DocParseNet: Advanced Semantic Segmentation and OCR Embeddings for Efficient Scanned Document Annotation

Ahmad Mohammadshirazi [1] **Ali Nosrati Firoozsalari** [1] **Mengxi Zhou** [1] **Dheeraj Kulshrestha** [2] **Rajiv Ramnath** [1]

Abstract

Automating the annotation of scanned documents is challenging, requiring a balance between computational efficiency and accuracy. DocParseNet addresses this by combining deep learning and multi-modal learning to process both text and visual data. This model goes beyond traditional OCR and semantic segmentation, capturing the interplay between text and images to preserve contextual nuances in complex document structures. Our evaluations show that DocParseNet significantly outperforms conventional models, achieving mIoU scores of 49.12 on validation and 49.78 on the test set. This reflects a 58% accuracy improvement over state-of-the-art baseline models and an 18% gain compared to the UNext baseline. Remarkably, DocParseNet achieves these results with only 2.8 million parameters, reducing model size by approximately 25 times and speeding up training by 5 times compared to other models. These metrics, coupled with a computational efficiency of 0.039 TFLOPs (BS=1), highlight DocParseNet's high performance in document annotation. The model's adaptability and scalability make it well-suited for real-world corporate document processing applications. The code is available at https://github.com/ahmad-shirazi/DocParseNet.

Keywords: Document AI, Multi-Modal Learning, Vision-and-Language, Document Annotation, document image understanding

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA [2]Flairsoft Company, Columbus, OH 43235, USA. Correspondence to: Ahmad Mohammadshirazi <mohammadshirazi.2@osu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Achieving efficient and precise annotation of data from scanned documents remains a formidable challenge in machine learning and computer vision. (Dutta et al., 2024; O'Mahony et al., 2020). These scanned images, rich with textual and graphical content, are crucial across various fields. Despite advancements in Optical Character Recognition (OCR) (Li et al., 2023b) and image processing, a significant gap persists between document digitization and true semantic understanding (Borchmann et al., 2021).

This gap is further widened by the complex nature of documents, where the structure of characters and their contextual meaning are intricately intertwined. Conventional methods like document layout analysis offer structural parsing (Shen et al., 2021; Zhou & Ramnath, 2022), while semantic analysis aims to uncover deeper meanings (Mercha & Benbrahim, 2023). Although template-based extraction efficiently harvests data from predefined fields in structured documents, these methods are insufficient for documents blending structured and unstructured data. The variability in document quality, format diversity, and layout complexity further challenges current methods, especially with documents exhibiting high variance in text density, font styles, and graphical elements, as well as poor scanning quality (Zhou et al., 2023). Additionally, the computational demand and need for large, annotated datasets for training models exacerbate these challenges, creating a bottleneck for efficient data extraction and processing (Lepcha et al., 2023; Liang et al., 2022).

A critical observation in current methodologies is their tendency to rely predominantly on either text or image data, often overlooking the comprehensive potential of utilizing both (Saharia et al., 2022). Solely text-based approaches can fail to appreciate the intricate layout and structural nuances of documents, while image-based methods might inadequately capture textual details. Multi-modal learning, which combines diverse data types such as images and text, significantly enhances the precision of automated annotation for complex tasks, albeit with higher computational costs. (Liu et al., 2024; Huang et al., 2022; Xu et al., 2021).

Recognizing this challenge, we introduce DocParseNet, a

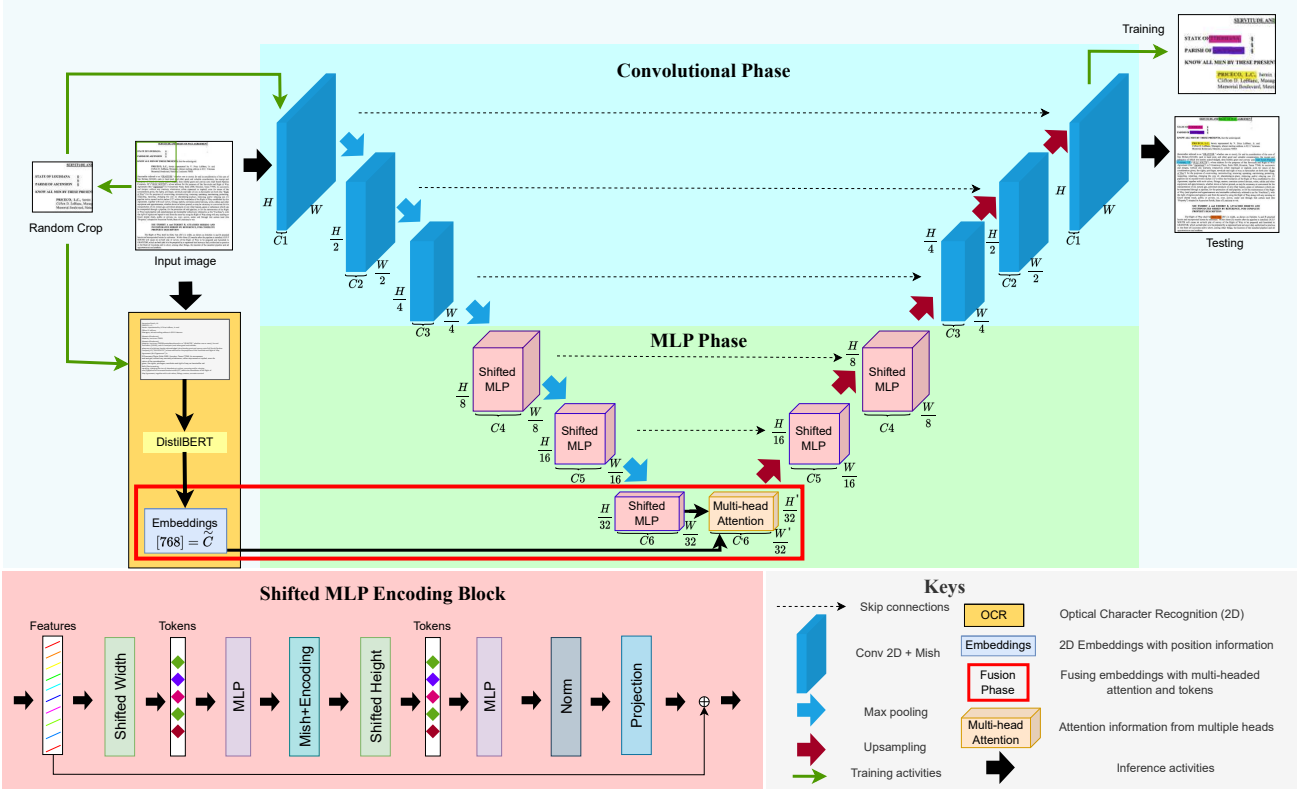


Figure 1. Diagram illustrating the DocParseNet architecture, emphasizing its four core modules: Pixel Level Module, utilizing convolutional layers for detailed visual feature extraction; MLP phase, distilling broader image features; Text Feature Module, employing OCR for textual data extraction and contextual understanding; and Fusion Module, synergistically integrating multi-modal features to enhance document parsing and prediction accuracy.

model specifically designed to integrate textual and visual data, leveraging the strengths of each modality to address the limitations inherent in single-type information processing. DocParseNet's architecture synergizes semantic segmentation with OCR embeddings to achieve comprehensive document annotation, not only increasing accuracy but also reducing computational costs.

The remainder of this paper is structured as follows: Section 2 reviews related work, Section 3 outlines the DocParseNet methodology, Section 4 presents experimental results and discussion, and Section 5 concludes with future work.

2. Related Work

Semantic segmentation, a cornerstone task in computer vision, necessitates a nuanced understanding of image details at multiple scales. Notable semantic segmentation models like SegFormer (Xie et al., 2021), LawinTransformer (Yan et al., 2022), and MaskFormer (Cheng et al., 2021), utilizing multi-scale feature extraction critical for intricate scene segmentation.

The UNet framework employs an encoder-decoder architecture with skip connections to preserve spatial information (Huang et al., 2020; Navard & Yilmaz, 2024). This approach benefits from convolutional neural networks' (CNNs') local feature extraction capabilities. Inspired by the UNet framework, UNeXt (Valanarasu & Patel, 2022) offers a less computationally intensive model by utilizing a Multi-Layer Perceptron (MLP) approach (Taud & Mas, 2018).

Building on the concept of global and local feature extraction exemplified by models like MedT (Valanarasu et al., 2021), VT-Unet (Peiris et al., 2022), and TransBTS (Wang et al., 2021), our model employs an efficient random window crop mechanism (Zhong et al., 2020). The integration of an innovative OCR stage for textual embeddings, combined with shifted MLPs and multi-head attention, enhances semantic comprehension (Fang et al., 2024; Huang et al., 2022).

Multi-modal learning, which merges vision and language, represents a vital research direction (Uppal et al., 2022). In document analysis, this approach enables systems to assimilate and interpret visual content alongside text, achieving a unified comprehension that surpasses uni-modal methods (Zhang et al., 2024).

Significant advancements include models like CLIP (Zhu et al., 2023), which uses a Vision Transformer (ViT) (Chen et al., 2023) for visual model learning from natural language supervision, and BLIP, which employs a ViT encoder for both understanding and generation tasks within vision-language pre-training (Li et al., 2023a). In document understanding, models like LayoutLM (Huang et al., 2022) and Docformer (Appalaraju et al., 2024) integrate text, layout, and image information to achieve state-of-the-art performance on visually-rich document tasks.

DocParseNet builds on these advancements by employing a modified UNet for visual feature extraction and DistilBERT (Sanh et al., 2019) for text embedding, ensuring efficiency without sacrificing language understanding. A specialized fusion module integrates these modalities seamlessly, providing an advanced multi-modal learning system with improved document interpretation capabilities.

3. DocParseNet

Our proposed model, DocParseNet (Figure 1), is a multi-modal deep learning architecture designed for parsing and annotating corporate agreement PDFs. It processes images derived from these PDFs, identifying key elements such as State, County, Agreement Title, Grantee Company, and Grantor Company. We trained DocParseNet on a dataset of 4,102 expert-annotated corporate agreement PDFs, comprising 19,821 images, split into training, validation, and testing sets using an 8-1-1 ratio. The non-static positioning of fields, influenced by the dynamic document structure, adds further complexity to the model's task. This heterogeneity, however, is essential, enriching the dataset and providing a robust foundation for efficient DocParseNet training.

3.1. DocParseNet_UNet

DocParseNet_UNet comprises a modified Convolutional Neural Network (CNN) for image processing and a Shifted MLP Encoding block for enhancing image features. This architecture is depicted in Figure 1.

CNNConv 2D+Mish: The CNN component of DocParseNet_UNet as demonstrated by equation 1 uses a series of 2D convolutional layers to extract hierarchical spatial features from input images, utilizing a 1x1 convolution (conv1x1) to refine feature maps between stages. The CNN employs the Mish activation function (Misra, 2019), which preserves more information compared to traditional activation functions like ReLU. Equation 2 takes the input image (I) and outputs a random crop (I′). The random crop is then passed onto the aforementioned CNN component, and the output is calculated.

$$X_{\text{shift}} = f_{\text{CNN}}(I') \tag{1}$$

$$I' = f_{\text{Random\_crop}}(I) \tag{2}$$

Shifted MLP Encoding Block: The Shifted MLP Encoding Block is designed to overcome the limitations of CNNs, particularly their localized receptive fields. By employing a shift operation in the MLP's receptive window across the feature map, it achieves cross-window connectivity, capturing broader contextual information efficiently. This structure enriches feature representation with contextual information spanning both local and distant regions of the input. The shifted MLP can be described as the following:

$$V_n = \text{Linear}\left(\text{Mish}\left(\text{DWConv}\left(W_1 \cdot X_{\text{shift}} + b_1\right)\right)\right), \tag{3}$$

where a Mish activation function is applied after Depth-Wise Convolution (DWConv) and a feed-forward neural network. Another feed-forward neural network is then applied to the output. Additional details are provided in Appendix A.1.

3.2. DocParseNet_OCR

The OCR phase in DocParseNet, shown in Figure 1, is designed for text recognition within document images. It uses optical character recognition to convert visual text into machine-encoded text, enabling the extraction of textual embeddings. This study utilizes a pre-trained DistilBERT model for its efficient text encoding capabilities. The OCR process includes Image-to-Text Conversion (f_{OCR}), text embedding using the Distilbert model (f_{BERT}), and [CLS] Token Embedding (E). For further details, please refer to Appendix A.2.

$$E = f_{\text{BERT}}(f_{\text{OCR}}(I')) \tag{4}$$

3.3. DocParseNet_Fusion

The fusion phase in DocParseNet (Figure 1) combines visual and textual information using multi-head attention. This mechanism aligns and merges features from the UNet and OCR streams, leveraging their complementary strengths for accurate document annotation.

In DocParseNet, the fusion process combines the last shifted MLP encoding block output with BERT embeddings via multi-head attention:

- Let $V_6 \in \mathbb{R}^{B \times C_6 \times \frac{H}{32} \times \frac{W}{32}}$ be the output from the last shifted MLP encoding block.

- Let $E \in \mathbb{R}^{B \times 768}$ be the embedding from the DistilBERT model.

The multi-head attention outputs a feature map $F \in \mathbb{R}^{B \times C'_6 \times \frac{H'}{32} \times \frac{W'}{32}}$ by:

Table 1. Test Set Performance and Efficiency of DocParseNet

Method	mIoU	AT	State	County	Grantor	Grantee	TFLOPs
Segformer-B5 (Xie et al., 2021)	34.81	37.36	38.72	44.28	8.81	44.87	0.39
UNext (Valanarasu & Patel, 2022)	42.04	36.16	53.81	52.01	24.48	43.73	0.06
DocParseNet	**49.78**	**43.06**	**65.66**	**53.90**	**36.14**	**50.12**	**0.04**

* AT: Agreement Title. For each field, we mark the highest score in bold.

$$F = f_{\text{MultiHead}}(V_6, E). \tag{5}$$

Placing the fusion at the UNet bottleneck integrates deep visual features with rich textual embeddings, creating a dense feature representation crucial for precise annotation, especially with limited training data.

DocParseNet combines UNet-based visual segmentation with OCR-based text recognition, enhanced by a fusion module for visual and textual modalities, to achieve high annotation precision. This process is detailed in Appendix A.3, Algorithm 1, demonstrating the model's robustness and efficiency (Figure 1). The hardware setup for experimentation is explained in Appendix A.4.

4. Result

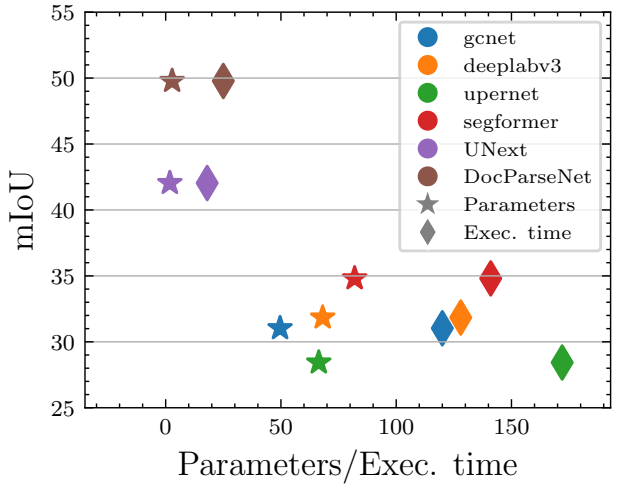


Figure 2. Parameters and Execution time

In our empirical study, we evaluated the proposed DocParseNet model by benchmarking it against leading semantic segmentation models, including GCNet (Cao et al., 2020), UPerNet (Xiao et al., 2018), DeeplabV3 (Chen et al., 2017), and Segformer (Xie et al., 2021), using the segmentation (Contributors, 2020; Perera et al., 2024) framework. These state-of-the-art models, optimized for image feature extraction, do not inherently account for text-based information.

The comparative analysis, detailed in Table 1 and Appendix B.2 (Table B1 and Figure B2), highlights DocParseNet's enhanced ability to incorporate text information, resulting in a significant performance boost. Our model's superior performance is evidenced by its mean Intersection over Union (mIoU) scores of 49.12 on the validation set and 49.78 on the test set. This demonstrates its robustness and superior generalization in handling the challenges of semantic segmentation and OCR embedding for scanned document annotation. Additionally, the training dynamics of DocParseNet, as shown in Appendix B.1, Figure B1, exhibit robust learning and generalization capabilities through a consistent upward trajectory in IoU scores and a decrease in loss, highlighting its resistance to overfitting.

As shown in Table 1, Table B1, and Figure B2 in Appendix B.1, DocParseNet achieved IoU scores of 39.77 on validation and 36.14 on testing for the 'Grantor' category, demonstrating a smaller discrepancy between validation and testing compared to baseline models. This robustness is due to our model's adept handling of varying boundary box locations across documents. The integration of OCR embeddings allows DocParseNet to maintain high accuracy despite spatial inconsistencies, showcasing its robustness and generalization capabilities. In the 'State' and 'County' categories, DocParseNet achieved exceptional performance with scores of 65.66 and 53.90, respectively. In the 'Agreement Title (AT)' and 'Grantee' categories, it maintained robust accuracy with scores of 43.06 and 50.12, respectively.

This adaptability is further enhanced by an empirically optimized channel configuration. Through extensive experimentation using Bayesian optimization, we determined the optimal channel sizes at each stage of the model to be [16, 32, 64, 128, 256, 320]. This configuration effectively captures the complex hierarchies of features necessary for accurate document segmentation.

Figure 2 and Table B2 in Appendix B.3 illustrate the performance evaluation of our baseline models and the proposed DocParseNet model. We utilize the HARP profiling framework (Vallabhajosyula & Ramnath, 2022) to gather model performance statistics, including Floating Point Operations (FLOPs) for a training step. A thorough examination consistently demonstrates DocParseNet's superior performance in accuracy and execution efficiency compared to all baselines. DocParseNet achieves this by efficiently leveraging reduced

parameters, leading to GPU memory savings, which allows for increased batch sizes or model fusion without compromising memory requirements or training latency. This efficiency enables our model to attain superior accuracies (Table 1) with reduced training latency, without incurring additional memory costs compared to competitive baselines.

5. Conclusion and Future

In summary, our work with DocParseNet establishes a new standard in document annotation, demonstrating exceptional performance and efficiency in semantic segmentation and OCR embedding for scanned documents. Our innovative framework, which utilizes deep learning and multi-modal learning, outperforms existing methods, evidenced by impressive mIoU scores. The accelerated training enhances reusability and facilitates frequent updates within a feedback loop, significantly reducing manual annotation efforts and improving accuracy. DocParseNet's economical computational resource usage and enhanced training speed highlight its potential for broad practical applications, particularly in domains requiring precise document analysis. Moving forward, we aim to expand the scope of DocParseNet to encompass additional domains in document analysis, leveraging the latest advancements in deep learning technology. Our focus will be on scaling up pre-trained models, enabling them to utilize extensive training data and further advancing SOTA results to enhance the applicability of our models in real-world business scenarios within the Document AI industry.

Acknowledgements

We extend our sincere gratitude to FlairSoft Company for their support. Additionally, we acknowledge the contributions of the OSU CETI Lab, with special thanks to Anuja Dixit and Manikya Swathi Vallabhajosyula, for their invaluable assistance.

References

Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., and Manmatha, R. Docformerv2: Local features for document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 709–718, 2024.

Borchmann, Ł., Pietruszka, M., Stanislawek, T., Jurkiewicz, D., Turski, M., Szyndler, K., and Graliński, F. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Cao, Y., Xu, J., Lin, S., Wei, F., and Hu, H. Global context networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Chen, M., Lin, M., Li, K., Shen, Y., Wu, Y., Chao, F., and Ji, R. Cf-vit: A general coarse-to-fine method for vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7042–7052, 2023.

Cheng, B., Schwing, A., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875, 2021.

Contributors, M. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Dutta, S., Adhikary, S., and Dwivedi, A. D. Visformers—combining vision and transformers for enhanced complex document classification. *Machine Learning and Knowledge Extraction*, 6(1):448–463, 2024.

Fang, H., Liang, D., and Xiang, W. Single-stage extensive semantic fusion for multi-modal sarcasm detection. *Array*, 22:100344, 2024.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1055–1059. IEEE, 2020.

Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4083–4091, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lepcha, D. C., Goyal, B., Dogra, A., and Goyal, V. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91:230–260, 2023.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102, 2023b.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., and Zou, J. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., and Bai, X. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

Mercha, E. M. and Benbrahim, H. Machine learning and deep learning for sentiment analysis across languages: A survey. *Neurocomputing*, 531:195–216, 2023.

Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.

Navard, P. and Yilmaz, A. A probabilistic-based drift correction module for visual inertial slams. *arXiv preprint arXiv:2404.10140*, 2024.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., and Walsh, J. Deep learning vs. traditional computer vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pp. 128–144. Springer, 2020.

Peiris, H., Hayat, M., Chen, Z., Egan, G., and Harandi, M. A robust volumetric transformer for accurate 3d tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 162–172. Springer, 2022.

Perera, S., Navard, P., and Yilmaz, A. Segformer3d: an efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4988, 2024.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., and Li, W. Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pp. 131–146. Springer, 2021.

Taud, H. and Mas, J.-F. Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, pp. 451–455, 2018.

Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., and Zadeh, A. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171, 2022.

Valanarasu, J. M. J. and Patel, V. M. Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 23–33. Springer, 2022.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pp. 36–46. Springer, 2021.

Vallabhajosyula, M. S. and Ramnath, R. Towards practical, generalizable machine-learning training pipelines to build regression models for predicting application resource needs on hpc systems. In *Practice And Experience In Advanced Research Computing*, pp. 1–5. Association for Computing Machinery, 2022.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pp. 109–119. Springer, 2021.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., and Wei, F. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*, 2021.

Yan, H., Zhang, C., and Wu, M. Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *arXiv preprint arXiv:2201.01615*, 2022.

Zhang, N., Cheng, H., Chen, J., Jiang, Z., Huang, J., Xue, Y., and Jin, L. M2doc: A multi-modal fusion approach for document layout analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7233–7241, 2024.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Zhou, H.-Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., and Li, W. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, 7(6):743–755, 2023.

Zhou, M. and Ramnath, R. A structure-focused deep learning approach for table recognition from document images. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 593–601. IEEE, 2022.

Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., and Gao, P. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2639–2650, 2023.

A. More Info-DocParseNet

A.1. Shifted MLP Encoding Block

Given an input feature map $x \in \mathbb{R}^{B \times C \times H \times W}$, where B is the batch size, C is the channel count, and H, W are the feature map's height and width, the Shifted MLP Encoding block executes the following operations:

1. A shift operation is applied along spatial dimensions (width) to rearrange the input tensor elements:

$$X_{\text{shift}} = \text{Shift}(x)$$

   where $\text{Shift}(\cdot)$ cyclically shifts elements within each channel.

2. The shifted output is processed through a linear layer:

$$x' = (W_1 \cdot X_{\text{shift}} + b_1)$$

   where $W_1 \in \mathbb{R}^{C \times C'}$, $b_1 \in \mathbb{R}^{C'}$ are the weights and biases of the first linear layer, and C' is the hidden feature dimension.

3. Integration of Depth-Wise Convolution (DWConv) and Mish Activation Function is applied for local spatial interaction:

$$x'' = \text{Mish}(\text{DWConv}(x'))$$

4. Dropout is then applied to the output of the Mish activation.

5. A shift operation is applied along spatial dimensions (height).

6. Finally, a second linear transformation is applied:

$$x''' = W_2 \cdot x'' + b_2$$

   where $W_2 \in \mathbb{R}^{C' \times C}$ and $b_2 \in \mathbb{R}^{C}$ are the weights and biases of the second linear layer.

A.2. OCR Component

The OCR component of DocParseNet converts the textual content within document images into a format suitable for embedding by a language model. This process includes the following steps:

1. The image-to-text conversion, typically performed by an OCR system such as Tesseract, can be denoted as a function $\text{OCR}(\cdot)$ that maps an image I to a string of text T:

$$T = \text{OCR}(I) \tag{6}$$

2. The extracted text T is encoded into embeddings by a pre-trained language model such as DistilBERT. Let the DistilBERT model be represented as $\mathcal{B}(\cdot)$. The process of obtaining the textual embeddings E is:

$$E = \mathcal{B}(T) \tag{7}$$

   which results in $E \in \mathbb{R}^{B \times 768}$, where B is the batch size.

3. The [CLS] token embedding, which serves as the aggregate representation of the input text, is extracted from the DistilBERT output to be used for downstream tasks:

$$E_{[\text{CLS}]} = E[:, 0, :] \tag{8}$$

   assuming that E is ordered such that the [CLS] token is the first in the sequence of embeddings.

A.3. DocParseNet Algorithm

To encapsulate, DocParseNet employs a cohesive methodology as delineated in Algorithm 1, which synergizes UNet-based visual feature segmentation with OCR-based text recognition. This integration is further enhanced by a fusion module adept at amalgamating the visual and textual modalities, yielding a harmonized system with heightened annotation precision. Such a synergistic approach is fundamental to DocParseNet's capability to deliver exceptional performance metrics, particularly notable given the limited training dataset. Algorithm 1 outlines the strategic process flow that enables DocParseNet to maintain high accuracy in document image analysis, demonstrating the model's robustness and efficacy, with its components and processes visualized in Figure 1.

Algorithm 1 DocParseNet Algorithm

Input: Document image dataset
Output: Set of annotated images with labels

```
Initialize DocParseNet with UNet and OCR components
for Each document image I in the dataset do
    // UNet-based Visual Feature Extraction
      I' ← Random crop(I)
      V ← CNN_Processing(I')
      V ← Apply Shifted MLP Encoding to V

    // OCR Text Extraction
      T ← OCR(I')
      E ← DistilBERT_Encoding(T)

    // Multi-Modal Fusion
      F ← MultiHeadAttention(V, E)
      R ← LayerNormalization(F + V)

    Output annotated image with labels from R
end
return Set of annotated images
```

A.4. Hardware Setup for Experimentation

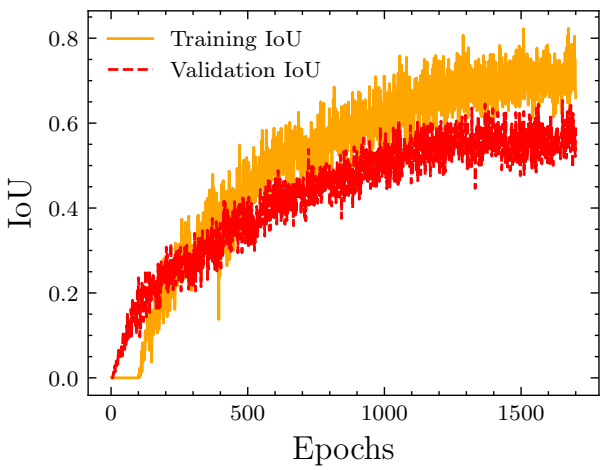
We conduct experiments on the Ascend cluster[1] at the Ohio Supercomputer Center (OSC) using single-node, single-GPU setups. Each node is equipped with 12 CPU cores, featuring AMD EPYC 7643 (Milan) processors clocked at 2.3 GHz, with a maximum memory utilization of 10GB per core (totaling 120GB across the cores). The GPU employed is the NVIDIA A100 with 80GB memory. Please refer to the Results section for insights into the performance of our model.

B. Additional Results

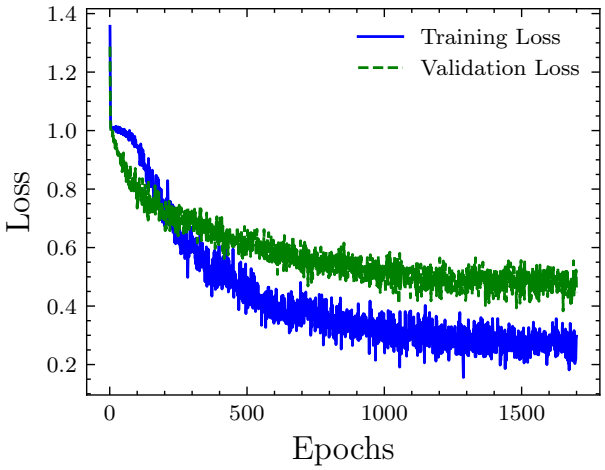
B.1. Training Dynamics

The training dynamics of DocParseNet, as shown in Figure B1, are characterized by a consistent upward trajectory in IoU scores for the 'State' category, affirming the model's capacity to learn and refine its segmentation capabilities over time. The convergence of training and validation IoU, alongside a corresponding decrease in loss, illustrates the model's ability to generalize beyond the training data. This performance is underscored by the minimal gap between training and validation metrics, highlighting the robustness of DocParseNet against overfitting and its adeptness in handling the variegated nature of document structures within the 'State' category.

[1] https://www.osc.edu/resources/technical_support/supercomputers/ascend



(a) IoU scores during training and validation phases.



(b) Training and validation loss curves over epochs.

Figure B1. Performance metrics of DocParseNet across 1700 epochs in the 'State' category. Subfigure (a) presents the IoU scores, while (b) illustrates the loss curves.

Table B1. Performance (IoU scores) of the proposed DocParseNet method against other competitive methods on the [training] / validation / (test set).

Method	mIoU	AT	State	County	Grantor	Grantee
GCNet (Cao et al., 2020)	[55.20] 30.18 (31.03)	[59.13] 26.29 (28.21)	[46.88] 32.72 (36.61)	[47.75] 34.63 (44.62)	[63.48] 22.47 (11.11)	[58.77] 34.78 (34.62)
UPerNet (Xiao et al., 2018)	[47.29] 25.98 (28.43)	[53.76] 26.22 (28.98)	[37.70] 25.65 (28.49)	[36.99] 25.94 (40.80)	[58.02] 19.28 (11.01)	[49.97] 32.83 (32.87)
DeeplabV3 (Chen et al., 2017)	[54.65] 30.16 (31.84)	[56.70] 28.36 (29.08)	[45.91] 31.36 (35.62)	[48.37] 30.82 (44.40)	[63.77] 23.98 (12.06)	[58.48] 36.26 (38.06)
Segformer-B5 (Xie et al., 2021)	[68.26] 35.23 (34.81)	[69.68] 34.53 (37.36)	[60.63] 34.78 (38.72)	[63.90] 39.69 (44.28)	[75.75] 30.67 (8.81)	[71.33] 36.48 (44.87)
UNext (Valanarasu & Patel, 2022)	[70.16] 41.65 (42.04)	[65.41] 35.53 (36.16)	[71.09] 51.47 (53.81)	[69.83] 47.69 (52.01)	[73.96] 31.51 (24.48)	[70.49] 42.04 (43.73)
DocParseNet	[71.79] **49.12** **(49.78)**	[69.89] **42.23** **(43.06)**	[78.13] **60.32** **(65.66)**	[71.27] **52.68** **(53.90)**	[68.06] **39.77** **(36.14)**	[71.59] **50.58** **(50.12)**

* AT: Agreement Title. For each field, we mark the highest score in bold.

Table B2. Performance metrics for our experiments in terms of memory usage and per-epoch execution time.

Model	**Params (M)**	**CPU Mem (GB)** max	**GPU Mem (GB)** max	**Model TFLOPs** (BS=1)	**Per Epoch (secs)** avg.
GCNet	49.62	2.61	12.20	1.1867	120
UPerNet	66.41	2.62	13.20	1.4207	172
DeeplabV3	68.1	2.62	12.60	1.6184	128
Segformer-B5*	81.97	2.85	11.5	0.3867	141
UNext	1.8	1.62	8.56	0.0546	18
DocParseNet	2.8	6.06	9.30	0.0391	25

All experiments utilize a batch size of 8, except for '*' (with a batch size of 4), and the maximum memory reflects this specific batch size requirement. TFLOPs are calculated for a batch size of 1.

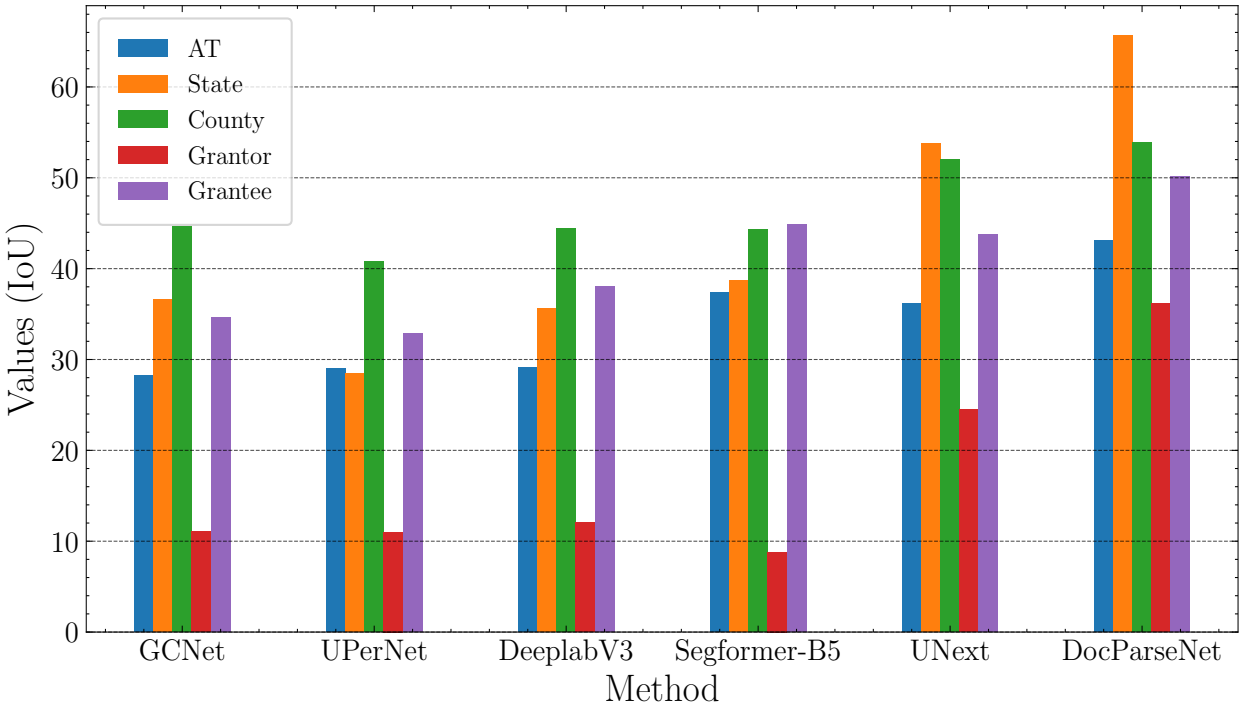


Figure B2. Assessment of DocParseNet compared to baseline methods across multiple datasets using IoU.

B.2. Comparative Analysis of DocParseNet's Superior IoU Performance

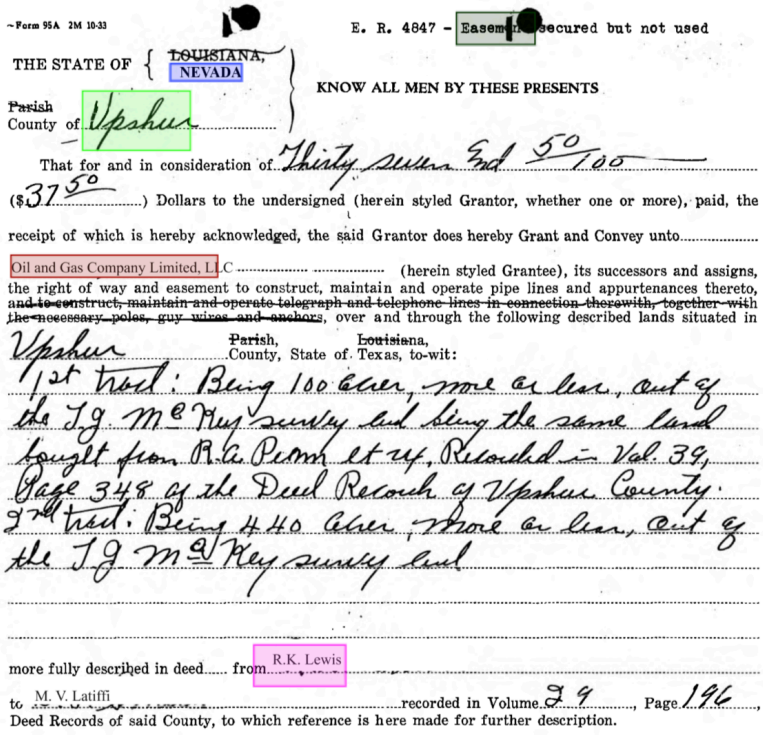
We provide a detailed comparative analysis of IoU scores, as shown in Table B1 and Figure B2, to evaluate the performance of DocParseNet relative to other advanced models. The results demonstrate that DocParseNet consistently outperforms the benchmarks set by GCNet, UPerNet, DeeplabV3, Segformer-B5, and UNext, with notable mIoU scores of 49.12 in validation and 49.78 in testing. Particularly in the 'State' and 'County' training categories, DocParseNet excels with scores of 65.66 and 53.90, respectively, showcasing its capability to effectively interpret the complex features of RoW documents. In the challenging 'Grantor' and 'Grantee' categories, DocParseNet maintains robust accuracy, with scores of 39.77 and 50.58, illustrating its proficiency in handling text location variability and document structural intricacies. The highest scores achieved by DocParseNet, highlighted in bold within the table, affirm its superior performance in extracting document information, substantiating the model's efficacy as detailed in the main body of this paper.

B.3. Computational Efficiency and Operational Metrics of DocParseNet

Table B2 delineates the computational efficiency and operational metrics of DocParseNet in relation to established semantic segmentation models. The table quantifies the models in terms of model parameters, memory usage, TFLOPs for one training

step, and execution time per epoch, underpinning the substantial computational advantages of DocParseNet. With only 2.8 million parameters, DocParseNet demonstrates an optimal balance between model complexity and performance latency, achieving a low average per-epoch execution time of 25 seconds. This is considerably faster than the other models, as DocParseNet uses 6.06 GB of CPU memory and 9.30 GB of GPU memory, which are within practical limits for deployment in real-world applications. These experiments, conducted with a batch size of 8, except for Segformer-B5, which uses a batch size of 4, showcase the model's ability to deliver high performance without the need for extensive computational resources.

B.4. Examples of Annotated Scanned Documents



~Form 95A 2M 10-33

E. R. 4847 - Easem[redacted]secured but not used

THE STATE OF { ~~LOUISIANA,~~ NEVADA

~~Parish~~ County of Upshur

KNOW ALL MEN BY THESE PRESENTS

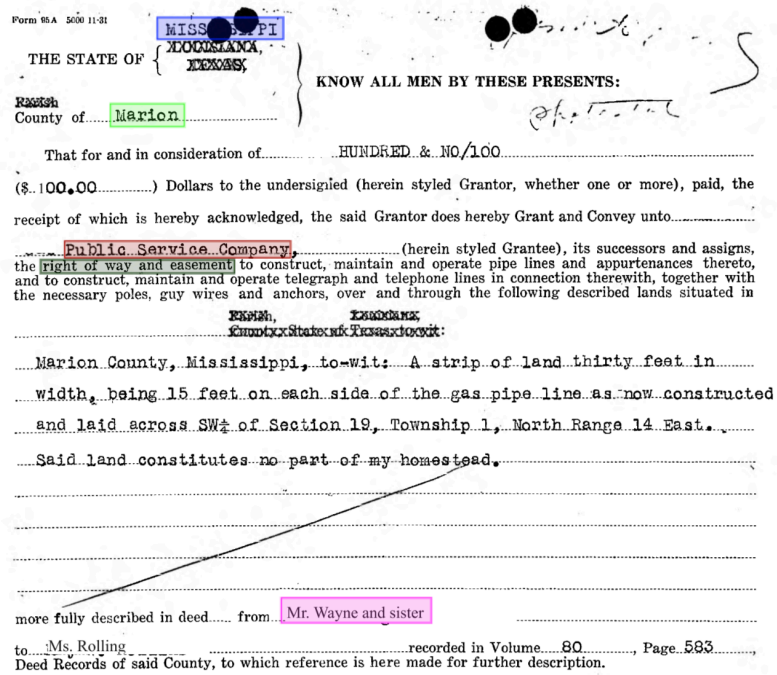
That for and in consideration of Thirty seven End 50/100 ($37.50) Dollars to the undersigned (herein styled Grantor, whether one or more), paid, the receipt of which is hereby acknowledged, the said Grantor does hereby Grant and Convey unto Oil and Gas Company Limited, LLC (herein styled Grantee), its successors and assigns, the right of way and easement to construct, maintain and operate pipe lines and appurtenances thereto, ~~and to construct, maintain and operate telegraph and telephone lines in connection therewith, together with the necessary poles, guy wires and anchors~~, over and through the following described lands situated in Upshur ~~Parish~~, County, State of ~~Louisiana~~, Texas, to-wit:

1st tract: Being 100 acres, more or less, out of the T.J. McKey survey and being the same land bought from R.A. Penn et al, Recorded in Vol. 39, Page 348 of the Deed Records of Upshur County.

2nd tract: Being 440 acres, more or less, out of the T.J. McKey survey and

more fully described in deed from R.K. Lewis to M. V. Latiffi recorded in Volume 29, Page 196, Deed Records of said County, to which reference is here made for further description.

(a) Annotated Sample 1



Form 95A 5000 11-31

THE STATE OF { MISS[redacted]PI ~~LOUISIANA,~~ ~~TEXAS,~~

~~Parish~~ County of Marion

KNOW ALL MEN BY THESE PRESENTS:

That for and in consideration of HUNDRED & NO/100 ($100.00) Dollars to the undersigned (herein styled Grantor, whether one or more), paid, the receipt of which is hereby acknowledged, the said Grantor does hereby Grant and Convey unto Public Service Company, (herein styled Grantee), its successors and assigns, the right of way and easement to construct, maintain and operate pipe lines and appurtenances thereto, and to construct, maintain and operate telegraph and telephone lines in connection therewith, together with the necessary poles, guy wires and anchors, over and through the following described lands situated in ~~Parish~~, ~~County, State of Texas, to-wit~~:

Marion County, Mississippi, to-wit: A strip of land thirty feet in width, being 15 feet on each side of the gas pipe line as now constructed and laid across SW¼ of Section 19, Township 1, North Range 14 East. Said land constitutes no part of my homestead.

more fully described in deed from Mr. Wayne and sister to Ms. Rolling recorded in Volume 80, Page 583, Deed Records of said County, to which reference is here made for further description.

(b) Annotated Sample 2

Figure B3. Annotated samples demonstrating different perspectives