REPRESENTATIVE SAMPLES FORMING OF URBAN AERIAL AND SATELLITE IMAGERY FOR BUILDING FOOTPRINT SEGMENTATION

Vit Roman
Ph.D. student
Molchanova Maryna
Ph.D., Senior lecturer
Mazurets Oleksandr
Ph.D in Engineering Science, Associate Professor
Computer Science Department
Khmelnytskyi National University, Ukraine

The rapid growth of urban areas and the increasing availability of high-resolution aerial and satellite imagery have made automated building footprint segmentation a critical task for urban planning, disaster management, infrastructure monitoring, and geospatial analytics [1]. However, the effectiveness of modern computer vision methods for this task strongly depends on the quality, representativeness, and balance of the training data [2]. Urban scenes are characterized by high variability in building shapes, scales, materials, roof structures, illumination conditions, sensor resolutions, and occlusions caused by vegetation or shadows, which makes naive dataset construction insufficient for robust model generalization. In this context, the formation of representative samples of urban aerial and satellite imagery becomes a key methodological problem that directly influences the performance and stability of convolutional neural network based segmentation models [3].

Representative sample forming addresses this challenge by explicitly accounting for the diversity of urban environments and imaging modalities during dataset construction [4]. By systematically selecting and balancing image tiles according to factors such as building density, footprint size distribution, acquisition altitude, spatial resolution, seasonal variability, and background complexity, it becomes possible to expose CNNs to a broader range of visual patterns during training [5, 6]. This, in turn, improves the robustness of learned feature representations and reduces overfitting to local or dataset-specific characteristics [7]. For building footprint segmentation, such representativeness is especially important because small residential buildings, dense historical quarters, and industrial complexes exhibit fundamentally different visual signatures that must all be accurately captured by a single model [8].

Within the scope of this study, CNN-based computer vision methods are considered not only as segmentation tools but also as implicit evaluators of dataset quality, since variations in validation stability and cross-domain performance directly reflect the adequacy of the formed samples [9]. The combination of representative dataset construction with deep convolutional models enables more reliable extraction of building footprints across heterogeneous urban scenes and supports scalable deployment of segmentation systems in real-world geospatial applications. Consequently, the proposed focus on representative sample forming constitutes an essential step toward improving the generalization capability, reproducibility, and practical value of CNN-based building footprint segmentation from urban aerial and satellite imagery [10, 11].

Building footprint segmentation from remote sensing imagery has become a foundational capability for urban analytics, post event monitoring, and rapid situational awareness. In operational settings, however, model quality is rarely limited by the availability of imagery as such [12]. The primary bottleneck is the mismatch between the imagery and annotations used for training and the conditions in which the model is expected to operate. Urban scenes exhibit high intra class variability, including strong texture changes across districts, heterogeneous roof materials, frequent occlusions by vegetation and infrastructure, and complex building adjacency that blurs boundaries between adjacent objects [13]. When the target scenario includes partially damaged or structurally altered buildings, additional ambiguity is introduced because the notion of a stable footprint competes with rubble, exposed interiors, and altered roof lines [14]. Under these conditions, the concept of a representative sample cannot be reduced to acquiring a large number of patches. Representativeness must be treated as a controlled coverage of the variability space that governs both the image formation process and the annotation process [15].

The thesis developed here is that a representative sample for footprint segmentation is achieved through a deliberate data formation methodology that aligns source diversity, annotation structure, and evaluation design. The methodology is anchored in satellite benchmarks that provide large scale coverage and consistent polygon labels [16], complemented by aerial imagery that introduces higher detail, perspective distortions, and more severe occlusion patterns. In the adopted setting, satellite data are exemplified by a large corpus of 1024 by 1024 pixel scenes paired across different temporal states, annotated with polygon footprints and damage semantics using an ordinal four level scheme [17].

This backbone is crucial because it supplies both the density of building instances and the formal label structure needed to establish a stable learning target. At the same time, aerial imagery from unmanned platforms alters the effective domain through viewing geometry, altitude choice, illumination variability, and transient artifacts such as dust and smoke, thereby testing the robustness of footprint boundary reconstruction beyond the assumptions of near nadir satellite acquisitions.

Data sources as complementary domains. The satellite benchmark foundation used in this line of work is notable for scale, geographic diversity, and explicit polygon footprint labeling. The dataset description indicates tens of thousands of image chips, a coverage measured in tens of thousands of square kilometers, and a very large number of building polygons, with imagery collected across multiple distinct events.

Such scale matters not because it automatically guarantees generalization, but because it enables the construction of meaningful strata for sampling and validation.

The availability of polygon footprints provides a natural segmentation target, while the presence of damage annotations creates the opportunity to connect footprint extraction to downstream post event assessment, even if the primary objective here remains footprint segmentation.

A further operational advantage of the satellite benchmark format is the organization of imagery into paired temporal states and the existence of labels that are simultaneously tied to pixel coordinates and to geographic reference.

It enables consistent projection between the segmentation mask domain and the geospatial domain, which is essential whenever model outputs must be fused with geographic information systems or compared across time. It also supports principled error analysis, because misalignments and systematic geometric distortions can be separated from semantic confusion.

It increases visual fidelity and boundary detail, but it also introduces factors that compromise the invariances learned from satellite data. The reported characteristics of aerial scenes include variable viewing angles, non uniform illumination, partial occlusions, and the presence of dust or smoke that can mask the signal of building edges. These conditions generate segmentation challenges that are qualitatively different from those in typical satellite benchmarks. Footprints may be partially hidden, edges may be indistinct, and the same roof may appear at different scales within a scene due to perspective. From the standpoint of representativeness, aerial data are therefore not an optional enhancement. They act as a domain stressor that exposes whether the training sample covers the variability that matters for deployment.

The existence of a domain shift between satellite and aerial data is explicitly acknowledged, and it is treated as a central factor influencing model behaviour.

A representative sample must therefore be designed not only to perform well within a single domain, but also to preserve acceptable performance when the acquisition regime changes. This requirement has immediate implications for how data should be partitioned, how augmentation should be interpreted, and how validation should be carried out.

Representativeness in this work is conceptualized as coverage of the joint distribution of scene factors and object factors. Scene factors include urban density, background texture complexity, shadow regimes, seasonal appearance, and imaging geometry. Object factors include footprint size, shape complexity, roof material patterns, adjacency patterns, and partial visibility. The practical point is that segmentation performance is determined by the hardest boundaries, not by the easiest. If the sample overrepresents clean roofs on homogeneous backgrounds, the model will display optimistic metrics while failing on dense urban blocks with heavy occlusion and attached buildings. Conversely, if the sample overrepresents severely damaged structures, the model may become conservative and under segment intact buildings in typical scenes. A representative sample must therefore maintain a controlled balance of difficulty.

The satellite benchmark statistics provide a basis for designing such coverage, because the large number of building polygons and the global variability of regions and events imply substantial diversity in building types and settlement morphology.

The key methodological choice is to treat this diversity as a sampling space rather than as a passive characteristic. In concrete terms, representativeness is pursued through a selection procedure that prioritizes diversity in building density and scene context and that explicitly avoids a training set dominated by repeated city patterns. The goal is to make it unlikely that the model succeeds through memorization of a narrow subset of architectural styles or roof textures.

At the annotation level, representativeness has a second dimension. Polygon footprint labels embed human decisions about what counts as a building boundary and how to treat ambiguous cases such as attached structures, complex roof overhangs, and partially collapsed buildings. These decisions are reflected in the four level ordinal damage scheme used in the benchmark.

Even though the present thesis focuses on footprint segmentation, the existence of the ordinal scheme is important because it indicates where boundary ambiguity becomes acute. Intermediate damage conditions often correspond to partial roof failure or wall damage where the footprint is still present but the roof appearance changes. A representative sample must include such cases, because they are precisely where segmentation failures propagate into downstream analytics [18].

A representative sample must be coherent, not merely diverse. Coherence here means that all images and labels are brought into a common operational space in which the segmentation objective is well defined and consistent across domains. The benchmark imagery provides fixed size chips with polygon footprints and damage labels.

From these polygons, segmentation masks are derived. The derivation step is not trivial: polygon filling must be handled carefully to preserve thin structures and avoid topological artifacts, and mask generation must maintain consistent conventions across scenes. Errors at this stage would mimic model errors and contaminate evaluation. A rigorous workflow therefore treats mask generation as a deterministic transformation with auditing on a subset of scenes representing different building densities and polygon complexities.

Aerial imagery introduces additional preprocessing requirements. Because aerial data are described as having variable geometry, viewpoint, and occlusions, it is necessary to ensure that the model sees comparable input scales during training. This is addressed through spatial normalization choices and through controlled augmentation that emulates expected acquisition variability without destroying the boundary cues needed for footprint reconstruction. The methodological stance is that augmentation should not be used to compensate for missing domains, but rather to densify coverage around the observed domains. In other words, augmentation is treated as a local perturbation tool rather than as a substitute for genuine aerial variability.

An important aspect of representativeness is the integrity of the train and validation split. If patches from the same geographic area or the same event context leak into both training and validation, evaluation will reflect memorization rather than generalization. The paired temporal structure of the benchmark, makes such leakage particularly risky because pre event and post event images of the same location share much of their appearance. A representativeness oriented protocol therefore treats event and region as grouping variables during splitting, ensuring that validation includes scenes that the model has not seen in either temporal state [19].

The central claim of this thesis is that the quality of a representative sample is revealed by model behaviour under controlled training and evaluation, not solely by descriptive dataset statistics. For this reason, the data formation process is validated through training segmentation detectors from the YOLO family on the assembled sample and examining both convergence properties and achieved metrics. The training dynamics reported for the YOLOv11n segmentation model indicate stable reduction of loss components over 100 epochs, with validation curves tracking training curves without abrupt divergence.

This is an important signal for dataset formation because unstable validation behaviour would suggest either excessive label noise, insufficient coverage, or leakage between splits. Stability suggests that the sample contains enough diversity to support learning while remaining coherent enough to avoid contradictory supervision.

The reported metric trajectories provide a more direct lens on representativeness. After training, detection precision and recall for bounding boxes reach approximately the mid range around 0.55 to 0.60, while mAP at an intersection over union threshold of 0.5 reaches roughly 0.35 to 0.40 and mAP averaged across stricter thresholds reaches roughly 0.12 to 0.15.

The corresponding mask metrics are lower, which is expected given that footprint segmentation demands accurate boundary localization rather than coarse object localization.

These values should not be interpreted as merely model performance numbers. They also encode properties of the dataset. The gap between mAP at 0.5 and the stricter mAP averaged across thresholds indicates that while the model often identifies buildings correctly, boundary precision remains challenging, especially under difficult urban conditions and domain shift. The dominance of precision over recall, where the model is more cautious and misses some difficult objects rather than producing many false positives, is consistent with a sample that contains hard negatives and ambiguous edges. If the sample were dominated by easy cases, the model would typically achieve higher recall with less conservative behaviour. The reported pattern therefore supports the conclusion that the sample includes a meaningful fraction of challenging urban scenes.

Figure 1 provides a qualitative view of the deployed inference setup used to interpret the metric results under domain shift. The interface enables controlled adjustment of confidence and NMS thresholds and visualizes predictions on an input patch, supporting error analysis of boundary localization and false detections in complex urban scenes.

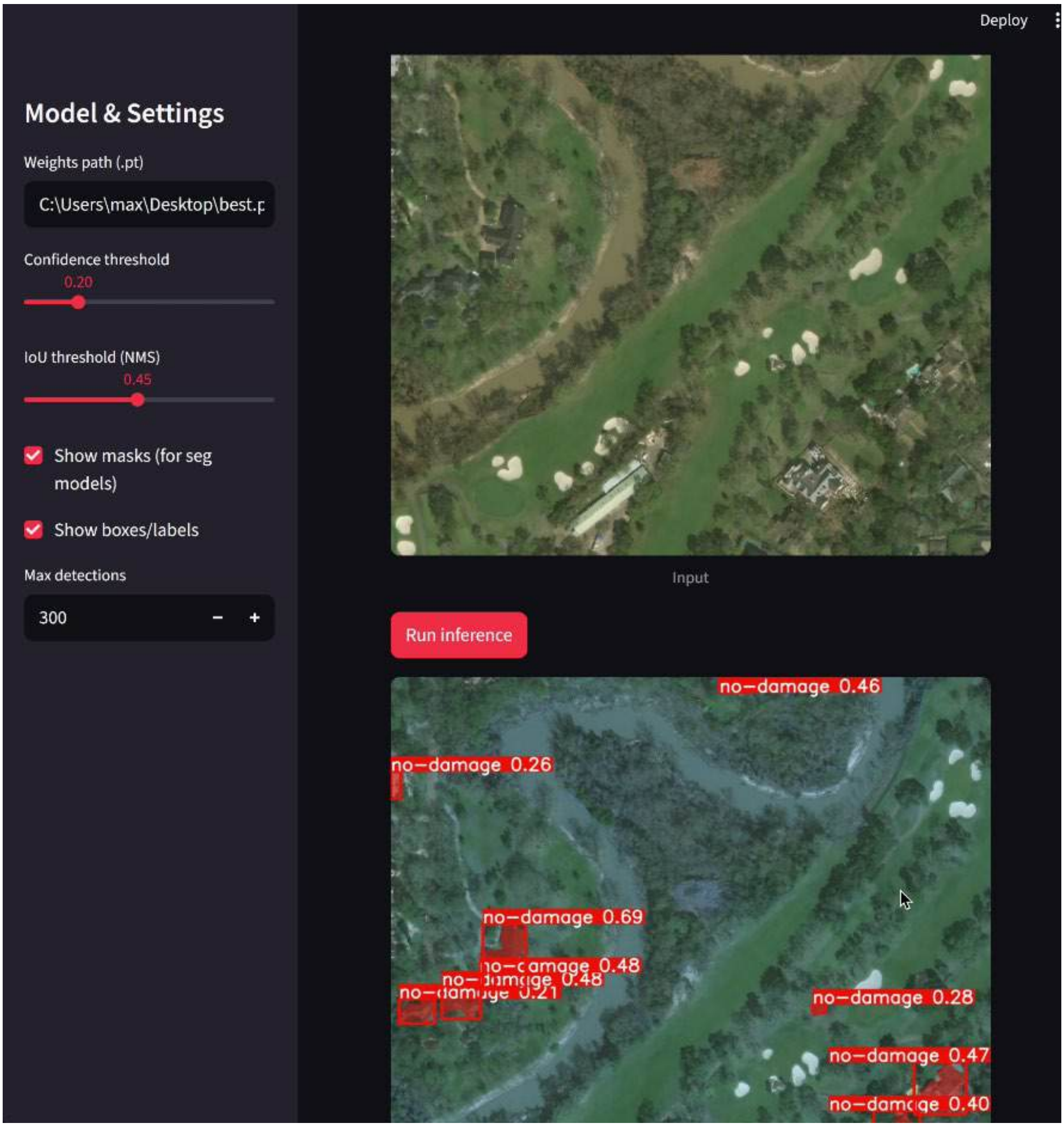

Figure 1. Inference and deployment interface of the trained building detection and footprint segmentation module for urban remote sensing imagery

The discussion explicitly attributes part of the observed performance profile to domain shift between satellite pretraining and aerial fine tuning. This attribution is important for the thesis argument. A representative sample should not eliminate domain shift by ignoring it. Rather, it should surface domain shift through validation and then enable iterative improvements through targeted sampling and fine tuning. In this sense, the model is used as an instrument for measuring whether the sample is representative of deployment conditions.

Urban imagery is notorious for producing segmentation errors that are localized at boundaries between adjacent buildings, at roof edges under strong shadows, and in areas where vegetation overlaps roof structures. A representative sample must capture these failure inducing contexts. The fact that the training results show lower mask metrics than box metrics is consistent with the claim that boundary reconstruction is the dominant difficulty, not object presence recognition. This observation motivates a data formation emphasis on cases that amplify boundary ambiguity. Such cases include high density blocks with shared walls, industrial zones with large continuous roofs and attached structures, and mixed scenes where buildings of different heights create shadow interactions.

The benchmark dataset properties support the ability to include these contexts, because buildings are described as appearing under varying densities and construction types and across variable imaging angles.

The sampling methodology therefore leverages the breadth of the benchmark to incorporate a wide range of density regimes, rather than focusing only on the most common medium density residential patterns. This is essential because in many cities, the operational use cases such as damage assessment and reconstruction planning are concentrated in dense urban cores where segmentation is hardest.

Aerial imagery further reinforces this focus by introducing occlusions and viewing angles that are difficult for satellite trained models. The dataset formation goal is to ensure that such aerial specific factors are not confined to a small corner of the training set. If they were, the model would treat them as outliers and would fail to generalize. Instead, they must appear frequently enough to shape the learned feature space, while remaining balanced so that the model does not overfit to aerial artifacts at the expense of general satellite performance.

Although the thesis topic emphasizes footprint segmentation, the adopted benchmark includes an ordinal damage labeling scheme with four levels. This structure has methodological implications because damage correlates with footprint appearance. Buildings labeled as destroyed may exhibit missing roof textures, fragmented edges, and heterogeneous debris patterns that disrupt the footprint signal. Minor and major damage often appear as subtle roof deformations and partial collapses that are difficult to distinguish from normal roof texture variability. As a result, a representative sample for segmentation should not only contain intact buildings. It must also include damaged buildings across the ordinal spectrum, because those cases determine whether the model can maintain boundary reconstruction when the roof appearance deviates from the intact pattern.

The dataset scale reported, including a very large number of polygons and many events, provides the opportunity to include such diversity without sacrificing coverage of intact buildings. The practical sampling stance is that damaged buildings should be present in sufficient quantity to shape the segmentation objective, but they should not dominate the sample, because the model must still segment intact buildings robustly in normal urban monitoring contexts. The balance is achieved through controlled inclusion of post event scenes, paired with pre event scenes, while enforcing geographic separation between train and validation.

The experiments performed with YOLO family segmentation models constitute the primary empirical evidence supporting the sampling strategy. Training over 100 epochs with stable convergence indicates that the sample supports learning without catastrophic overfitting, which would be expected if the sample were either too small, too noisy, or too redundant. The achieved precision and recall levels around the mid range and the moderate mAP values suggest that the model captures the dominant building signal but continues to struggle with fine boundary localization. This is consistent with the target problem and with the presence of challenging scenes.

A key interpretive point is that moderate strict mAP does not necessarily signal methodological failure. In footprint segmentation, strict intersection over union thresholds penalize small boundary offsets heavily, particularly for small buildings and for fragmented shapes. In dense urban contexts, small buildings are plentiful and are often partially occluded. Therefore, strict mAP acts as a proxy for whether the sample truly contains small scale and high ambiguity footprints. The observed strict mAP range around 0.12 to 0.15 indicates that while the model can identify and segment many buildings, high precision boundary alignment remains difficult across the full variability of the sample. From a representativeness perspective, this outcome is expected and even desirable at the stage of building a robust sample, because it shows that validation is not artificially easy.

The cautious model behaviour indicated by precision slightly exceeding recall supports a similar conclusion. In datasets dominated by easy positives, recall often exceeds precision, because the model labels many regions as buildings without being heavily penalized. Here, the conservative profile suggests that the sample includes a substantial number of hard negatives and ambiguous backgrounds such as rubble fields, vegetation overlays, and shadowed regions that could be mistaken for buildings. This is exactly the type of variability that a representative sample should capture.

A representative sample designed in this manner supports several deployment pathways. In a monitoring context, the footprint segmentation module can serve as the first stage of a pipeline that aggregates building level statistics over space. The benchmark's linkage between pixel and geographic coordinates makes it feasible to translate predicted masks into geospatial features suitable for mapping and for integration with planning systems. In a post event context, the ordinal damage scheme provides a natural extension toward damage estimation. Even if damage classification is handled by a separate module, robust footprint segmentation is a prerequisite because it defines the spatial support over which damage evidence is aggregated.

A second deployment implication concerns domain shift management. The explicit recognition that aerial imagery differs materially from satellite imagery implies that any real system must include a mechanism for periodic sample refresh and model update. Urban environments change, sensors evolve, and reconstruction alters the appearance of damaged areas. A representative sample is therefore not a static artifact but a maintained asset. The methodology argued here, in which model performance is used as a diagnostic for coverage gaps, naturally extends to this lifecycle view.

The reported metrics indicate that boundary precision under strict thresholds remains the primary limitation. This is consistent with the inherent difficulty of the task, but it also points to concrete opportunities for strengthening the sample. One opportunity lies in increasing the density of examples that contain small buildings and tight adjacency, because these cases contribute disproportionately to strict intersection over union penalties. Another opportunity lies in targeted inclusion of aerial scenes with pronounced perspective and occlusion, because these factors were identified as characteristic of unmanned platform imagery and as drivers of domain shift. The ordinal damage structure also suggests a need for more targeted coverage of intermediate damage states.

In many datasets, extreme classes are visually distinctive, while intermediate classes are ambiguous and underrepresented. For segmentation, these intermediate states are particularly important because they exhibit partial roof failure that modifies boundary cues without removing the building entirely. A representative sample that aims to support reliable footprint extraction under damage conditions should therefore ensure sufficient representation of these ambiguous states, not only for downstream classification but for segmentation itself.

Finally, representativeness would benefit from explicit geographic and morphological stratification criteria, implemented in a way that reduces redundancy. The large scale benchmark statistics make this feasible, and the stable training behaviour observed indicates that the current sample is already coherent enough to support further targeted refinement.

So, this thesis argues that forming a representative sample for building footprint segmentation in urban aerial and satellite imagery is a methodological problem in its own right, not a preliminary data collection step. Representativeness is defined as coverage of the variability space induced by urban morphology, imaging conditions, and annotation conventions, and it is validated through model behaviour rather than assumed from dataset size. Satellite benchmark data provide scale, polygon footprints, paired temporal organization, and an ordinal damage labeling structure that supports systematic sampling and evaluation.

Aerial imagery complements this foundation by introducing higher detail and challenging acquisition conditions such as variable viewpoints and occlusions, thereby exposing domain shift that must be reflected in a truly representative training sample.

The empirical results obtained with YOLO family segmentation models support the adequacy of the formed sample in the sense of learnability and robustness: training over 100 epochs exhibits stable convergence without pronounced overfitting, and the achieved metrics demonstrate a cautious detection profile with moderate mAP and lower mask performance consistent with the boundary sensitive nature of footprint segmentation in complex urban scenes.

The remaining gap under strict thresholds is interpreted not as a contradiction of representativeness, but as evidence that the sample indeed includes the hard cases that dominate real world errors. This aligns the data formation process with the stated objective: building a dataset that does not merely enable optimistic validation, but that supports trustworthy model behaviour under the variability expected in operational monitoring.

References

1. Caro D., Lodato C., Damgaard A., Cristóbal J., Foster G., Flachenecker F., Tonini D. Environmental and socio-economic effects of construction and demolition waste recycling in the European Union // Science of the Total Environment. – 2024. – Vol. 908. – Article 168295. – DOI: 10.1016/j.scitotenv.2023.168295.
2. Yevle D. V., Mann P. S. Artificial intelligence-based classification for waste management: A survey based on taxonomy, classification & future direction // Computer Science Review. – 2025. – Vol. 56. – Article 100723. – DOI: 10.1016/j.cosrev.2024.100723.
3. Jha S., Nanda S., Zapata O., Acharya B., Dalai A. K. A review of systems thinking perspectives on sustainability in bioresource waste management and circular economy // Sustainability. – 2024. – Vol. 16. – Article 10157. – DOI: 10.3390/su162310157.
4. Kazmi R., Chakraborty M. Use of different forms of waste promoting a circular economy in the construction industry: A systematic literature review // International Conference on Environmental Geotechnology, Recycled Waste Materials and Sustainable Engineering. – Singapore : Springer, 2025. – P. 193–204. – DOI: 10.1007/978-981-97-9777-6_17.
5. Pokhytun A., Mazurets O., Molchanova M., Tyschenko O. Method for neural network detecting changed images of people's faces using CNN // New Horizons in Scientific Research: Challenges and Solutions : proceedings of the 1st Int. Sci. and Pract. Conf. (Marseille, France, Oct. 21–23, 2024). – Marseille, 2024. – P. 35–40.
6. Kharysh I., Sobko O., Mazurets O. Designing CNN neural network model for detecting fractures of lower extremities by X-ray images // The Impact of Scientific Research on the Development of the Modern World : proceedings of the XLIV Int. Sci. and Pract. Conf. (Dubrovnik, Croatia, Oct. 23–25, 2024). – Dubrovnik, 2024. – P. 91–96.
7. Mazurets O., Molchanova M., Klimenko V., Klopotivskyi D. Datalogic model for image recognition by convolutional neural network using cloud services // Modern Scientific Research: Theoretical and Practical Aspects : proceedings of the XXII Int. Sci. and Pract. Conf. (Oslo, Norway, May 8–10, 2024). – Oslo, 2024. – P. 64–68.
8. Molchanova M., Mazurets O., Klimenko V., Kuflevsky E. Object-oriented model for neural network damage detection of mail packages // Solving Scientific Problems Using Innovative Concepts : proceedings of the XIV Int. Sci. and Pract. Conf. (Copenhagen, Denmark, Mar. 13–15, 2024). – Copenhagen, 2024. – P. 58–62.
9. Novak Y., Mazurets O. Practical application of method of automated personal identification by fingerprints using convolution neural networks // Modern Strategies

of Global Scientific Solutions : proceedings of the V Int. Sci. and Pract. Conf. (Stockholm, Sweden, Dec. 27–29, 2023). – Stockholm, 2023. – P. 136–140.
10. Zharnovskyi O., Mazurets O., Sobko O. Approach to identification of artificial intelligence-generated people images by means of machine learning // Key Aspects of the Development of Scientific Research in Modern Conditions : proceedings of the XLV Int. Sci. and Pract. Conf. (Constanta, Romania, Oct. 30 – Nov. 1, 2024). – Constanta, 2024. – P. 69–73.
11. Mazurets O., Zalutska O., Tyschenko O., Bohdanova A. An approach to using MobileNet CNN-model for gesture recognition // Problems of Science and Technology: the Search for Innovative Solutions : proceedings of the XXIII Int. Sci. and Pract. Conf. (Munich, Germany, May 15–17, 2024). – Munich, 2024. – P. 59–64.
12. Malaydakh V., Molchanova M., Shevchuk P., Mazurets O. Deep learning neural network architecture for determining sunflower growth stage from visual data // Modern Scientific Research: Theoretical and Practical Aspects : proceedings of the II Int. Sci. and Pract. Conf. (Riga, Latvia, May 26–28, 2025). – Riga, 2025. – P. 143–148.
13. Ostapchenko N., Tyschenko O., Denysenko B., Mazurets O. Semantic search of relevant images using vector databases // Modern Scientific Research: Theoretical and Practical Aspects : proceedings of the II Int. Sci. and Pract. Conf. (Riga, Latvia, May 26–28, 2025). – Riga, 2025. – P. 161–165.
14. Bohdanova A., Mazurets O., Sobko O. Gesture recognition using a neural network in real time // Black Sea Science 2023 : proceedings of the International Competition of Student Scientific Works. – Odesa : Odesa National University of Technology, 2023. – P. 556–566.
15. Mazurets O. V., Klimenko V. I., Shurypa M. O. Neural network assessment of buildings condition based on visual data // Innovations of Modern Science and Education : proceedings of the IV Int. Sci. and Pract. Conf. (Vancouver, Canada, Dec. 25–27, 2025). – Vancouver, 2025. – P. 120–129.
16. Derzhak V. V., Mazurets O. V. Approach to image preprocessing for household waste classification in circular economy // Resource-Saving Technologies of Apparel, Textile & Food Industry : proceedings of the Int. Sci. and Pract. Conf. (Khmelnytskyi, Ukraine, Nov. 20, 2025). – Khmelnytskyi, 2025. – P. 319–323.
17. Mazurets O., Ovcharuk O., Vit R. Investigation of neural network detecting of construction waste by photos // Information Technology and Implementation (Satellite) : proceedings of the 12th Int. Conf. (Kyiv, Ukraine, Nov. 21, 2025). – Kyiv, 2025. – P. 96–97.
18. Molchanova M., Didur V., Mazurets O., Sobko O., Zakharkevich O. Method for construction and demolition waste classification using two-factor neural network image analysis // CEUR Workshop Proceedings. – 2025. – Vol. 3970. – P. 168–182.
19. Mazurets O., Sobko O., Dydo R., Zalutska O., Molchanova M. Augmented reality audiostream creation using CNN: Boosting inclusion and safety for visually impaired people // CEUR Workshop Proceedings. – 2025. – Vol. 4004. – P. 347–361.