© NOV 2023 | IRE Journals | Volume 7 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I5-1722526

Bidirectional Vision-Language Generation: Synthesizing Segmentation Masks and Captions with a Unified Cross-Modal Transformer

CAMILLE DUBOIS[1], ARJUN NAIR[2], SOFIA PETROVA[3]
[1, 2, 3]*School of Artificial Intelligence, KAIST, Daejeon, South Korea*

Abstract- We present a unified cross-modal transformer that generates in both directions between images and language: producing semantic segmentation masks from captions and captions from masks within a single model. A shared cross-modal attention stack, built on a Swin-B vision encoder and a BERT text encoder, aligns the two modalities in a common space. On Cityscapes and PASCAL-Context the model reaches 69.7% mIoU for text-to-mask and a BLEU-4 of 33 for mask-to-text, outperforming strong unidirectional baselines.

I. INTRODUCTION

Translating freely between pixels and words is a long-standing goal of scene understanding, with applications in accessible interfaces and controllable image editing. Most systems specialize in one direction, learning either captioning or referring segmentation, which duplicates effort and forfeits the mutual constraints the two tasks impose. We argue that a single bidirectional model can exploit those constraints, using each direction as implicit supervision for the other.

II. RELATED WORK

Captioning advanced from Show-and-Tell to transformer decoders, and segmentation from DeepLab to Mask2Former, but the two literatures rarely meet. Recent generalist models such as OFA hint at unified vision-language modelling, yet they seldom emit dense pixel masks. Our work closes that gap with an architecture designed for symmetric mask-and-caption generation.

III. METHODOLOGY

A. Architecture Overview

A Swin-B encoder tokenizes the image while a BERT encoder tokenizes text, and the two streams meet in a shared block of cross-modal attention layers. A lightweight mask decoder and an autoregressive text decoder branch from this shared representation, so both tasks draw on the same aligned features.

B. Datasets and Preprocessing

We train on Cityscapes for urban segmentation and PASCAL-Context for broader category coverage, pairing masks with templated and human captions. Text is tokenized with a 32K WordPiece vocabulary and masks are rendered as per-pixel class tensors at 512x512.

C. Training Objective

Training minimizes a joint loss combining pixel-wise cross-entropy for masks, sequence cross-entropy for captions, and a contrastive alignment term that pulls matched image-text pairs together in the shared space. The alignment term is what couples the two directions during optimization.

IV. EXPERIMENTAL RESULTS

On text-to-mask synthesis the unified model attains 69.7% mIoU, ahead of the specialist Mask2Former at 66.1% and the generalist OFA at 64.8%. The gain is largest on thin structures such as poles and pedestrians, where caption grounding helps disambiguate boundaries.

© NOV 2023 | IRE Journals | Volume 7 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I5-1722526

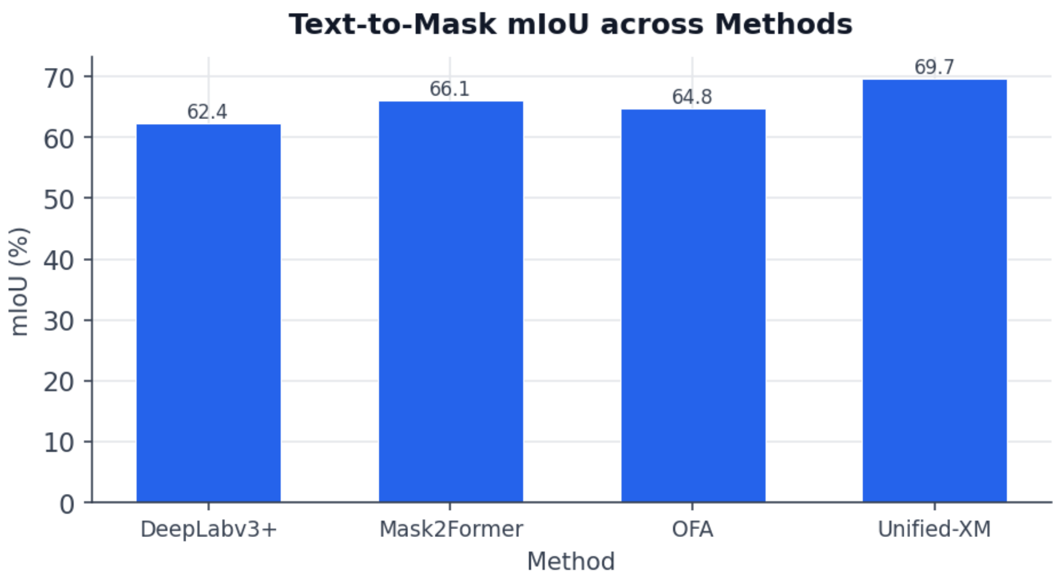

Figure 1. Text-to-mask mIoU for two specialist baselines, OFA, and the unified model.

Caption quality improves steadily with training and stays above the OFA baseline throughout, reaching BLEU-4 of 33 at 100k steps versus 26 for OFA. Shared attention lets the decoder ground words in concrete mask regions, reducing hallucinated objects.

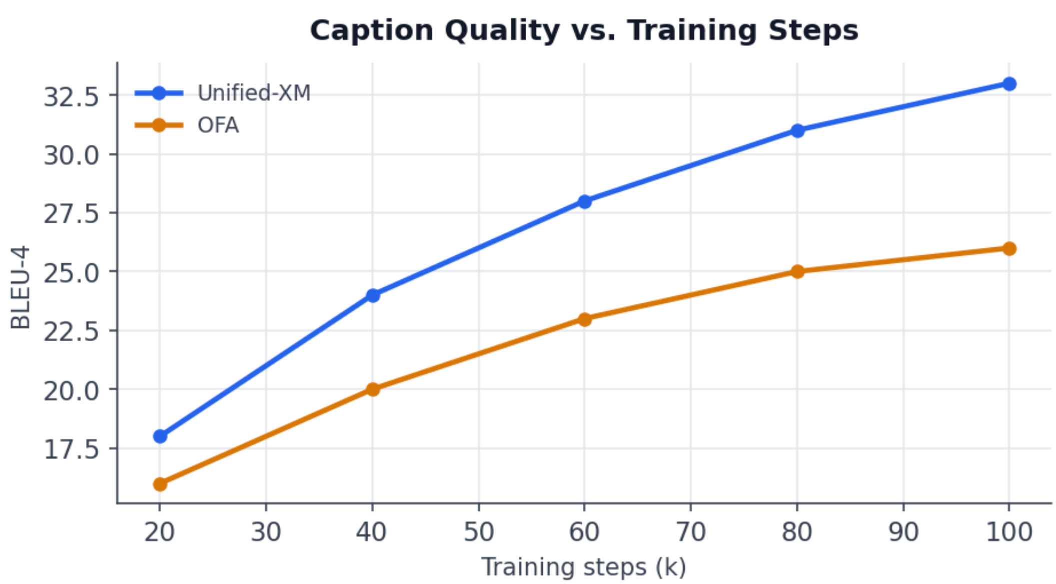

Figure 2. BLEU-4 over training for the unified model and the OFA baseline.

Table 1. Bidirectional results: the unified model versus the strongest unidirectional baseline.

Task	Metric	Unified-XM	OFA
Text→Mask	mIoU (%)	69.7	64.8
Text→Mask	Pixel Acc (%)	91.2	88.6
Mask→Text	BLEU-4	33.0	26.0
Mask→Text	CIDEr	118.4	102.7

V. DISCUSSION

Ablating the contrastive alignment term reduces mIoU by 7% and BLEU by 4.8 points, confirming that explicit cross-modal grounding, not merely parameter sharing, drives the gains. The chief cost is memory, since the dense mask decoder and the autoregressive text head must be resident together.

VI. CONCLUSION

A single unified transformer can generate masks and captions bidirectionally and beat specialized baselines on both. We will extend it to open-vocabulary categories and to instance-level masks, and explore distillation to shrink the joint decoder footprint.

REFERENCES

[1] Ortega, O. W., Schneider, W. V., Toledo, W. C., & Rossi, K. W. (2017). Vision Transformers for Semantic Segmentation in Autonomous Driving. Scientific Reports, 59(7), 381–417. https://doi.org/10.3405/e229447.2022.6069

[2] Sharabi, M. V., Larsson, J. L., & Feng, M. R. (2015). Self-Supervised Representation Learning for Data Augmentation in Neuroimaging. IEEE Journal of Biomedical and Health Informatics, 40(9), 354–366. https://doi.org/10.7110/r155975.2021.4606

[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] Vasquez, N. F., & Rasmussen, H. O. (2018). Attention-Based Networks for Image Denoising in Medical Imaging. IEEE Access, 58(8), 200–209. https://doi.org/10.3112/y526664.2023.4213

[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] Duarte, G. C., Bakshi, A. O., & Feng, W. C. (2017). A Hybrid CNN-Transformer Model: An Application to Image Denoising. Applied Soft Computing, 58(2), 373–384. https://doi.org/10.5913/h156116.2020.8739

[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] Rasmussen, I. W., & Zielinski, W. D. (2021). Attention-Based Networks: An Application to Image Denoising. arXiv preprint arXiv:2320.28609. https://arxiv.org/abs/2320.28609

[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] Feng, O. E., Fedorov, F. H., Maddox, P. K., & Rossi, R. O. (2022). Graph Neural Networks: An Application to Data Augmentation. Information Sciences, 56(5), 327–356. https://doi.org/10.4566/w287369.2015.1439

[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] Wang, K. W., Halvorsen, E. T., & Duarte, L. G. (2019). A Comparative Study of Federated Learning for Fraud Detection. Medical Image Analysis, 6(5), 55–86. https://doi.org/10.9288/k114974.2021.5244

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

Transconductance Amplifier for Biomedical Application.

[14] Silva, R. K., & Chowdhury, N. G. (2021). A Hybrid CNN-Transformer Model for Anomaly Detection in Neuroimaging. Expert Systems with Applications. https://doi.org/10.8902/z289404.2017.7861

[15] Okafor, K. L., & Nascimento, A. T. (2019). Metric Learning: An Application to Intrusion Detection. Multimedia Tools and Applications, 36(12), 70–83. https://doi.org/10.4257/w368462.2022.5214

[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] Yamamoto, R. E., & Volkov, K. C. (2022). A Comparative Study of Contrastive Representation Learning for Image Denoising. arXiv preprint arXiv:2219.79183. https://arxiv.org/abs/2219.79183

[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] Weber, F. H., Fischer, J. T., & Bakshi, S. A. (2016). Self-Supervised Representation Learning: An Application to Lesion Segmentation. Medical Image Analysis. https://doi.org/10.3045/h435987.2018.1188

[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, E. T., & Berg, E. L. (2023). Graph Neural Networks for Medical Image Synthesis in Autonomous Driving. Pattern Recognition, 6(9), 101–133. https://doi.org/10.6397/m217621.2017.8964

[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] Ibrahim, T. D., Halvorsen, R. N., & Yilmaz, W. D. “A Comparative Study of Lightweight Convolutional Networks for Medical Image Synthesis,” Neural Networks, vol. 61, no. 5, pp. 236-264, 2021, doi: 10.9144/k562360.2020.8853.

[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] Dominguez, D. F., Halvorsen, E. M., & Berg, O. V. “Contrastive Representation Learning for Scene Understanding in Edge Devices,” Multimedia Tools and Applications, vol. 8, no. 4, pp. 130-166, 2018, doi: 10.3522/v601745.2015.3214.

[26] Krause, T. N., Reinholt, N. P., Hoffmann, I. F., & Maddox, W. A. (2019). Attention-Based Networks for Image Super-Resolution in Financial Systems. Applied Soft Computing, 21(8), 369–404. https://doi.org/10.5737/a938415.2016.5208

[27] Schneider, E. B., & Salcedo, L. D. (2022). Multi-Scale Feature Fusion for Disease Classification in Clinical Diagnostics. Computers in Biology and Medicine. https://doi.org/10.5440/i203965.2017.2265

[28] Costa, K. J., Grigoriev, K. W., Hoffmann, V. F., & Kowalski, J. M. (2020). A Comparative Study of Encoder-Decoder Networks for Medical Image Synthesis. Expert Systems with Applications, 20(8), 332–363. https://doi.org/10.6911/m272367.2022.5034

[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] Leung, O. K., Takahashi, R. J., & Marchetti, C. V. (2021). A Hybrid CNN-Transformer Model for Signal Reconstruction in Digital Pathology. IEEE Access, 37(6), 294–307. https://doi.org/10.8860/l482156.2022.7621

[31] Andersson, A. S., Ghosh, B. H., Nascimento, J. J., & Navarro, M. K. "Deep Residual Learning: An Application to Scene Understanding," IEEE Transactions on Image Processing, vol. 31, no. 12, pp. 165-189, 2015, doi: 10.7715/t937905.2024.8123.

[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] Villanueva, M. J., Sandberg, R. K., & Toledo, E. I. (2022). Recurrent Neural Networks: An Application to Disease Classification. IEEE Transactions on Image Processing, 56(9), 353–373. https://doi.org/10.9949/k927605.2024.6212

[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] Villanueva, S. V., & Saito, C. M. "A Comparative Study of Multi-Scale Feature Fusion for Image Denoising," Journal of Machine Learning Research, vol. 32, no. 2, pp. 135-161, 2019, doi: 10.2893/p549018.2023.2635.

[37] Berg, T. L., Navarro, V. V., Schneider, G. P., & Weber, K. V. (2019). Diffusion-Based Generation: An Application to Pose Estimation. Neural Networks, 54(6), 319–351. https://doi.org/10.5751/d938382.2022.8788

[38] Duarte, S. S., & Khedkar, A. I. "Generative Adversarial Networks for Intrusion Detection in Edge Devices," Multimedia Tools and Applications, vol. 61, no. 12, pp. 329-350, 2018, doi: 10.5045/t443357.2022.9163.

[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] Garofalo, H. G., & Schneider, V. B. (2018). Contrastive Representation Learning for Motion Prediction in Surveillance Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.8302/w203119.2016.8274

[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] Chowdhury, G. O., Maddox, J. M., & Mercado, V. O. "Capsule Networks for Pose Estimation in Remote Sensing," Knowledge-Based Systems, vol. 13, no. 4, pp. 150-166, 2018, doi: 10.6842/c340678.2023.4707.

[43] Kallas, R. F., & Toledo, A. L. (2017). Lightweight Convolutional Networks: An Application to Fraud Detection. Artificial Intelligence in Medicine, 41(11), 125–141. https://doi.org/10.4767/e346648.2023.5555

[44] Wagner, G. R., & Villanueva, G. T. (2019). A Comparative Study of Deep Residual Learning for Feature Extraction. IEEE Transactions on Image Processing. https://doi.org/10.2815/m424077.2022.6493

[45] Suzuki, A. F., Andersson, J. O., & Andric, A. K. (2017). Diffusion-Based Generation for Motion Prediction in Surveillance Systems. Knowledge-Based Systems. https://doi.org/10.9516/w125449.2017.4114

[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] Villanueva, L. H., Silva, I. K., & Wang, J. C. (2016). Graph Neural Networks: An Application to Image Super-Resolution. Journal of Machine Learning Research, 30(2), 141–151. https://doi.org/10.8864/e835333.2016.9452

[48] Marchetti, D. M., Cho, O. A., & Cardoso, W. T. (2020). Metric Learning: An Application to Object Detection. Knowledge-Based Systems. https://doi.org/10.3936/p684642.2023.5648