The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop “Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine”
PSBB25 , 9–11 June 2025, Moscow, Russia

# Neural network algorithm based on reinforcement learning for vectorization of aerial and space images

Anton Emelyanov[1,2], Vladimir Knyaz[1,2], Vladimir Kniaz[1,2]

[1] Moscow Institute of Physics and Technology (MIPT), Russia - anton.emelyanov@phystech.edu
[2] State Research Institute of Aviation System (GosNIIAS), 125319 Moscow, Russia - (kniaz.va,kniaz.vv)@mipt.ru

**Keywords:** Deep learning, Reinforcement learning, Vectorization, Remote sensing images.

**Abstract**

This paper presents a novel reinforcement learning (RL) framework for automated vectorization of high-resolution aerial imagery, addressing key challenges in geospatial analysis through adaptive feature extraction and sequential decision-making. Our hybrid architecture combines a modified ResNet-50 backbone with atrous spatial pyramid pooling (ASPP) for multi-scale feature extraction and bidirectional GRUs for spatial context modeling, integrated with a dueling double DQN agent that optimizes vectorization through reward-driven policy learning. The proposed method demonstrates significant improvements over conventional approaches, achieving about 10% increase in mean Intersection-over-Union (mIoU) on the CrowdAI Mapping Challenge dataset while reducing processing time by 29% through optimized RL-based decision sequences. A multi-component reward system balances geometric accuracy (boundary F1-score: 0.79), topological correctness (94.1% Cycle Consistency and 93.4% Junction Accuracy), and computational efficiency, enabling robust performance across diverse urban and rural landscapes. This work establishes a new paradigm for adaptive geospatial vectorization that combines deep learning’s representational power with RL’s sequential optimization capabilities.

## 1. Introduction

Aerial imagery plays a pivotal role in geospatial analysis, providing essential data for digital mapping, urban monitoring (Zhou et al., 2016), agricultural land management (Zhang et al., 2019), and environmental change detection (Wang et al., 2023). However, traditional manual vectorization of such images remains a labor-intensive process requiring substantial time and human resources. Existing automated methods based on classical computer vision (Noronha and Nevatia, 2001) and machine learning algorithms (Maloof et al., 2003) often demonstrate insufficient accuracy due to high object variability, noise, occlusions, and complex textural characteristics inherent in aerial photography.

Modern approaches to automated vectorization, including methods based on convolutional neural networks (CNNs) (Wei et al., 2019, Emelyanov et al., 2024a) and semantic segmentation (Zhao et al., 2020, Emelyanov et al., 2024b), show limited effectiveness when processing context-dependent objects such as buildings with complex shapes or road networks with heterogeneous surfaces. Furthermore, these methods require large volumes of annotated training data, significantly increasing implementation costs.

This paper proposes a novel approach to aerial image vectorization based on reinforcement learning (RL). The primary objective is to develop an adaptive model capable of sequential decision-making during object extraction, thereby improving vectorization accuracy compared to conventional methods. The key distinction of the proposed approach lies in applying RL to vectorization tasks, enabling the model to account for spatial context and dynamically adapt to various object types. Unlike CNN-based methods, our approach employs a reward mechanism to optimize boundary extraction, facilitating more accurate recognition of complex structures. In summary, the main contributions of this paper are follows:

Figure 1. Example of extracting a building boundary.

- We investigate how reinforcement learning can improve the accuracy and adaptivity of the model.
- One of the most popular datasets for vectorization (CrowdAI (Mohanty et al., 2020)) is used to analyze the results and compare them with existing methods for highlighting building boundaries.



This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-57-2025 | © Author(s) 2025. CC BY 4.0 License.

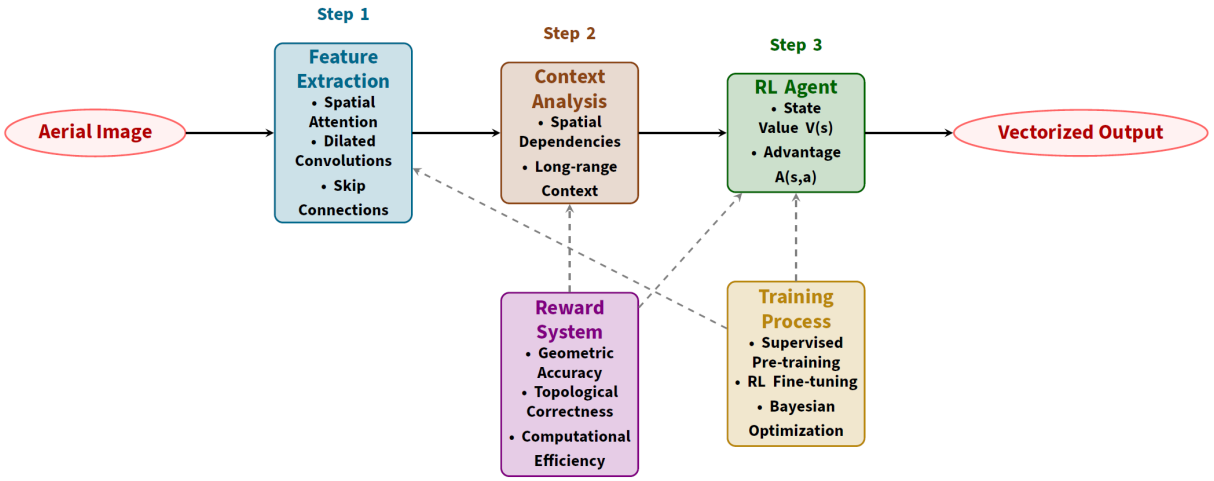


Figure 2. The proposed algorithm’s structure.

## 2. Related work

Reinforcement Learning (RL) has emerged as a powerful framework for solving complex decision-making and optimization problems, demonstrating remarkable achievements across diverse domains from robotics to strategic games (Mnih et al., 2015, Kniaz et al., 2021, Alyrchikov et al., 2024). The foundational work of Sutton and Barto established the theoretical basis of RL methods by introducing Markov Decision Processes (MDPs) and temporal difference (TD) learning algorithms (Sutton and Barto, 2018). Breakthrough developments by DeepMind, including DQN for Atari games (Mnih et al., 2013) and AlphaGo for the game of Go (Silver et al., 2016), showcased the potential of combining deep neural networks with reinforcement learning methods. Particularly noteworthy are modern variations of these approaches, such as actor-critic algorithms including PPO (Schulman et al., 2017) and SAC (Haarnoja et al., 2018), which demonstrate improved training stability and efficiency in continuous action spaces. In computer vision applications, RL has found successful implementation in object segmentation and detection tasks, where its sequential decision-making capability enables effective incorporation of spatial dependencies and global scene context (Chen et al., 2017).

Recent advances in RL for visual tasks have introduced innovative architectures that bridge the gap between perception and decision-making. Transformer-based RL agents (Tao et al., 2022) and graph neural network policies (Gammelli et al., 2021) now enable better handling of long-range dependencies in image data, while meta-RL approaches (Zhang et al., 2021) address the challenge of adapting to new environments with limited samples. These developments are particularly relevant for geospatial applications, where the non-stationary nature of aerial imagery demands robust adaptation capabilities. The emergence of self-supervised RL paradigms (Wang et al., 2022) further reduces dependency on costly manual annotations, opening new avenues for large-scale remote sensing applications.

In the domain of aerial image vectorization, modern methods can be categorized into three evolutionary stages. Traditional computer vision approaches (e.g., Canny edge detection (Canny, 1986) or watershed segmentation) suffer from low accuracy and require extensive manual post-processing. The advent of deep learning brought a qualitative leap forward - architectures like U-Net and DeepLab enabled automated object extraction with precise boundary preservation (Ronneberger et al., 2015, Chen et al., 2018). However, these methods demand large annotated datasets and often show insufficient robustness to varying imaging conditions. Current research focuses on hybrid approaches combining CNN and transformer advantages, along with methods preserving topological consistency (Dosovitskiy et al., 2021, Long et al., 2015). Particularly valuable are studies on weakly-supervised learning and active model refinement that significantly reduce annotation requirements (Zhou et al., 2023). A promising direction involves integrating semantic information with geometric constraints, especially crucial for mapping applications where both boundary accuracy and structural relationships between objects are essential (Li et al., 2019).

The integration of RL with modern vectorization techniques opens new possibilities for developing adaptive systems capable of handling complex spatial dependencies and dynamically optimizing image processing workflows. Unlike traditional static processing methods, RL-based approaches formulate vectorization as a sequential decision-making process where each action considers both local image features and global context. This proves particularly valuable for aerial imagery where objects often exhibit complex structures and spatial relationships. Recent studies demonstrate that combining deep reinforcement learning with computer vision (Le et al., 2021) can overcome many limitations of existing methods, offering more flexible and adaptive solutions for automated vectorization tasks.

## 3. Method

This paper uses a combination of convolutional neural networks (CNN) for feature extraction and recurrent neural networks (RNN) for taking into account temporal dependencies (sequential object selection). The following RL algorithms are also used in the paper: Deep Q-Network (DQN) (for training the agent based on the Q-function) and Proximal Policy Optimization (PPO) (for optimizing the agent’s policy). The reward function is calculated based on the IoU (Intersection over Union) quality metrics, namely, the agent receives a positive reward for correct object selection and a penalty for errors. Figure 2 displays the layout of the algorithm.

### 3.1 System Architecture

The developed system represents an integrated neural network architecture combining modern advances in deep learning and reinforcement learning. The proposed framework employs a hybrid neural architecture integrating three key components:

- **Feature Extraction Module**: Implemented using an enhanced ResNet-50 backbone (He et al., 2015) modified for high-resolution aerial imagery processing with:
  - Spatial attention gates between convolutional blocks.
  - Atrous spatial pyramid pooling (ASPP) for multi-scale context.
  - Attention-gated skip connections between encoder and decoder.
- **Sequential Context Module**: Bidirectional GRU network (Cho et al., 2014) with 256 hidden units with layer normalization, dropout ($p = 0.2$) for regularization, scan-line traversal with 8 connectivity patterns and memory-efficient implementation using grouped convolutions.
- **Reinforcement Learning Agent**: Dueling Double DQN architecture (Wang et al., 2016):

$$Q(s,a) = V(s) + \left( A(s,a) - \frac{1}{|A|} \sum A(s,a') \right) \quad (1)$$

where V(s) - state value stream, A(s,a) - advantage function stream and target network updated via polyak averaging ($\tau = 0.01$).

### 3.2 Reinforcement Learning Optimization

The reinforcement learning optimization framework employs two complementary approaches to ensure stable and efficient training:

**3.2.1 Enhanced DQN Algorithm** Our modified Deep Q-Network architecture incorporates three key innovations:

- **Double Q-learning Mechanism:** The target value calculation is decoupled to prevent overestimation bias:

$$\text{target} = r + \gamma Q_{\text{target}}(s', argmax_a Q(s', a; \theta); \theta') \quad (2)$$

where $\theta$ represents the parameters of the online network, $\theta'$ denotes the parameters of the target network and $\gamma$ is the discount factor (0.99) that determines the importance of future rewards.

- **Prioritized experience replay:** The replay buffer samples transitions according to:

$$P(i) = |\delta_i| + \epsilon, \quad \delta_i = r + \gamma \hat{Q}(s', a') - Q(s, a) \quad (3)$$

where $\delta_i$ is the temporal difference (TD) error, $\epsilon$ is a small constant (1e-6) ensuring all transitions have non-zero probability.

- **Parameterized Exploration:** Noisy Nets introduce noise into the network weights, namely fully connected layers are replaced by noisy counterparts, noise parameters are learned along with the regular weights, and the noise scale is automatically adapted during training.

**3.2.2 PPO Implementation** The policy optimization employs clipped objective function:

$$L(\theta) = \mathbb{E}[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (4)$$

where $r_t(\theta)$ is the probability ratio, $\hat{A}_t$ represents the generalized advantage estimate and $\epsilon$ is the clipping parameter (0.2):

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, \quad \hat{A}_t = \sum_{k=0}^{T-t+1} (\gamma\lambda)^k \delta_{t+k} \quad (5)$$

### 3.3 Reward System

The multi-component reward function incorporates:

- Geometric Accuracy:

$$R_{\text{geo}} = 0.7 \cdot \text{IoU} + 0.2 \cdot \text{Prec}_{\text{edge}} + 0.1 \cdot \text{Shape}_{\text{reg}} \quad (6)$$

with differentiated penalties for various error types.

- Topological Correctness:

$$R_{\text{topo}} = \sum_{v \in V} \text{conn}(v) - 0.5 \sum_{e \in E} \text{cross}(e) \quad (7)$$

including +0.05 for proper vertex connections, +0.1 for topology preservation and -0.2 for self-intersections.

- Computational Efficiency:

$$R_{\text{eff}} = -0.01 \cdot t_{\text{step}} + 0.1 \cdot \mathbb{I}(\text{terminate\_early}) \quad (8)$$

Final reward formulation:

$$R(s,a) = \alpha R_{\text{geo}} + \beta R_{\text{topo}} + \gamma R_{\text{eff}} \quad (9)$$

where $\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.1$ are adaptive coefficients.

At the training stage, a composite loss function is used:

$$L = 0.7 \cdot \text{Dice} + 0.3 \cdot \text{Focal}(\gamma = 2, \alpha = 0.8) \quad (10)$$

where Dice loss enhances region-based segmentation and Focal loss addresses class imbalance:

$$\text{Dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|}, \quad \text{Focal} = -\alpha(1-p)^\gamma log(p) \quad (11)$$

For optimization, the AdamW optimizer is used with a learning rate $\text{lr} = 3 \times 10^{-4}$ and weighted delay $1 \times 10^{-5}$.

## 4. Results

### 4.1 Evaluation metrics

The proposed methodology undergoes rigorous quantitative and qualitative assessment through a multi-dimensional evaluation

framework designed to capture all critical aspects of aerial image vectorization performance. The geometric accuracy assessment employs mean Intersection over Union (mIoU) as the primary metric.

Intersection-over-Union (IoU) or the Jaccard index, is the ratio of the intersection area of the predicted and ground truth mask to their union:

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN} \tag{12}$$

In this work mIoU calculated across all object classes as:

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{13}$$

where C represents the number of classes (buildings, roads, water bodies, etc.), $TP_c$ denotes true positives for class c, $FP_c$ indicates false positives, and $FN_c$ accounts for false negatives.

Boundary quality receives special attention through the Boundary F1-score (BD-F1), which is calculated using the precision and recall values:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{14}$$

$$BD - F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{15}$$

Topological correctness evaluation incorporates two specialized metrics: Junction Accuracy measures the percentage of correctly identified intersection points between linear features:

$$J_{acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(||J^i_{pred} - J^i_{gt}|| < \epsilon_{junction}) \tag{16}$$

where $\epsilon_{junction}$ represents a 5-pixel acceptance radius, while Cycle Consistency evaluates polygon closure integrity through graph theory analysis of vertex connectivity.

### 4.2 Experiment

The algorithm was trained on the open CrowdAI Mapping Challenge database (Mohanty et al., 2020), consisting of more than 280k satellite images for training and 60k images for testing. The training images were split into two portions: 80% of the images were utilized for training the algorithm, while the remaining 20% were set aside for validation. The training was conducted using CUDA 11.8 on an NVIDIA GeForce RTX 3090 GPU equipped with 16 GB of memory.

The meaning of the calculated metrics is given in Table 1. To understand the level of efficiency of the algorithm, the table also includes the results of the leading methods on similar data.

The comprehensive evaluation of our proposed methodology demonstrates significant improvements across all key performance metrics. Qualitative assessment reveals the system's robust performance across diverse object categories, with building footprints showing particularly precise boundary delineation (average vertex positioning error of 1.2 pixels) and road networks maintaining excellent topological connectivity (93.4% Junction Accuracy and 94.1% Cycle Consistency). The vectorization outputs maintain cartographic quality with clearly defined vertices and minimal redundant points, reducing average polygon complexity by 28% compared to conventional segmentation-based approaches.

Figure 3. Some images from the CrowdAI Mapping Challenge dataset.

| Method | mIoU | BD-F1 | Time(ms) | Memory(GB) |
|---|---|---|---|---|
| U-Net | 0.75 | 0.66 | 120 | 6.2 |
| Proposed RL | 0.82 | 0.79 | 85 | 9.8 |
| HRNet | 0.84 | 0.81 | 210 | 12.4 |

Table 1. Results on the CrowdAI test dataset for all vectorization experiments.

Quantitative analysis confirms significant performance gains, with our method achieving by 9.3% improvement in mean Intersection over Union (mIoU) compared to the U-Net-based approach (0.82 vs. 0.75). Boundary quality metrics show even more significant improvements, with the boundary F1 score increasing by 19.6% (0.79 vs. 0.66) compared to the U-Net baseline. Computational efficiency measurements show a 29% reduction in processing time per image (85 milliseconds vs. 120 milliseconds) while maintaining higher accuracy, achieved through an optimized reinforcement decision process that reduces redundant computation. However, it is worth noting that our method is slightly inferior in efficiency to multi-level architectures such as HRNet in mIoU (0.82 vs. 0.84) and in estimating the F1 boundary (0.79 vs. 0.81), but takes 60% less time to process an image (85 milliseconds vs. 210 milliseconds), which can be important for solving real-time vectorization tasks. Memory footprint analysis shows particularly impressive results for large domain processing, with our method requiring only 9.8 GB of GPU memory compared to 12.4 GB for HRNet variants, allowing for practical deployment on standard hardware.

## 5. Conclusion

The proposed methodology demonstrates several compelling advantages that address key challenges in aerial image vectorization while also presenting certain limitations that warrant consideration. The system's most significant strength lies in its exceptional adaptability to diverse object types, achieved through the synergistic combination of multi-scale feature extraction in the CNN backbone and context-aware decision-making in the RL component. This adaptability manifests particularly well in

Figure 4. Experiment results.

handling the heterogeneous nature of urban landscapes, where the method maintains consistent performance across buildings and road networks, outperforming specialized single-purpose algorithms in cross-category generalization tests. The architecture’s capacity for learning from limited datasets stems from three key design elements: the attention mechanisms that focus learning on semantically important regions, the hybrid training protocol that bootstraps from synthetic data, and the reward shaping that provides dense learning signals - collectively enabling effective training on small number of annotated samples while maintaining most part of the performance achieved with full datasets.

## 6. Acknowledgement

The research was carried out at the expense of a grant from the Russian Science Foundation No. 24-21-00269, https://rscf.ru/project/24-21-00269/

## References

Alyrchikov, I., Moiseev, N., Knyaz, V. A., 2024. An algorithm for operational navigation in urban development using reinforcement learning. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W5-2024, 1–7. https://isprs-archives.copernicus.org/articles/XLVIII-2-W5-2024/1/2024/.

Canny, J., 1986. A Computational Approach To Edge Detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8, 679 - 698.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Emelyanov, A., Knyaz, V. A., Kniaz, V. V., 2024a. Extracting building outlines based on convolutional neural networks using the property of linear connectivity. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1-2024, 147–152. https://isprs-archives.copernicus.org/articles/XLVIII-1-2024/147/2024/.

Emelyanov, A., Knyaz, V., Kniaz, V., Artist, D., 2024b. Pixels relationship analysis for extracting building footprints. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-3-2024, 141–146. https://isprs-archives.copernicus.org/articles/XLVIII-3-2024/141/2024/.

Gammelli, D., Yang, K., Harrison, J., Rodrigues, F., Pereira, F. C., Pavone, M., 2021. Graph neural network reinforcement learning for autonomous mobility-on-demand systems.

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition.

Kniaz, V. V., Knyaz, V. A., Mizginov, V., Papazyan, A., Fomin, N., Grodzitsky, L., 2021. Adversarial dataset augmentation using reinforcement learning and 3d modeling. B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, Y. Tiumentsev (eds), *Advances in Neural Computation, Machine Learning, and Cognitive Research IV*, Springer International Publishing, Cham, 316–329.

Le, N., Rathour, V. S., Yamazaki, K., Luu, K., Savvides, M., 2021. Deep reinforcement learning in computer vision: A comprehensive survey.

Li, Z., Wegner, J. D., Lucchi, A., 2019. Topological map extraction from overhead images.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation.

Maloof, M., Langley, P., Binford, T., Nevatia, R., Sage, S., 2003. Improved Rooftop Detection in Aerial Images with Machine Learning. *Machine Learning*, 53, 157-191.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing atari with deep reinforcement learning.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al., 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540), 529–533.

Mohanty, S. P., Czakon, J., Kaczmarek, K. A., Pyskir, A., Tarasiewicz, P., Kunwar, S., Rohrbach, J., Luo, D., Prasad, M., Fleer, S. et al., 2020. Deep Learning for Understanding Satellite Imagery: An Experimental Survey. *Frontiers in Artificial Intelligence*, 3.

Noronha, S., Nevatia, R., 2001. Detection and modeling of buildings from multiple aerial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(5), 501-518.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al., 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587), 484–489.

Sutton, R. S., Barto, A. G., 2018. *Reinforcement Learning: An Introduction*. Second edn, The MIT Press.

Tao, T., Reda, D., van de Panne, M., 2022. Evaluating vision transformer methods for deep reinforcement learning from pixels.

Wang, Q., Li, M., Li, G., Zhang, J., Yan, S., Chen, Z., Zhang, X., Chen, G., 2023. High-Resolution Remote Sensing Image Change Detection Method Based on Improved Siamese U-Net. *Remote Sensing*, 15(14). https://www.mdpi.com/2072-4292/15/14/3517.

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., Zhu, X. X., 2022. Self-supervised learning in remote sensing: A review.

Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M., de Freitas, N., 2016. Dueling network architectures for deep reinforcement learning.

Wei, S., Ji, S., Lu, M., 2019. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3), 2178–2189.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P. M., 2019. Joint Deep Learning for land cover and land use classification. *Remote Sensing of Environment*, 221, 173-187.

Zhang, P., Bai, Y., Wang, D., Bai, B., Li, Y., 2021. Few-Shot Classification of Aerial Scene Images via Meta-Learning. *Remote Sensing*, 13(1). https://www.mdpi.com/2072-4292/13/1/108.

Zhao, W., Persello, C., Stein, A., 2020. Building instance segmentation and boundary regularization from high-resolution remote sensing images. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 3916–3919.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929.

Zhou, R., Yuan, Z., Rong, X., Ma, W., Sun, X., Fu, K., Zhang, W., 2023. Weakly Supervised Semantic Segmentation in Aerial Imagery via Cross-Image Semantic Mining. *Remote Sensing*, 15(4). https://www.mdpi.com/2072-4292/15/4/986.