Cross-Modal Synthesis: Generating Semantic Segmentation Masks from Image Captions and Vice-Versa using Multi-Modal Transformers

SUBHAM SAHOO[1], RAJAT GUPTA[2], CHINTAN TIBREWALA[3]
[1, 2, 3]College of Engineering, Amity University, Noida, India

Abstract- The integration of visual and linguistic modalities has transformed computer vision and natural language processing research. Cross-modal synthesis, which seeks to generate segmentation masks from text and captions from images, presents significant opportunities for understanding visual scenes semantically. In this work, we propose a unified transformer-based framework that performs bi-directional synthesis between image captions and semantic segmentation masks. The model, built on top of a multi-modal transformer encoder-decoder, learns shared latent representations enabling seamless translation between visual regions and linguistic tokens. Extensive experiments on COCO-Stuff and ADE20K datasets demonstrate that our method outperforms baseline models by 16% in mIoU for caption-to-mask synthesis and 14% BLEU improvement for mask-to-caption generation, establishing a new benchmark for multi-modal reasoning.

I. INTRODUCTION

The ability to translate between visual and linguistic modalities has gained prominence with the advent of transformer-based architectures. Tasks like image captioning, visual question answering, and image segmentation have traditionally been treated independently, yet they share a common semantic space. Cross-modal synthesis bridges this gap by enabling bi-directional reasoning—generating pixel-level semantic maps from textual descriptions and conversely, generating textual descriptions from structured visual cues.

This study explores a unified framework leveraging a multi-modal transformer that performs both text-to-mask and mask-to-text generation. Our motivation arises from medical and autonomous driving domains where captions or textual prompts can describe scenes requiring semantic decomposition. Unlike prior unidirectional models (e.g., CLIP, BLIP, and Mask2Former), our model performs simultaneous dual translation by sharing cross-attention layers between image and text embeddings.

II. RELATED WORK

Early works such as Show-and-Tell (Vinyals et al., 2015) and Show-Attend-and-Tell (Xu et al., 2016) established attention-based captioning frameworks. For segmentation, architectures like DeepLab (Chen et al., 2017) and Mask R-CNN (He et al., 2018) became standard. However, the convergence of visual-language models was spearheaded by CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023), which demonstrated shared latent spaces between modalities. Recent research, including Segment Anything (Kirillov et al., 2023) and OFA (Wang et al., 2022), revealed the potential of transformer-based cross-modal learning for multi-task applications. Nevertheless, these frameworks primarily focus on representation alignment rather than generative synthesis. Our work advances this by achieving true bi-directional cross-modal synthesis, aligning linguistic semantics with pixel-level visual content.

III. METHODOLOGY

A. Architecture Overview

Our architecture comprises a dual-stream transformer with shared cross-modal attention layers. The encoder processes inputs in both modalities—image embeddings via Vision Transformer (ViT-B/16) and tokenized captions via BERT embeddings—while the decoder performs the reverse synthesis task. Positional encodings ensure spatial alignment between word tokens and segmentation patches.

B. Dataset and Preprocessing

We utilize the COCO-Stuff (164K images) and ADE20K (25K scenes) datasets. Captions are tokenized using WordPiece with 30K vocabulary size. Segmentation masks are converted into one-hot pixel-level tensors of 512×512 resolution. We also synthetically augment captions with paraphrased variants to improve linguistic robustness.

C. Training Objective

Our model employs a joint training objective combining pixel-wise cross-entropy loss for segmentation and sequence-level cross-entropy for text generation:

$L\_total = \lambda_1 * L\_mask + \lambda_2 * L\_text + \lambda_3 * L\_align$
where L_align minimizes the cosine distance between latent embeddings of text and mask representations. The λ values were empirically set to 0.6, 0.3, and 0.1 respectively.

D. Evaluation Metrics

For text-to-mask synthesis, we report mean Intersection-over-Union (mIoU), pixel accuracy, and visual fidelity score (VFS). For mask-to-text synthesis, we evaluate BLEU-4, CIDEr, and METEOR metrics. All experiments were run on 8×A100 GPUs for 100 epochs with AdamW optimizer (learning rate 5e-5).

IV. EXPERIMENTAL RESULTS

Quantitative comparisons demonstrate significant improvements over existing baselines. Diffusion-based multimodal models such as StableDiffusion and OFA were used as benchmarks. Our approach achieves higher accuracy and textual coherence across both synthesis directions.

Model	Dataset	mIoU ↑	Pixel Acc. ↑	VFS ↑
OFA	COCO-Stuff	58.7	85.3	0.72
BLIP-2	COCO-Stuff	61.2	87.1	0.74
Proposed	COCO-Stuff	70.9	91.4	0.81
Proposed	ADE20K	68.4	90.6	0.80

Table 2 shows performance in mask-to-caption generation compared to recent baselines.

Model	Dataset	BLEU-4 ↑	CIDEr ↑	METEOR ↑
OFA	COCO-Stuff	27.6	110.5	21.8
BLIP-2	COCO-Stuff	31.4	118.9	23.6
Proposed	COCO-Stuff	35.9	126.3	25.2
Proposed	ADE20K	33.1	122.4	24.1

Figure 1 visually demonstrates caption-to-mask synthesis quality improvements. The proposed model accurately delineates semantic regions such as 'road', 'car', and 'person', matching caption content with high spatial precision.

V. DISCUSSION

The shared latent representation achieved via multi-modal attention facilitates superior context preservation during translation. When translating from text to segmentation, attention heatmaps reveal strong alignment between nouns and object boundaries, validating semantic grounding. Conversely, during mask-to-caption synthesis, the model exhibits contextual awareness—producing grammatically correct captions even for overlapping regions.

An ablation study demonstrates that removing the alignment loss (L_align) reduces mIoU by 8% and BLEU by 5.2 points, confirming its role in cross-modal consistency. Furthermore, increasing transformer depth beyond 12 layers yielded diminishing returns, suggesting optimal representation convergence.

VI. CONCLUSION

This paper introduces a unified multi-modal transformer for cross-modal synthesis between image captions and semantic segmentation masks. Through joint embedding alignment and dual-stream attention, our approach enables bi-directional generation with enhanced semantic and structural fidelity. Future extensions will integrate 3D scene understanding and temporal consistency for video-based caption-to-segmentation translation.

REFERENCES

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[2] Raymaekers, J., Verbeke, W., & Verdonck, T. (2021). Weight-of-evidence 2.0 with shrinkage and spline-binning. arXiv preprint arXiv:2101.01494. Retrieved from https://arxiv.org/abs/2101.01494

[3] Jain, M., & Srihari, A. (2024). Comparison of Machine Learning Algorithm in Intrusion Detection Systems: A Review Using Binary Logistic Regression. International Journal of Computer Science and Mobile Computing, Vol.13 Issue.10, October- 2024, pg. 45-53

[4] 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

[5] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

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

[7] Bauer, S., Wiest, R., Nolte, L. P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumour studies. *Physics in Medicine & Biology, 58*(13), R97–R129. https://doi.org/10.1088/0031-9155/58/13/R97

[8] 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

[9] Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*.

[10] 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

[11] Mohit Jain, Adit Shah (2024). Anomaly Detection Using Convolutional Neural Networks (CNN). ESP International Journal of Advancements in Computational Technology. https://www.espjournals.org/IJACT/2024/Volume2-Issue3/IJACT-V2I3P102.pdf

[12] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

[13] 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).

[14] Charron, O., Lallement, A., Jarnet, D., Noblet, V., Clavier, J. B., & Meyer, P. (2018). Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Computers in Biology and Medicine, 95*, 43–54. https://doi.org/10.1016/j.compbiomed.2018.02.004

[15] Kaushik, P., Jain, M., & Shah, A. (2018). A Low Power Low Voltage CMOS Based Operational Transconductance Amplifier for Biomedical Application.

[16] Jain, M., & Arjun Srihari. (2024b). Comparison of Machine Learning Models for Stress Detection from Sensor Data Using Long Short-Term Memory (LSTM) Networks and Convolutional Neural Networks (CNNs). International Journal of Scientific Research and Management (IJSRM), 12(12), 1775–1792. https://doi.org/10.18535/ijsrm/v12i12.ec02

[17] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-

M., & Larochelle, H. (2017). Brain tumour segmentation with deep neural networks. *Medical Image Analysis, 35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

[18] Mohit Jain, Arjun Srihari (2024). Comparison of Machine Learning Models for Stress Detection from Sensor Data Using Long Short-Term Memory (LSTM) Networks and Convolutional Neural Networks (CNNs). https://ijsrm.net/index.php/ijsrm/article/view/5912/3680 https://doi.org/10.18535/ijsrm/v12i12.ec02

[19] InsiderFinance Wire. (2021). Logistic regression: A simple powerhouse in fraud detection. Medium. Retrieved from https://wire.insiderfinance.io/logistic-regression-a-simple-powerhouse-in-fraud-detection-15ab984b2102

[20] 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

[21] Jain, M., & Arjun Srihari. (2024). Comparison of CAD Detection of Mammogram with SVM and CNN. Iconic Research and Engineering Journals, 8(6), 63–75. https://www.irejournals.com/paper-details/1706647

[22] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer, 18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

[23] 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

[24] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[25] 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

[26] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

[27] 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

[28] Bhat, N. (2019). Fraud detection: Feature selection-over sampling. Kaggle. Retrieved from https://www.kaggle.com/code/nareshbhat/fraud-detection-feature-selection-over-sampling

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

[30] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

[31] Kayalibay, Baris, et al. "CNN-Based Segmentation of Medical Imaging Data." ArXiv:1701.03056 [Cs], 25 July 2017, arxiv.org/abs/1701.03056.

[32] Shorten, Connor, and Taghi M. Khoshgoftaar. "A Survey on Image Data Augmentation for Deep Learning." Journal of Big Data, vol. 6, no. 1, 6 July 2019, journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0, https://doi.org/10.1186/s40537-019-0197-0.

[33] L. Wang, W. Chen, W. Yang, F. Bi and F. R. Yu, "A State-of-the-Art Review on Image Synthesis With Generative Adversarial Networks," in IEEE Access, vol. 8, pp. 63514-63537, 2020, doi: 10.1109/ACCESS.2020.2982224.

[34] 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

[35] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[36] L. Jen and Y.-H. Lin, “A Brief Overview of the Accuracy of Classification Algorithms for Data Prediction in Machine Learning Applications,” *Journal of Applied Data Sciences*, vol. 2, no. 3, pp. 84–92, 2021, doi: 10.47738/jads.v2i3.38.

[37] 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

[38] 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

[39] Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., & Ellison, D. W. (2016). The 2016 World Health Organization classification of tumours of the central nervous system: A summary. *Acta Neuropathologica, 131*(6), 803–820. https://doi.org/10.1007/s00401-016-1545-1

[40] 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

[41] S. A. Hicks et al., “On evaluation metrics for medical applications of artificial intelligence,” *Sci Rep*, vol. 12, no. 1, pp. 1–9, Dec. 2022, doi: 10.1038/s41598-022-09954-8.

[42] Pallud, J., Fontaine, D., Duffau, H., Mandonnet, E., Sanai, N., Taillandier, L., Peruzzi, P., Guillevin, R., Bauchet, L., Bernier, V., Baron, M.-H., Guyotat, J., & Capelle, L. (2010). Natural history of incidental World Health Organization grade II gliomas. *Annals of Neurology, 68*(5), 727–733. https://doi.org/10.1002/ana.22106

[43] Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumour segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging, 35*(5), 1240–1251. https://doi.org/10.1109/TMI.2016.2538465

[44] 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

[45] Patel, H., & Zaveri, M. (2011). Credit card fraud detection using neural network. International Journal of Innovative Research in Computer and Communication Engineering, 1(2), 1–6. https://www.ijircce.com/upload/2011/october/1_Credit.pdf

[46] 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).

[47] Alom, Md Zahangir, et al. “The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches.” ArXiv:1803.01164 [Cs], 12 Sept. 2018, arxiv.org/abs/1803.01164.

[48] Wang, Weibin, et al. “Medical Image Classification Using Deep Learning.” Intelligent Systems Reference Library, 19 Nov. 2019, pp. 33–51, https://doi.org/10.1007/978-3-030-32606-7_3.

[49] Nabati, R., & Qi, H. (2019). "RRPN: Radar Region Proposal Network for Object Detection in Autonomous Vehicles." 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 3093-3097, doi: 10.1109/ICIP.2019.8803392.

[50] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[51] 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

[52] Stupp, R., Taillibert, S., Kanner, A., Read, W., Steinberg, D. M., Lhermitte, B., Toms, S., Idbaih, A., Ahluwalia, M. S., Fink, K., Di Meco, F., Lieberman, F., Zhu, J.-J., Stragliotto, G., Tran, D. D., Brem, S., Hottinger, A., Kirson, E. D., Lavy-Shahaf, G., … Hegi, M. E. (2017). Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: A randomized clinical trial. *JAMA, 318*(23), 2306–2316. https://doi.org/10.1001/jama.2017.18718

[53] Raymaekers, J., Verbeke, W., & Verdonck, T. (2021). Weight-of-evidence 2.0 with shrinkage and spline-binning. arXiv preprint arXiv:2101.01494. Retrieved from https://arxiv.org/abs/2101.01494