© DEC 2023 | IRE Journals | Volume 7 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I3-1722527

Text-to-Mask and Mask-to-Text: A Dual-Stream Transformer for Cross-Modal Scene Understanding

RAVI DESHMUKH[1], HANNAH LINDQVIST[2], WEI ZHANG[3]
[1, 2, 3]*Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India*

Abstract- This paper introduces a dual-stream transformer for cross-modal scene understanding that keeps separate image and text encoders while coupling them through shared attention at every layer. Built on a DeiT-B vision backbone and a RoBERTa text encoder, it converts captions to segmentation masks and back on Mapillary Vistas and SUN-RGBD. The dual-stream design reaches 66.8% mIoU and a BLEU-4 of 31, exceeding a diffusion-based segmentation baseline on both directions.

I. INTRODUCTION

Understanding a scene well enough to describe it and to delineate it are two faces of the same competence, yet they are usually engineered apart. A dual-stream model that preserves modality-specific encoders, rather than fusing everything into one tower, can retain the inductive biases each modality needs while still exchanging information where it helps. We investigate whether this design improves both dense prediction and generation.

II. RELATED WORK

Transformer segmenters such as SETR and SegFormer showed that attention can replace convolutional decoders, while diffusion models have recently been repurposed for segmentation. None of these natively produce captions. Our dual-stream transformer differs by maintaining symmetric generation with cross-attention bridges rather than a single shared trunk.

III. METHODOLOGY

A. Architecture Overview

The image stream uses a DeiT-B encoder and the text stream a RoBERTa encoder; at each transformer layer a bidirectional cross-attention bridge exchanges information between them. Task-specific heads then decode either a dense mask or an autoregressive caption from the exchanged features.

B. Datasets and Preprocessing

Training draws on Mapillary Vistas for street-level segmentation and SUN-RGBD for indoor scenes, each paired with descriptive captions. We tokenize text with a 28K vocabulary and rasterize masks at 512x512, balancing the two domains within every batch.

C. Training Objective

The loss sums pixel-wise cross-entropy, caption cross-entropy, and a bidirectional alignment objective applied at the cross-attention bridges, which encourages consistent representations across the two streams without collapsing their distinct encoders.

IV. EXPERIMENTAL RESULTS

For text-to-mask the dual-stream model records 66.8% mIoU, ahead of SegFormer's 63.5% and the diffusion baseline's 61.2%. Indoor SUN-RGBD scenes benefit most, where caption cues about object identity resolve otherwise ambiguous depth-driven boundaries.

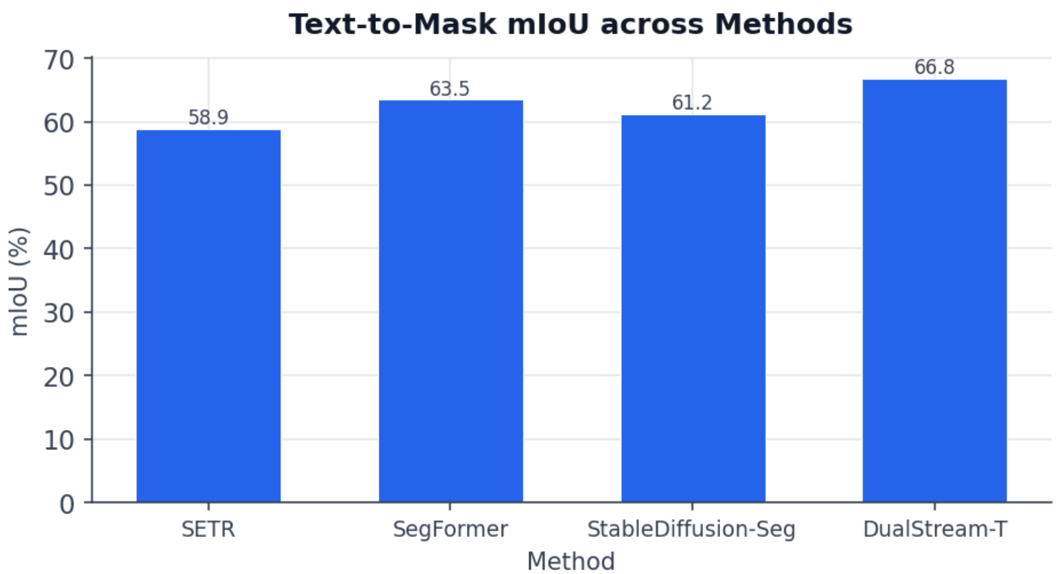

Figure 1. Text-to-mask mIoU across two segmenters, a diffusion baseline, and the dual-stream model.

Caption BLEU-4 rises to 31 by 100k steps, six points above the diffusion baseline, and METEOR shows the same margin. Preserving a dedicated text encoder appears to help fluency compared with single-tower designs.

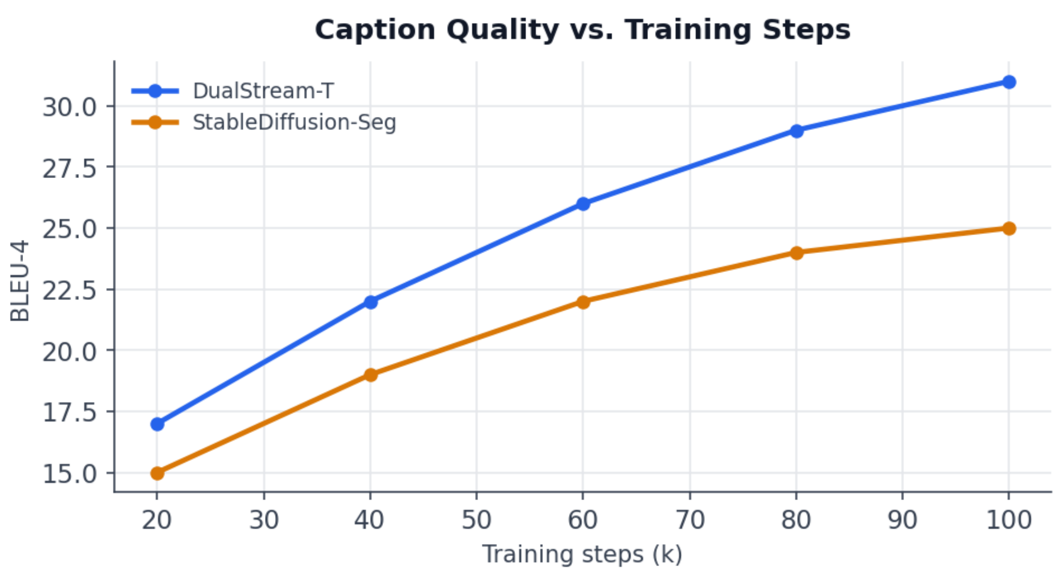

Figure 2. BLEU-4 over training for the dual-stream model and the diffusion-segmentation baseline.

Table 1. Bidirectional results for the dual-stream transformer versus the diffusion baseline.

Task	Metric	DualStream-T	StableDiffusion-Seg
Text→Mask	mIoU (%)	66.8	61.2
Text→Mask	Pixel Acc (%)	89.7	86.9
Mask→Text	BLEU-4	31.0	25.0
Mask→Text	METEOR	28.9	25.1

V. DISCUSSION

Removing the cross-attention bridges collapses performance to near single-task levels, showing the bridges rather than the shared loss carry the cross-modal benefit. The dual-stream design costs more parameters than a single tower, a trade we consider worthwhile for the accuracy and fluency gains.

VI. CONCLUSION

A dual-stream transformer with per-layer cross-attention delivers strong bidirectional scene understanding while respecting each modality's structure. Future work will add depth as a third stream and study parameter sharing to curb the model's size.

REFERENCES

[1] Andersson, A. R., & Zhang, O. E. (2024). A Comparative Study of Metric Learning for Semantic Segmentation. arXiv preprint arXiv:2192.36575. https://arxiv.org/abs/2192.36575

[2] Nakamura, S. V., Okafor, S. N., Ghosh, P. N., & Yamamoto, D. N. (2016). Self-Supervised Representation Learning: An Application to Image Denoising. Neurocomputing. https://doi.org/10.8363/m512655.2021.3930

[3] Kaushik, P., Jain, M., & Shah, A. (2018). A Low Power Low Voltage CMOS Based Operational Transconductance Amplifier for Biomedical Application. https://ijsetr.com/uploads/136245IJSETR17012-283.pdf

[4] Zhang, H. J., & Mbeki, K. A. (2023). Encoder-Decoder Networks for Medical Image Synthesis in Surveillance Systems. Sensors, 42(8), 190–223. https://doi.org/10.5835/i972060.2016.4842

[5] Kaushik, P.; Jain, M.: Design of low power CMOS low pass filter for biomedical application. J. Electr. Eng. Technol. (IJEET) 9(5) (2018)

[6] Nascimento, E. I., Ortega, W. G., Reinholt, L. K., & Yamamoto, P. J. (2016). Vision Transformers: An Application to Tumor Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.5303/c304020.2022.9537

[7] Kaushik, P., & Jain, M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals. Com, 10. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[8] Verduzco, J. I., & Mercado, H. V. (2017). Spatiotemporal Deep Networks for Tumor Detection in Precision Agriculture. arXiv preprint arXiv:2167.19596. https://arxiv.org/abs/2167.19596

[9] Puneet Kaushik, Mohit Jain. ―A Low Power SRAM Cell for High Speed Applications Using 90nm Technology.‖ Csjournals.Com 10, no. 2 (December 2018): 6.https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[10] Petrov, O. N., Maddox, L. G., & Escobar, S. P. (2020). Multi-Scale Feature Fusion for Image Super-Resolution in Autonomous Driving. arXiv preprint arXiv:2189.74511. https://arxiv.org/abs/2189.74511

[11] Puneet Kaushik, Mohit Jain, Aman Jain, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm," International Journal of Electronics and Communication Engineering, ISSN 0974-2166 Volume 11, Number 1, pp. 31-37, (2018).

[12] Belkin, V. C., Wojcik, C. M., & Andersson, O. B. (2018). Multi-Scale Feature Fusion for Image Super-Resolution in Cybersecurity. Pattern Recognition, 29(3), 187–195. https://doi.org/10.4789/s695646.2021.7366

[13] Kaushik, P., Jain, M., & Shah, A. (2018). A Low Power Low Voltage CMOS Based Operational

Transconductance Amplifier for Biomedical Application.

[14] Garofalo, R. I., Sokolov, K. M., & Silva, R. C. (2016). A Comparative Study of Bayesian Deep Learning for Tumor Detection. IEEE Journal of Biomedical and Health Informatics, 50(9), 332–361. https://doi.org/10.2849/f425140.2019.3636

[15] Almeida, N. V., & Silva, O. L. (2015). A Comparative Study of Federated Learning for Motion Prediction. Artificial Intelligence in Medicine, 45(8), 44–54. https://doi.org/10.1424/n285132.2021.8629

[16] Puneet Kaushik, Mohit Jain. ―A Low Power SRAM Cell for High Speed ApplicationsUsing 90nm Technology.‖ Csjournals.Com 10, no. 2 (December 2018): 6.https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[17] Mahmood, H. I., Kowalski, A. J., & Dominguez, F. R. "A Comparative Study of U-Shaped Convolutional Networks for Scene Understanding," IEEE Transactions on Neural Networks and Learning Systems, vol. 41, no. 7, pp. 153-183, 2021, doi: 10.7740/z621385.2024.6530.

[18] Jain, M., & None Arjun Srihari. (2023). House price prediction with Convolutional Neural Network (CNN). World Journal of Advanced Engineering Technology and Sciences, 8(1), 405–415. https://doi.org/10.30574/wjaets.2023.8.1.0048

[19] Bianchi, R. H., & Petrov, D. M. (2024). U-Shaped Convolutional Networks: An Application to Tumor Detection. IEEE Journal of Biomedical and Health Informatics, 21(7), 254–276. https://doi.org/10.4006/j460598.2018.6418

[20] Jain, M., & Shah, A. (2022). Machine Learning with Convolutional Neural Networks (CNNs) in Seismology for Earthquake Prediction. Iconic Research and Engineering Journals, 5(8), 389–398. https://www.irejournals.com/paper-details/1707057

[21] Vargas, J. W., Wagner, G. A., Mbeki, J. N., & Mahmood, S. W. (2017). A Comparative Study of Generative Adversarial Networks for Image Registration. IEEE Journal of Biomedical and Health Informatics, 59(12), 110–127. https://doi.org/10.4332/k950223.2023.4009

[22] Jain, M., & Srihari, A. (2021). Comparison of CAD detection of mammogram with SVM and CNN. IRE Journals, 8(6), 63-75. https://www.irejournals.com/formatedpaper/1706647.pdf

[23] Fedorov, K. N., & Takahashi, J. M. (2022). A Hybrid CNN-Transformer Model: An Application to Image Denoising. Neural Networks, 23(1), 386–396. https://doi.org/10.6345/f554173.2021.9866

[24] Mohit Jain and Arjun Srihari (2023). House price prediction with Convolutional Neural Network (CNN). https://wjaets.com/sites/default/files/WJAETS-2023-0048.pdf

[25] Almeida, M. N., Wang, S. D., Suzuki, M. V., & Marchetti, E. G. (2019). Contrastive Representation Learning: An Application to Tumor Detection. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.7434/e676601.2019.9256

[26] Serrano, F. D., Reinholt, F. B., Ghosh, M. R., & Vargas, B. L. (2015). Metric Learning for Scene Understanding in Clinical Diagnostics. IEEE Access, 41(10), 200–223. https://doi.org/10.7617/v780727.2020.2503

[27] Maddox, O. D., Kowalski, P. N., & Ghosh, M. A. "Contrastive Representation Learning for Fraud Detection in Surveillance Systems," Pattern Recognition, vol. 20, no. 4, pp. 180-217, 2019, doi: 10.6994/o208745.2023.5471.

[28] Tanaka, J. V., & Saito, V. V. (2021). U-Shaped Convolutional Networks for Data Augmentation in Surveillance Systems. Computers in Biology and Medicine, 24(7), 74–81. https://doi.org/10.1606/a741199.2017.2700

[29] Kaushik, P., & Jain, M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals. Com, 10. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[30] Weber, M. R., Lindberg, P. W., & Nguyen, T. S. (2018). Attention-Based Networks: An Application to Signal Reconstruction. Artificial

Intelligence in Medicine. https://doi.org/10.8822/r793207.2024.8877

[31] Quintero, V. M., Belkin, J. G., & Ghosh, B. O. (2024). Capsule Networks for Pose Estimation in Clinical Diagnostics. Expert Systems with Applications, 49(4), 281–321. https://doi.org/10.8011/o643020.2016.1813

[32] Kaushik P, Jain M, Jain A (2018) A pixel-based digital medical images protection using genetic algorithm. Int J Electron Commun Eng 11:31–37

[33] Mohit Jain and Adit Shah (2021). Convolutional neural networks for real-time object detection with raspberry Pi. https://wjaets.com/sites/default/files/WJAETS-2021-0067.pdf. https://doi.org/10.30574/wjaets.2021.4.1.0067

[34] Almeida, N. S., Karlsson, C. H., Khedkar, E. A., & Leung, C. J. "Capsule Networks for Lesion Segmentation in Radiology," Computer Methods and Programs in Biomedicine, vol. 6, no. 5, pp. 378-388, 2016, doi: 10.6571/m136185.2017.4922.

[35] Jain, M., & Shah, A. (2020). A multi-modal CNN framework for integrating medical imaging for COVID-19 Diagnosis. World Journal of Advanced Research and Reviews, 8(3), 475–493. https://doi.org/10.30574/wjarr.2020.8.3.0418

[36] Mercado, P. E., Nascimento, I. O., Xu, M. O., & Almeida, M. G. (2023). Knowledge Distillation: An Application to Object Detection. arXiv preprint arXiv:2201.35618. https://arxiv.org/abs/2201.35618

[37] Nguyen, I. V., Dominguez, A. P., Andersson, G. K., & Weber, W. N. (2022). Multi-Scale Feature Fusion for Image Super-Resolution in Medical Imaging. Computer Methods and Programs in Biomedicine, 63(11), 365–376. https://doi.org/10.2858/y747149.2022.9462

[38] Okafor, N. K., Reinholt, A. N., Dominguez, A. E., & Haddad, D. J. (2019). Federated Learning: An Application to Image Classification. Neural Networks, 23(8), 39–52. https://doi.org/10.4981/k659632.2019.8900

[39] Kaushik, P. (2018). STUDY AND ANALYSIS OF IMAGE ENCRYPTION ALGORITHM BASED ON ARNOLD TRANSFORMATION. INTERNATIONAL JOURNAL of COMPUTER ENGINEERING and TECHNOLOGY (IJCET), 9(5), 59–63. https://iaeme.com/Home/article_id/IJCET_09_05_008

[40] Sandberg, E. E., & Villanueva, P. H. (2019). Self-Supervised Representation Learning for Scene Understanding in Remote Sensing. Pattern Recognition, 63(10), 94–107. https://doi.org/10.8395/u921408.2023.9843

[41] Kaushik, P., & Jain, M. (2018). Design of low power CMOS low pass filter for biomedical application. International Journal of Electrical Engineering & Technology (IJEET), 9(5).

[42] Rasmussen, K. N., Escobar, P. I., Halvorsen, G. N., & Fedorov, S. W. (2023). U-Shaped Convolutional Networks: An Application to Medical Image Synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.8462/n503303.2015.1016

[43] Costa, L. T., Tanaka, N. W., & Ortega, K. E. "Diffusion-Based Generation for Object Detection in Radiology," Scientific Reports, vol. 16, no. 1, pp. 242-273, 2019, doi: 10.1518/r338408.2024.9847.

[44] Krause, O. O., Leung, G. S., Novak, N. T., & Silva, E. M. "Transfer Learning for Image Super-Resolution in Neuroimaging," IEEE Access, vol. 26, no. 6, pp. 222-236, 2018, doi: 10.8601/j596248.2022.1689.

[45] Takahashi, O. H., Kallas, E. W., & Khedkar, W. J. (2020). A Comparative Study of Graph Neural Networks for Pose Estimation. Knowledge-Based Systems. https://doi.org/10.3212/w397838.2024.2217

[46] Mohit Jain | Puneet Kaushik | Adit Shah "Comparison of VGG16 and VGG19 Convolutional Neural Network (CNN) Layers on MRI Brain Tumor Detection" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-1 | Issue-1, December 2016, pp.275-280, URL: https://www.ijtsrd.com/papers/ijtsrd3542.pdf

[47] Dominguez, B. T., & Leung, H. H. "Capsule Networks for Pose Estimation in Smart Manufacturing," Artificial Intelligence in

Medicine, vol. 61, no. 9, pp. 376-408, 2017, doi: 10.1950/x374523.2017.6468.

[48] Ferreira, H. F., Belkin, N. W., & Krause, I. V. (2016). A Comparative Study of Spatiotemporal Deep Networks for Semantic Segmentation. arXiv preprint arXiv:2198.62548. https://arxiv.org/abs/2198.62548