Table I
Accuracy and Predictive Values

	UNet Segmentation				CRF Segmentation			
	Acc.	CPV	DPV	NPV	Acc.	CPV	DPV	NPV
Binary Segmentation, Dice Loss								
Minimal	0.2629	0.0108	-	0.3144	0.2556	0.0086	-	
Average	0.7427	0.3962	-	0.8984	0.7440	0.3978	-	0.8987
Median	0.7580	0.3933	-	0.9464	0.7589	0.3939	-	0.9479
Maximal	0.9243	0.8968	-	1	0.9217	0.9001	-	1
Binary Segmentation, Focal Loss								
Minimal	0.3379	0	-	0.3330	0.3389	0	-	0.3322
Average	0.8308	0.4023	-	0.8513	0.8098	0.4185	-	0.8292
Median	0.8671	0.3960	-	0.8891	0.8412	0.4110	-	0.8635
Maximal	0.9785	0.9897	-	0.9965	0.9780	1	-	0.9950
Ternary Segmentation, Dice Loss								
Minimal	0.0685	0.1310	0	0	0.0622	0.1271	0	0
Average	**0.3175**	0.6830	**0.3746**	0.5127	**0.4498**	0.6819	**0.5169**	0.5132
Median	**0.3067**	0.7412	**0.3245**	0.5017	**0.4386**	0.7447	**0.5616**	0.5132
Maximal	**0.7331**	0.9992	0.9992	1	**0.9168**	0.9983	0.9947	1
Ternary Segmentation, Focal Loss								
Minimal	0.0053	0.0466	0	0	0.0534	0.0899	0.0002	0
Average	**0.3178**	0.7028	**0.2437**	0.4929	**0.5026**	0.7014	**0.4890**	0.4765
Median	**0.3012**	0.8016	**0.0771**	0.4913	**0.4952**	0.7785	**0.4911**	0.4693
Maximal	**0.7775**	0.9998	1	1	**0.9381**	1	0.9911	1

Sometimes the rates $N_1/\hat{N}_1$ and $N_2/\hat{N}_2$ are greater than 1. It may indicate situations where separate persons are detected as crowds and sparse crowds are recognized as dense crowd regions. According to our calculations, from 15% to 25% images have such values. Hence, the model might both under- and overestimate the crowd density. Also, data from Table II prove again using the CRF is crucial for ternary segmentation prediction.

Table II
Crowd Detection Rates

	UNet segmentation		CRF segmentation	
	Dice	Focal	Dice	Focal
Binary Segmentation ($N_1/\hat{N}_1$)				
Minimal	0.0097	0	0	0
Average	0.7896	0.1719	0.7909	0.1687
Median	0.8454	0.1189	0.8473	0.0872
Maximal	1.3235	0.8868	1.3235	0.9057
Ternary Segmentation ($N_2/\hat{N}_2$)				
Minimal	0	0.0402	0	0
Average	**0.5298**	0.0199	**0.8785**	0.7896
Median	**0.5215**	0	**0.9704**	0.8517
Maximal	1.2971	1	1.2971	1.2971

B. Crowd Semantics

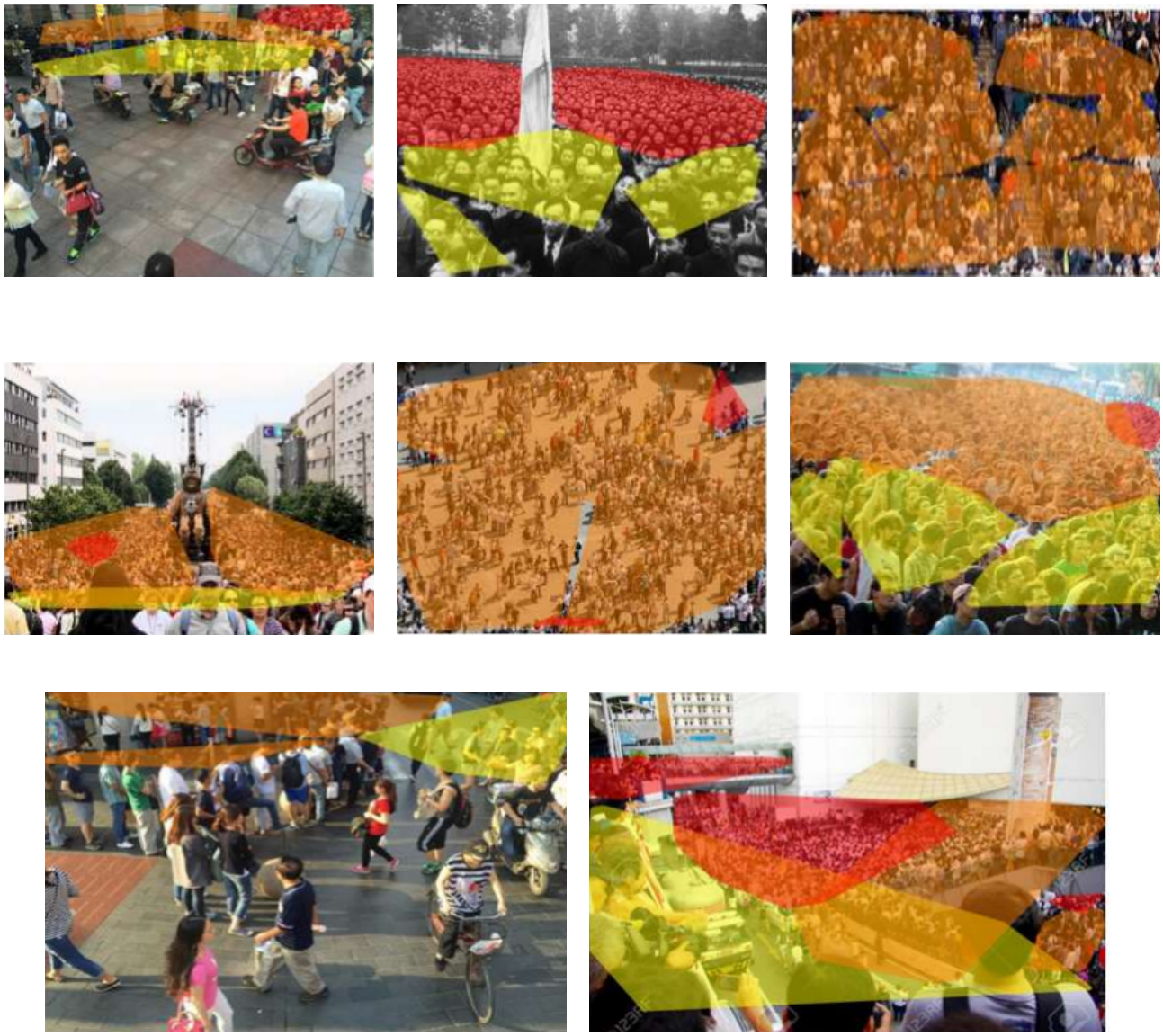
Most of the ShanghaiTech datasets images are captured by a camera observing a nearby scene from above. Hence, the typical clusterization consists of distant dense clusters, closer regular clusters, and near sparse clusters (1). However, the crowd in an image can be divided vertically if there is a tall object like a pole or a flag (Fig. 4b).

Sometimes the pattern doesn't hold, which could indicate a group with interest. Some examples where we detect a people's interest include:

- Multiple clusters with equal density spanning most of the image (Fig. 4c). Those usually present a uniform crowd with regular attention.
- A small cluster within or near a bigger one of a different type (Fig. 4, d-e). Those situations usually present a concentration of interest in particular groups within or near the crowd.
- A significant overlapping between clusters of different types (Fig. 4f). This one can indicate a spreading interest or joining the people groups. Real-time surveillance systems must detect such actions to prevent any dire situations.
- Elongated clusters presenting regular or dense crowds may indicate the presence of a queue in the region (Fig. 4g, the bottom orange cluster). If the density is high enough, some extraordinary situations might take place like queue crushes or evacuation panic which must be dealt with immediately.
- A wide sparse cluster at the bottom of the image might indicate a group of people that is very close to the observer (Fig. 4, d, h). Overlapping between the close cluster and other, distant ones is another feature of such a situation. Depending on the people's behavior, such a close group might be considered an outlier or an interest group, especially when it grows or approaches the observer.

v. Conclusions

This paper presents an approach for semantic segmentation of dense and sparse crowd images, addressing the critical



Semantic segmentation of crowds of different types: regular crowds with no attention (a, b), uniform crowd with regular attention (c),diverse crowd containing groups with increased interest (d, e), diverse crowd with a spreading group with interest (f), crowd with a queue (g), crowd with a close cluster (h)

need for accurate crowd analysis in various applications such as crowd management, surveillance, and urban planning. Our proposed method leverages a combination of UNet and CRF networks, augmented by a semi-automatic labeling technique based on Gaussian blur and thresholding methods to generate ground truth maps. Furthermore, we highlight some typical crowd behavior patterns based on clustering the people groups by their density and interconnections between them. Indicating those patterns is important for understanding crowd structures and dynamics as well as establishing crowd management and safety.

Through extensive experimentation and evaluation, we have demonstrated the effectiveness of our approach in accurately segmenting crowd images, particularly in binary segmentation tasks distinguishing crowded from non-crowded regions. While our model excels in binary segmentation, we acknowledge the challenges encountered in ternary segmentation tasks involving dense crowds, sparse crowds, and non-crowded areas. Despite this, our model shows promising results in crowd detection regardless of crowd density. Besides, we prove the necessity of CRF refinement to get better results.

References

[1] K. Khan et al., "Crowd Counting Using End-to-End Semantic Image Segmentation," *Electronics*, 2021, vol. 10, no. 11, #1293, doi: 10.3390/electronics10111293.

[2] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, L. Lin, "Crowd counting with deep structured scale integration network," in *2019 IEEE International Conference on Computer Vision (CVF)*, pp. 1774-1783.

[3] S. Sholtanyuk, A. Leuniaku, "Lightweight Deep Neural Networks for Dense Crowd Counting Estimation," in *Pattern Recognition and Information Processing (PRIP'2021)*, United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, 2021, pp. 61–64.

[4] S. Sholtanyuk, "Finding The Optimal Segmentation of a Crowd Image with Watershed Method," in *Information Systems and Technologies (CSIST'22)*, Part 2, Belarusian State University, Minsk, 2022, pp. 217-223. Available at: https://elib.bsu.by/handle/123456789/288544.

[5] F. Abdullah, A. Jalal, "Semantic Segmentation Based Crowd Tracking and Anomaly Detection via Neuro-Fuzzy Classifier in Smart Surveillance System," *Arabian Journal for Science and Engineering*, 2023, vol. 48, no. 2, pp. 2173-2190.

[6] M. Gruosso, N. Capece, U. Erra, "Human Segmentation in Surveillance Video with Deep Learning," *Multimedia Tools and Applications*, 2021, vol. 80, no. 1, pp. 1175-1199.

[7] A. Kroschanka, E. Mikhno, M. Kovalev, V. Zaharieva, A. Zagorskiy, "Semantic Analysis of the Video Stream Based on Neuro-Symbolic Artificial Intelligence," in *Open Semantic Technologies for Intelligent Systems*.

[8] OSTIS-2021. Belarusian State University of Informatics and Radioelectronics, Minsk, 2021, pp. 193-204.

[9] A. Kroshchanka, V. Golovko, E. Mikhno, M. Kovalev, V. Zaharie, and A. Zagorskij, "A Neural-Symbolic Approach to Computer Vision," in *Open Semantic Technologies for Intelligent Systems*, eds.: V. Golenkov, V. Krasnoproshin, V. Golovko, and B. Shunkevich, Cham: Springer International Publishing, 2022, pp. 282-309, doi: 10.1007/978-3-031-15882-7_15.

[10] L. Greco, P. Ritrovato, M. Vento, "On the use of semantic technologies for video analytics," *J Ambient Intell Human Comput*, 2021, vol. 12, pp. 567-587, doi: 10.1007/s12652-020-02021-y.

[11] P. Anderson, B. Fernando, M. Johnson, S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *Computer Vision–ECCV 2016*, eds.: B. Leibe, J. Matas, N. Sebe, M. Welling, Cham: Springer International Publishing, 2016, pp. 382-398, doi: 10.1007/978-3-319-46454-1_24.

[12] M. H. T. De Boer, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo, W. Kraaij, "Semantic Reasoning in Zero Example Video Event Retrieval," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2017, vol. 13, no. 4, Article 60, 17 p., doi: 10.1145/3131288.

[13] Z. Feng, Z. Zeng, C. Guo, Z. Li, "Exploiting Visual Semantic Reasoning for Video-Text Retrieval," *arXiv preprint*, arXiv:2006.08889, 2020.

[14] S. Munir, S.I. Jami, S. Wasi, "Towards the Modelling of Veillance based Citizen Profiling using Knowledge Graphs," *Open Computer Science*, 2021, vol. 11, no. 1, pp. 294-304, doi: 10.1515/comp-2020-0209.

[15] X. Guo, M. Gao, G. Zou, A. Bruno, A. Chehri, G. Jeon, "Object Counting via Group and Graph Attention Network," in *IEEE Transactions on Neural Networks and Learning Systems*, 2023, pp. 1-12, doi: 10.1109/TNNLS.2023.3336894.

[16] L. Greco, P. Ritrovato, A. Saggese, M. Vento, "Improving reliability of people tracking by adding semantic reasoning," in *IEEE conference on advanced video and signal based surveillance (AVSS)*, IEEE, 2016, pp. 194-199.

[17] K. Humphrey, G. Underwood, "Domain knowledge moderates the influence of visual saliency in scene recognition," *British Journal of Psychology*, 2008, vol. 100, no. 2, pp. 377-398, doi: 10.1348/000712608X334780.

[18] B. Chen, Z. Yan, K. Li, P. Li, B. Wang, W. Zuo, L. Zhang, "Variation Attention: Propagating Semantic-Specific Knowledge for Multi-Domain Learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16065-16075.

[19] V. A. Sindagi, V. M. Patel, "HA-CCN: Hierarchical Attention-Based Crowd Counting Network," in *IEEE Transactions on Image Processing*, 2020, vol. 29, pp. 323-335, doi: 10.1109/TIP.2019.2928634.

[20] W. Liu, M. Salzmann, P. Fua, "Context-Aware Crowd Counting," in *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5099-5108.

[21] J. Wang, Z. Chen, Y. Wu, "Action recognition with multiscale spatiotemporal contexts," in *CVPR 2011*. IEEE, June 2011, pp. 3185-3192.

[22] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234-241.

[23] V. Badrinarayanan, A. Kendall, R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 12, pp. 2481-2495.

[24] F.I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, "ResUNet-a: a Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data," in *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, vol. 162, pp. 94-114.

[25] J. Long, E. Shelhamer, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440.

[26] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, J. Malik, "Semantic Segmentation Using Regions and Parts," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2012, pp. 3378-3385.

[27] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[28] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint*, arXiv:1409.1556, 2014.

[29] P. Krähenbühl, V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," *Advances in Neural Information Processing Systems*, 2011, vol. 24, pp. 109-117.

[30] X. He, R.S. Zemel, M.A. Carreira-Perpiñán, "Multiscale Conditional Random Fields for Image Labeling," in *2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 695-702, doi: 10.1109/CVPR.2014.135223.

[31] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886-893.

[32] D. Helbing, P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, 1995, vol. 51, no. 5, pp. 4282-4286.

[33] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, "Single-Image Crowd Counting via a Multi-Column Convolutional Neural Network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589-597.

[34] M. Tan, Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International conference on machine learning*. PMLR, May 2019, pp. 6105-6114.

[35] S. He et al., "An Image Inpainting-Based Data Augmentation Method for Improved Sclerosed Glomerular Identification Performance with The Segmentation Model EfficientNetB3-UNet," *Scientific Reports*, 2024, vol. 14, no. 1, 1033.

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520.

[37] P. Yakubovskiy, "Segmentation Models," GitHub repository. Available at: `https://github.com/qubvel/segmentation_models` (accessed 2024, Mar).

[38] S. Jadon, "A Survey of Loss Functions for Semantic Segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, Oct. 2020, pp. 1-7.

[39] F. Milletari, N. Navab, S.A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 International Conference on 3D Vision (3DV)*. IEEE, Oct. 2016, pp. 565-571.

[40] M.S. Hossain, J.M. Betts, A.P. Paplinski, "Dual Focal Loss to Address Class Imbalance in Semantic Segmentation," *Neurocomputing*, 2021, vol. 426, pp. 69-87.

[41] L. Chen, "PyDenseCRF," GitHub repository. Available at: `https://github.com/lucasb-eyer/pydensecrf` (accessed 2024, Mar).

[42] M. Chen, S. Banitan, M. Maleki, Y. Li, "Pedestrian Group Detection with K-Means and DBSCAN Clustering Methods," in *2022 IEEE International Conference on Electro Information Technology (eIT)*. Mankato, MN, USA, 2022, pp. 1-6, doi: 10.1109/EIT53591.2022.9813918.

[43] A. Bouhmidi, J. Paquet, E. Bocher, "Using a Clustering Method to Detect Spatial Events in a Smartphone-Based Crowd-Sourced Database for Environmental Noise Assessment," *Sensors*, 2022, vol. 22, 8832, doi: 10.3390/s22228832.

АВТОМАТИЗАЦИЯ ОЦЕНКИ ВНИМАНИЯ СКОПЛЕНИЙ ЛЮДЕЙ НА ОСНОВЕ ПОЛУАВТОМАТИЧЕСКОЙ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ СЕТЕЙ UNET И CRF

Шолтанюк С. В., Малёнкин Я. О., Лэй Б., Недзьведь А. М.

Семантическая сегментация изображений скоплений людей играет ключевую роль в различных приложениях, таких как управление толпой, наблюдение и городское планирование. В данной статье предложен подход к семантической сегментации изображений с плотной и разреженной толпой на основе полуавтоматической разметки, используя комбинацию UNet и условных случайных полей (CRF).

Представляется новый метод семантической сегментации для изображений скоплений людей. Сеть UNet используется для первоочередной сегментации, после которой необходимо следует её уточнение с использованием CRF. Результаты экспериментов показали, что модель лучше выявляет бинарную маску (области, занятые людьми, и фона), свободную от ошибок, чем без использования CRF. Кроме того, улучшение точности, показанной с помощью количественных и качественных результатов, превосходит существующие методы сегментации.Таким образом, предложенный подход даёт лучшие результаты по сегментации толпы в целом (без учёта типа плотности толпы), показав значительное улучшение сегментации при помощи CRF в задаче генерации сегментации толпы.

Также на основе предложенной модели сегментации выделены некоторые закономерности поведения скоплений людей. Они связаны как с плотностью внимания людей, связанным с восприятием людей и между ними, а также вероятностью возникновения чрезвычайных ситуаций.

Received 25.03.2024