PROCEEDINGS OF SCIENCE

Data-driven approximation of downward solar radiation flux based on all-sky optical imagery using machine learning models trained on DASIO dataset

Vasilisa S. Koshkina,[a,*] **Mikhail A. Krinitskiy,**[b,a] **Nikita N. Anikin,**[b] **Mikhail A. Borisov**[a] **and Sergey K. Gulev**[b]

[a] *Moscow Institute of Physics and Technology, Moscow, Russia*

[b] *Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia*

E-mail: koshkina.vs@phystech.edu, krinitsky@sail.msk.ru

Cloud cover is the main physical factor limiting the downward shortwave (SW) solar radiation flux. In modern models of climate and weather forecasts, physical models describing radiative transfer through clouds may be used. However this option computationally expensive. Instead, one may use parameterizations which are simplified schemes for approximating environmental variables. The purpose of our study is to assess the capabilities of machine learning models of approximating radiation flux based on all-sky optical imagery in order to assess the links between observed cloud cover properties with the flux. We applied various machine learning (ML) models: classic ML models and convolutional neural networks (CNN). These models were trained using the dataset of all-sky optical imagery accompanied by SW radiation flux measurements. The Dataset of All-Sky Imagery over the Ocean (DASIO) is collected in Indian, Atlantic and Arctic oceans during several expeditions from 2014 till 2021. When training our CNN, we applied heavy source data augmentation in order to force the CNN to become invariant to brightness variations and, thus, approximating the relationship between the visual structure of clouds and SW flux. We demonstrate that the CNN supersedes existing parameterizations known from literature in terms of RMSE. Our results allow us to assume that one may acquire downward shortwave radiation flux directly from all-sky imagery. We also demonstrate that CCNs are capable of estimating downward SW radiation flux based on clouds' visible structure.

**** The 6th International Workshop on Deep Learning in Computational Physics (DLCP2022) ****
**** 6-8 July 2022 ****
**** JINR, Dubna, Russia ****

[*] Speaker

PoS(DLCP2022)022

1. Introduction

Solar radiation is the main source of energy on Earth [12]. Cloud cover, in turn, is the main physical factor limiting the downward solar radiation flux. Cloud cover during the day reduces the influx of solar radiation to the Earth's surface, and significantly weakens its outgoing longwave radiation at night due to backscattering [13]. This entails corresponding changes in other meteorological quantities. The functioning of agriculture, transport, aviation, resorts, alternative energy enterprises and other sectors of the economy, in one way or another, depends on the amount and shape of clouds.

There are two options for flux estimation in modern models of climate and weather forecasts. First is a physics-based modeling of radiation transfer through two-phase medium (clouds) which includes modeling of multi-scattering taking into account the microphysics of cloud water drops [11] and aerosols. This option is computationally extremely expensive. Alternatively, one may use parameterizations which are simplified schemes for approximating environmental variables using only routinely observed cloud properties, such as Total Cloud Cover (TCC), cloud types and cloud cover per height layer. The existing parametrizations are empirical and were proposed years and decades ago based on observations and expert-based assumptions [10, 14]. As a result, they may not take into account the entire variety of cloud situations occurring in nature, which may lead to a reduced quality of approximation of downward SW solar radiation flux.

Our goals are to get computationally cheaper estimations of downward solar radiation flux and to study flux dependence on structural characteristics of clouds. The aim of this study is to improve the accuracy of existing parameterizations of downward SW radiation flux. In this study, we assess the capability of machine learning models in the scenario of statistical approximation of radiation flux from all-sky optical imagery. We solve the problem using various machine learning (ML) models within the assumption that an all-sky photo contains complete information about the downward SW radiation.

2. Data

In this section, we present source data for our study. The problem we tackle is to map all-sky imagery to SW radiation flux using state of the art statistical models (a.k.a. machine learning models). We use a high-resolution fish-eye cloud-camera «SAIL cloud v.2» [15] to collect all-sky images, and the radiometer Kipp&Zonen CNR-1 to measure SW flux. In fig. 1, we present the equipment used to collect the data.

The source data we use in our study is the Dataset of All-sky Imagery over the Ocean (DASIO) [1] which we collect in marine expeditions starting from 2014. The regions covered in these missions include Indian and Atlantic oceans, Mediterranean sea and Arctic ocean. In this dataset, the exhaustive set of cloud types is present. DASIO contains over 1 500 000 images of skydome over the ocean accompanied with downward SW radiation flux measurements. SW solar flux is averaged in 10-second intervals, and the all-sky images are registered every 20 seconds. The viewing angle of the Kipp&Zonen CNR-1 sensors is 180° in both vertical planes. The viewing angle of the cloud-camera is similar. Photos taken from the fisheye cloud-camera have high enough resolution

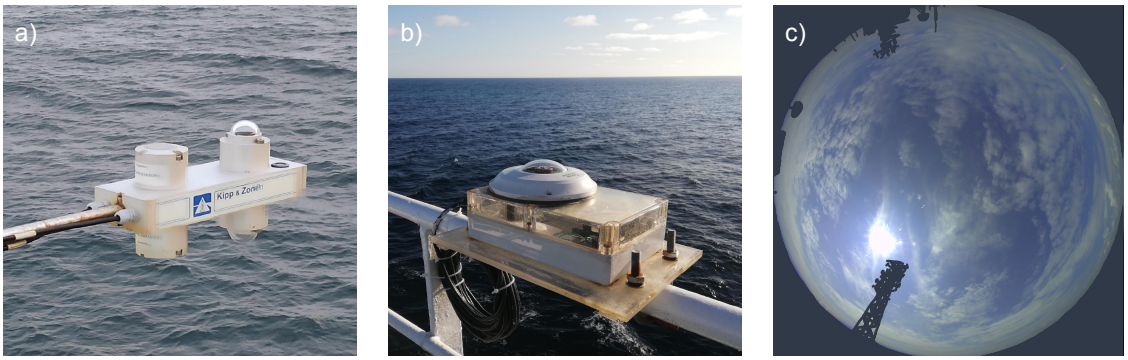

Figure 1: Equipment we use to collect the data, and an example of all-sky optical imagery over the ocean: a) radiometer Kipp&Zonen CNR-1; b) cloud-camera «SAIL cloud v.2»; c) all-sky photo with mask.

to resolve fine clouds structural details (1920*1920 px). White balance and brightness of photos are adjusted automatically for the most comfortable visual experience.

In our study, we employed a subset of DASIO. The size of the training subset was more than 1 000 000, and the size of the test subset was more than 350 000 images. In other words, the ratio of the volumes of test and training subsets is 1:3.

Fig. 1c also demonstrates a mask we applied to each photo, which filters our visual objects that are not related to the subject of our study. In addition, to train the models, we used only data obtained during daylight hours, when the Sun altitude exceeded 5 °, and the radiation flux exceeded 5 W/m^2.

In terms of machine learning, the problem is formulated as follows: objects are all-sky images of the visible hemisphere of the sky, and target values are the measurements of the downward shortwave radiation flux, in W/m^2. With this problem statement, it is a regression problem, thus, we exploited mean square error (MSE) as a loss function for all the ML models in our study.

3. Dataset re-weighting

In target value distribution, one may notice a strong predominance of data points with low SW flux. In order to improve the approximation skills of our models, we balance the train dataset using inverse-frequency re-weighting. We make the weights w_i of individual examples of train dataset to be inversely proportional to the frequency of target values:

$$w_i = \frac{d_i \cdot N_p}{\sum_{i=1}^{N_p} d_i} \sim d_i,$$

where i enumerates inter-percentile intervals from 0-th to 99-th; d_i are the inter-percentile intervals of empiric target value distribution, and $N_p = 100$ is a number of inter-percentile intervals. Here, the less target frequency, the more inter-percentile interval d_i, thus, the more are the weights w_i of the examples. We also propose the scheme for controlling the re-weighting strength using α

coefficient:

$$w'_i = (w_i - 1) \cdot \alpha + 1.$$

Here, one may notice, that the closer α to 1, the stronger re-weighting is applied. In case of $\alpha = 0$, there is no re-weighting, meaning $w'_i = 1$. Given the form of the weights w_i and w'_i, one may notice that their expected value is exactly 1.0. Coefficient α is a hyperparameter of our re-weighting scheme which is optimized during hyperparameters optimization stage.

4. Methods

In our study, we used two approaches: the classic approach and the so-called end-to-end approach with convolutional neural network employed.

Within the classic approach, we examined the following ML models: multilinear regression and non-parametric ensemble models Random Forests (RF) [2] and Gradient Boosting (GB) [3–5]. In this approach, we built a real-valued feature space of images consisting of 163 features. In particular, the following statistics were calculated for each color channel (Red, Green, Blue, Hue, Saturation and Brightness) of an image: maximum and minimum values; mean and variance; skewness, kurtosis and percentile set from 5 to 95 with increment of 5, as well as 1 and 99 percentiles. We also used the feature of sun altitude calculated using geographic position and UTC time of images.

Within the end-to-end approach, we did not computed any of expert-designed features. In contrast, we applied Convolutional Neural Network (CNN) [6] directly to the images. We also applied heavy images augmentation meaning strong alterations of average brightness. We also added spatially correlated gaussian noise. We applied these color-wise augmentations in order to increase the generalization ability of the CNN, and also in order to train the network to infer a flux link to the spatial structure of cloudiness instead of average brightness fo an image. Within this end-to-end neural networks-based approach, we also used the feature of sun altitude. The structure of the CNN exploited in our study is shown in Fig. 2. As one may see in this figure, input data are the all-sky RGB imagery resized to the resolution of 512x512 px. In order to speed up the learning process and to improve the quality of the approximation, we employed Transfer learning approach [7]: a pre-trained version of the ResNet50 [22] network is used, which was pre-trained on the ImageNet [23] dataset. The output of the ResNet50 convolutional sub-network is a 2048-dimensional vector. We concatenate sun altitude to this vector, thus the resulting vector is 2049-dimensional. We then apply a fully-connected sub-network to it. The structure of this sub-network is presented in Figure 2. The output of this subnet is real scalar value approximating the flux. We used Adam algorithm [17] for optimizing our CNN. Training and inference of the CNN we presented was implemented with Python programming language [21] using Pytorch [20], OpenCV [19] for Python and other high-level computational libraries for Python.

In both the ensemble models, RF and GB, we exploited in our study, there are hyperparameters besides the α re-weighting coefficient we presented above. Among them are the number of ensemble members in RF and GB; the maximum depth of the trees of the ensemble, *etc*. The CNN is also characterized by a number of hyperparameters: its depth, the width of fully-connected layers in fully-connected subnet, the hyperparameters of Adam optimization procedure, and also the magnitude of data augmentation transformations. We employed Optuna framework [18] for hyperparameters

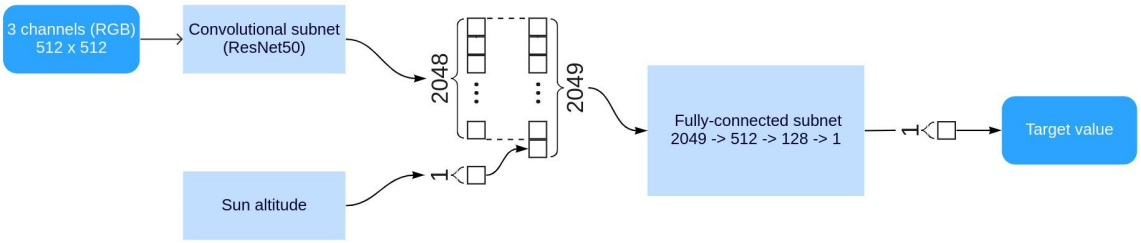

Figure 2: Architecture of a CNN we exploited in our study. Here, with numbers, we present the shapes of input data or activation maps.

optimization (HPO). During the HPO stage, the quality of each model initialized with a sampled hyperparameters set is assessed within K-fold cross-validation (CV) approach with $K = 5$. Due to strongly correlated examples (all-sky images) that are close in temporal domain, we ensured the independence of train and validation CV-subsets using Group K-fold cross-validation approach where groups are hourly subsets of all-sky images. In case of RF and GB models, we assess the mean RMSE measure as well as its uncertainty within the Group K-fold CV approach.

5. Results and discussion

In this section, we present the current results of our study. To assess the quality of our models, we used root mean square error (RMSE) measure. Also, the visual representation of the results is given in the form of value mapping diagrams (Fig. 3), where the correspondence between approximated and measured flux values is presented in a form of points density. In figure 3, one may see that the models generally underestimate high fluxes and overestimate low fluxes. It is also clear that multilinear model approximates the flux worse than other models, which is supported by the RMSE measures in table 1. The results of CNN are the best among others in terms of formal RMSE measure as well as approximated-to-measured values mapping diagram.

In our study, we built and trained four ML models to approximate the downward shortwave radiation flux. We found, that the quality of the CNN, which was built within the end-to-end approach, is the best compared to other models and also to existing SW radiation parameterizations known from the literature [10, 14]. In table 1, we present the quality of our models assessed after the hyperparameters optimization. We also provide RMSE estimates of the parametrizations as a reference. One may also see that parameterizations error strongly depends on the amount of cloudiness: the higher Total Cloud Cover (TCC) the higher a parameterization error. We provide the errors range in brackets for parameterizations known from the literature.

In Fig. 4, we also demonstrate error distributions for each ML models of our study. In the CNN error distribution (Fig. 4d), one may see that the neural network is prone to slight underestimation of SW flux. Also, it is clear that erors distribution tails are pretty heavy for both RF and GB models, and are light for CNN. These features of errors distribution for our models are also in agreement with the variance of errors that are presented in Tab. 1 in a form of RMSE (taking into account that the errors are zero-centered, thus RMSE is the square root of variance in this case.)

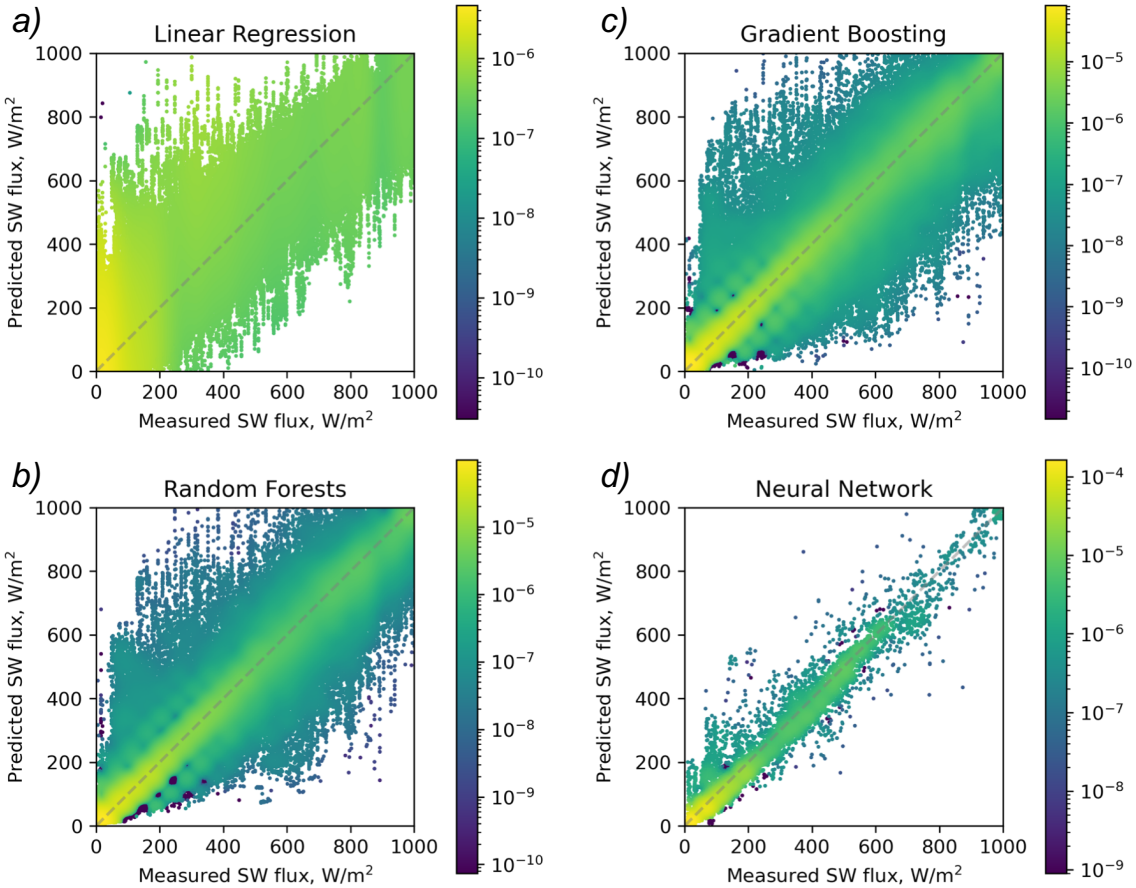

Figure 3: Value mapping diagrams for: a) Linear Regression, b) Random Forests (re-weighted), c) Gradient Boosting (re-weighted), d) CNN. Density colormap is logarithmic for presentation purpose.

The models we present demonstrate some issues. Multilinear regression is a fast model, however, it has the worst quality. RF and GB demonstrate comparable quality and are relatively fast at inference time. At the same time, one may note non-smooth errors distribution in diagrams in Fig. 3(b,c). We suppose that the regular drops of points density may be explained by decision-tree-based nature of these two ensemble models. One may also notice the outliers in these diagrams that may be of interest in forthcoming study. In this study, we did not filter the outliers comprehansively, thus there may be irrelevant imagery in the dataset that represent photographs of birds, glass dome cleaning operator, *etc.*

Table 1: Quality metrics of (1-4) ML models used in this study and (5-6) parameterizations of SW radiation known from the literature.

No	Model	Parameterization	RMSE, W/m^2
1	Linear Regression		84 ± 23.5
2	Gradient Boosting		68.0 ± 23.5
3	Random Forests		68.5 ± 18.7
4	CNN		**39.2**
5		Dobson–Smith [14]	78.2(38 – 116)
6		LVOAMKI [10]	61.9(26 – 115)

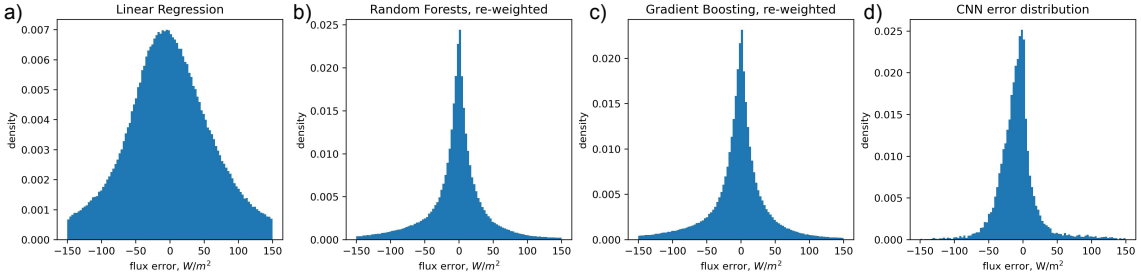

Figure 4: Error distribution figures for: a) Linear Regression, b) Random Forests (re-weighted), c) Gradient Boosting (re-weighted), d) CNN

6. Conclusions and outlook

In this study, we presented the approach for the approximation of short-wave solar radiation flux over the Ocean from all-sky optical imagery using state of the art machine learning algorithms including multilinear regression, Random Forests, Gradient Boosting and convolutional neural networks. We trained our models using the data of DASIO dataset [1]. The quality of our models was assessed in terms of RMSE, approximated-vs.-measured flux diagrams and errors histograms. The results allow us to conclude that one may estimate downward SW radiation fluxes directly from all-sky imagery taking some well-known uncertainty into account. We also demonstrate that our CNN trained with strong data augmentations is capable of estimating downward SW radiation flux based on clouds' visible structure mostly. At the same time, the CNN is shown to be superior in terms of flux RMSE compared to other ML models in our study.

Our method of flux estimation may be especially useful in the tasks of low-cost monitoring of downward fluxes of shortwave solar radiation, as well as in exploratory studies of territories for their placement. The solution of the presented problem makes it possible to obtain estimates of the downward SW solar flux based on model atmospheric data containing clouds characteristics, which may reduce the computational load of the radiation subroutine.

Our results suggest that there are outliers in DASIO dataset that may be filtered in forthcoming studies. The results also suggest that hyperparameters optimization of our CNN and ensemble models may help discovering better configurations including proper dataset re-weighting as well as

more suitable CNN architecture. In further studies, we plan to approximate downward longwave solar radiation flux using the approach similar to the one presented in this paper. Also, modern statistical models of Machine Learning class provide opportunity for short-term forecasting of fluxes which may be useful in forecasting the generation of solar power plants.

Aknowledgments This work is supported by the grant of Russian Foundation for Basic Research (20-05-00244). Development, training and inference of the CNN presented in the study is supported by the program FMWE-2022-0002.

References

[1] Krinitskiy, M., Aleksandrova, M., Verezemskaya, P., Gulev, S., Sinitsyn, A., Kovaleva, N. & Gavrikov, A. On the generalization ability of data-driven models in the problem of total cloud cover retrieval. *Remote Sensing*. **13**, 326 (2021)

[2] Breiman, L. Random forests. *Machine Learning*. **45**, 5-32 (2001)

[3] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. & Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances In Neural Information Processing Systems*. **31** (2018)

[4] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. Lightgbm: A highly efficient gradient boosting decision tree. *Advances In Neural Information Processing Systems*. **30** (2017)

[5] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining*. pp. 785-794 (2016)

[6] O'Shea, K. & Nash, R. An introduction to convolutional neural networks. *ArXiv Preprint ArXiv:1511.08458*. (2015)

[7] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. & Liu, C. A survey on deep transfer learning. *International Conference On Artificial Neural Networks*. pp. 270-279 (2018)

[8] Trenberth, K., Fasullo, J. & Kiehl, J. Earth's Global Energy Budget. *Bulletin Of The American Meteorological Society*. **90**, 311-324 (2009,3), Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society

[9] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 770-778 (2016)

[10] Aleksandrova, M., Gulev, S. & Sinitsyn, A. An improvement of parametrization of short-wave radiation at the sea surface on the basis of direct measurements in the Atlantic. *Russian Meteorology And Hydrology*. **32**, 245-251 (2007)

[11] Stephens G. L. Radiation Profiles in Extended Water Clouds. I: Theory // Journal of the Atmospheric Sciences. — 1978. — . — Vol. 35, no. 11. — Pp. 2111–2122.

[12] Trenberth Kevin E., Fasullo John T., Kiehl Jeffrey. Earth's Global Energy Budget // Bulletin of the American Meteorological Society. — 2009. — . — Vol. 90, no. 3. — Pp. 311–324.

[13] Ming-Dah Chou, Kyu-Tae Lee, Si-Chee Tsay, Qiang Fu . Parameterization for Cloud Longwave Scattering for Use in Atmospheric Models // Journal of Climate. — 1999. — . — Vol. 12, no. 1. — Pp. 159–169.

[14] Dobson Fred W, Smith Stuart D. Bulk models of solar radiation at sea // Quarterly Journal of the Royal Meteorological Society. — 1988. — Vol. 114, no. 479. — Pp. 165–182.

[15] Krinitskiy Mikhail A, Sinitsyn Alexey V. Adaptive algorithm for cloud cover estimation from all-sky images over the sea // Oceanology. — 2016. — Vol. 56, no. 3. — Pp. 315–319.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — Pp. 770–778.

[17] Kingma Diederik P, Ba Jimmy. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. — 2014.

[18] Takuya Akiba, Shotaro Sano, Toshihiko Yanase et al. . Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2019.

[19] Bradski G. The OpenCV Library // Dr. Dobb's Journal of Software Tools. — 2000.

[20] Adam Paszke, Sam Gross, Francisco Massa et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems 32. — Curran Associates, Inc., 2019. — Pp. 8024–8035. — URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[21] Van Rossum Guido, Drake Fred L. Python 3 Reference Manual. — Scotts Valley, CA: CreateSpace, 2009.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun Deep Residual Learning for Image Recognition. — in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2016. — Pp. 770–778.

[23] Jia Deng, Wei Dong, Richard Socher et al. Imagenet: A large-scale hierarchical image database // IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009. — Ieee, 2009. — Pp. 248–255.