DiT

A Preprint

Lately, Image Transformers became SOTA in field of detection, segmentation etc. However, for document image understanding, there is no commonly used large-scale human-labeled benchmark like ImageNet, which makes large-scale supervised pre-training impractical so models for working with documents. Also, there is a problem that in real world documents ofter have format that differs from ones used in these datasets, so models lack domain adaptation. The architecture is inspired by BEiT(BERT Pre-Training of Image Transformers). First of all, the input is resized to 224×224 image, then it is split into a sequence of 16×16 patches which are used as the input to the image Transformer. Distinct from BEiT wher discrete VAE in DALL-E is used, authors suggest using d-VAE, so that the generated visual tokens are more domain relevant to the Document AI tasks. The pre-training objective is to recover visual tokens from dVAE based on the corrupted input document images using the Masked Image Modeling (MIM) in BEiT. In this way, the DiT model does not rely on any human-labeled document images, but only leverages large-scale unlabeled data to learn the global patch relationship within each document image. Sequence of patches is fed into a ViT. Later, output of Transformer encoder is used as the representation of image patches. During pre-training, DiT accepts the image patches as input and predicts the visual tokens with the output representation. To effectively pre-train authors randomly mask a subset of inputs with a special token [MASK] given a sequence of image patches. The DiT encoder embeds the masked patch sequence by a linear projection with added positional embeddings, and then contextualizes it with a stack of Transformer blocks. The model is required to predict the index of visual tokens with the output from masked positions. Instead of predicting the raw pixels, the masked image modeling task requires the model to predict the discrete visual tokens obtained by the image tokenizer.

After that during fine tuning authors split all tasks in two: classification and detection. For classification they use average pooling to aggregate the representation of image patches and after that they pass the global representation into a simple linear classifier. For detection:

The detection frameworks employed are Mask R-CNN and Cascade R-CNN, with ViT-based models serving as the backbone. The implementation is built upon the Detectron2 framework. To bridge the gap between the single-scale nature of ViT and the multi-scale requirements of the FPN, resolution-modifying modules are incorporated at four distinct transformer blocks.

Let d denote the total number of transformer blocks. At the $\frac{d}{3}$-th block, the feature map is upsampled by a factor of 4 using two consecutive stride-two 2×2 transposed convolution layers. At the $\frac{d}{2}$-th block, a single stride-two 2×2 transposed convolution is applied to achieve a $2\times$ upsampling. The output from the $\frac{2d}{3}$-th block is passed directly without modification. Lastly, the output from the final $\frac{3d}{3}$-th block is downsampled by a factor of 2 using a stride-two 2×2 max pooling operation.

State-of-the-Art Performance

DiT consistently achieves new state-of-the-art (SOTA) results across a wide range of vision-based Document AI tasks. This demonstrates the effectiveness of its self-supervised pre-training strategy in learning robust and generalizable document image representations.

Performance Across Key Document AI Tasks

- Document Image Classification (RVL-CDIP Dataset): DiT-B (the base version) significantly outperforms all selected single-model baselines. Moreover, DiT-L (the larger version) achieves scores

that are comparable to or better than previous SOTA ensemble models. These results highlight DiT's strong modeling capabilities for classifying document types based solely on image content.

- Document Layout Analysis (PubLayNet Dataset): DiT demonstrates superior performance in analyzing structural document elements. Both DiT-B and DiT-L achieve substantially higher mAP scores compared to strong baselines such as ResNeXt, DeiT, BEIT, and MAE, particularly in challenging categories like List and Figure. The use of advanced detection architectures such as Cascade R-CNN further enhances the results.

- Table Detection (ICDAR 2019 cTDaR Dataset): DiT outperforms most baseline models across both archival and modern subsets of the dataset. This showcases its strong few-shot learning ability in low-resource scenarios. On the combined dataset, DiT-L achieves the highest weighted F1 score among all Mask R-CNN-based approaches, confirming its versatility across document categories and its capability for fine-grained object detection.

- Text Detection for OCR (FUNSD Dataset): When used as a backbone for text detection, DiT models establish new SOTA performance levels. These results are further improved when DiT is augmented with synthetic training data.

In summary, DiT's self-supervised pre-training on large-scale unlabeled data proves highly effective, enabling strong generalization and state-of-the-art performance across diverse Document AI benchmarks. The public availability of code and pre-trained models further contributes to the advancement of the field by facilitating reproducibility and follow-up research.

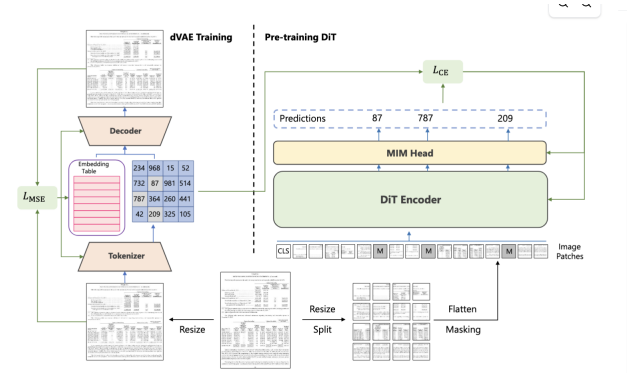
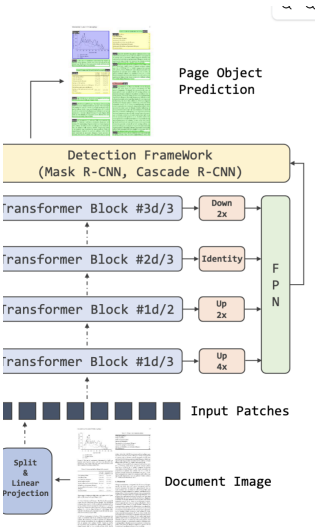


Figure 2: The model architecture of DiT with MIM pre-training.



Model	IoU@0.6	IoU@0.7	IoU@0.8	IoU@0.9	WAvg. F1
1st place in cTDaR	96.97	95.99	95.14	90.22	94.23
ResNeXt-101-32×8d	96.42	95.99	95.15	91.36	94.46
DeiT-B	96.26	95.56	94.57	90.91	94.04
BEiT-B	96.82	96.40	95.41	92.44	95.03
MAE-B	96.86	96.31	95.05	91.57	94.66
DiT-B	96.75	96.19	95.62	93.36	95.30
DiT-L	**97.83**	**97.41**	96.29	92.93	95.85
ResNeXt-101-32×8d (Cascade)	96.54	95.84	95.13	92.87	94.90
DiT-B (Cascade)	97.20	96.92	96.78	94.26	96.14
DiT-L (Cascade)	97.68	97.26	**97.12**	**94.74**	**96.55**

(a) Table detection accuracy on ICDAR 2019 cTDaR (combined: archival+modern)

Model	IoU@0.6	IoU@0.7	IoU@0.8	IoU@0.9	WAvg. F1
1st place in cTDaR	97.16	96.41	95.27	91.12	94.67
ResNeXt-101-32×8d	96.60	96.60	95.09	91.70	94.73
DeiT-B	97.54	97.16	96.41	92.63	95.68
BEiT-B	**98.10**	**98.10**	95.82	94.30	96.35
MAE-B	97.54	97.54	96.03	94.14	96.12
DiT-B	97.53	97.15	96.02	94.88	96.24
DiT-L	97.53	97.15	96.39	95.26	96.46
ResNeXt-101-32×8d (Cascade)	96.76	96.38	95.24	93.71	95.35
DiT-B (Cascade)	96.97	96.97	96.97	95.83	96.63
DiT-L (Cascade)	97.34	97.34	**97.34**	**96.20**	**97.00**

(b) Table detection accuracy on ICDAR 2019 cTDaR (archival)

Model	IoU@0.6	IoU@0.7	IoU@0.8	IoU@0.9	WAvg. F1
1st place in cTDaR	96.86	95.74	95.07	89.69	93.97
ResNeXt-101-32×8d	96.30	95.63	95.18	91.15	94.30
DeiT-B	95.51	94.61	93.48	89.89	93.07
BEiT-B	96.06	95.39	95.16	91.34	94.25
MAE-B	96.47	95.58	94.48	90.07	93.81
DiT-B	96.29	95.61	95.39	92.46	94.74
DiT-L	**98.00**	**97.56**	96.23	91.57	95.50
ResNeXt-101-32×8d (Cascade)	96.41	95.52	95.07	92.38	94.63
DiT-B (Cascade)	97.33	96.89	96.67	93.33	95.85
DiT-L (Cascade)	97.89	97.22	**97.00**	**93.88**	**96.29**

(c) Table detection accuracy on ICDAR 2019 cTDaR (modern)