Supplementary –
One-Shot Doc Snippet Detection: Powering Search in Document Beyond Text

Abhinav Java*, Shripad Deshmukh*, Milan Aggarwal, Surgan Jandial, Mausoom Sarkar, and Balaji Krishnamurthy

Adobe Media and Data Science Research Labs, Noida, India
{ajava, shdeshmu, milaggar, jandial, msarkar, kbalaji}@adobe.com

1. Additional Architectural Details

1.1. Feature Extractors.

In this section, we provide additional details about the backbones in the proposed MONOMER. We discuss the hyperparameter choices in the *Image Encoder*, followed by the *Text Encoder* and finally the *Layout Encoder*. Please note that the hyperparameters are kept consistent across experiments on all the datasets.

1.1.1 Image Features

The image encoder is a DiT-backbone with encoder-only architecture having 4 layers, each containing 4 attention heads with model dimension of 512. The encoder takes 3 channel document image resized (using bi-cubic interpolation) to 224×224 resolution which is further cut into 16×16 sized patches and outputs a token sequence of length 197. The 197 tokens are formed as follows – $\frac{224 \times 224}{16 \times 16} + 1$, where the additional token corresponds to the CLS token as in the original BEiT [1]. We choose a pretrained DiT base model for our experiments that has a hidden dimension of 768. Since both query image Q_i^{inp} and target image T_i^{inp} are preprocessed to the same dimension, we obtain two feature vectors Q_v, T_v each of size $BS \times 197 \times 1024$, where 1024 is the maximum sequence length and BS denotes the batch size. Note that the maximum sequence length is a hyperparameter choice that is chosen based on the maximum number of text-blocks in the target document. The encodings are then padded to final vectors Q_v, T_v of size $BS \times 1024 \times 1024$ each. The rationale behind doing so is to conveniently be able to perform the subsequent cross-attention with different modalities. We can summarize the sequence of operations as follows -

$$Q_v = pad(D(Q_i^{inp})) \in \mathcal{R}^{1024 \times 1024} \tag{1}$$

$$T_v = pad(D(T_i^{inp})) \in \mathcal{R}^{1024 \times 1024} \tag{2}$$

where D is the DiT image encoder and pad is the padding operation.

1.1.2 Text Features

We use a pretrained BeRT-based sentence transformer [5] that generates 768 dimensional embedding for a given block of text. The continuous blocks of text in the document are fed into this encoder to generate token sequence T_t^{inp}, Q_t^{inp} of dimension $BS \times text_t \times 768$ and $BS \times text_q \times 768$ respectively, where $text_t$ is the number of text-blocks in the target document and $text_q$ is the number of text-blocks in the query patch. Additionally, we pad both T_t^{inp}, Q_t^{inp} to a constant size of $BS \times 1024 \times 768$. Unlike all the other MONOMER parameters, the text encoder weights are kept *frozen*. Mathematically, text encoding is represented as follows –

$$Q_t = B(pad(Q_t^{inp})) \in \mathcal{R}^{1024 \times 768} \tag{3}$$

$$T_t = B(pad(T_t^{inp})) \in \mathcal{R}^{1024 \times 768} \tag{4}$$

where Q_t, T_t are the final query and target text features and B is the BeRT text encoder.

1.1.3 Bounding Box Features

We leverage a ViT-like [6] architecture to encode the bounding box (spatial) information in the target document and query patch. We implement an encoder-only transformer architecture with 4 layers, 4 heads and hidden dimension of 1024. It takes bounds of the target T_s^{inp} and query Q_s^{inp} of size $BS \times box_t \times 4$, $BS \times box_q \times 4$, where box_t, box_q are the

*These authors contributed equally to this work number of bounding boxes in target and query respectively. Similar to the text-encoder, box_t and box_q are padded to the maximum sequence length of 1024. Weights of this encoder are initialized randomly. We denote the bounding box encoding as follows –

$$Q_s = V(pad(Q_s^{inp})) \in \mathcal{R}^{1024\times1024} \tag{5}$$

$$T_s = V(pad(T_s^{inp})) \in \mathcal{R}^{1024\times1024} \tag{6}$$

where V is the ViT-like bounding box encoder and Q_s, T_s are the final feature sets corresponding the query and target respectively.

1.2. Feature Fusion

1.2.1 Symmetric Attention Module.

The symmetric attention module consists of 2 multi-head attention (MHA) modules each containing 4 heads and embedding dimension of 512. To ensure that the input token feature dimension matches with MHA's specifications, the input sequences are passed through fully-connected layers to project feature dimension onto dimension of 512. The outputs of the MHA blocks are concatenated (along last dimension) to obtain final token sequence with feature dimension of $BS \times 1024 \times 1024$.

1.2.2 Co-Attention and Cross-Attention Modules.

Co-Attention Module contains 3 symmetric attention modules one for each modality, outputting sequences VV, TT and SS of length 1024 and token size 1024.

$$VV = SA(Q_v, T_v) \in \mathcal{R}^{1024\times1024} \tag{7}$$

$$TT = SA(Q_t, T_t) \in \mathcal{R}^{1024\times1024} \tag{8}$$

$$SS = SA(Q_s, T_s) \in \mathcal{R}^{1024\times1024} \tag{9}$$

where SA is the Symmetric Attention operation.

Similarly, the Cross-Attention Module consists of 2 symmetric attention modules for generating spatio-visual features and 2 for attending text over those generated features. It generates $S_qV_tT_t$ and $S_tV_qT_q$ the dimensions of which are, once again, length of 1024 and token size of 1024. Finally we concatenate the outputs of Co-Attention and Cross-Attention blocks to create feature volume F_{sim}. Obtain F_{sim} as follows –

$$S_qV_t = SA(Q_s, T_v) \in \mathcal{R}^{1024\times1024} \tag{10}$$

$$S_qV_tT_t = SA(S_qV_t, T_t) \in \mathcal{R}^{1024\times1024} \tag{11}$$

$$S_tV_q = SA(T_s, Q_v) \in \mathcal{R}^{1024\times1024} \tag{12}$$

$$S_tV_qT_q = SA(S_tV_q, Q_t) \in \mathcal{R}^{1024\times1024} \tag{13}$$

$$F_{sim} = concat(VV, TT, SS, S_qV_tT_t, S_tV_qT_q) \tag{14}$$

where $F_{sim} \in \mathcal{R}^{1024\times1024}$ is the final set of features and are processed as described in the main paper.

1.3. Bounding Box Detection

The F_{sim} is passed through a linear layer followed by a sequence of 4 convolutional layers to produce features which are reshaped to give outputs at 4 different levels as described in the main paper. Then we apply a standard FPN [3] to obtain features at a common representation size of 1024. Finally, we generate proposals most similar to the query through an RPN [2] and subsequently detect bounding boxes using RoI Heads. We choose the default parameters for the RPN and RoI heads (from [4]) –

- RPN NMS threshold = 0.7
- RPN IOU threshold = 0.7 (FG), 0.30 (BG)
- RPN Score Threshold = 0
- ROI NMS Threshold = 0.40
- ROI Score Threshold = 0.05
- Detections per image = 200
- ROI IOU threshold = 0.50 (FG), 0.50 (BG)

where FG, BG are the foreground and background respectively. Note that the bounding box detection for LayoutLMv3 baseline is kept exactly the same as the proposed MONOMER whereas for BHRL, the github implementation is used.

1.4. Model size comparison vs baselines

On comparing the model sizes of the selected baselines with our proposed model, we found that BHRL has $48M$ parameters but is considerably worse at task performance compared to LayoutLMv3 ($126M$) and MONOMER ($146M$). Further, despite the comparable number of parameters in LayoutLMv3, our method outperforms it significantly ($+20\%$ in mAP).

2. Human Evaluation

In this section, we delineate the human evaluation conducted on the generated dataset through the proposed technique. A summary of the results have been tabulated in Table 1. We create 4 dataset split containing 40 samples each and share each split with 3 human evaluators to report the metrics. The high recall and precision over all the splits indicates that not only does our method generate high quality ground truths (87.96%) but is also able to find most of the target regions (81.07%) in a given document corresponding to a particular query. This saves a considerable amount of human annotation costs while maintaining reliability. Further, we also note that a substantial number of samples (48.12%) over all splits are complex and hard to search for in a document. While this metric is largely subjective, its

Metrics	Split-1	Split-2	Split-3	Split-4	Average over Splits
Recall	87.25	71.93	80.74	84.38	81.07
Precision	83.06	87.83	91.41	89.58	87.96
F1	84.28	78.56	85.55	86.48	83.71
% Complex Pattern	43.33	39.16	80.0	30.0	48.12
% non exact but similar matches over correct matches highlighted.	79.61	93.07	91.09	86.17	87.48

Table 1: Human Evaluation of the proposed dataset over 4 different splits. The last row indicates the non-trivial nature of dataset generation through the percentage of examples that are not exactly the same as the query.

	AP50	*AR50*	*AP75*	*AR75*	*mAP*
Heuristic	**89.94**	41.70	64.81	29.89	44.39
MONOMER	81.59	**59.88**	**73.17**	**52.78**	**56.49**

Table 2: Performance comparison between data generation heuristic as a baseline and MONOMER on human-annotated data

consistency over multiple splits verifies our claim. Finally, the last row shows that query matching within targets is non-trivial such that only 12.42% cases where snippet and highlighted similar regions are exact matches with the rest of the non-trivial matches containing the same layout but possibly different variation, text, fonts etc which allows the model to learn ”advanced one-shot search capabilities”.

Hard samples (question 3 in human evaluation) The evaluators were asked to decide if a snippet is hard based on whether its structure is complex - a) the snippet comprises of a significant number of elements or that the elements are arranged in a complex layout. Further, they were instructed to judge a snippet as hard b) if they would find it difficult to search a given query snippet in some target document.

3. Dataset Generation

Link with real world data We asked the human evaluators to annotate 160 test samples (used in human study) to compare the performance of our data generation heuristic (as baseline) to that of our proposed MONOMER. The results in Table 2 indicate that MONOMER performs better than heuristic and demonstrates its ability to perform well on a real dataset, highlighting the effectiveness of the proposed data generation technique to enable MONOMER to generalize better. It is apparent in comparison with the heuristic-based baseline in Table 2 that training our model on this dataset enables us to obtain a more generalized one-shot detector. We hypothesize that despite our heuristic not finding an exhaustive set of matches (81% recall) because it is more stringent and based on layout-based matching, the data generated by our dataset generation heuristic contains enough accurate samples for the model to learn snippet similarity. We also observe in Table 2 that the average precision (@50) of the heuristic remains high, and recall is slightly lower, verifying our hypothesis. Further, at a higher threshold of 0.75, MONOMER performs better in AP75, AR75, and mAP.

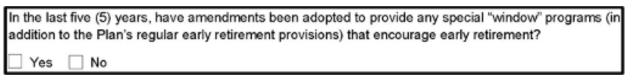
Limitations of dataset generation It is noteworthy to mention that the dataset generation method relies on a heuristic (metric Eq. 1 in main paper) and consequently encounters some failure cases. In this section, we discuss some of these failure cases using qualitative examples. The quantitative efficacy of our method is already elucidated in Table 1. A summary of failure cases is presented in Fig 1-Fig 2. Particularly, in Fig 1 A the crop only contains a single ”Yes/No” choice option, however it considers elements in the target with multiple structures within it. This happens as a result of the threshold penalty and a less strict size heuristic. The additional elements are considered as single blocks of text or fillable areas which our metric fails to capture. Further, in Fig 1 B the heuristic is unable to discern the difference between a block of comma separated text and option-like text. This can be attributed to lack of detailed annotations (of low level elements) in the dataset, typically ”option-like” text is annotated as a list which wasn't correctly done in this instance. Lastly, in Fig 2 C the heuristic fails to consider the length of the text and ends up matching a large block with a smaller target. This can be regulated using the size penalty. However, we reiterate that these examples occur sporadically and the same is justified through a quantitative analysis in Table 1.

4. Additional Qualitative Results

We add more results produced by the proposed MONOMER in this section. Please refer to Fig. 3 to Fig. 7 for the predictions.

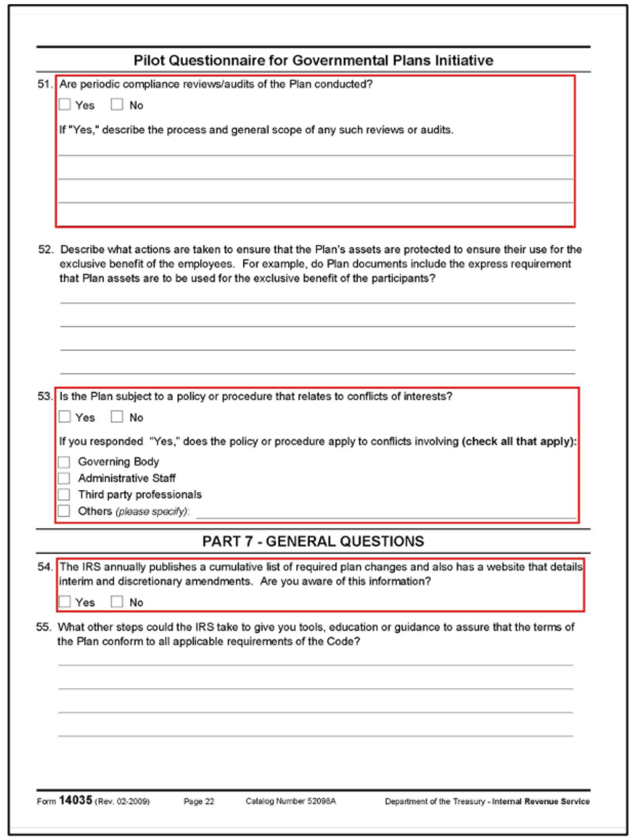
References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*,


In the last five (5) years, have amendments been adopted to provide any special "window" programs (in addition to the Plan's regular early retirement provisions) that encourage early retirement?

☐ Yes ☐ No

(a) Crop A


Pilot Questionnaire for Governmental Plans Initiative

51. Are periodic compliance reviews/audits of the Plan conducted?

☐ Yes ☐ No

If "Yes," describe the process and general scope of any such reviews or audits.

52. Describe what actions are taken to ensure that the Plan's assets are protected to ensure their use for the exclusive benefit of the employees. For example, do Plan documents include the express requirement that Plan assets are to be used for the exclusive benefit of the participants?

53. Is the Plan subject to a policy or procedure that relates to conflicts of interests?

☐ Yes ☐ No

If you responded "Yes," does the policy or procedure apply to conflicts involving **(check all that apply)**:

☐ Governing Body
☐ Administrative Staff
☐ Third party professionals
☐ Others *(please specify)*:

PART 7 - GENERAL QUESTIONS

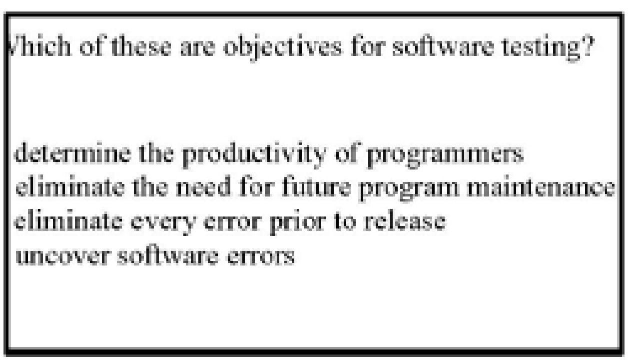
54. The IRS annually publishes a cumulative list of required plan changes and also has a website that details interim and discretionary amendments. Are you aware of this information?

☐ Yes ☐ No

55. What other steps could the IRS take to give you tools, education or guidance to assure that the terms of the Plan conform to all applicable requirements of the Code?

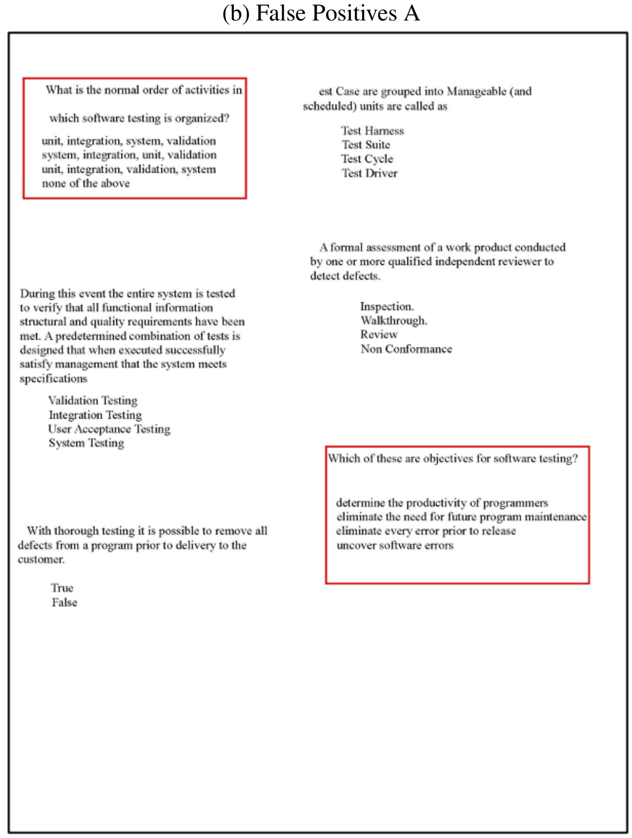
Form **14035** (Rev. 02-2009) Page 22 Catalog Number 52098A Department of the Treasury - **Internal Revenue Service**

(b) False Positives A


Vhich of these are objectives for software testing?

determine the productivity of programmers
eliminate the need for future program maintenance
eliminate every error prior to release
uncover software errors

(c) Crop B


What is the normal order of activities in which software testing is organized?

unit, integration, system, validation
system, integration, unit, validation
unit, integration, validation, system
none of the above

est Case are grouped into Manageable (and scheduled) units are called as

Test Harness
Test Suite
Test Cycle
Test Driver

A formal assessment of a work product conducted by one or more qualified independent reviewer to detect defects.

Inspection.
Walkthrough.
Review
Non Conformance

During this event the entire system is tested to verify that all functional information structural and quality requirements have been met. A predetermined combination of tests is designed that when executed successfully satisfy management that the system meets specifications

Validation Testing
Integration Testing
User Acceptance Testing
System Testing

Which of these are objectives for software testing?

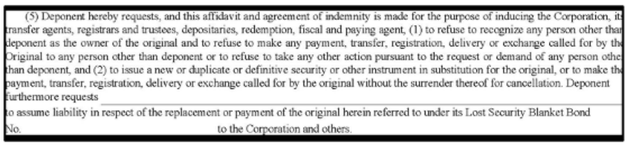
determine the productivity of programmers
eliminate the need for future program maintenance
eliminate every error prior to release
uncover software errors

With thorough testing it is possible to remove all defects from a program prior to delivery to the customer.

True
False

(d) False Positives B

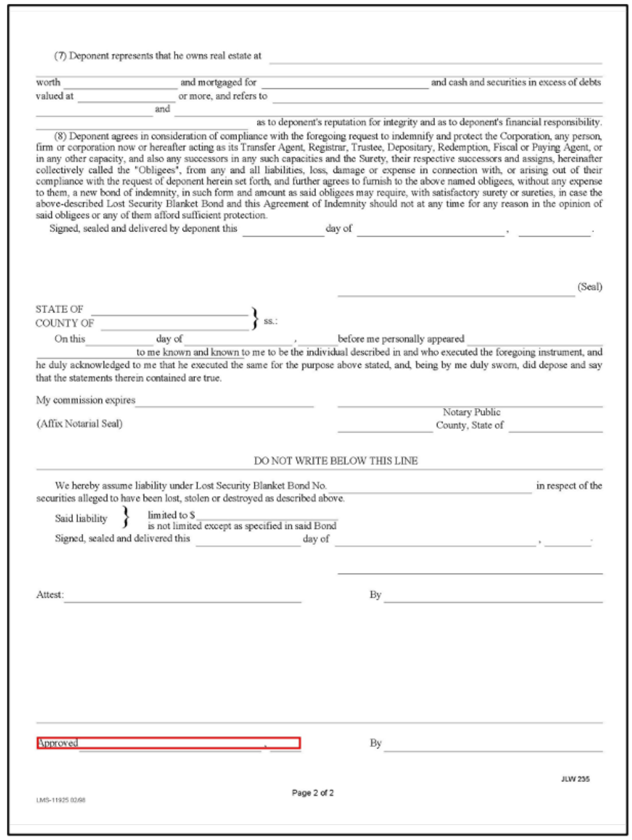
Figure 1: False positive examples (1)



(5) Deponent hereby requests, and this affidavit and agreement of indemnity is made for the purpose of inducing the Corporation, its transfer agents, registrars and trustees, depositaries, redemption, fiscal and paying agent, (1) to refuse to recognize any person other than deponent as the owner of the original and to refuse to make any payment, transfer, registration, delivery or exchange called for by the Original to any person other than deponent or to refuse to take any other action pursuant to the request or demand of any person other than deponent, and (2) to issue a new or duplicate or definitive security or other instrument in substitution for the original, or to make the payment, transfer, registration, delivery or exchange called for by the original without the surrender thereof for cancellation. Deponent furthermore requests

to assume liability in respect of the replacement or payment of the original herein referred to under its Lost Security Blanket Bond No. to the Corporation and others.

(a) Crop C



(7) Deponent represents that he owns real estate at

worth and mortgaged for and cash and securities in excess of debts valued at or more, and refers to and as to deponent's reputation for integrity and as to deponent's financial responsibility.

(8) Deponent agrees in consideration of compliance with the foregoing request to indemnify and protect the Corporation, any person, firm or corporation now or hereafter acting as its Transfer Agent, Registrar, Trustee, Depositary, Redemption, Fiscal or Paying Agent, or in any other capacity, and also any successors in any such capacities and the Surety, their respective successors and assigns, hereinafter collectively called the "Obligees", from any and all liabilities, loss, damage or expense in connection with, or arising out of their compliance with the request of deponent herein set forth, and further agrees to furnish to the above named obligees, without any expense to them, a new bond of indemnity, in such form and amount as said obligees may require, with satisfactory surety or sureties, in case the above-described Lost Security Blanket Bond and this Agreement of Indemnity should not at any time for any reason in the opinion of said obligees or any of them afford sufficient protection.

Signed, sealed and delivered by deponent this day of , .

(Seal)

STATE OF
COUNTY OF } ss.:

On this day of , before me personally appeared to me known and known to me to be the individual described in and who executed the foregoing instrument, and he duly acknowledged to me that he executed the same for the purpose above stated, and, being by me duly sworn, did depose and say that the statements therein contained are true.

My commission expires

Notary Public

(Affix Notarial Seal) County, State of

DO NOT WRITE BELOW THIS LINE

We hereby assume liability under Lost Security Blanket Bond No. in respect of the securities alleged to have been lost, stolen or destroyed as described above.

Said liability } limited to $
is not limited except as specified in said Bond

Signed, sealed and delivered this day of , .

Attest: By

Approved , By

JLW 235

Page 2 of 2

LMS-11925 02/98

(b) False Positives C

Figure 2: False positive examples (2)

2021.

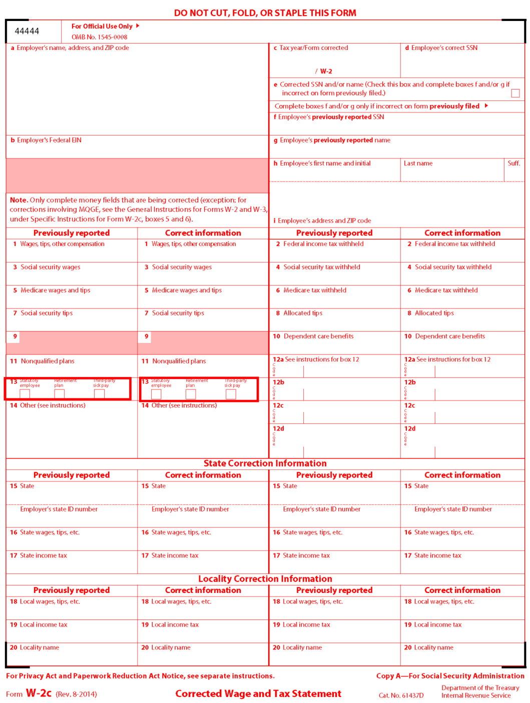
[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

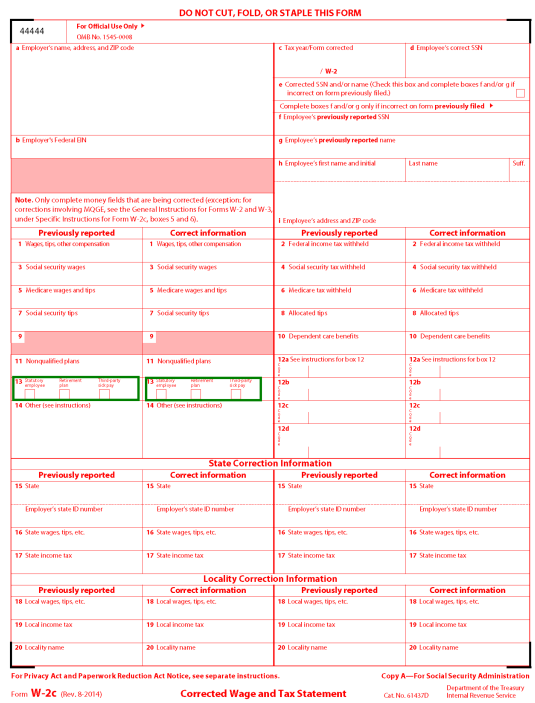
[4] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery.

[5] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

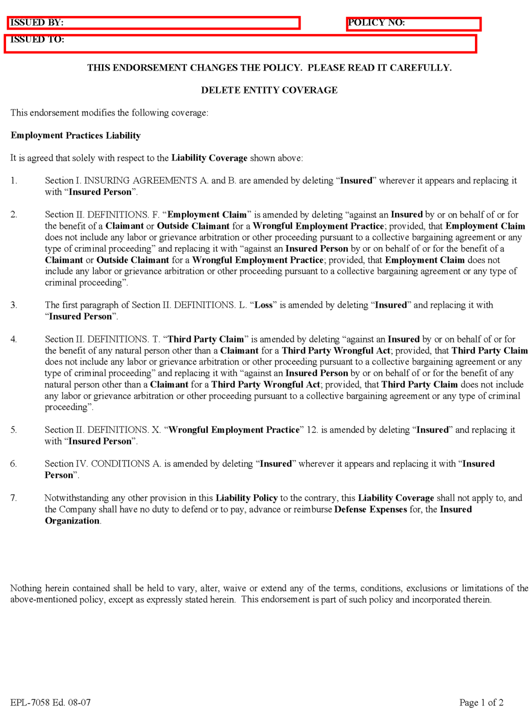
[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.



(a) Pred A



(b) Target A



ISSUED BY: **POLICY NO:**

ISSUED TO:

THIS ENDORSEMENT CHANGES THE POLICY. PLEASE READ IT CAREFULLY.

DELETE ENTITY COVERAGE

This endorsement modifies the following coverage:

Employment Practices Liability

It is agreed that solely with respect to the **Liability Coverage** shown above:

1. Section I. INSURING AGREEMENTS A. and B. are amended by deleting "**Insured**" wherever it appears and replacing it with "**Insured Person**".

2. Section II. DEFINITIONS. F. "**Employment Claim**" is amended by deleting "against an **Insured** by or on behalf of or for the benefit of a **Claimant** or **Outside Claimant** for a **Wrongful Employment Practice**; provided, that **Employment Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding" and replacing it with "against an **Insured Person** by or on behalf of or for the benefit of a **Claimant** or **Outside Claimant** for a **Wrongful Employment Practice**; provided, that **Employment Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding".

3. The first paragraph of Section II. DEFINITIONS. L. "**Loss**" is amended by deleting "**Insured**" and replacing it with "**Insured Person**".

4. Section II. DEFINITIONS. T. "**Third Party Claim**" is amended by deleting "against an **Insured** by or on behalf of or for the benefit of any natural person other than a **Claimant** for a **Third Party Wrongful Act**; provided, that **Third Party Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding" and replacing it with "against an **Insured Person** by or on behalf of or for the benefit of any natural person other than a **Claimant** for a **Third Party Wrongful Act**; provided, that **Third Party Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding".

5. Section II. DEFINITIONS. X. "**Wrongful Employment Practice**" 12. is amended by deleting "**Insured**" and replacing it with "**Insured Person**".

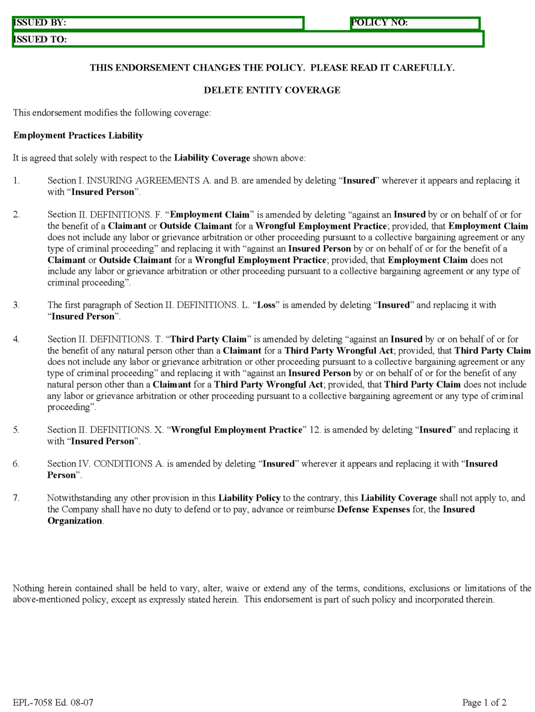
6. Section IV. CONDITIONS A. is amended by deleting "**Insured**" wherever it appears and replacing it with "**Insured Person**".

7. Notwithstanding any other provision in this **Liability Policy** to the contrary, this **Liability Coverage** shall not apply to, and the Company shall have no duty to defend or to pay, advance or reimburse **Defense Expenses** for, the **Insured Organization**.

Nothing herein contained shall be held to vary, alter, waive or extend any of the terms, conditions, exclusions or limitations of the above-mentioned policy, except as expressly stated herein. This endorsement is part of such policy and incorporated therein.

EPL-7058 Ed. 08-07 Page 1 of 2

(c) Pred B



ISSUED BY: **POLICY NO:**

ISSUED TO:

THIS ENDORSEMENT CHANGES THE POLICY. PLEASE READ IT CAREFULLY.

DELETE ENTITY COVERAGE

This endorsement modifies the following coverage:

Employment Practices Liability

It is agreed that solely with respect to the **Liability Coverage** shown above:

1. Section I. INSURING AGREEMENTS A. and B. are amended by deleting "**Insured**" wherever it appears and replacing it with "**Insured Person**".

2. Section II. DEFINITIONS. F. "**Employment Claim**" is amended by deleting "against an **Insured** by or on behalf of or for the benefit of a **Claimant** or **Outside Claimant** for a **Wrongful Employment Practice**; provided, that **Employment Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding" and replacing it with "against an **Insured Person** by or on behalf of or for the benefit of a **Claimant** or **Outside Claimant** for a **Wrongful Employment Practice**; provided, that **Employment Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding".

3. The first paragraph of Section II. DEFINITIONS. L. "**Loss**" is amended by deleting "**Insured**" and replacing it with "**Insured Person**".

4. Section II. DEFINITIONS. T. "**Third Party Claim**" is amended by deleting "against an **Insured** by or on behalf of or for the benefit of any natural person other than a **Claimant** for a **Third Party Wrongful Act**; provided, that **Third Party Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding" and replacing it with "against an **Insured Person** by or on behalf of or for the benefit of any natural person other than a **Claimant** for a **Third Party Wrongful Act**; provided, that **Third Party Claim** does not include any labor or grievance arbitration or other proceeding pursuant to a collective bargaining agreement or any type of criminal proceeding".

5. Section II. DEFINITIONS. X. "**Wrongful Employment Practice**" 12. is amended by deleting "**Insured**" and replacing it with "**Insured Person**".

6. Section IV. CONDITIONS A. is amended by deleting "**Insured**" wherever it appears and replacing it with "**Insured Person**".

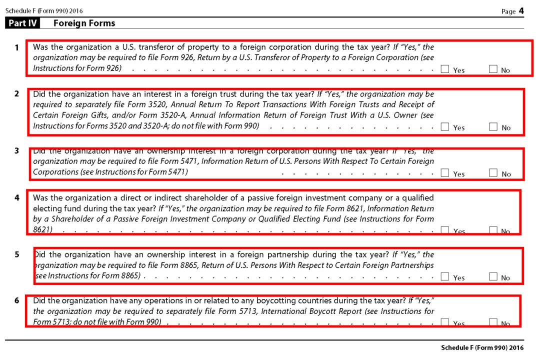
7. Notwithstanding any other provision in this **Liability Policy** to the contrary, this **Liability Coverage** shall not apply to, and the Company shall have no duty to defend or to pay, advance or reimburse **Defense Expenses** for, the **Insured Organization**.

Nothing herein contained shall be held to vary, alter, waive or extend any of the terms, conditions, exclusions or limitations of the above-mentioned policy, except as expressly stated herein. This endorsement is part of such policy and incorporated therein.

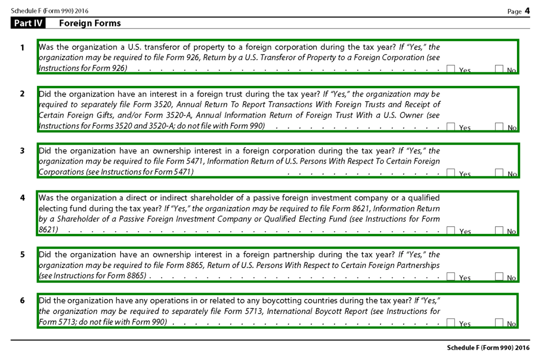
EPL-7058 Ed. 08-07 Page 1 of 2

(d) Target B

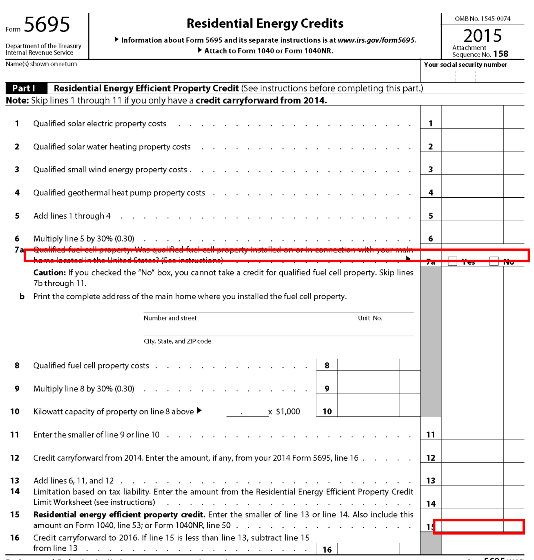
Figure 3: Flamingo Forms Examples (1)



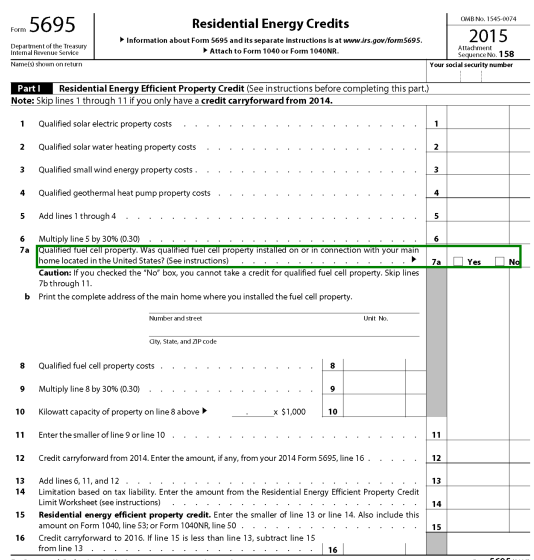
(a) Pred C



(b) Target C

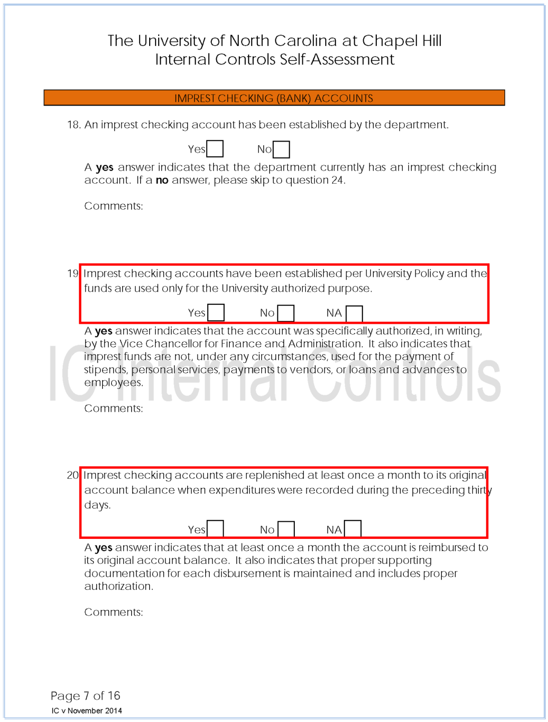


(c) Pred D

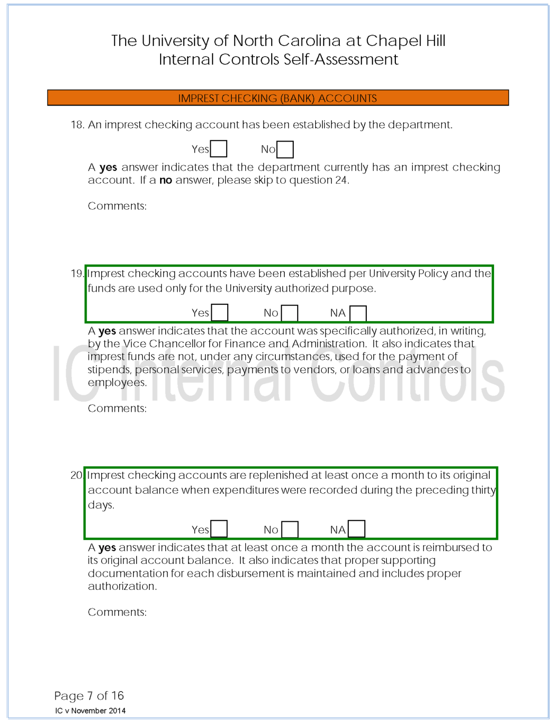


(d) Target D

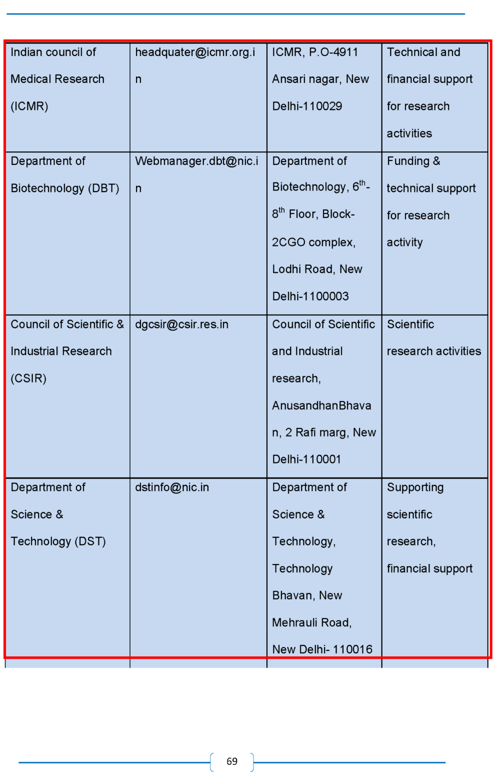
Figure 4: Flamingo Forms Examples (2)



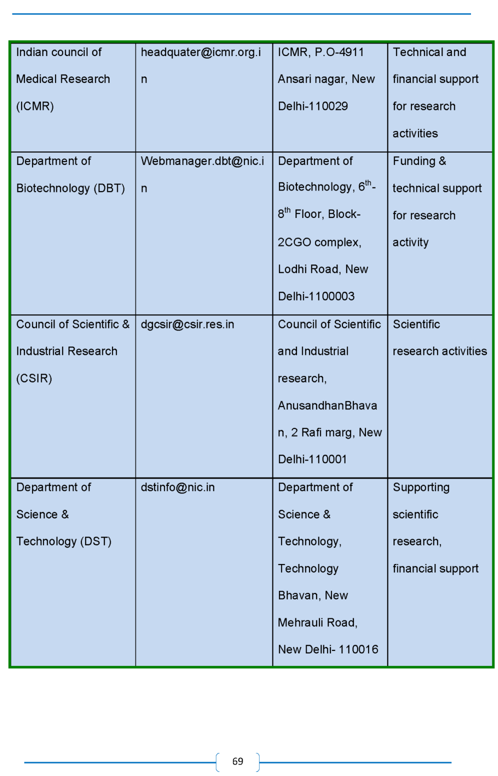
(a) Pred E



(b) Target E

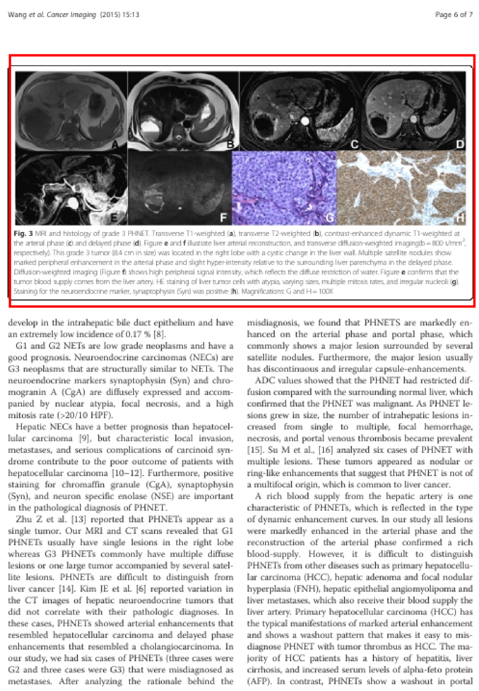


(c) Pred F

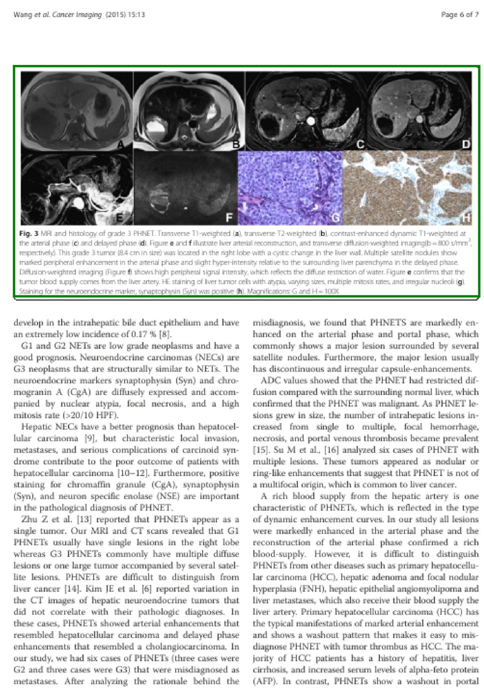


(d) Target F

Figure 5: Flamingo Forms Examples (3)



(a) Pred G

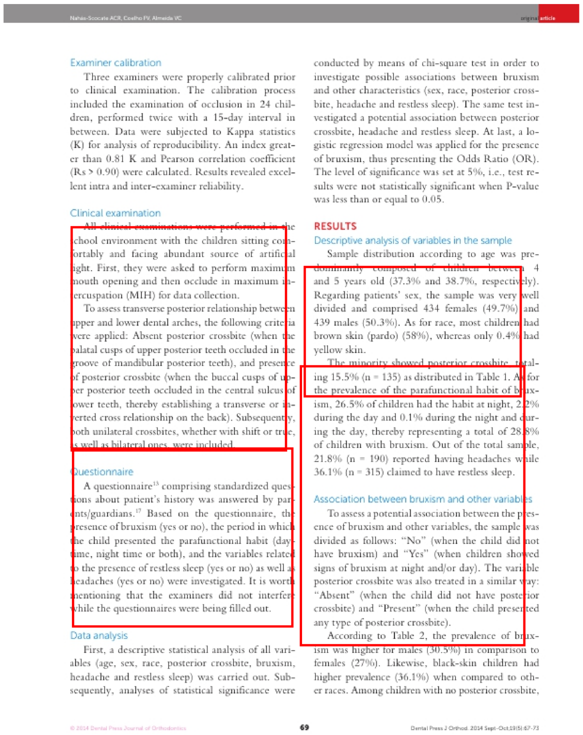


(b) Target G

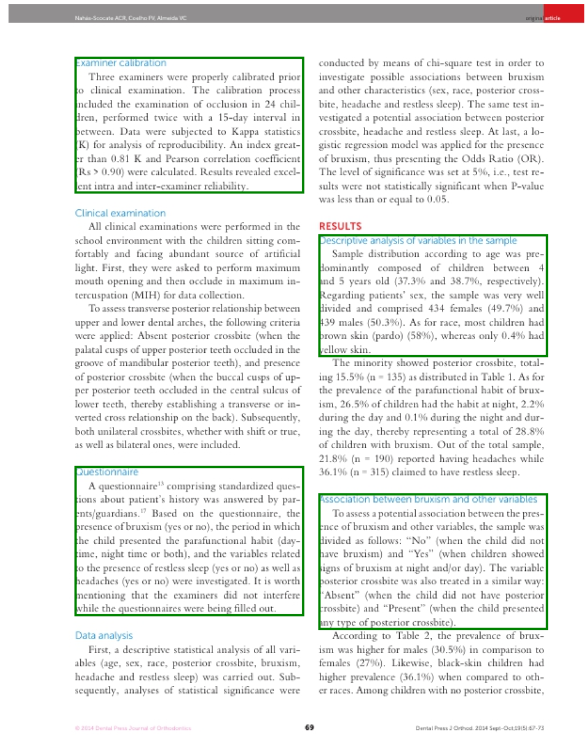
(c) Pred H

(d) Target H

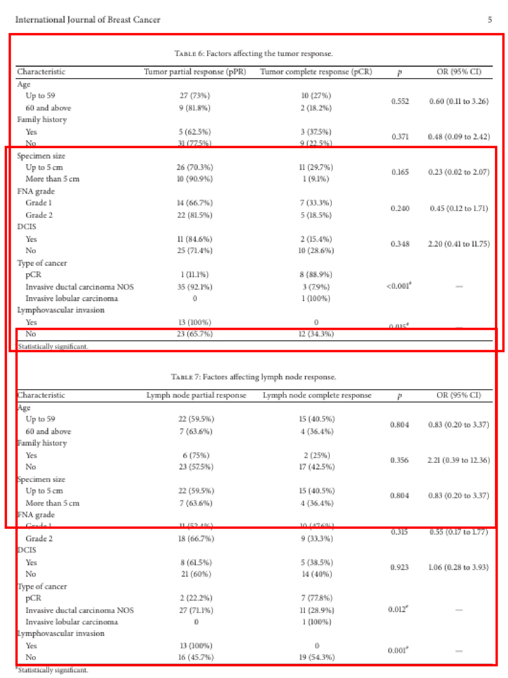
Figure 6: PubLayNet Examples (1)



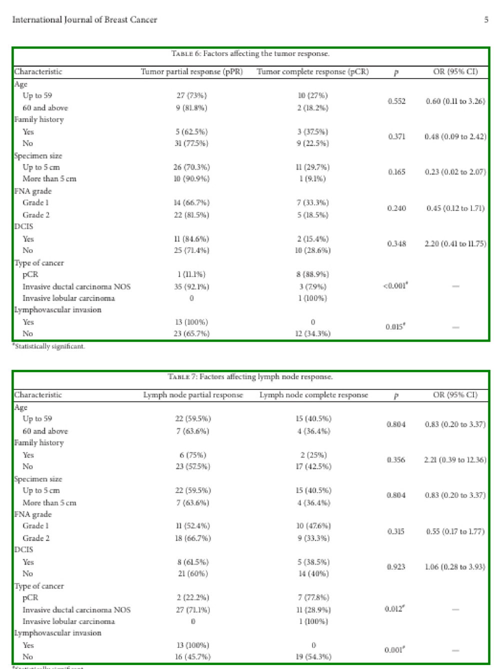
(a) Pred I



(b) Target I



(c) Pred J



(d) Target J

Figure 7: PubLayNet Examples (2)