DOI 10.34185/1562-9945-2-163-2026-01
UDC 004:83

A.R. Kovtunenko, S.V. Mashtalir

EXPERIMENTAL STUDY OF DISTANCE MAP DECODER ARCHITECTURAL CONFIGURATIONS FOR INSTANCE SEGMENTATION BY TEXT QUERY

Anotation. This article presents an experimental study of the architectural configurations of distance map decoder in the InstanceCLIPSeg model for instance segmentation by text query. We investigate the influence of various mechanisms for restoring spatial resolution (bilinear interpolation, PixelShuffle), the use of coordinate convolutions (CoordConv), and multi-level feature fusion strategies. Based on the evaluation of nine configurations on the LVIS and PhraseCut datasets, it was found that a hybrid architecture with PixelShuffle and single-stage feature fusion from transformer layers achieves the best results (mean Dice 0.2374), outperforming baseline approaches. The redundancy of coordinate channels in the presence of multi-level fusion was revealed.
Keywords: instance segmentation, distance decoder, PixelShuffle, coordinate convolution, feature fusion, CLIP, open-vocabulary segmentation, InstanceCLIPSeg.

Problem statement. Modern computer vision tasks increasingly require a transition from a closed set of classes to an open-vocabulary paradigm. Instance segmentation based on text queries requires neural network architectures to be able to process arbitrary descriptions in natural language while maintaining high accuracy of spatial predictions. Our InstanceCLIPSeg [1] model solves this problem using a bottom-up approach. A key feature of the method is the simultaneous prediction of a heatmap of object centers and a four-channel map of distances from each pixel of an object to the boundaries of its bounding box. The results are then post-processed. In this scheme, the distance decoder is a critically important component. It needs to restore the spatial resolution from 22×22 (encoder output) to 352×352 pixels, preserving the positional information and structure of the objects. The goal of this study is to compare architectural solutions for this task and find the optimal configuration.

Analysis of recent studies and publications. Existing approaches to segmentation can be divided into two-stage (top-down) and one-stage (bottom-up) methods. Two-stage methods, such as Mask R-CNN, first detect objects and then segment them, which can be computationally expensive. Single-stage bottom-up methods offer an alternative by first identifying individual features (centers, corners), which are then combined.

Among multimodal architectures, the CLIP [2] model forms the foundation, creating a shared space for visual and textual features. CLIPSeg [3] extends this approach to pixel-level segmentation, but is limited to semantic tasks without separating instances. InstanceCLIPSeg

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

integrates the concepts of CLIPSeg, Panoptic DeepLab [4], and PRN [5] to solve the task of instance segmentation.

We analyze the mechanisms of resolution enhancement (upsampling). Transposed convolutions often create "checkerboard" artifacts [6], which is critical for distance maps that require smoothness. Bilinear interpolation does not have this drawback, but it is non-learnable. The PixelShuffle [7] method redistributes channels into spatial dimensions, but also requires special initialization (ICNR) [8] to avoid artifacts. Additionally, CoordConv [9] is often used to account for the position of objects in space.

In the context of feature fusion, there are various strategies: hierarchical skip connections (U-Net [10], FPN [11]) and one-time fusion (SegFormer [12]). Since the authors of SegFormer have demonstrated the effectiveness of simple MLP decoders for transformer encoders, it is necessary to verify the applicability of this approach to our task. Since this is crucial for our architecture, additional analysis of the impact of positional encoding and multi-level fusion methods for ViT decoders is required.

Research objective. The objective is to experimentally investigate the architectural configurations of the distance map decoder in the InstanceCLIPSeg model to determine the optimal combination of resolution recovery approaches, the impact of coordinate convolutions, and feature fusion strategies.

Presentation of the main research material. Decoder architectural configurations. The decoder restores image resolution through four stages of 2x upscaling (from 22×22 to 352×352). Each stage consists of an upsampling block and a refinement block (3×3 convolution + BatchNorm + ReLU). For the study, we implemented two fundamentally different feature fusion strategies:

1. hybrid fusion (similar to SegFormer): features from transformer layers L3, L7, and L9 are integrated at the decoder input. Each source is projected into a common dimension, concatenated, and processed by the fusion block;

2. hierarchical fusion (similar to U-Net): skip connections are introduced in stages. Features from L7 are added in the first stage (22×22 → 44×44), and from L3 in the second. Since ViT retains a resolution of 22×22, skip features are scaled by interpolation.

Figures 1-4 show the diagrams of the architectures under study.

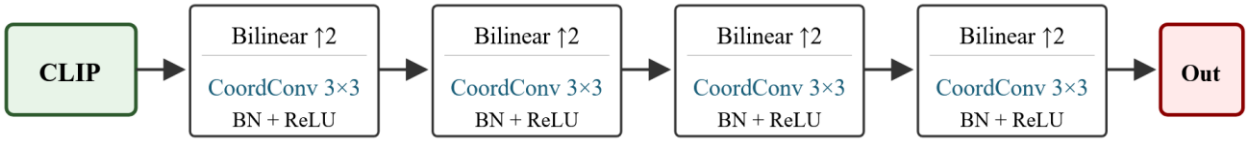


Figure 1 – Architecture configuration: basic decoder

Figure 2 – Architecture configuration: PixelShuffle

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

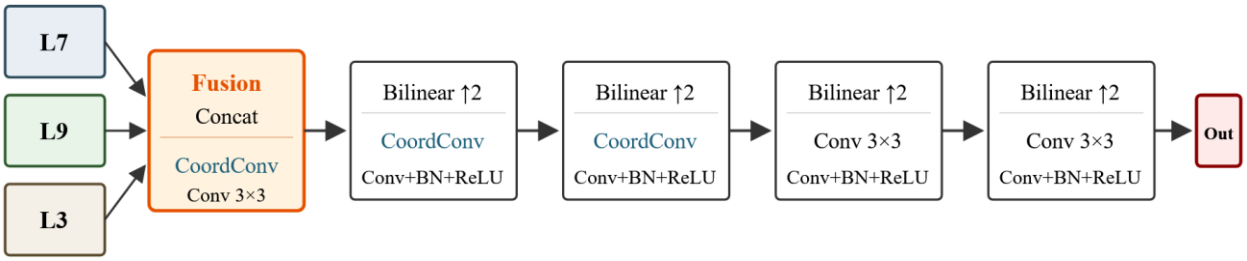


Figure 3 – Architecture configuration: hybrid with bilinear interpolation

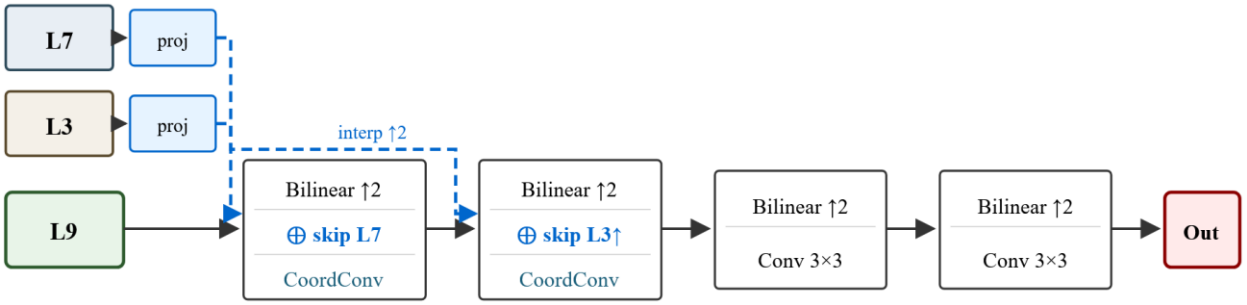


Figure 4 – Architecture configuration: hierarchical with bilinear interpolation

A total of nine configurations (C1–C9) were implemented and trained, differing in the upsampling method (bilinear interpolation or PixelShuffle + ICNR), the presence of CoordConv, and the type of feature fusion. The configurations are described in Table 1.

Table 1

Experimental decoder configurations

Config	Resolution enhancement	CoordConv	Feature fusion
C1	Bilinear	All stages	No
C2	PixelShuffle	No	No
C3	PixelShuffle+ICNR	Stages 1-2	No
C4	Bilinear	Merge + stages 1-2	Hybrid
C5	PixelShuffle+ICNR	Merge block	Hybrid
C6	PixelShuffle+ICNR+CoordConv	Refinement	Hybrid
C7	Bilinear	Stages 1-2	Hierarchical
C8	PixelShuffle+ICNR	Refinement	Hierarchical
C9	PixelShuffle+ICNR+CoordConv	Refinement	Hierarchical

Experimental conditions. Training was performed on the LVIS [13] dataset, and testing was performed on PhraseCut [14]. This approach allows us to evaluate the model's ability to generalize knowledge to new text descriptions. All models were trained for 3 epochs with a frozen CLIP encoder. Weighted L1-loss was used as the loss function for the distance decoder. Mean Dice (mDice) was chosen as the main quality metric.

The comparative results of the experiments are presented in Table 2.

Table 2

Comparison of decoder configurations

Configuration	mDice
C1: Bilinear	0.2170
C2: PixelShuffle	0.1867
C3: PixelShuffle+ICNR+CoordConv	0.2232
C4: Hybrid+Bilinear	0.2198
C5: Hybrid+PixelShuffle	**0.2374**
C6: Hybrid+PixelShuffle+CoordConv	0.2058
C7: Hierarchical+Bilinear	0.2127
C8: Hierarchical+PixelShuffle	0.2257
C9: Hierarchical+PixelShuffle+CoordConv	0.2022

Analysis of resolution enhancement mechanisms. Experiments have shown that simple bilinear interpolation (C1, mDice=0.2170) outperforms standard PixelShuffle (C2, mDice=0.1867). This is because, without special initialization, PixelShuffle creates block artifacts that disrupt the smoothness of distance fields. However, the use of ICNR initialization in combination with CoordConv (C3) allows PixelShuffle to outperform the bilinear method (mDice=0.2232). In complex configurations with feature merging (C5 vs C4, C8 vs C7), PixelShuffle consistently shows an advantage (+6-8%), confirming the effectiveness of the learning mechanism when sufficient context is available.

A visualization of the best configuration (C5) is shown in Figure 5. The model correctly predicts the centers and generates distance maps for four directions.

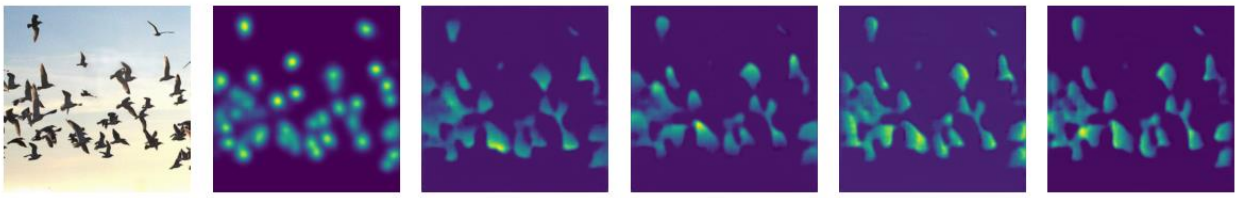


Figure 5 – Result of network output for hybrid decoder with PixelShuffle

In configurations without feature fusion, CoordConv provides a significant gain (+19.5% when transitioning from C2 to C3). However, when features from the transformer are added (configurations C6 and C9), CoordConv degrades the metrics (a 10-13% drop relative to counterparts without it). This indicates that multi-level fusion already provides sufficient positional context through ViT features, making explicit coordinate channels redundant and complicating training. This is consistent with [15, 16], indicating that superfluous processing can be detrimental to model accuracy.

The process of restoring bounding boxes based on predicted maps is illustrated in Figure 6.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

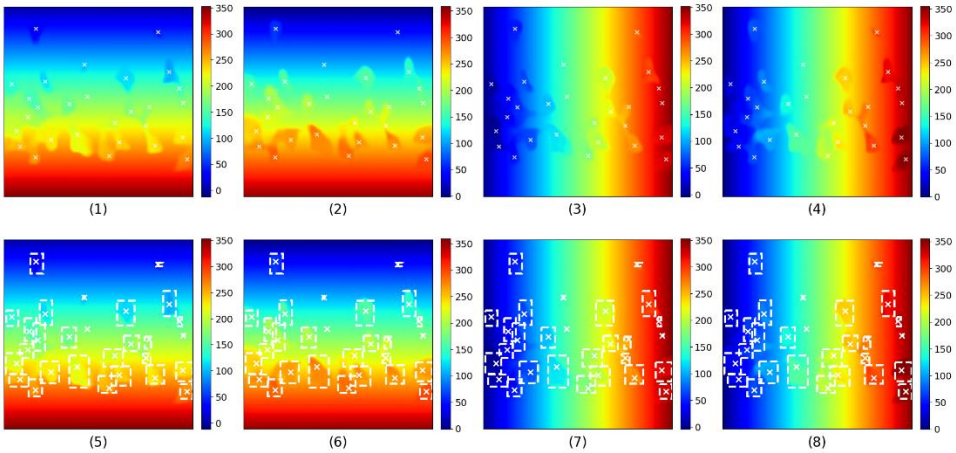
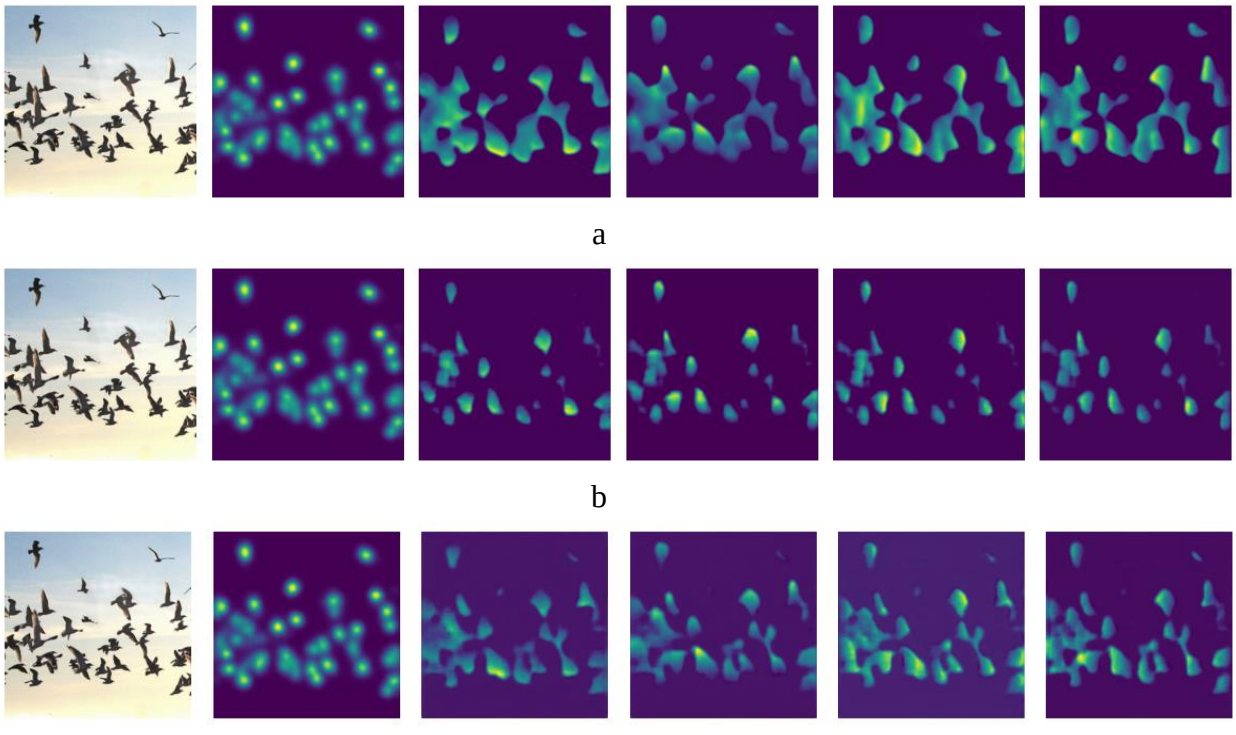


Figure 6 – Finding objects by their predicted distances to their edges (1-4) and calculating their bounding boxes (5-8) for a hybrid decoder with PixelShuffle

Feature fusion strategies. The hybrid approach (C5) showed the best result (mDice=0.2374), outperforming the hierarchical approach (C8, mDice=0.2257) by 5.2%. The advantage of the hybrid scheme is explained by the specifics of Vision Transformer – the encoder outputs features of fixed resolution (22×22) on all layers. Hierarchical fusion requires additional interpolation to match the sizes in the pyramid, which introduces noise. Hybrid fusion performs integration at the encoder's "native" resolution, which is consistent with the conclusions of the SegFormer authors.

The qualitative difference in the predicted values is shown in Figure 7. The basic configuration (C1) produces objects that are difficult to distinguish, C2 detects significantly fewer objects and suffers from artifacts, while C5 provides smooth and clear object boundaries.



a

b

c

Figure 7 – Comparison of distance maps: (a) C1; (b) C2; (c) C5

Figures 8 and 9 show the result of the final post-processing. If post-processing relies strictly on detected centers, the result may prove inferior to using clustering based on the obtained distance maps.

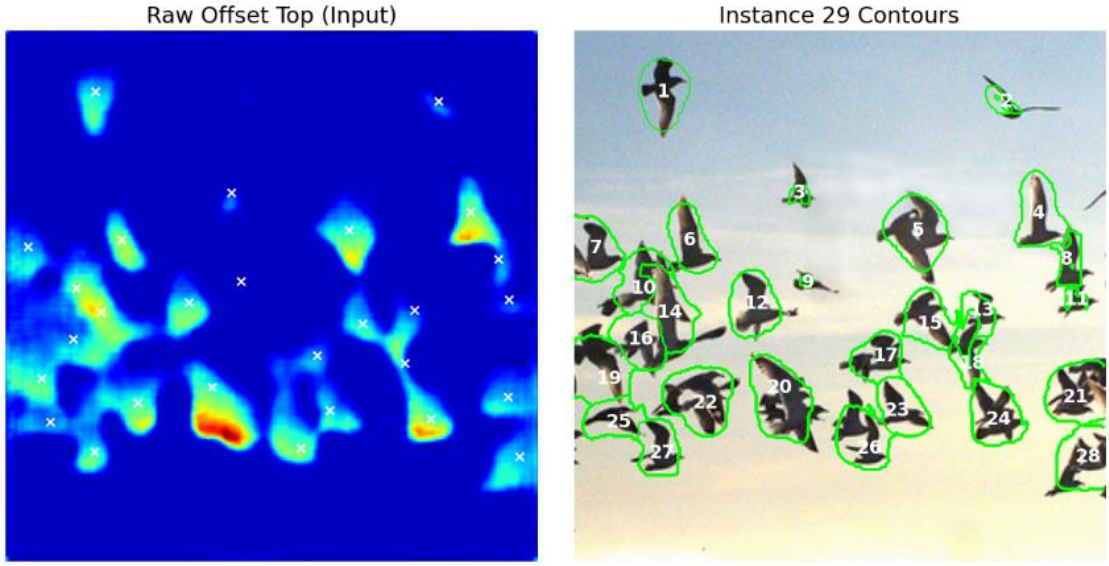


Figure 8 – Post-processing result using centers for the hybrid decoder with PixelShuffle

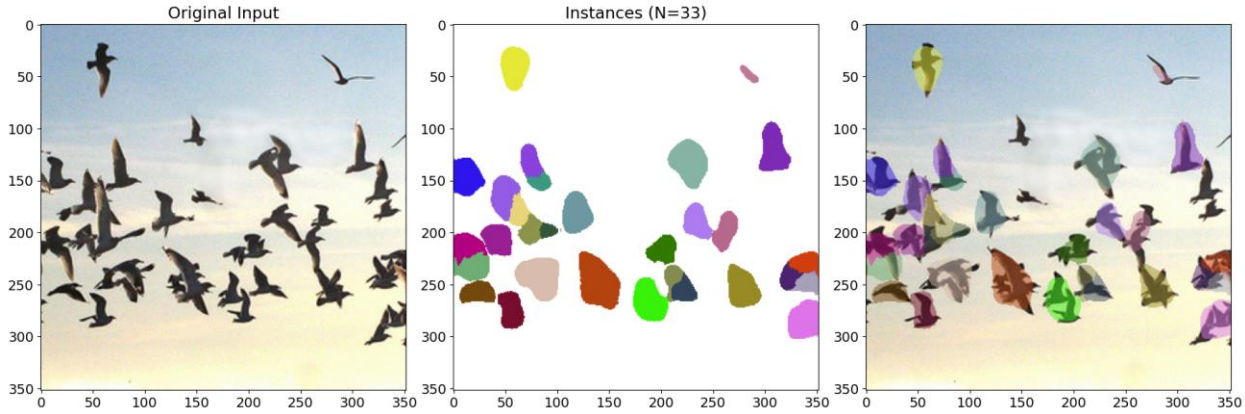


Figure 9 – Post-processing result without using centers
for the hybrid decoder with PixelShuffle

Conclusions. This paper investigates distance decoder architectures for the InstanceCLIPSeg model.

1. It was found that the optimal configuration is a hybrid architecture with PixelShuffle, ICNR initialization, and single feature merging (L3, L7, L9) without using CoordConv in refinement blocks. It outperforms the baseline method by 9.4% on the mean Dice metric.

2. CoordConv is only effective in simple architectures. In the presence of rich context from the transformer encoder, it becomes redundant and degrades segmentation quality.

3. For ViT encoders, hybrid one-time feature fusion is preferable to classic hierarchical skip connections (U-Net), as it avoids the artifacts of interpolating intermediate feature maps.

Future work involves extending the training (more than 3 epochs) to investigate the impact of training duration on the integration of positional information.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ЛІТЕРАТУРА

1. Kovtunenko A.R., Mashtalir S.V. Improved segmentation model to identify object instances based on textual prompts. Herald of advanced information technology. 2025. Т. 8, № 1. С. 54–66. URL: https://doi.org/10.15276/hait.08.2025.4 (дата звернення: 10.02.2026).

2. Learning transferable visual models from natural language supervision / A. Radford та ін. Proceedings of the 38th International Conference on Machine Learning (ICML), 18–24 лип. 2021 р. PMLR, 2021. Т. 139. С. 8748–8763.
URL: http://proceedings.mlr.press/v139/radford21a.html (дата звернення: 10.02.2026)

3. Luddecke T., Ecker A. Image segmentation using text and image prompts. 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), м. New Orleans, LA, USA, 18–24 черв. 2022 р. 2022. URL: https://doi.org/10.1109/cvpr52688.2022.00695 (дата звернення: 10.02.2026).

4. Panoptic-DeepLab: a simple, strong, and fast baseline for bottom-up panoptic segmentation / B. Cheng та ін. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), м. Seattle, WA, USA, 13–19 черв. 2020 р. 2020.
URL: https://doi.org/10.1109/cvpr42600.2020.01249 (дата звернення: 10.02.2026).

5. PRN: panoptic refinement network / B. Sun та ін. 2023 IEEE/CVF winter conference on applications of computer vision (WACV), м. Waikoloa, HI, USA, 2–7 січ. 2023 р. 2023. URL: https://doi.org/10.1109/wacv56688.2023.00395 (дата звернення: 10.02.2026).

6. Odena A., Dumoulin V., Olah C. Deconvolution and checkerboard artifacts. Distill. 2016. Т. 1, № 10. URL: https://doi.org/10.23915/distill.00003 (дата звернення: 10.02.2026).

7. Real-Time single image and video super-resolution using an efficient sub-pixel convolutional neural network / W. Shi та ін. 2016 IEEE conference on computer vision and pattern recognition (CVPR), м. Las Vegas, NV, USA, 27–30 черв. 2016 р. 2016.
URL: https://doi.org/10.1109/cvpr.2016.207 (дата звернення: 10.02.2026).

8. Checkerboard artifact free sub-pixel convolution / A. Aitken та ін. arXiv preprint arXiv:1707.02937. 2017. URL: https://doi.org/10.48550/arXiv.1707.02937 (дата звернення: 10.02.2026).

9. An intriguing failing of convolutional neural networks and the CoordConv solution / R. Liu та ін. Advances in Neural Information Processing Systems 31 (NeurIPS), м. Montréal, Canada, 3–8 груд. 2018 р. 2018. Т. 31.
URL: https://proceedings.neurips.cc/paper/2018/hash/60106888f8977b71e1f15db7bc9a88d1-Abstract.html (дата звернення: 10.02.2026).

10. Ronneberger O., Fischer P., Brox T. U-Net: convolutional networks for biomedical image segmentation. Lecture notes in computer science. Cham, 2015. С. 234–241. URL: https://doi.org/10.1007/978-3-319-24574-4_28 (дата звернення: 10.02.2026).

11. Feature pyramid networks for object detection / T.-Y. Lin та ін. 2017 IEEE conference on computer vision and pattern recognition (CVPR), м. Honolulu, HI, 21–26 лип. 2017 р. 2017. URL: https://doi.org/10.1109/cvpr.2017.106 (дата звернення: 10.02.2026).

12. SegFormer: simple and efficient design for semantic segmentation with transformers / E. Xie та ін. Advances in Neural Information Processing Systems 34 (NeurIPS), 6–14 груд. 2021 р. 2021. Т. 34. С. 12077–12090.

URL: https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html (дата звернення: 10.02.2026).
13. Gupta A., Dollar P., Girshick R. LVIS: a dataset for large vocabulary instance segmentation. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), м. Long Beach, CA, USA, 15–20 черв. 2019 р. 2019.
URL: https://doi.org/10.1109/cvpr.2019.00550 (дата звернення: 10.02.2026).
14. PhraseCut: language-based image segmentation in the wild / C. Wu та ін. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), м. Seattle, WA, USA, 13–19 черв. 2020 р. 2020. URL: https://doi.org/10.1109/cvpr42600.2020.01023 (дата звернення: 10.02.2026).
15. Alternate encoder and dual decoder CNN-Transformer networks for medical image segmentation / L. Zhang та ін. Scientific reports. 2025. Т. 15, № 1.
URL: https://doi.org/10.1038/s41598-025-93353-2 (дата звернення: 10.02.2026).
16. Chen J., Liang Z., Lu X. A dual attention and cross layer fusion network with a hybrid CNN and transformer architecture for medical image segmentation. Scientific reports. 2025. Т. 15, №1. URL: https://doi.org/10.1038/s41598-025-19563-w (дата звернення: 10.02.2026).

REFERENCES

1. Mashtalir, S. V., & Kovtunenko, A. R. (2025). Improved segmentation model to identify object instances based on textual prompts. Вісник сучасних інформаційних технологій, 8(1), 54-66.
2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.
3. Lüddecke, T., & Ecker, A. (2022). Image segmentation using text and image prompts. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7086-7096).
4. Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., & Chen, L. C. (2020). Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12475-12485).
5. Sun, B., Kuen, J., Lin, Z., Mordohai, P., & Chen, S. (2023). PRN: panoptic refinement network. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3963-3973).
6. Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. Distill, 1(10), e3.
7. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1874-1883).

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

8. Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., & Shi, W. (2017). Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. arXiv preprint arXiv:1707.02937.
9. Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., & Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the coordconv solution. Advances in neural information processing systems, 31.
10. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.
11. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).
12. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 34, 12077-12090.
13. Gupta, A., Dollar, P., & Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5356-5364).
14. Wu, C., Lin, Z., Cohen, S., Bui, T., & Maji, S. (2020). Phrasecut: Language-based image segmentation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10216-10225).
15. Zhang, L., Guo, X., Sun, H., Wang, W., & Yao, L. (2025). Alternate encoder and dual decoder CNN-Transformer networks for medical image segmentation. Scientific Reports, 15(1), 8883.
16. Chen, J., Liang, Z., & Lu, X. (2025). A dual attention and cross layer fusion network with a hybrid CNN and transformer architecture for medical image segmentation. Scientific Reports, 15(1), 35707.

Received 02.02.2026
Accepted 06.02.2026
Published 31.03.2026

Експериментальне дослідження архітектурних конфігурацій декодера карт відстаней для сегментації екземплярів за текстовим запитом

Сучасні завдання комп'ютерного зору все частіше вимагають переходу від закритого набору класів до парадигми відкритого словника. Сегментація екземплярів на основі текстових запитів вимагає від архітектур нейронних мереж можливості обробляти довільні описи природною мовою, зберігаючи при цьому високу точність просторових прогнозів. Модель InstanceCLIPSeg вирішує цю проблему за допомогою висхідного підходу. Ключовою особливістю методу є одночасне прогнозування теплової карти центрів об'єктів та чотириканальної карти відстаней від кожного пікселя об'єкта до меж його обмежувальної рамки.

Мета полягає в експериментальному дослідженні архітектурних конфігурацій декодера карт відстаней у моделі InstanceCLIPSeg для визначення оптимальної комбі-

нації підходів до відновлення роздільної здатності, впливу згорток координат та стратегій об'єднання ознак.

Проведено експериментальне порівняння дев'яти архітектурних конфігурацій декодера карт відстаней у моделі InstanceCLIPSeg. Досліджено вплив методів підвищення роздільної здатності (білінійна інтерполяція, PixelShuffle), координатних згорток та стратегій злиття ознак. Встановлено, що гібридна архітектура з PixelShuffle та однорівневим злиттям ознак є найефективнішою (mean Dice 0,2374), перевершуючи базовий підхід на 9,4%. Виявлено надлишковість CoordConv при наявності контексту трансформера та перевагу гібридного злиття над ієрархічним.

У цій статті досліджуються архітектури декодера відстаней для моделі InstanceCLIPSeg.

1. Було виявлено, що оптимальною конфігурацією є гібридна архітектура з PixelShuffle, ініціалізацією ICNR та об'єднанням окремих ознак (L3, L7, L9) без використання CoordConv у блоках уточнення. Вона перевершує базовий метод на 9,4% за середньою метрикою Dice.

2. CoordConv ефективний лише в простих архітектурах. За наявності багатого контексту від трансформаторного кодера він стає надлишковим та погіршує якість сегментації.

3. Для ViT-кодерів гібридне одноразове об'єднання ознак є кращим за класичні ієрархічні пропускні з'єднання (U-Net), оскільки дозволяє уникнути артефактів інтерполяції проміжних карт ознак.

Подальша робота передбачає розширення навчання (більше ніж на 3 епохи) для дослідження впливу тривалості навчання на інтеграцію інформації про положення.

Ковтуненко Андрій Романович – аспірант кафедри Інформатики Харківського національного університету радіоелектроніки, Україна.
ORCID: https://0009-0004-9072-7779
Машталір Сергій Володимирович – доктор технічних наук, професор кафедри Інформатики Харківського національного університету радіоелектроніки, Україна.
ORCID: https://orcid.org/0000-0002-0917-6622

Kovtunenko Andrii Romanovych – PhD student, Informatics Department. Kharkiv National University of Radio Electronics, Ukraine.
ORCID: https://0009-0004-9072-7779
Mashtalir Sergii Volodymyrovych – Doctor of Engineering Science, Professor, Informatics Department. Kharkiv National University of Radio Electronics, Ukraine.
ORCID: https://orcid.org/0000-0002-0917-6622

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)