UDC 004.93

URBAN SCENE SEGMENTATION USING HOMOGENEOUS U-NET ENSEMBLE: A STUDY ON THE CITYSCAPES DATASET

Hmyria I. O. – Post-graduate student of the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Kravets N. S. – Associate Professor, Associate Professor of the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. Semantic segmentation plays a critical role in computer vision tasks such as autonomous driving and urban scene understanding. While designing new model architectures can be complex, improving performance through ensemble techniques applied to existing models has shown promising potential. This paper investigates ensemble learning as a strategy to enhance segmentation accuracy without modifying the underlying U-Net architecture.

Objective. The aim of this work is to develop and evaluate a homogeneous ensemble of U-Net models trained with distinct initialization and data augmentation techniques, and to assess the effectiveness of various ensemble aggregation strategies in improving segmentation performance on complex urban dataset.

Method. The proposed approach constructs an ensemble of five structurally identical U-Net models, each trained with unique weight initialization and augmentation schemes to ensure prediction diversity. Several ensemble strategies are examined, including softmax averaging, max voting, proportional weighting, exponential weighting, and optimized weighted voting. Evaluation is conducted on the Cityscapes dataset using a range of segmentation metrics.

Results. Experimental findings demonstrate that ensemble models outperform individual U-Net instances and the baseline in terms of accuracy, mean IoU, and specificity. The optimized weighted ensemble achieved the highest accuracy (87.56%) and mean IoU (0.6504), exceeding the best individual model by approximately 3%. However, these improvements come with a notable increase in inference time, highlighting a trade-off between accuracy and computational efficiency.

Conclusions. The ensemble-based approach effectively enhances segmentation accuracy while leveraging existing model architectures. Although the increased computational cost presents a limitation for real-time applications, the method is well-suited for high-precision tasks. Future research will focus on reducing inference time and extending the ensemble methodology to other architectures and datasets.

KEYWORDS: convolutional neural network, semantic segmentation, U-Net, ensemble learning, data augmentation techniques, model initialization, Cityscapes, urban scenes.

ABBREVIATIONS

CNN is a convolutional Neural Network;
U-Net is a U-shaped network architecture;
ELU is an Exponential Linear Unit;
ReLU is a Rectified Linear Unit;
IoU is a Intersection over Union;
Mean IoU / mIoU is a Mean Intersection over Union;
TP is a True Positive;
FP is a False Positive;
FN is a False Negative;
TN is a True Negative;
RGB is a Red Green Blue.

NOMENCLATURE

$F(X)$ is a convolutional neural network performing semantic segmentation;
X is an input image space;
Y is an output segmentation map (label space);
x is an input image;
y is a ground truth segmentation map;
H is a height of the input image;
W is a width of the input image;
C is a number of channels of the input image;
$\hat{y}$ is a predicted segmentation output;
$\arg\max_k$ is a predicted class;
c_i is a segmentation class;
$F_i(x)$ is a prediction of the U-Net model in the ensemble;
w_i is a weight assigned to the model in the ensemble;
$E(x)$ is an ensemble output;
$\hat{y}_{final}$ is a final predicted class map from the ensemble;
$Metric$ is a segmentation metric;
$accuracy$ is an accuracy of a model;
$T(F_i)$ is an inference time of the model;
$T(E)$ is a total inference time of the ensemble;
IoU is an intersection over union;
N is a normal distribution with mean;
μ is a variance;
α is a displacement intensity in elastic deformation;
σ is a standard deviation of Gaussian noise;
w is an ensemble weight vector;
fan_in is a number of input units;
fan_out is a number of output units.

INTRODUCTION

Deep learning has made impressive strides in image segmentation, especially when it comes to parsing

© Hmyria I. O., Kravets N. S., 2025
DOI 10.15588/1607-3274-2025-3-7

complex urban scenes. This task is essential for technologies like autonomous vehicles, smart traffic systems, and overall city infrastructure management. A widely recognized benchmark in this domain is the Cityscapes dataset [1], known for its high-resolution, finely annotated road scene images that have become standard for evaluating model performance.

Architectures like U-Net have shown strong results in semantic segmentation, yet they still struggle with generalizing across varying conditions – think changes in lighting, weather, or city layouts. To mitigate these issues, ensemble learning has emerged as a valuable approach. By merging predictions from multiple models, it boosts accuracy and stabilizes results. While traditional ensemble methods often mix different types of models [2], a homogeneous ensemble – where multiple U-Net models are trained separately – offers a balance between performance and efficiency. This diversity, introduced through unique initializations and data augmentations, helps improve outcomes without the added complexity of mixing architectures.

In this work, we investigate several homogeneous ensembling techniques – such as averaging, max pooling, and weighted voting – to see how each impacts segmentation results. We propose a U-Net-based ensemble tailored for urban image segmentation, using distinct augmentation and initialization variations, and run experiments to evaluate its effectiveness. Our analysis includes comparisons of accuracy, speed, and computational overhead to weigh the trade-offs of each method.

The object of study in this research is semantic segmentation of urban scenes, specifically focusing on the challenges posed by varying environmental conditions, such as changes in lighting, weather, and occlusions. The study is conducted using the Cityscapes dataset, which provides high-resolution images of complex urban environments with 34 semantic classes.

The subject of study is the method of constructing homogeneous U-Net ensembles to improve the accuracy and robustness of semantic segmentation in urban environments. This includes exploring different ensemble strategies to enhance generalization across diverse scenes while maintaining computational efficiency.

The purpose of this work is to **i**mprove the generalization ability of U-Net models for urban scene segmentation by leveraging networks homogeneous ensembling. The study aims to demonstrate that a homogeneous ensemble of multiple U-Net models can achieve higher segmentation accuracy and robustness compared to a single U-Net, particularly in conditions found in real-world urban environments.

1 PROBLEM STATEMENT

Formally, the semantic segmentation task can be described as a pixel-wise classification problem, where the goal is to assign each pixel of an input image to one of the predefined semantic classes.

Let the input image be denoted as $x \in X$, where $X \subset R^{H \times W \times C}$. The output segmentation map is denoted as $y \in Y$, where $Y \subset Z^{H \times W}$, and each pixel value corresponds to a class label from the set of predefined classes $C = \{c_1, c_2, ..., c_K\}$.

The model is a function $F : X \to Y$, implemented using a deep convolutional neural network architecture, particularly U-Net. The output of the model is a probability tensor $\hat{y} = F(x) \in [0,1]^{H \times W \times K}$, and the predicted class for each pixel is obtained by:

$$\hat{y}_{i,j} = \arg\max_{k \in C} F(x)_{i,j,k} .$$

Let's define a set of n trained models $\{F_1, F_2, ..., F_n\}$ each producing a prediction $\hat{y}_i = F_i(x)$. We define the ensemble function E as a weighted combination of the model outputs $E(x) = \sum_{i=1}^{n} w_i F_i(x)$, subject to $\sum_{i=1}^{n} w_i = 1$, $w_i \geq 0$. The final prediction is obtained by taking the argmax over the ensembled output:

$$\hat{y}_{final} = \arg\max_{k \in C} E(x)_{i,j,k} .$$

The objective is to find the optimal weight vector $w = (w_1, w_2, ..., w_n)$ such that the ensemble prediction maximizes a chosen segmentation quality metric, such as the mean Intersection over Union (mIoU):

$$\max_{w} Metric(E(x), y) , \ \sum_{i=1}^{n} w_i = 1 , \ w_i \geq 0.$$

Additionally, due to hardware limitations, inference must be performed on a CPU-based system, where parallel execution is not available, and models are evaluated sequentially. Let $T(F_i)$ denote the execution time of model F_i. The ensemble execution time is therefore:

$$T(E) = \sum_{i=1}^{n} T(F_i).$$

The constraint is to keep inference time within an acceptable range $T_{\max}$, defined based on application requirements:

$$T(E) \leq T_{\max} .$$

DOI 10.15588/1607-3274-2025-3-7

2 REVIEW OF THE LITERATURE

Image segmentation [3] plays a key role in computer vision, and in robot vision, aiming to identify and outline objects within an image. Traditional methods – like thresholding [4], region growing [5], and edge detection [6] – have largely given way to deep learning-based approaches [7–9], which excel at learning layered features and handling complex textures.

Among deep learning models, Convolutional Neural Networks (CNNs) [10] have become the backbone of many segmentation tasks. Fully Convolutional Networks (FCNs) [11] were among the earliest deep models to perform pixel-level classification, showing the potential of CNNs in segmentation. Yet, U-Net [12] – a fully convolutional encoder-decoder design with skip connections – has emerged as the preferred choice, especially in biomedical applications. Its symmetric structure allows it to retain spatial details, making it particularly suitable for use cases like autonomous driving. Enhanced versions, such as Attention U-Net [13] and Residual U-Net [14], add mechanisms for better feature focus and improved performance in complex datasets.

Despite its strengths, U-Net and other single-model architectures often face challenges in generalizing across varied datasets. Differences in image quality, noise, and structural variations can lead to inconsistent results. These issues have encouraged the adoption of ensemble learning to boost consistency and resilience.

Ensemble learning has long been explored as a way to enhance model reliability by combining multiple learners. Techniques like bagging [15], boosting [16], and stacking [17] have shown success in improving generalization through model diversity. Ensemble learning has demonstrated strong performance improvements across a variety of machine learning tasks even beyond computer vision. For instance, in time series forecasting [18], authors proposed an ensemble of adaptive predictors capable of real-time learning on multivariate non stationary sequences. In segmentation, deep learning ensembles are typically either heterogeneous or homogeneous.

Heterogeneous ensembles [19], which mix various model types, can improve accuracy by capturing different feature perspectives. However, this comes at the cost of greater computational demands and system complexity. Homogeneous ensembles [20], on the other hand, use multiple instances of the same architecture, each trained under varied conditions – such as different initializations, hyperparameters, or data augmentations. Research [21] suggests that such homogeneous setups can match or even surpass heterogeneous ensembles, all while remaining more efficient.

Several studies illustrate the promise of ensembling in segmentation. One work combined 3D CNNs for brain lesion detection [22], demonstrating reduced uncertainty through model fusion. Another leveraged a U-Net ensemble trained with diverse loss functions to improve lung nodule segmentation [23]. These examples underline the benefits of ensembles in reducing prediction variance and improving robustness.

The importance of adaptivity in visual systems has been emphasized not only in segmentation architectures but also in image preprocessing approaches. For example, Smelyakov et al. [24] developed an adaptive image enhancement model for robotic vision systems, enabling real-time responsiveness to variable environmental conditions.

The existing literature consistently shows that ensemble learning enhances both the accuracy and stability of segmentation models. While many studies have tested different ensemble strategies, few have taken a detailed look at the trade-offs between segmentation accuracy and scalability. Building on previous findings, this paper proposes a homogeneous ensemble of U-Net models, each trained with unique weight initializations and augmentation schemes, using optimized voting strategy. Various inference methods are evaluated to better understand how ensemble design choices affect segmentation performance and efficiency.

3 MATERIALS AND METHODS

The Cityscapes dataset serves as a large-scale benchmark tailored for urban scene understanding, particularly focusing on tasks such as semantic segmentation, instance segmentation, and depth estimation. It features high-resolution imagery (2048×1024 pixels) captured from a vehicle-mounted camera as it navigates through 50 cities across Germany, Switzerland, and France. These images encompass a wide range of environmental conditions, including various weather scenarios and lighting settings throughout the day, thereby providing a comprehensive dataset for evaluating and training deep learning models used in autonomous driving and urban analysis.

This dataset contains 5.000 finely annotated images, distributed across 2.975 for training, 500 for validation, and 1.525 for testing, with annotations for the test set not publicly available. Additionally, it offers 20.000 coarsely annotated images as a supplementary resource. The annotation schema spans 34 semantic categories, including classes such as roads, buildings, vegetation, vehicles, pedestrians, and traffic signs. Each pixel in the finely annotated set is labeled with a semantic class, allowing for precise pixel-wise learning. Due to its detailed labeling, high resolution, and inherent class imbalance, the Cityscapes dataset has become a gold standard for evaluating segmentation models like U-Net. Given these challenges, leveraging a homogeneous ensemble of U-Net models offers a promising approach to improving segmentation performance by reducing variance and enhancing generalization, particularly in urban environments with fine-grained structures, dynamic lighting, and frequent occlusions.

To bolster the generalization capacity of the homogeneous U-Net ensemble, a diverse range of data

DOI 10.15588/1607-3274-2025-3-7

augmentation strategies was applied, with each of the five networks in the ensemble trained using a distinct transformation method. This approach promotes the learning of unique and complementary feature representations across models, thereby reducing overfitting and enhancing robustness on the Cityscapes dataset. Augmentations were chosen to simulate real-world visual variability while preserving the fundamental structure and semantics of objects within the scene. Urban environments naturally involve variations in object distance, camera perspective, noise, occlusion, and image distortion. Therefore, the selected augmentations were designed to reflect these real-world variations while maintaining semantic integrity.

Scaling was applied by randomly resizing input images within a predefined range. This helped the model develop scale-invariant features, which are critical for segmenting objects appearing at varying distances from the camera. Rotation was used to introduce random angular transformations, enhancing the model's ability to recognize and segment objects regardless of orientation – a common challenge in dynamic urban settings. Affine transformations, including shearing, translation, and reflection, were incorporated to introduce spatial diversity without disrupting the essential spatial structure of objects, thereby encouraging the model to generalize better under changes in viewpoint or alignment. One network in the ensemble was trained using elastic deformation, a technique adapted from medical imaging applications. This method simulates local, nonlinear distortions within the image, which is particularly useful for modeling real-world deformations in classes like pedestrians or vehicles, which often exhibit variable shapes and poses. Gaussian noise was added to simulate sensor noise, compression artifacts, and environmental distortions. This augmentation made the model more resilient to unpredictable visual noise and inconsistencies present in real-world imagery.

By assigning a unique augmentation strategy to each model, the ensemble was exposed to a broad spectrum of visual conditions. This diversity in learning experiences encouraged the models to acquire distinct yet complementary internal representations. Consequently, the ensemble could capture a wider range of features and generalize more effectively across complex urban scenes with challenging visual variability.

To improve training stability and ensure convergence across the homogeneous ensemble, each U-Net model was initialized using a distinct weight initialization technique. The importance of proper initialization in deep neural networks is well-established, particularly in preventing vanishing or exploding gradients, enhancing learning efficiency, and improving generalization performance. In this work, five different initialization strategies were employed: Glorot Normal, He Uniform, Orthogonal Initialization, LeCun Normal, and Random Normal.

The Glorot Normal method [25], also referred to as Xavier Normal, initializes weights from a truncated normal distribution centered at zero, with variance scaled based on both the number of incoming and outgoing connections. Weights are initialized by sampling from a truncated normal distribution centered at 0 with a standard deviation of:

$$\sigma = \sqrt{\frac{2}{\text{fan\_in} + \text{fan\_out}}} .$$

This technique helps maintain a balanced variance of activations across layers, which is particularly beneficial when using sigmoid or tanh activation functions.

He Uniform initialization [26], designed for networks employing ReLU activations, samples weights from a uniform distribution scaled by the number of input units. This ensures that activations are well-scaled during forward propagation, improving training stability in deep architectures. Weights are initialized by sampling from a uniform distribution within [–limit, limit], where:

$$\text{limit} = \sqrt{\frac{6}{\text{fan\_in}}} .$$

Orthogonal Initialization involves generating weight matrices that form an orthogonal basis, typically achieved through QR decomposition of randomly generated matrices. This approach helps preserve information flow during both forward and backward passes, making it especially effective for deep convolutional models. Weights are initialized by generating a random matrix and applying QR decomposition to obtain an orthogonal matrix. Specifically, for a weight matrix $\mathrm{W} = \mathrm{Q} \times \mathrm{R}$.

LeCun Normal initialization [27] is similar in concept to Glorot Normal but scales weights based solely on the number of input units, offering improved stability for tanh and sigmoid-based networks of moderate depth. Weights are initialized by sampling from a truncated normal distribution centered at 0 with a standard deviation of:

$$\sigma = \sqrt{\frac{1}{\text{fan\_in}}} .$$

Finally, one network was initialized using a Random Normal distribution with manually specified mean and standard deviation, providing a baseline for comparing the effectiveness of more sophisticated initializers. Weights are initialized by sampling from a normal distribution:

$$W \sim N(\mu, \sigma^2) .$$

The assignment of initialization methods to specific augmentation strategies was done purposefully to enhance model diversity and learning dynamics. For

DOI 10.15588/1607-3274-2025-3-7

augmentations that alter spatial characteristics – such as scaling or affine transformations – initialization techniques like Glorot Normal and Orthogonal Initialization were chosen, as they preserve activation variance even under substantial input variation. Rotation-based augmentations, which introduce directional shifts without distorting spatial structure, were paired with He Uniform initialization due to its suitability for ReLU-based networks and its ability to facilitate rapid early learning. Elastic deformation, which applies localized and nonlinear distortions, was combined with LeCun Normal initialization, providing a low-variance starting point that helps avoid overfitting in early training phases. The combination of Gaussian noise augmentation and Random Normal initialization introduced variability both at the data and model initialization level, offering a useful control scenario for measuring the effects of structured randomness.

This strategic pairing of augmentations and initializations promoted heterogeneity in feature representations and error patterns across the ensemble, which is essential for achieving high segmentation accuracy through ensemble learning. The result is a more resilient and generalizable model, capable of handling the diverse challenges inherent in urban scene segmentation.

To establish a baseline for evaluating the effectiveness of the proposed homogeneous U-Net ensemble, a standard U-Net model was implemented. This model, widely recognized in semantic segmentation tasks, is particularly well-suited for applications involving urban scenes, such as those found in the Cityscapes dataset. The U-Net architecture (Fig. 1) adopts a symmetric encoder-decoder structure, which enables accurate pixel-level classification – an essential capability for high-resolution urban segmentation.

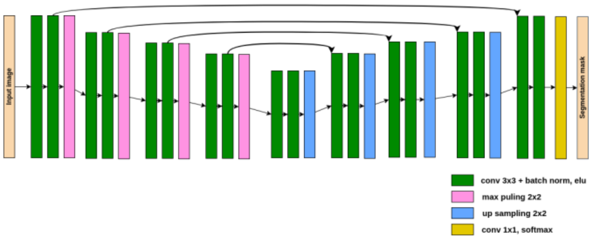

Figure 1 – Unet baseline acrhitecture

The network's architecture is composed of two main components. The encoder, also known as the contracting path, systematically reduces the spatial dimensions of the input image while extracting progressively higher-level features. Each block in the encoder includes two convolutional layers followed by Batch Normalization and ReLU activations, with Max Pooling layers applied between blocks to downsample the feature maps. At each stage of downsampling, the number of feature channels doubles, beginning from 64 and reaching up to 1024, allowing the model to learn increasingly abstract representations of the input scene.

In the decoder, or expanding path, the spatial resolution of the feature maps is gradually restored using transposed convolutions. To retain fine-grained spatial information lost during the encoding process, skip connections linking corresponding layers between the encoder and decoder. After each upsampling step, the upsampled feature map is concatenated with its encoder counterpart, followed by two convolutional layers and Batch Normalization, which further refine the segmentation outputs. The network concludes with a 1×1 convolutional layer that projects the final feature map to the desired number of segmentation classes, and a softmax activation function is applied to produce the class probabilities for each pixel.

The baseline model is configured to accept input images of size 256×256×3, with a total parameter count of approximately 35.8 million. Training was conducted using categorical cross-entropy as the loss function, optimized with the Adam algorithm and an initial learning rate of 0.001. The training set was processed in mini-batches of 16 images, and the network was trained for up to 30 epochs, with early stopping triggered based on the validation loss to prevent overfitting.

This baseline U-Net model serves as a reference point against which the ensemble approach is assessed. By comparing its performance with that of the ensemble – composed of multiple U-Net variants trained with different augmentation strategies and initialization schemes – it becomes possible to quantify the benefits of ensemble learning in enhancing segmentation accuracy and robustness.

To improve segmentation accuracy and enhance the generalization capabilities beyond what a single U-Net model can offer, an ensemble of five U-Net networks was constructed. While all five models shared the same architectural design as the baseline U-Net, they differed in their training setup through distinct combinations of weight initialization and data augmentation strategies. This intentional diversification enabled the ensemble to learn a wider array of feature representations, ultimately leading to stronger performance on complex urban segmentation tasks within the dataset.

The diversity within the ensemble was introduced through two complementary mechanisms. The first involved using different weight initialization schemes for each model, which encouraged unique learning dynamics by altering the starting conditions of training. The initializers applied – Glorot Normal, He Uniform, Orthogonal, LeCun Normal, and Random Normal – each influenced the convergence path in different ways, thereby promoting model independence and reducing the risk of all networks settling into similar local minima. The second mechanism of diversification relied on data augmentation. Each U-Net model was trained using a specific transformation technique – ranging from scaling and rotation to affine transformation, elastic deformation,

DOI 10.15588/1607-3274-2025-3-7

and Gaussian noise. These augmentations simulated a variety of real-world conditions found in urban environments, compelling each network to adapt to distinct types of variability, which in turn increased the ensemble's robustness to unseen data.

To combine the outputs from the ensemble, multiple prediction aggregation strategies were explored, each offering a different method of consolidating the networks' decisions. The first approach involved averaging the softmax probability outputs of each model on a pixel-wise basis. This method smoothed out individual inconsistencies and allowed the final segmentation map to reflect a balanced consensus across all predictions. An alternative strategy employed a maxing operation, selecting the highest softmax probability across the ensemble for each pixel. This method emphasized high-confidence predictions by giving more weight to confident outputs from any individual model.

Beyond these basic ensemble strategies, a more refined weighted voting method was developed to optimize how each model contributed to the final output. Here, the influence of each network was proportional to its validation accuracy, ensuring that more reliable models had a stronger impact on the final segmentation results. To further refine this weighting scheme, an exponential scaling mechanism was introduced, amplifying the contributions of the top-performing models while still allowing all ensemble members to participate in the decision-making process. This balance maintained the diversity benefits of ensembling while increasing the precision of the final predictions.

To optimize the distribution of weights among the models, a grid search procedure was performed. Rather than assigning equal weights, the goal was to identify the optimal weight vector w=[w1,w2,...,wN] that would yield the highest segmentation accuracy, as measured by the Dice score across the entire validation set. This optimization process ensured that the ensemble not only leveraged the strengths of individual models but also fine-tuned their contributions to achieve maximal overall performance.The ensemble prediction is computed as a weighted sum of the individual model predictions:

$$\hat{Y} = \sum_{i=1}^{N} w_i \cdot P_i \,.$$

To ensure model contributions remain meaningful and balanced, we enforce the following constraints on the weights:

$$0 \le w_i \le 1 \,, \ \sum_{i=1}^{N} w_i \approx 1 \,.$$

This formulation prevents any single model from dominating the ensemble while allowing flexibility for weight adjustments.

The optimization process seeks to maximize the dice score:

$$Dice(Y, \hat{Y}) = \frac{2\sum_{i=1}^{N}(Y_i \cdot \hat{Y}_i) + \varepsilon}{\sum_{i=1}^{N} Y + \sum_{i=1}^{N} \hat{Y} + \varepsilon} \,.$$

We define the objective function as:

$$\max_{w} \frac{2\sum_{i=1}^{N}(Y_i \cdot \hat{Y}_i) + \varepsilon}{\sum_{i=1}^{N} Y_i + \sum_{i=1}^{N} \hat{Y}_i + \varepsilon} \,.$$

We use constrained numerical optimization or Powell's method [28] to solve for the optimal weight vector.

4 EXPERIMENTS

The experiment was conducted using the Cityscapes dataset, which was divided into 2,975 training images and 500 for validation. To ensure consistency, all input images were resized and normalized prior to training, enhancing numerical stability and model convergence. The dataset was also shuffled randomly to avoid any learning bias, and mini-batches of size 16 were used to optimize computational efficiency.

For the baseline, a standard U-Net model was deployed without explicit weight initialization – weights were set to zero by default. The Exponential Linear Unit (ELU) activation function was used throughout the network to support better gradient flow and accelerate convergence in deeper layers. An early stopping strategy was applied, halting training automatically once the validation loss ceased to improve, thus preventing overfitting and reducing computational overhead.

Performance was evaluated using a suite of metrics, including Mean Intersection over Union (Mean IoU), pixel accuracy, precision, sensitivity, and specificity. These metrics offered a comprehensive view of model performance, capturing both pixel-level accuracy and class-level segmentation effectiveness. This baseline served as a critical reference point for assessing the effectiveness of the proposed homogeneous ensemble approach.

The first ensemble model maintained the baseline architecture but introduced Glorot Normal initialization to ensure balanced activation variance across layers. The activation function was switched to sigmoid to produce smoother probability maps suitable for segmentation tasks. Additionally, scaling augmentation was applied, randomly zooming input images to simulate changes in object size and distance. These modifications aimed to improve stability, generalization, and robustness to scale

DOI 10.15588/1607-3274-2025-3-7

variance while preserving compatibility with the overall ensemble structure.

The second U-Net also retained the base architecture but employed He Normal initialization, tailored for ReLU activations, to facilitate deeper gradient flow. Rotation-based augmentation was introduced, with input images randomly rotated up to 30 degrees to simulate real-world changes in camera angle. Nearest-neighbor interpolation was used to maintain pixel quality. This configuration allowed the model to develop rotation-invariant features, enhancing its performance in dynamic urban environments.

The third model applied Orthogonal initialization to promote stable training by preserving variance throughout deep layers, in conjunction with ELU activation to support gradient propagation. Affine transformations – including translation, scaling, shearing, and minor rotations – were used as augmentations to introduce spatial diversity. This combination encouraged the model to learn features invariant to subtle spatial distortions typical in real-world imagery.

The fourth model employed LeCun Normal initialization, optimized for tanh activations, and was paired with Elastic Deformation as the augmentation technique. By introducing smooth, localized warping through parameterized displacement fields ($\alpha = 10$, $\sigma = 4$), the model became better equipped to generalize across irregular object shapes and occlusions. This setup enabled the model to develop fine-grained sensitivity to structural deformations commonly seen in urban environments.

The final U-Net used Random Normal initialization to introduce variability in early learning trajectories. Gaussian noise was added to the input during training to simulate sensor-level imperfections, using a standard deviation of $\sigma = 0.05$. The ELU activation function was retained to aid in stable convergence. This model served to improve robustness under noisy conditions, rounding out the ensemble with additional stochastic diversity.

Upon training the five U-Net models, they were integrated into a homogeneous ensemble to capitalize on their individual strengths and improve segmentation accuracy, robustness, and generalization. To achieve this, three distinct ensemble strategies were explored: averaging, maxing, and weighted voting – each offering a different method for aggregating pixel-wise predictions.

In the averaging ensemble, the probability distributions generated by each model were averaged for every pixel. This approach mitigated the noise and uncertainty present in individual model outputs, yielding smoother and more balanced segmentation maps. It was particularly effective at improving generalization by consolidating diverse prediction patterns across the ensemble.

The maxing ensemble took a different approach, selecting the highest softmax probability across all five models for each pixel. This strategy emphasized confident predictions, allowing the most certain model to determine the final class decision per pixel. While this method enhanced decisiveness, it also introduced the risk of amplifying isolated high-confidence errors, depending on the reliability of individual networks.

To further refine prediction quality, a weighted voting ensemble was implemented. Here, each model's prediction was weighted according to its validation performance. The first weighting scheme assigned weights proportional to each model's validation accuracy, allowing higher-performing models to contribute more significantly to the final segmentation output.

The second approach used exponential scaling, amplifying differences between strong and weak models by applying an exponential function to the accuracy scores. This method increased the influence of top performers while still preserving the diversity contributed by other networks. Finally, a grid search optimization was conducted to identify the optimal weight vector $w = [w_1, w_2, ..., w_N]$. This involved evaluating different weight configurations on a subset of 10 validation images. The aim was to maximize the Dice score across the ensemble, ensuring that the final weighted output delivered the highest possible segmentation accuracy.

5 RESULTS

Once the ensemble models were constructed and integrated using the proposed aggregation strategies, a comprehensive evaluation was carried out to compare their performance against the baseline U-Net. This analysis focused on measuring segmentation accuracy, generalization, and robustness across both individual and ensemble models. All models were tested on the same validation set using consistent evaluation metrics, which included Mean Intersection over Union (Mean IoU), pixel accuracy, precision, sensitivity, specificity, and execution time measured in seconds per image. This consistent methodology ensured a fair comparison and provided a granular understanding of how each configuration performed. The training accuracy graph (Fig. 2) illustrates the learning progression of multiple U-Net models compared to the baseline over 25 epochs.

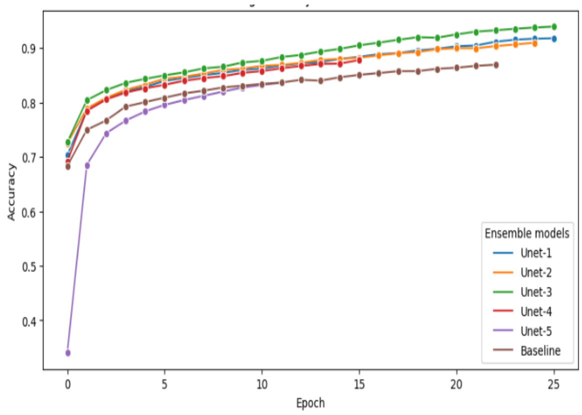

Figure 2 – Training accuracy across networks

Training accuracy trends revealed that all models experienced a rapid increase in performance within the

© Hmyria I. O., Kravets N. S., 2025
DOI 10.15588/1607-3274-2025-3-7

first five epochs, reflecting effective initial feature learning from the dataset. The baseline U-Net, though following a similar pattern, consistently trailed behind the other models. Among the ensemble components, U-Net-2 and U-Net-3 achieved the highest accuracy throughout training, indicating their ability to extract and generalize critical features. U-Net-5, on the other hand, consistently recorded the lowest accuracy, suggesting challenges in learning effective feature representations. By epoch 15, most models began to converge, with accuracy improvements tapering off and stabilizing near the 90% mark – except for U-Net-5, which continued to underperform. The baseline model remained consistently below the performance of all U-Net variants, reaffirming the benefits introduced by tailored augmentation and initialization strategies in the ensemble.

The graph Fig. 3 illustrates the validation accuracy of different U-Net models and the baseline over 25 epochs.

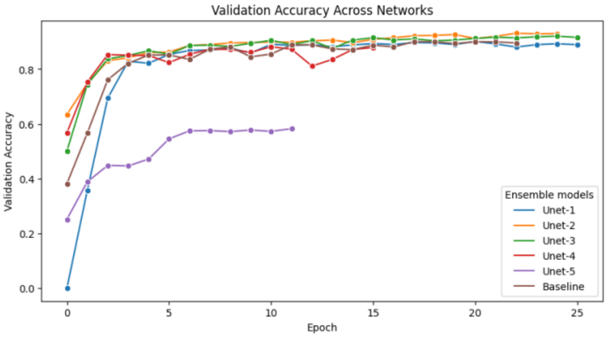

Figure 3 – Validation accuracy across networks

Validation accuracy followed a similar trajectory, offering further insight into the generalization capabilities of each model on unseen data. All models showed sharp improvements in validation accuracy during the initial training phase, mirroring their training performance. U-Net-2, U-Net-3, and U-Net-4 achieved the highest early-stage validation scores, indicating robust learning dynamics and generalization from augmented and well-initialized architectures. In contrast, U-Net-5 lagged significantly behind, maintaining a noticeably lower accuracy curve throughout training. The baseline model started with low initial accuracy but gradually improved, though by the fifth epoch, all U-Net variants had surpassed it. This confirmed the value of ensemble diversification strategies in enhancing model generalization.

By the end of training, the validation accuracy of most U-Net models converged between 82% and 84%, while the baseline plateaued slightly below this range. U-Net-5 remained a notable outlier, stabilizing around 57%, which suggests either insufficient regularization or overfitting to the training data. After epoch 10, most models displayed stable accuracy with minimal variation, indicating convergence. U-Net-2 and U-Net-3 maintained superior performance throughout, reflecting their consistency across both training and validation phases. The observed gap between training accuracy (approaching 90%) and validation accuracy (around 80%) across models points to potential overfitting – a likely result of dataset limitations and model complexity.

To assess the performance of the models, several evaluation metrics were employed, each designed to capture different aspects of segmentation accuracy and classification quality. One of the core metrics used was sparse categorical accuracy, which is particularly suitable when ground truth labels are provided as integer-encoded class indices rather than one-hot encoded vectors. This metric computes the proportion of correctly classified pixels by comparing the predicted class index – determined by the highest predicted probability – with the actual class label for each pixel:

$$accuracy = \frac{1}{\mathrm{N}}\sum_{i=1}^{N} 1(\arg\max_{\mathrm{c}} p_{i,c} = y_i)\ .$$

Another key metric is the Mean Intersection over Union (Mean IoU), a standard in semantic segmentation tasks. Mean IoU quantifies the average overlap between predicted and ground truth segmentation masks across all considered classes. However, given the class imbalance inherent to the Cityscapes dataset – where some classes dominate the dataset while others are infrequently represented – Mean IoU was computed over a targeted subset of six representative classes: 7 (road), 11 (building), 20 (traffic sign), 21 (vegetation), 23 (sky), and 26 (car). These selected categories encompass both large structural elements and smaller, yet semantically important, urban objects. This focused evaluation offers a more meaningful representation of model performance in real-world scenarios, rather than being skewed by rare or less relevant classes. IoU is calculated as following:

$$IoU = \frac{TP}{TP + FP + FN}\ .$$

Precision was also utilized to evaluate how reliable the model's positive predictions were. It measures the proportion of pixels that were correctly predicted as belonging to a particular class out of all pixels the model assigned to that class:

$$precision = \frac{TP}{TP + FP}\ .$$

In contrast, Sensitivity, also referred to as Recall, measures the model's ability to detect all relevant pixels that belong to a given class. This is calculated as the ratio of True Positives to the sum of True Positives and False Negatives:

$$sensitivity = \frac{TP}{TP + FN}\ .$$

DOI 10.15588/1607-3274-2025-3-7

Finally, Specificity was included to assess how well the model avoids false alarms. It evaluates the proportion of correctly identified negative pixels – those that do not belong to a particular class – relative to all true negatives and false positives. In this case, True Negatives refer to pixels correctly classified as not part of the target class, and False Positives indicate pixels that were incorrectly predicted as belonging to it:

$$specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}.$$

Result values of metrics are displayed in Table 1.

The results outlined in the table highlight the clear advantage of ensemble strategies over both the baseline and individual U-Net models. The most effective configuration – the ensemble with optimized weights – achieved the highest accuracy, reaching 0.8756. This marks an approximate 4.7% improvement over the baseline model, which recorded an accuracy of 0.8360. These findings are consistent with trends observed in the training and validation accuracy curves, where ensemble methods consistently surpassed the performance of individual networks, particularly in the later stages of training.

In terms of segmentation quality, the mean Intersection over Union (Mean IoU) also shows a notable boost. The optimized weight ensemble attained a Mean IoU of 0.6504, outperforming the baseline's 0.6145 by a margin of 3.6%. Beyond overall accuracy and IoU, additional evaluation metrics such as precision, sensitivity, and specificity provide deeper insight into the segmentation behavior of each model. Precision, which quantifies the correctness of positive pixel classifications, varied across configurations. The highest precision was observed in U-Net-2 at 0.4050, while the optimized ensemble closely followed with a precision of 0.3980, demonstrating its ability to maintain segmentation accuracy while effectively limiting false positives.

Specificity, which measures how accurately negative pixels are classified, remained consistently high across all configurations. The optimized ensemble achieved the highest specificity at 0.9953, indicating its strong capacity to reduce false positive classifications without compromising performance. This reliability in identifying background or non-target areas is especially valuable in high-precision segmentation tasks.

While ensemble approaches deliver substantial gains in accuracy and segmentation quality, these improvements come with increased computational demands. The baseline U-Net offered the fastest inference speed, processing an image in 0.1604 seconds. In contrast, the optimized ensemble required 0.4135 seconds per image – roughly 2.6 times longer.

Among the individual models, U-Net-4 exhibited the lowest execution time at 0.1512 seconds, making it a compelling option for applications that prioritize speed over marginal gains in accuracy. Nevertheless, the superior accuracy and segmentation fidelity achieved by ensemble configurations justify their use in domains where precision is paramount and computational cost is secondary.

Overall, the findings clearly demonstrate that ensemble methods offer meaningful improvements in both accuracy and IoU compared to standalone models. The ensemble with optimized weights emerges as the most effective approach, achieving the best overall balance: high accuracy (0.8756), strong IoU (0.6504), and leading specificity (0.9953).

Table 1 – Networks metrics

	Baseline	Unet-1	Unet-2	Unet-3	Unet-4	Unet-5	Ensemble (max)	Ensemble (avg)	Ensemble optimized weights	Ensemble proportional weights	Ensemble exponential
accuracy	0.8360	0.8265	0.8462	0.8365	0.8223	0.5754	0.8458	0.8672	0.8756	0.8620	0.8622
mean IoU	0.6145	0.6103	0.6284	0.6324	0.5898	0.3463	0.5931	0.6346	0.6504	0.6380	0.6410
precision	0.3783	0.3368	0.4050	0.3820	0.3297	0.1650	0.2922	0.2991	0.3980	0.3005	0.3003
sensitivity	0.3094	0.2893	0.3269	0.3047	0.3056	0.1623	0.2409	0.2480	0.2576	0.2485	0.2485
specificity	0.9944	0.9940	0.9947	0.9944	0.9940	0.9852	0.9946	0.9952	0.9953	0.9952	0.9952
time per image, s	0.1604	0.1524	0.1618	0.1570	0.1512	0.1627	0.3611	0.3710	0.4135	0.3947	0.4584

6 DISCUSSION

In this study, we explored the effect of ensemble methods on convolutional neural networks applied to semantic segmentation tasks. The proposed method integrates multiple U-Net networks and aggregates their outputs using an optimized weighting technique, aiming to enhance segmentation accuracy while keeping computational demands within practical limits.

Our research began with a literature review, examining established techniques for improving semantic segmentation, particularly those focused on single-model refinement and ensemble learning. While individual model optimizations can yield modest improvements, the

DOI 10.15588/1607-3274-2025-3-7

reviewed studies consistently highlight ensemble learning as a more effective approach for increasing model robustness and generalization. However, these benefits are often accompanied by a notable rise in computational cost.

To assess the proposed approach, we implemented and evaluated several U-Net models, each combined through different ensembling strategies – namely max voting, simple averaging, optimized weighting, proportional weighting, and exponential weighting. Across all configurations, the ensemble models outperformed standalone networks in terms of both accuracy and mean Intersection over Union (IoU). The ensemble using optimized weights delivered the best results, achieving an accuracy of 87.56% and a mean IoU of 0.6504, outperforming the top-performing individual U-Net by roughly 3%. These gains, however, came at the cost of increased inference time, a factor that becomes particularly relevant in time-sensitive or real-time applications, even though it stays within acceptable limits.

Our findings further underscore that ensemble performance is most effective when constituent models produce diverse yet complementary predictions. Variability among the individual U-Net models was evident, with some excelling in precision and others in sensitivity. Through ensembling, these strengths were combined, effectively balancing the trade-offs inherent in each individual model and producing a more stable and consistent segmentation output.

Despite these advantages, the study also sheds light on the limitations of ensemble learning. Running multiple networks in sequence substantially increases computational requirements, especially on systems without hardware acceleration. This poses challenges for deployment in scenarios where real-time inference is critical. Moreover, ensemble models did not show significant gains in specificity, suggesting that some segmentation errors are systemic and may persist regardless of the aggregation strategy.

Overall, the results demonstrate that ensemble techniques offer meaningful improvements in semantic segmentation performance and model generalization across diverse classes. Yet, the balance between performance gains and computational efficiency remains a key consideration. Future research should focus on optimizing ensemble methodologies to reduce overhead, potentially through model distillation, parallel inference strategies, or lightweight ensembling techniques, all while preserving segmentation quality.

CONCLUSIONS

The paper analyses the effectiveness of ensemble methods for convolutional neural networks in solving the semantic segmentation task.

The scientific novelty of the presented work lies in the development of a weighted ensemble approach based on five U-Net models sharing the same architecture, but each trained using distinct augmentation strategies and weight initialization techniques. This design improves segmentation accuracy and consistency without altering the network structure itself. By applying an optimized weighting mechanism during ensemble prediction, the proposed method achieves notable improvements in both accuracy and mean IoU when compared to individual models, while maintaining a high level of specificity. These results demonstrate that ensembling is a viable and efficient strategy for enhancing semantic segmentation performance using existing architectures.

The practical significance of the research is underscored by the fact that the ensemble models were trained and evaluated on a real-world dataset, validating their relevance for practical deployment. The findings support the recommendation of this ensemble strategy for applications that demand high segmentation accuracy, such as autonomous driving systems. However, the increased computational overhead introduced by ensemble methods should be carefully considered, particularly in scenarios requiring real-time processing.

Prospects for further research include refining the computational efficiency of the ensemble to reduce inference time while preserving segmentation quality. Future investigations may also explore the effectiveness of the proposed ensembling strategy when applied to alternative network architectures and larger, more diverse datasets, thereby broadening its applicability across different domains and use cases.

ACKNOWLEDGEMENTS

We thank the management of Kharkiv National University of Radioelectronics for the opportunity to conduct scientific research.

REFERENCES

1. Cordts M., Omran M., Ramos S. et al. The Cityscapes Dataset for Semantic Urban Scene Understanding, *In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 27–30 June 2016 : proceedings*. Las Vegas, IEEE, 2016, pp. 3213–3223. DOI: 10.1109/cvpr.2016.350.
2. A Comprehensive Survey on Ensemble Methods / Suyash Kumar, Prabhjot Kaur, Anjana Gosain, *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*. Mumbai, India, 7–9 April 2022. [S. l.], 2022. DOI: 10.1109/i2ct54291.2022.9825269.
3. Hao S., Zhou Y., Guo Y. A brief survey on semantic segmentation with deep learning, *Neurocomputing,* 2020, Vol. 406, pp. 302–321. DOI: 10.1016/j.neucom.2019.11.118.
4. Pare S. [et al.]Image Segmentation Using Multilevel Thresholding: A Research Review, *Iranian Journal of Science and Technology, Transactions of Electrical Engineering,* 2019, Vol. 44, No. 1, pp. 1–29. DOI: 10.1007/s40998-019-00251-1.
5. Tang Jun A color image segmentation algorithm based on region growing, *2010 2nd International Conference on Computer Engineering and Technology*. Chengdu, China, 16–18 April 2010. [S. l.], 2010. DOI: 10.1109/iccet.2010.5486012.

DOI 10.15588/1607-3274-2025-3-7

6. Jeyalaksshmi S., Prasanna S. A Review of Edge Detection Techniques for Image Segmentation, *International Journal of Data Mining Techniques and Applications,* 2016, Vol. 5, No. 2, pp. 140–142. DOI: 10.20894/ijdmta.102.005.002.008.
7. Liu Xiangbin et al. A Review of Deep-Learning-Based Medical Image Segmentation Methods, *Sustainability,* 2021, Vol. 13, No. 3, P. 1224. DOI: 10.3390/su13031224.
8. Guo Zhe et al. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging, *IEEE Transactions on Radiation and Plasma Medical Sciences,* 2019, Vol. 3, No. 2, pp. 162–169. DOI: 10.1109/trpms.2018.2890359.
9. Mzoughi O., Mzoughi Olfa, Yahiaoui Itheri Deep learning-based segmentation for disease identification, *Ecological Informatics,* 2023, pp. 102000. DOI: 10.1016/j.ecoinf.2023.102000.
10. Sultana Farhana, Sufian Abu, Dutta Paramartha Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey, *Knowledge-Based Systems,* 2020, Vol. 201–202, P. 106062. DOI: 10.1016/j.knosys.2020.106062.
11. Shelhamer E. et al. Fully Convolutional Networks for Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2017, Vol. 39, No. 4, pp. 640–651. DOI: 10.1109/tpami.2016.2572683.
12. Li Xiaojin et al. Image Segmentation Based on Improved Unet, *Journal of Physics: Conference Series,* 2021, Vol. 1815, No. 1, P. 012018. DOI: 10.1088/1742-6596/1815/1/012018.
13. Mobarakol Islam et al. Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham, 2020, pp. 262–272. DOI: 10.1007/978-3-030-46640-4_25.
14. Karaali A. et al. DR-VNet: Retinal Vessel Segmentation via Dense Residual UNet, *Pattern Recognition and Artificial Intelligence*. Cham, 2022, pp. 198–210. DOI: 10.1007/978-3-031-09037-0_17.
15. Ngo G. et al. Evolutionary bagging for ensemble learning, *Neurocomputing,* 2022. DOI: 10.1016/j.neucom.2022.08.055.
16. Drucker Harris et al. Boosting and Other Ensemble Methods, *Neural Computation,* 1994, Vol. 6, No. 6, pp. 1289–1301. DOI: 10.1162/neco.1994.6.6.1289.
17. Verma Anurag Kumar, Pal Saurabh Prediction of Skin Disease with Three Different Feature Selection Techniques Using Stacking Ensemble Method, *Applied Biochemistry and Biotechnology,* 2019, Vol. 191, No. 2, pp. 637–656. DOI: 10.1007/s12010-019-03222-8.
18. Bodyanskiy Y. V., Lipianina-Honcharenko K. V., Sachenko A. O. Ensemble of Adaptive Predictors for Multivariate Nonstationary Sequences and its Online Learning, *Radio Electronics, Computer Science, Control,* 2024, No. 4, P. 91. DOI: 10.15588/1607-3274-2023-4-9.
19. Ahmad Numan, Behram Wali, Khattak Asad J. Heterogeneous ensemble learning for enhanced crash forecasts – A frequentist and machine learning based stacking framework, *Journal of Safety Research,* 2022, DOI: 10.1016/j.jsr.2022.12.005.
20. Lo Hung-Yi, Wang Ju-Chiang, Wang Hsin-Min Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval, *2010 IEEE International Conference on Multimedia and Expo (ICME), Singapore.* Singapore, 19–23 July 2010. [S. l.], 2010. DOI: 10.1109/icme.2010.5583009.
21. Bian Shun, Wang Wenjia Investigation on Diversity in Homogeneous and Heterogeneous Ensembles, *The 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada, 16–21 July 2006.* [S. l.], 2006. DOI: 10.1109/ijcnn.2006.247268.
22. Kamnitsas Konstantinos et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, *Medical Image Analysis,* 2017, Vol. 36, pp. 61–78. DOI: 10.1016/j.media.2016.10.004.
23. Gautam Nandita et al. An Ensemble of UNet Frameworks for Lung Nodule Segmentation, *Current Problems in Applied Mathematics and Computer Science and Systems.* Cham, 2023, pp. 450–461. DOI: 10.1007/978-3-031-34127-4_44.
24. Smelyakov K. et al. Adaptive Image Enhancement Model for the Robot Vision System, *ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference,* 2023, Vol. 3, pp. 246–251.
25. Abdullahi A. M. et al. A comparison of weight initializers in deep learning, *2023 IEEE 21st Student Conference on Research and Development (SCOReD); 2023 Dec 13–14.* Kuala Lumpur, Malaysia, [S.l.], 2023. DOI: 10.1109/scored60679.2023.10563215.
26. Lee H. et al. Improved weight initialization for deep and narrow feedforward neural network, *Neural Networks,* 2024, P. 106362. DOI: 10.1016/j.neunet.2024.106362.
27. LeCun Y. et al. Efficient BackProp, *In: Lecture Notes in Computer Science*. Berlin, Heidelberg, 1998, pp. 9–50. DOI: 10.1007/3-540-49430-8_2.
28. Kramer O. Iterated local search with Powell's method: a memetic algorithm for continuous global optimization, *Memetic Computing*, 2010, Vol. 2, No. 1, pp. 69–83. DOI: 10.1007/s12293-010-0032-9.

Received 27.03.2025.
Accepted 21.05.2025.

DOI 10.15588/1607-3274-2025-3-7

УДК 004.93

СЕГМЕНТАЦІЯ МІСЬКИХ СЦЕН ЗА ДОПОМОГОЮ ОДНОРІДНОГО АНСАМБЛЮ U-NET: ДОСЛІДЖЕННЯ НА ДАТАСЕТІ CITYSCAPES

Гмиря І. О. – аспірант кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна.

Кравець Н. С. – канд. техн. наук, доцент, доцент кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. Семантична сегментація є ключовим завданням комп'ютерного зору, зокрема в таких сферах, як автономне водіння та аналіз міських сцен. Створення нових архітектур є складним і трудомістким процесом, однак поліпшення точності за допомогою ансамблевих методів на основі вже існуючих моделей показує високий потенціал. У даній роботі досліджується застосування ансамблевого навчання як стратегії підвищення точності сегментації без модифікації архітектури U-Net.

Мета роботи – розробка та оцінка однорідного ансамблю моделей U-Net, навчання яких здійснюється із використанням різних методів ініціалізації ваг та збільшення обсягу даних, а також вивчення ефективності різних стратегій агрегації ансамблю для підвищення якості сегментації на складних урбаністичних даних.

Метод. Запропоновано ансамбль з п'яти моделей U-Net з однаковою архітектурою, але різною ініціалізацією ваг та підходами до збільшення обсягу даних, що забезпечує різноманітність прогнозів. Розглянуто кілька стратегій об'єднання вихідних даних: середнє по softmax, максимум, пропорційне зважування, експоненціальне зважування та оптимізоване вагове голосування. Оцінювання виконано на датасеті Cityscapes із використанням стандартних метрик сегментації.

Результати. Результати експериментів показують, що ансамблеві моделі стабільно перевищують точність окремих моделей U-Net та базової моделі за такими показниками, як точність, середній IoU та специфічність. Ансамбль із оптимізованим зважуванням досяг найвищої точності (87,56%) та середнього IoU (0,6504), перевищивши найкращу окрему модель приблизно на 3%. Водночас покращення якості супроводжується збільшенням часу виведення результату, що вказує на необхідність компромісу між точністю та обчислювальною ефективністю.

Висновки. Запропонований підхід на основі ансамблю ефективно покращує результати сегментації без зміни архітектури моделі. Незважаючи на збільшення обчислювальних витрат, метод є придатним для задач, де критично важлива точність сегментації. Подальші дослідження будуть зосереджені на зменшенні часу виведення результату та поширенні ансамблевого підходу на інші архітектури та датасети.

КЛЮЧОВІ СЛОВА: згорткова нейронна мережа, семантична сегментація, U-Net, ансамблеве навчання, методи збільшення обсягу даних, ініціалізація ваг, Cityscapes, урбаністичні сцени.

ЛІТЕРАТУРА

1. The Cityscapes Dataset for Semantic Urban Scene Understanding / [M. Cordts, M. Omran, S. Ramos, et al.] – In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 27–30 June 2016 : proceedings. – Las Vegas : IEEE, 2016. – P. 3213–3223. DOI: 10.1109/cvpr.2016.350.
2. A Comprehensive Survey on Ensemble Methods / Suyash Kumar, Prabhjot Kaur, Anjana Gosain // 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India, 7–9 April 2022. – [S. l.], 2022. DOI: 10.1109/i2ct54291.2022.9825269.
3. Hao S. A brief survey on semantic segmentation with deep learning / S. Hao, Y. Zhou, Y. Guo // Neurocomputing. – 2020. – Vol. 406. – P. 302–321. DOI: 10.1016/j.neucom.2019.11.118.
4. Image Segmentation Using Multilevel Thresholding: A Research Review / S. Pare [et al.] // Iranian Journal of Science and Technology, Transactions of Electrical Engineering. – 2019. – Vol. 44, No. 1. – P. 1–29. DOI: 10.1007/s40998-019-00251-1.
5. A color image segmentation algorithm based on region growing / Jun Tang // 2010 2nd International Conference on Computer Engineering and Technology, Chengdu, China, 16–18 April 2010. – [S. l.], 2010. DOI: 10.1109/iccet.2010.5486012.
6. Jeyalaksshmi S. A Review of Edge Detection Techniques for Image Segmentation / S. Jeyalaksshmi, S. Prasanna // International Journal of Data Mining Techniques and Applications. – 2016. – Vol. 5, No. 2. – P. 140–142. DOI: 10.20894/ijdmta.102.005.002.008.
7. A Review of Deep-Learning-Based Medical Image Segmentation Methods / Xiangbin Liu [et al.] // Sustainability. – 2021. – Vol. 13, No. 3. – P. 1224. DOI: 10.3390/su13031224.
8. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging / Zhe Guo [et al.] // IEEE Transactions on Radiation and Plasma Medical Sciences. – 2019. – Vol. 3, No. 2. – P. 162–169. DOI: 10.1109/trpms.2018.2890359.
9. Mzoughi O. Deep learning-based segmentation for disease identification / Mzoughi O., Mzoughi Olfa, Yahiaoui Itheri // Ecological Informatics. – 2023. – P. 102000. DOI: 10.1016/j.ecoinf.2023.102000.
10. Sultana Farhana Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey / Farhana Sultana, Abu Sufian, Paramartha Dutta // Knowledge-Based Systems. – 2020. – Vol. 201–202. – P. 106062. DOI: 10.1016/j.knosys.2020.106062.
11. Fully Convolutional Networks for Semantic Segmentation / E. Shelhamer et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. –

DOI 10.15588/1607-3274-2025-3-7

Vol. 39, No. 4. – P. 640–651. DOI: 10.1109/tpami.2016.2572683.

12. Image Segmentation Based on Improved Unet / Xiaojin Li et al. // Journal of Physics: Conference Series. – 2021. – Vol. 1815, No. 1. – P. 012018. DOI: 10.1088/1742-6596/1815/1/012018.
13. Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet / Mobarakol Islam et al. // Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. – Cham, 2020. – P. 262–272. DOI: 10.1007/978-3-030-46640-4_25.
14. DR-VNet: Retinal Vessel Segmentation via Dense Residual UNet / A. Karaali et al. // Pattern Recognition and Artificial Intelligence. – Cham, 2022. – P. 198–210. DOI: 10.1007/978-3-031-09037-0_17.
15. Evolutionary bagging for ensemble learning / G. Ngo et al. // Neurocomputing. – 2022. DOI: 10.1016/j.neucom.2022.08.055.
16. Boosting and Other Ensemble Methods / Harris Drucker et al. // Neural Computation. – 1994. – Vol. 6, No. 6. – P. 1289–1301. DOI: 10.1162/neco.1994.6.6.1289.
17. Verma Anurag Kumar Prediction of Skin Disease with Three Different Feature Selection Techniques Using Stacking Ensemble Method / Anurag Kumar Verma, Saurabh Pal // Applied Biochemistry and Biotechnology. – 2019. – Vol. 191, No. 2. – P. 637–656. DOI: 10.1007/s12010-019-03222-8.
18. Bodyanskiy Y. V. Ensemble of Adaptive Predictors for Multivariate Nonstationary Sequences and its Online Learning / Y. V. Bodyanskiy, K. V. Lipianina-Honcharenko, A. O. Sachenko // Radio Electronics, Computer Science, Control. – 2024. – No. 4. – P. 91. DOI: 10.15588/1607-3274-2023-4-9.
19. Ahmad Numan Heterogeneous ensemble learning for enhanced crash forecasts – A frequentist and machine learning based stacking framework / Numan Ahmad, Behram Wali, Asad J. Khattak // Journal of Safety Research. – 2022. DOI: 10.1016/j.jsr.2022.12.005.
20. Lo Hung-Yi Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval / Hung-Yi Lo, Ju-Chiang Wang, Hsin-Min Wang // 2010 IEEE International Conference on Multimedia and Expo (ICME), Singapore, Singapore, 19–23 July 2010. – [S. l.], 2010. DOI: 10.1109/icme.2010.5583009.
21. Bian Shun Investigation on Diversity in Homogeneous and Heterogeneous Ensembles / Shun Bian, Wenjia Wang // The 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada, 16–21 July 2006. – [S. l.], 2006. DOI: 10.1109/ijcnn.2006.247268.
22. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation / Konstantinos Kamnitsas et al. // Medical Image Analysis. – 2017. – Vol. 36. – P. 61–78. DOI: 10.1016/j.media.2016.10.004.
23. An Ensemble of UNet Frameworks for Lung Nodule Segmentation / Nandita Gautam et al. // Current Problems in Applied Mathematics and Computer Science and Systems. – Cham, 2023. – P. 450–461. DOI: 10.1007/978-3-031-34127-4_44.
24. Adaptive Image Enhancement Model for the Robot Vision System / K. Smelyakov et al. // ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference. – 2023. – Vol. 3. – P. 246–251.
25. A comparison of weight initializers in deep learning / A. M. Abdullahi et al. // 2023 IEEE 21st Student Conference on Research and Development (SCOReD); 2023 Dec 13–14; Kuala Lumpur, Malaysia. – [S.l.]: 2023. DOI: 10.1109/scored60679.2023.10563215.
26. Improved weight initialization for deep and narrow feedforward neural network / H. Lee et al. // Neural Networks. – 2024. – P. 106362. DOI: 10.1016/j.neunet.2024.106362.
27. Efficient BackProp / Y. LeCun et al. – In: Lecture Notes in Computer Science. – Berlin, Heidelberg: 1998. – P. 9–50. DOI: 10.1007/3-540-49430-8_2.
28. Kramer O. Iterated local search with Powell's method: a memetic algorithm for continuous global optimization / O. Kramer // Memetic Computing. – 2010. – Vol. 2, No. 1. – P. 69–83. DOI: 10.1007/s12293-010-0032-9.

DOI 10.15588/1607-3274-2025-3-7