Vision for Uncooperative Targets through AI-based Methods and Sensor Fusion.* PhD thesis, Politecnico di Milano, 2024.

[27] James R Janesick. *Scientific Charge-coupled Devices*, volume 83. SPIE Press, Bellingham, WA, 2001.

[28] Margherita Piccinin, Stefano Silvestrini, Giovanni Zanotti, Andrea Brandonisio, Paolo Lunghi, and Michéle Lavagna. Argos: calibrated facility for image based relative navigation technologies on ground verification and testing. In 72nd *International Astronautical Congress (IAC 2021), International Astronautical Federation, IAF, Dubai, United Arab Emirates*, pages 1–12, 10 2021.

[29] William C. Porter, Bradley Kopp, Justin C. Dunlap, Ralf Widenhorn, and Erik Bodegom. Dark current measurements in a CMOS imager. In Morley M. Blouke and Erik Bodegom, editors, *Sensors, Cameras, and Systems for Industrial/Scientific Applications IX*, volume 6816, page 68160C. International Society for Optics and Photonics, SPIE, 2008.

[30] Bo Chen, Jiewei Cao, Alvaro Parra, and Tat-Jun Chin. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[31] Massimo Piazza, Michele Maestrini, and Pierluigi Di Lizia. Monocular relative pose estimation pipeline for uncooperative resident space objects. *Journal of Aerospace Information Systems*, 19(9):613–632, 2022.

[32] Michele Bechini, Geonmo Gu, Paolo Lunghi, and Michèle Lavagna. Robust spacecraft relative pose estimation via cnn-aided line segments detection in monocular images. *Acta Astronautica*, 215:20–43, 2024.

[33] Sumant Sharma and Simone D'Amico. Neural network-based pose estimation for noncooperative spacecraft rendezvous. *IEEE Transactions on Aerospace and Electronic Systems*, 56(6):4638–4658, 2020.