CoD: Coherent Detection of Entities from Images with Multiple Modalities

Vinay Verma, Dween Sanny, Abhishek Singh, Deepak Gupta
International Machine Learning (IML) Amazon India
vinayugc@gmail.com, drsanny@amazon.com, p15abhisheks@iima.ac.in, deepakgupta.cbs@gmail.com

Abstract

Object detection is a fundamental problem in computer vision, whose research has primarily focused on unimodal models, solely operating on visual data. However, in many real-world applications, data from multiple modalities may be available, such as text accompanying the visual data. Leveraging traditional models on these multi-modal data sources may lead to difficulties in accurately delineating object boundaries. For example, in a document containing a combination of text and images, the model must encompass the images and texts pertaining to the same object in a single bounding box. To address this, we propose a model that takes in multi-scale image features, text extracted through OCR, and 2D positional embeddings of words as inputs, and returns bounding boxes that incorporate the image and associated description as single entities. Furthermore, to address the challenge posed by the irregular arrangement of images and their corresponding textual descriptions, we propose the concept of a "Negative Product Bounding Box" (PBB). This box encapsulates instances where the model faces confusion and tends to predict incorrect bounding boxes. To enhance the model's performance, we incorporate these negative boxes into the loss function governing matching and classification. Additionally, a domain adaptation model is proposed to handle scenarios involving a domain gap between training and test samples. In order to assess the effectiveness of our model, we construct a multi-modal dataset comprising product descriptions from online retailers' catalogs. On this dataset, our proposed model demonstrates significant improvements of 27.2%, 4.3%, and 1.7% in handling hard negative samples, multi-modal input, and domain shift scenarios, respectively.

1. Introduction

Object detection [3, 4, 6, 10, 21–23], extensively studied in the computer vision field, has traditionally been regarded as a unimodal problem, primarily focusing on visual objects. However, given the expansion of the advertising industry, the proliferation of social media content, and the growing prevalence of multimodal interactions, there arises a necessity for detecting objects that amalgamate information from diverse sources. For instance, catalogs featuring advertisements, social media posts discussing events or endorsing products, and interactive news often encompass multimodal details, encompassing images, graphs, descriptions, and attribute-value pairs. While humans can effortlessly discern boundaries between distinct products or articles, this becomes a formidable task for machines, particularly as modalities increase and models grow more intricate.

Multimodal product or content detection finds applications across various real-world domains, including robot interaction [20], information retrieval [2, 27], targeted advertising, and attribute extraction [5, 17]. However, contemporary search engines predominantly rely on text for information retrieval and extraction. Predicting information bounding boxes can yield more accurate outcomes. Similarly, predicting product bounding boxes from catalogs or news articles can offer more detailed information, beneficial for advertising and attribute extraction. This capacity can also enhance document-machine interaction, as multimodal information enriches content comprehensiveness and enables machines to follow more precise commands. In the subsequent sections, we use the term *entity* to denote the multimodal bounding box.

While there is a growing interest in multi-modal object detection [16, 21, 22], these approaches have distinct objectives and do not address our specific problem. Specifically, they employ one modality to search within the other modality (e.g., utilizing textual descriptions to locate specific objects), whereas our objective is to encompass the entire multimodal entity present in the input.

In this paper, we present a multi-modal approach for detecting entities or information within article or product catalogs. We collected the dataset from a variety of products and brands catalogs and manually annotated it. The annotation process involved labeling both product (positive entity) and non-product (negative entity) bounding boxes (BB). The negative entity would either contain only text, or image. It may also contain multiple products or partial product information. For entity detection, we commence with DDETR [31] as the foundational model. By introducing BB of negative samples and training the model in a discriminative manner, we observe noteworthy performance enhancement. Additionally, we integrate multi-modality into our model by extracting accompanying textual descriptions for each entity through OCR. Furthermore, we incorporate 2D positional embedding corresponding to the positional coordinates of each word, thus preserving the spatial location of words within the image. This multi-modal approach also contributes to the model's performance improvement. Another challenge we encounter pertains to domain shift, given the substantial diversity among brand catalogs. Catalog pages for the same items from different sellers can differ significantly. To address this, we leverage domain adaptation models within the multi-modal transformer architecture. This strategy tackles the domain shift issue and further elevates model performance. Our key contributions are as follows:

- We introduce a multi-modal entity detection problem where each entity incorporates multi-modal information within a single bounding box.
- To address the aforementioned problem, we propose the **Co**herent **D**etection model which (a) explicitly penalizes negative samples if they yield high scores for the positive class, (b) integrates multi-modal information by utilizing word and 2D positional embeddings (to preserve spatial location) and learning deformable attention on the text, and (c) employs a gradient reversal-based domain adaptation model at each layer of the transformer encoder-decoder architecture. This approach handles the domain shift problem and enhances model performance.
- Our model demonstrates significant improvements of 27.2%, 4.3%, and 1.7% in scenarios involving negative samples, multi-modal information, and domain shift, respectively. The ablation study validates the significance of the proposed components.

2. Related Work

Object detection is a widely explored researched topic and has yielded promising outcomes across various real-world scenarios. Pioneering object detection models, including R-CNN [11], FasterRCNN [10], and YOLO [23], harness the power of convolutional architectures, delivering compelling results. Notably, recent advancements in transformer architecture, particularly the vision transformer, have prompted researchers to adopt transformers for object detection. DETR [4] represents an end-to-end, transformer-based [26] object detection model, employing set prediction loss for precise bounding box matching. Impressively, DETR sidesteps conventional object detection components such as region proposals, rule-based train-target assignment, and non-maximal suppression (NMS). However, the training process for DETR is comparatively slow when contrasted with object detectors like YOLO and FasterRCNN, and its performance diminishes concerning small objects. To mitigate these challenges, Deformable-DETR [31] harnesses deformable transformers, leading to accelerated convergence. This architecture integrates multi-scale features and deformable attention, showcasing favorable outcomes in object detection tasks. Notably, these methodologies focus on resolving the unimodal object detection problem. However, the real-world landscape is intricate, frequently involving multimodal contexts. Directly applying these techniques yields suboptimal performance.

Given the growing complexity of real-world data and its diverse forms, multimodal object detection [16, 21, 22] has gained traction. MDETR [16] proposes a transformer-based object detection model tailored to provided object descriptions, albeit incurring substantial inference costs. Lite-DETR [21] streamlines parameters, enhancing training and inference efficiency. Furthermore, MAVL [22] extends multimodal object search to encompass open classes, suitable for dynamically evolving environments. Notably, these strategies differ significantly in objectives from ours. These approach target object search through object descriptions in the unimodal domain. In contrast, our approach accommodates complex multimodal data, regardless of spatial location, as objects. Hence, these endeavors fall short in addressing our specific problem.

Furthermore, substantial literature tackles domain adaptation in object detection, with numerous studies employing conventional methods such as FasterRCNN and YOLO [13, 15, 19, 28]. In the realm of transformer architecture, DA-DETR [29] introduces hybrid attention for domain adaptation in DETR, while Gong et al. [12] concentrate on enhancing transferability in DETR for domain adaptation. However, the application of domain adaptation in the context of multi-modal object detection via transformer architecture remains relatively unexplored. To the best of our knowledge, this study marks the inaugural approach tackling domain adaptation in multi-modal object detection. Subsequent sections will elaborate on the specifics of the proposed model.

3. Notations and Background

3.1. Notations

Assume that we have a dataset $\mathcal{D} = \{d_i\}_{i=1}^{N}$, where d_i represents a sample containing n_i product bounding boxes, which can be either positive or negative. The positive BB contains the image of the product and its description, however in the negative same BB contains the image and description of two different product. The i^{th} sample is represented as a triplet $d_i = (I_i, \{c_j^i, b_j^i\}_{j=1}^{n_i})$, where $c_j^i \in \{0, 1\}$ indicates whether the bounding box is positive or negative, and $b_j^i \in \mathbb{R}^4$ denotes the four-dimensional coordinates of the bounding box. The backbone network of our model is a convolutional neural network (CNN), and the output of the CNN for an image I_i is a multi-scale feature map represented as

$\{\boldsymbol{x}^l\}_{l=1}^L$, where L is the number of scales. The input image also contains a description for each associated product image. We extract the text for each image using optical character recognition (OCR[1]). Let T_i and P_i be the extracted text and 2D positional embeddings for each word in the image. For a given input image I_i, the model is expected to predict $\{\hat{c}_j^i, \hat{b}_j^i\}_{j=1}^{n_i}$, i.e. the bounding box and label containing all the entities in the image.

3.2. Background

This section provides the background for the proposed model on which CoD is built upon. Recently, DDETR [31] proposed the object detection model that leverages the deformable attention [7, 30] in transformers and multi-scale feature to capture the various image resolutions. Let for the image I_i, we have a multi-scale image feature map $\{\boldsymbol{x}^l\}_{l=1}^L$, extracted text T_i and 2D positional embedding for each word P_i. Here $\boldsymbol{x}^l \in \mathbb{R}^{C \times H_l \times W_l}$, (C, H_l, W_l) represents channel, height and width respectively. Let $\hat{\boldsymbol{p}}_q \in [0,1]^2$ be the normalized coordinates of the reference point for each query element q and content feature z_q. The multi-scale deformable attention [31] is defined as:

$$\text{MSDA}_I(\boldsymbol{z}_q, \boldsymbol{p}_q, \{\boldsymbol{x}^l\}_{l=1}^L) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{l=1}^{L} \sum_{k=1}^{K} A_{mlqk} \boldsymbol{W}'_m \boldsymbol{x}^l (\phi_l(\boldsymbol{p}_q) + \Delta \boldsymbol{p}_{mlqk}) \big] \quad (1)$$

Where M, L and K are number of multi-head, multi-scale and sampled keys, respectively. DDETR model uses the encoder-decoder transformer architecture and set prediction loss for object detection. The scalar attention weight A_{mlqk} is normalized by $\sum_{l=1}^{L} \sum_{k=1}^{K} A_{mlqk} = 1$, also $\boldsymbol{W}_m$ and $\boldsymbol{W}'_m$ is the learnable weight of the m^{th} attention head.

The purpose of our multi-scale multi-modal deformable transformer module is to improve upon the standard deformable attention layer found in the deformable transformer model. This module takes in multi-modal input and produces output of equal dimensions. The decoder model includes self-attention and cross-attention modules. These attentions are focused on object queries, which are specialized to identify objects within various regions of the image. In the cross-attention module, object queries obtain features from the feature maps, while the keys are the output feature maps from the encoder. In the self-attention module, object queries interact with each other, using the object queries as the key elements. The decoder output is passed to the prediction network that contains two fully connected layer to predict the bounding box and class label. DDETR model leverage the matching loss objective [4], however, the key challenge in defining the matching between the predicted and ground truth BBs is to determine an optimal bipartite matching with minimal cost, denoted as $\hat{\sigma}$. Here, let y represents the ground truth set of objects, and $\hat{y} = \{\hat{y}_i\}_{i=1}^N$ represents the set of N objects predicted by the model. Since N is larger than the actual number of objects in an image, y is padded with $\varnothing$ (no object) to calculate the matching loss. The optimal bipartite matching with minimal cost($\hat{\sigma}$), can be defined as: $\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_i^N \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$, where $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is a pair-wise *matching cost* between ground truth y_i and prediction with index $\sigma(i)$, it is given as:

$$\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = -\mathbb{1}_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\sigma(i)}) \quad (2)$$

The optimal assignment is computed efficiently with the Hungarian algorithm [18]. We have positive, negative, and $N - n_i$ no-object BB for each image. Once we have optimal assignment $\hat{\sigma}$, we can compute the classification and BB prediction loss. Each type of object (positive, negative, and no-object) is considered as a separate class. The DETR [4] loss can be defined as a linear combination of a negative log-likelihood for class prediction and a BB prediction loss, and it is given as:

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^{N} \Big[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \Big] \quad (3)$$

where $\hat{\sigma}$ is the optimal assignment computed in the first step (2). The BB loss is defined as: $\mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}(i)}) = \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(b_i, \hat{b}_{\hat{\sigma}(i)}) + \lambda_{\text{L1}} ||b_i - \hat{b}_{\hat{\sigma}(i)}||_1$ Where $\lambda_{\text{iou}}, \lambda_{\text{L1}} \in \mathbb{R}$ are hyperparameters.

4. Proposed Model

4.1. Overview

Figure 1 shows the architecture of our proposed model. The proposed approach CoD leverages the DDETR approach and extends it for the entity detection in a multi-modal scenario. The model uses OCR to extract each word from the image (to find its association with the image) and its relative 2D coordinates (to understand location of the word) and this pair is considered as input along with the image. We observe that using only the positive samples for the entity detection does not contain the discriminative information leading to model confusion. The addition of the negative samples and minimizing the loss for the incorrect entities significantly improves the model performance. Further, to handle the issue of domain shift, we leverage the gradient reversal method in the transformer encoder-decoder architecture which further helps to improve the model performance.

4.2. Multi-modal Fusion

In the Section-3.2 we discussed about the object detection using the DDETR model, this section extends this for the

[1] https://aws.amazon.com/textract/

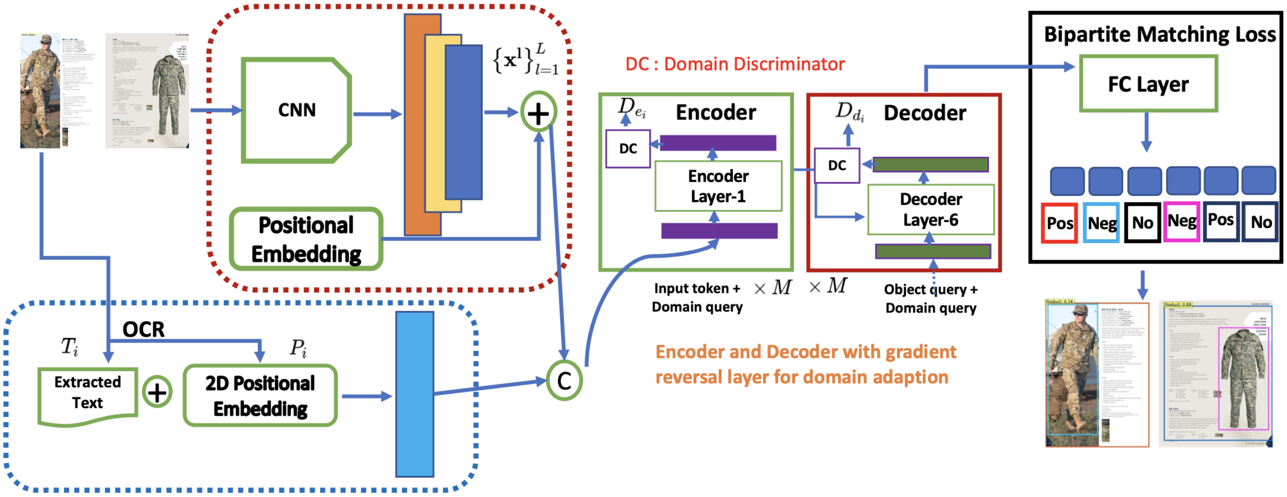

Figure 1. The model leverages the multi-scale feature map of the CNN backbone and extracted text with 2D positional embedding. The joint input is sent to multi-scale deformable transformer-based encoder and decoder architecture and final loss is calculated using the bipartite based matching consistency.

multi-modal. We leverage the BERT ($\mathbf{B}$) tokenizer [8] to tokenize the extracted text T_i and append the 2D positional embedding (P_i). On the concatenated features, we apply a linear projection to maintain the dimension with the image feature, which is given as

$$S_i = f([\mathbf{B}(T_i), P_i]) \tag{4}$$

On the output S_i, a separate deformable attention block is applied, which is given as:

$$\text{MSDA}_T(\boldsymbol{z}_q, \boldsymbol{p}_q, S_i) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m S_i(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \big] \tag{5}$$

Note that in the equations [5] and [1], weights are shared and we are using the same reference point. Here m, l and k are the indices for multi-head, multi-scale and sampled keys respectively. $\Delta \boldsymbol{p}_{mlqk}$ and A_{mlqk} denote the sampling offset and attention weight of the k^{th} sampling point in the l^{th} feature level and the m^{th} attention head, respectively. The joint deformable attention over the image and text are combined as follows:

$$\text{MSDA}_{IT} = \text{MSDA}_I(\boldsymbol{z}_q, \boldsymbol{p}_q, \{\boldsymbol{x}^l\}_{l=1}^L) + \text{MSDA}_T(\boldsymbol{z}_q, \boldsymbol{p}_q, T_i) \tag{6}$$

MSDA$_{IT}$ joins the deformable attention from both modalities and final value is used to predict the bounding box of each entity. Here we are not using the cross-attention instead we are using the concatenated features of the two modalities as we observed that cross-attention degrades the model performance. The cross-attention mostly helps when we have two modes and our objective is to align the both i.e. projecting into a joint embedding space. Here the objective is to expand the boundary of two models and learn a single bounding box when both are from the same entity.

4.3. Negative Sample Aware Set prediction Loss

We observe that during the entity prediction the model mostly confuses in the following scenario: 1) The image of one entities combines with the description of the other entities, 2) Two or more entities are combined into a single one and 3) only text or image are considerd as entities. To handle the above scenario we annotated the hard negative samples it contains the BB of the above three scenario. In our approach we consider it is negative class also we add the loss such that the negative BB has the minimal prediction score. Let id_n is the index of the negative boxes in the complete set prediction and positive samples are considers as label one. To overcome the confusion if negative BB assigned as positive we put here more penalty on it, to achieve the same we have

$$\mathcal{L}_{neg} = \sum_{\forall \hat{y}_i} abs(\hat{y}[:, 1][I_{neg}]) \tag{7}$$

The joint loss for the image and text is:

$$\mathcal{L}_{IT} = \mathcal{L}(y, \hat{y}) + \delta \mathcal{L}_{neg} \tag{8}$$

Where δ is used to control the weight to the negative samples and we call this approach as NSA-CoD.

4.4. Auxiliary Loss

It is helpful to use an auxiliary loss [1] in the decoder during training. It primarily helps the model to predict the correct number of objects in each image, since it stabilizes the training procedure. Auxiliary loss is a multistage loss and is added after each decoder layer.

4.5. Domain Adaptive CoD

The multi-modal catalogs or articles vary significantly across the sellers. This results in a poor generalization of the

	IoU	AP@75	AP@80	AP@90	AR@75	AR@80	AR@90
DDETR w/o NS [31]	-19.1	-18.9	-21.9	-6.1	-12.2	-11.3	-9.8
FasterRCNN [25]	0.0	0.0	0.0	0.0	0.0	0.0	0.0
YOLO [24]	-0.8	-0.9	-0.8	-2.	-2.1	-1.8	-4.7
DETR [4]	-0.1	-0.3	0.2	-0.2	-0.5	-0.4	-4.1
Lite-MDETR [21]	2.3	4.1	0.5	0.4	0.7	0.9	-3.8
MAVL [22]	2.8	4.	1.2	1.5	1.3	1.2	-0.7
DDETR [31]	3.1	4.5	2.4	2.4	1.7	1.7	0.1
CoD \(NS,MM,DA)	-19.1	-18.9	-21.9	-6.1	-12.2	-11.3	-9.8
CoD \(DA)	6.4	7.3	7.7	4.2	-1.0	0.1	0.5
CoD \(MM)	3.8	5.0	2.9	1.7	-0.1	0.2	-0.3
CoD (Ours)	**8.1**	**7.6**	**7.1**	**7.7**	**4.1**	**3.2**	**3.4**

Table 1. Results of our model on the proposed multi-modal dataset and its comparison with the standard object detection algorithm. Here AP@K and AR@K show the average precision and recall when IoU is greater than $K\%$ ground truth bounding box. Note that all models use the negative sample aware dataset. Here \ is set difference operator and X\Y shows model X without Y component. Note that all the baseline models are trained including the negative samples.

model when novel samples from different sellers or companies are used for inference. The poor generalization occurs because of the high variance in the data and domain shift. To address this issue, we propose domain adaption for the multi-modal entity detection model. Our approach leverages the gradient reversal method that is applied to each encoder and decoder layer of the transformer. The model leverages both labeled source domain data (for training) and unlabeled target domain data (for testing) to adapt the distribution of the test domain.

In the transformer-based object detection model, domain adaptation differs from the standard approach used in models, such as FasterRCNN or YOLO. In this model, we have a set of encoder and decoder blocks, and the output is received at each block. Therefore, directly applying domain adaptation to the final output does not yield good results, leading to degraded performance. To address this issue in the proposed NSA-CoD architecture, we apply layer-wise feature alignment. This process leverages a domain discriminator between the source and target domains, and the model attempts to learn features that cannot be discriminated by the domain discriminator. We use the gradient reversal layer approach proposed by Ganin et. al. [9] to deceive the domain discriminator, which is applied to each encoder and decoder layer. Let $\boldsymbol{e}_0, \boldsymbol{e}_1, \dots \boldsymbol{e}_M$ are the features obtained on M encoder layers. We also have layer-wise domain discriminators, represented by $D_1, D_2, \dots$ for the encoder layers. The encoder objective for feature alignment is:

$$\mathcal{L}_E(\theta,\phi) = \sum_{i=1}^{M} [c\log(D_i(e_i)) + (1-c)\log(1-D_i(e_i))] \quad (9)$$

Similarly, we can define the decoder objective $\mathcal{L}_D(\theta,\phi)$ for feature alignment. Here c represents the domain i.e. $c=0$ represents the source and $c=1$ represents the target. Now the complete training objective for the CoD is the combination of the equations 8 and 9 and decoder objective which is given as:

$$\mathcal{L}_{joint}(\theta,\phi) = \mathcal{L}_{I_i}(\theta) - \lambda_E\mathcal{L}_E(\theta,\phi) - \lambda_D\mathcal{L}_D(\theta,\phi) \quad (10)$$

Our final objective is $\min_\theta \max_\phi \mathcal{L}_{joint}(\theta,\phi)$

The model parameter, denoted as θ, comprises the backbone, encoder, decoder, and BERT tokenizer. Meanwhile, the domain discriminator parameter is represented as ϕ. In optimizing the final objective, the gradient reversal approach proposed by Ganin et. al. [9] is used.

4.6. Post-processing

Although the DDETR model does not require the post processing, but in our case we observe that sometime small BB are detected as entities with high score. Most of the time the entities are not too small therefore we can safely discard the small entities detection. Also, applying the non-maximal suppression further improve the model performance.

5. Experiment and Results

In this study, we conducted a thorough set of experiments to evaluate the performance of the proposed model.

5.1. Multi-Modal Datasets

The object detection approach typically focuses on unimodal objects. To the best of our knowledge, there are currently no available multimodal product detection datasets. In this study, we created a multi-modal dataset by gathering catalogs from various online retailers, provided by sellers, featuring a variety of products. These products include chairs, clothing, bed sheets, toys, books, electronic items, game accessories, and more. We exclusively considered catalog pages that pertain to product descriptions. Each of these pages was converted into an image file and subsequently annotated using an annotation tool. We used 85% of the data for training and validation, and 15% for testing. The annotations included positive and negative classes. Positive annotations consisted of bounding boxes drawn around the product images, descriptions, tables, or any relevant information. In contrast, negative annotations identified areas in which the model might encounter confusion. These negative annotations included bounding boxes encompassing two or more products, an image of one product accompanied by the description of another, or instances containing solely images or descriptions without a clear correlation. The training dataset was composed of $52,947$ bounding boxes, with $26,751$ being positive and $26,196$ being negative bounding boxes, respectively. On average, each dataset page contained 3.75 annotations for positive bounding boxes and 3.83 annotations for negative bounding boxes. Illustrative examples of the dataset can be observed in Figure-2, which showcases both positive and negative annotations. Additional examples demonstrating the dataset's diversity are provided in the supplementary materials.

5.2. Implementation Details

The proposed model for this study uses a CNN architecture, specifically ResNet50, as the base network. The input image, represented by $I \in \mathbb{R}^{3\times H\times W}$, is passed through the base CNN to produce a low-resolution feature map with various channels. For multi-scale feature extraction, the feature maps from the last three blocks are collected. Each feature map has a fixed shape of HW in terms of channels, but with varying numbers of channels, denoted as C_0, C_1, C_2. Optical Character Recognition (OCR) is also utilized for text feature extraction, which is then tokenized using the BERT tokenizer. A 2D spatial positional embedding is appended to each word to preserve the spatial location in the image. The image pixel features are represented by a 256-dimensional feature in the image feature map, while each word token is projected to a 256-dimensional feature using a linear layer. The joint flattened feature of the text and image is then sent to the encoder layer. The decoder uses object queries, with $N = 300$. Domain classifiers are added to the encoder and decoder layers to discriminate between the source and target domains. The domain classifier loss is used to align the source and target domains. Further details regarding the hyper-parameters and training can be found in the supplementary material.

5.3. Evaluation Metric

We report the results over the standard evaluation metrics [25] ($AP@75$, $AP@80$, $AP@90$, $AR@75$, $AR@80$ and $AR@90$) for the object detection. Apart from the standard metric, we also report the average intersection over union (IoU) for the detected bounding box w.r.t. ground truth. The average IoU is defined as:

$$IoU = \frac{1}{M}\sum_{i=1}^{K}\frac{area(g_i \cap p_i)}{area(g_i \cup p_i)} \tag{11}$$

Where g_i and p_i are the ground truth and predicted bounding box, respectively. The area is calculated as the number of pixels in that region. Note that because of the privacy reason instead of reporting the absolute IoU we report the relative value wrt. baseline model FasterRCNN. The positive or negative value shows the improvement of degradation in compared to FasterRCNN baselines.

We also evaluated the model performance for the page level and product level correctness. Empirically, we observe if the predicted and ground truth BB have IoU $\geq \mathcal{T} = 0.80\%$ then we have minimal information loss. Therefore, we define the product correctness as:

$$c_i = \begin{cases} 1 & IoU(g_i, p_i) \geq \mathcal{T} \\ 0 & Otherwise \end{cases} \tag{12}$$

Here g_i and p_i are the ground truth and predicted BB for the i^{th} product, and $\mathcal{T}$ is the threshold value where $\mathcal{T} = \{0.75, 0.90\}$. The product level correctness over the test samples is defined as: $\frac{1}{K}\sum_{i=1}^{K} c_i$, where k is total product BB in the test data. Also we define the page level correctness as:

$$pc_j = \begin{cases} 1 & \forall g_i, p_i \in G : IoU(g_i, p_i) \geq \mathcal{T} \ \& \ |G| == |P| \\ 0 & Otherwise \end{cases} \tag{13}$$

The overall page-level correctness is defined as the mean of pc_i across the dataset.

5.4. Baseline

We compare our model on the proposed dataset over various object detectors like FasterRCNN [25], YOLO [24], DETR [4], DDETR [31]. Also, we incorporated the modified multi-modal approach MAVL [22] and Lite-MDETR [21] as recent multi-modal work. Note that we are reporting all the results using the hard negative samples as without using the negative samples, the model performance significantly drops. DDETR is the most competitive approach therefore we show the results without hard negative samples for the DDETR model only.

5.5. Results

The results of the proposed model CoD on the multimodal dataset are presented in Table-1. The results are shown using the IoU metric, the average precision of correct bounding boxes (BB) when overlap is greater than K (i.e. $AP@K$), and the average recall when IoU is greater than K (i.e. $AR@K$). In the baseline models, both faster-RCNN [25] and DETR [4] display similar performance across most metrics. However, vanilla DDETR [31] shows improved performance because of its multi-scale features and deformable transformer. The multimodal approach MAVL [22] and Lite-MDETR [21] do not perform well since

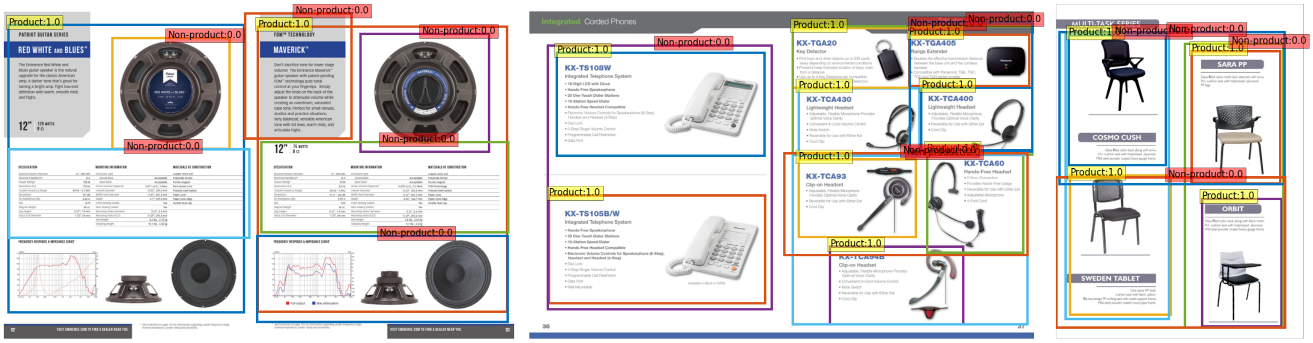

Figure 2. Positive and negative annotations for the product and incorrect product. The negative annotation contains the incorrect BB where the model mostly confuses.

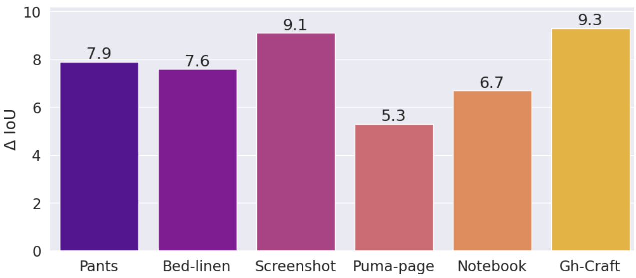

Figure 3. IoU score for category-wise evaluation for the few product classes, here Puma, Notebook and Craft are the novel category that are not available during training. We can observe that model shows high generalization ability to the novel classes.

they are designed for the matching of two spaces and hence do not consider two spaces as a single bounding box. These result are reported when hard negative samples are taken into consideration as without negative samples, the performance of compared models drops significantly (that can be seen in results of DDETR w/o NS and CoD$\backslash$(NS,MM,DA)).

Our proposed negative sample aware multi-modal approach with domain adaptation (CoD) shows significant improvement compared to the state-of-the-art baseline models. Compared to the best DDETR model, our approach demonstrates a 5.0% absolute gain in IoU, as well as similar performance gains in $AP@K$ and $AR@K$. Empirical observations have shown that when IoU is greater than 80%, there is minimal loss of information in the predicted BB. Therefore, $AP@80$ and $AR@80$ are key metrics. In these metrics, we see an absolute improvement of 4.7% and 1.5%, respectively. The supplementary contains qualitative results for the diverse dataset. Also, we observe the page correctness for CoD is 33.2% for the threshold $\mathcal{T} = 0.8$.

5.5.1 Generalization to Novel category

The dataset used includes various subcategories, ranging from relatively simple to highly complex. To evaluate the performance of our model, we selected a range of subcategories and analyzed the results on a category-by-category basis. The results ΔIoU represents the gain/loss in compared to the base model FasterRCNN. As shown in Figure-3, we found the model performed well on all categories, including those that were not present during training (Puma, Computer Notebook, and Craft). Despite encountering novel categories during the inference phase, the model could generalize and accurately detect bounding boxes.

5.6. Ablation Study

To better understand the contribution of each component in the proposed model CoD, we conducted ablation over the various components. Specifically, we evaluated the performance of the model with and without hard negative bounding boxes in the annotation process, as well as the impact of the CoD without multimodal (MM) and Domain Adaptation (DA) scenarios. The ablation is shown in the Table-1. Our results showed that including hard negative bounding boxes (CoD$\backslash$(MM,DA)) significantly improved the model's performance achieving a gain of 22.2% absolute IoU compared to the vanilla model. The final model shows the IoU gain of 1.7% wrt. the without domain adaption (CoD$\backslash$DA). The MM information is another key factor and without incorporating MM data (CoD$\backslash$MM) we observe a performance drop of 2.3% in absolute IoU. The standard best object detection approach DDETR without negative sample and multi-modal information shows the significant IoU drop, however the proposed model CoD has a performance gain of 27.2% in absolute IoU. Therefore various proposed components play a significant role to improve the model performance. Please refer to Table-1 for detailed ablation analysis of the discussed components. Furthermore, in Figure-5, we present qualitative results obtained by utilizing various components. It is evident that the incorporation of negative samples and multimodal information substantially aids in mitigating model confusion and enhancing accurate entity prediction. Also, in the Figure-4 we have shown the qualitative result over the various multi-model approaches.

We have previously highlighted that the inclusion of negative samples yields significant improvements compared to the multimodal component alone. To validate this, we con-

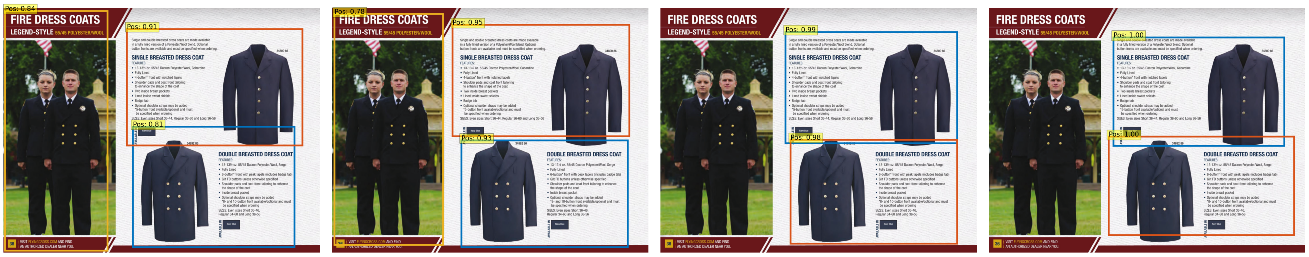

Figure 4. The qualitative analysis of the proposed model in compared to the Lite-DETR and MAVL approach. The images from left to right are using the model: Lite-DETR, MAVL and Ours. The right image contains the ground truth bounding box.

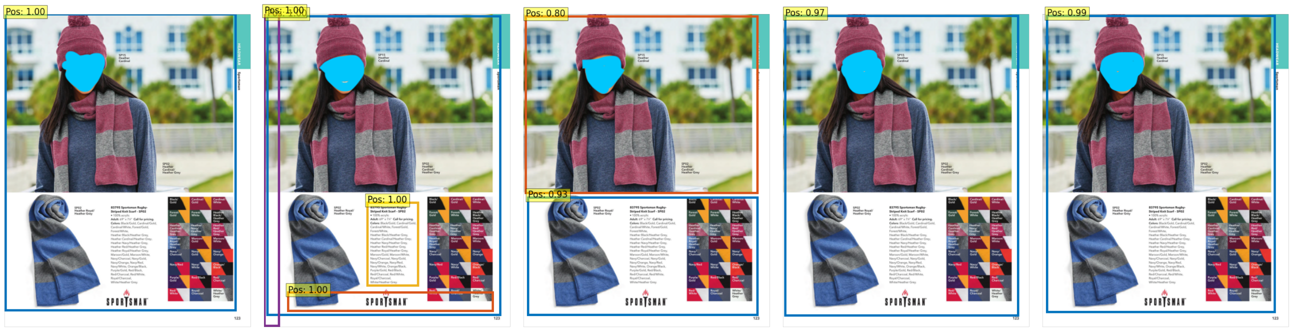

Figure 5. The sequence from left to right contains ground truth, CoD \ (NS,MM,DA), CoD \ (MM,DA), CoD\ DA and proposed CoD results. Here \ is set difference operator and X\Y shows model X without Y component.

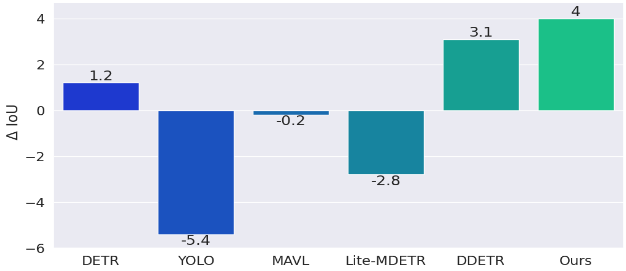

Figure 6. Ablations over various model without using the negative samples. The proposed model (Ours) includes DA in the DDETR.

ducted experiments using various approaches without the utilization of negative samples. The outcomes are showcased in Figure-6, revealing a notable performance degradation across all approaches in the absence of negative samples. Notably, DDETR demonstrates high competitiveness in comparison to our approach, while YOLO performs as the least effective method.

6. Conclusions

The aim of this work is to develop an automated system for detecting multi-modal entity or information. To the best of our knowledge, this is the first dataset and study to detect the bounding box of entities in multi-modal information scenario. Our approach has many potential applications in various fields, such as multi-modal information search, advertising, and product detection. One of the main challenges in this field is preserving the spatial location and interactions between different modes of information. To address this challenge, we maintained 2D positional coordinates and implemented self attention modules to facilitate communication between the different models. We also applied a domain adaptation model, which aligns the source and target embeddings on each decoder and encoder layer using the gradient reversal approach. The domain adaption applied address the diversity between various information sources and the resulting domain shifts. Although our model showed significant improvement over the baseline, it suffered from high rates of false positive and negative predictions. To address this issue, we used hard negative annotations. We also proposed various metrics for evaluating the model's performance, which could be useful for evaluating other multi-modal models in the future. To understand the impact of the proposed components, we evaluated model with different proposed modules. In future work, it would be interesting to explore the model's performance in low data regimes and to better understand the interactions between different modes of information. Given the high cost of annotation, it may also be worthwhile to investigate self or semi-supervised approaches. Overall, our study presents a promising approach for automating the detection of multi-modal entity or information.

Acknowledgements: We are highly grateful to Srujana Merugu, Happy Mittal and Prateek Sircar for their constructive feedback, fruitful discussion and paper editing.

References

[1] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *AAAI'19: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* , pages 3159–3166. AAAI Press, Jan. 2019. 4

[2] Ashish Bagwari, Anurag Sinha, NK Singh, Namit Garg, and Jyotshana Kanti. Cbir-dss: Business decision oriented content-based recommendation model for e-commerce. *Information*, 13(10):479, 2022. 1

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 3, 5, 6, 10

[5] Giovanna Castellano and Gennaro Vessio. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Computing and Applications*, 33(19):12263–12282, 2021. 1

[6] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 1

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 5

[10] Ross Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 7–13. IEEE, 2015. 1, 2

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv*, Nov 2013. 2

[12] Kaixiong Gong, Shuang Li, Shugang Li, Rui Zhang, Chi Harold Liu, and Qiang Chen. Improving transferability for domain adaptive detection transformers. *arXiv preprint arXiv:2204.14195*, 2022. 2

[13] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 24:2502–2514, 2021. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 10

[15] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 749–757, 2020. 2

[16] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 1, 2, 10

[17] Boeun Kim, Young Han Lee, Hyedong Jung, and Choongsang Cho. Distinctive-attribute extraction for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[18] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics*, 2(1-2):83–97, Mar 1955. 3

[19] Vinod Kumar Kurmi, Shanu Kumar, and Vinay P Namboodiri. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 491–500, 2019. 2

[20] Kai Lin, Yihui Li, Jinchuan Sun, Dongsheng Zhou, and Qiang Zhang. Multi-sensor fusion for body sensor network in medical human–robot interaction scenario. *Information Fusion*, 57:15–26, 2020. 1

[21] Qian Lou, Yen-Chang Hsu, Burak Uzkent, Ting Hua, Yilin Shen, and Hongxia Jin. Lite-mdetr: A lightweight multi-modal detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12206–12215, 2022. 1, 2, 5, 6, 10

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 512–531. Springer, 2022. 1, 2, 5, 6, 10

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *arXiv*, Jun 2015. 1, 2

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5, 6

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5, 6

[26] 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. 2

[27] Xin-Jing Wang, Wei-Ying Ma, Gui-Rong Xue, and Xing Li. Multi-model similarity propagation and its application for web image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 944–951, 2004. 1

[28] Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3273–3282, 2021. 2

[29] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, and Shijian Lu. Da-detr: Domain adaptive detection transformer by hybrid attention. *arXiv preprint arXiv:2103.17084*, 2021. 2

[30] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 3

[31] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 5, 6, 10

A. Implementation Details

The implementation details are provided in the main paper here we are providing more details about the implementation and model complexity. The proposed model for this study uses a convolutional architecture, specifically ResNet50, as the base network. The input image, represented by $I \in \mathbb{R}^{3\times H\times W}$, is passed through the base CNN to produce a low-resolution feature map with various channels. For multi-scale feature extraction, the feature maps from the last three blocks are collected. Each feature map has a fixed shape of HW in terms of channels, but with varying numbers of channels, denoted as C_0, C_1, C_2. Optical Character Recognition (OCR) is also utilized for text feature extraction, which is then tokenized using the BERT tokenizer. A 2D spatial positional embedding is appended to each word. The image pixel features are represented by a 256-dimensional feature in the image feature map, while each word token is projected to a 256-dimensional feature using a linear layer. The joint flatten feature of the text and image is then sent to the encoder layer. The decoder uses object queries, with $N = 300$ in this study. Domain classifiers are added to the encoder and decoder layers to discriminate between the source and target domains. The domain classifier loss is used to align the source and target domains. We have the initial learning rate of $1e^{-4}$ and it reduced to $1e^{-5}$ after 40 epoch of training. The model is trained for the 50 epoch with the above given step size learning rate. The weight to domain adaption loss is 0.1 and weight decay 0.001 is used the complete training. The model contains $\sim 65M$ parameter, where $24M$ parameters are in the BERTs word embedding and $\sim 23M$ parameter are in the backbone ResNet50 model.

A.1. Deformable Encoder

The purpose of our multi-scale multi-modal deformable transformer module is to improve upon the standard deformable attention layer found in the deformable transformer model. This module takes in multi-modal input and produces output of equal dimensions. To extract the multi-scale feature map, we use the last three stages of the ResNet50 [14]. These stages yield feature maps of varying resolutions, with the final stage producing the lowest resolution map. As the number of channels in each stage's feature map is relatively large, we apply a $d = 256$, 1×1 convolution filter to reduce the number of channels to a manageable size. We also reduce the text size and its corresponding 2D positional information to a dimension of $d = 256$. We set the maximum token size to 256 and convert it to a feature map of size 16×16. The image and text yield four feature maps of size $\{W_i \times H_i \times d\}_{i=1}^{4}$. To use these feature maps as input for the encoder, we reshape them to size $W_iH_i \times d$ and concatenate them along the dimension of H_iW_i. The key and query elements for this module are derived from the pixels in the Multi-Modal Multi-Scale (MMMS) feature maps. To determine the feature level of each query pixel, we employ a method similar to that utilized in the DDETR [31].

A.2. Deformable Decoder

The decoder in our model is similar to the one proposed in DDETR [4], which includes self-attention and cross-attention modules. These attentions are focused on object queries, which are specialized to identify objects within various regions of the image. In the cross-attention module, object queries obtain features from the feature maps, while the keys are the output feature maps from the encoder. In the self-attention module, object queries interact with each other, using the object queries as the key elements. For more details on this, please refer to [4, 31].

A.3. Prediction Network (PN)

The Prediction Network (PN) is a fully connected neural network that is used to predict the bounding box coordinates, including the box center, height, and width. There is also a linear projection layer that is utilized to predict the class label, with an output size of three for predicting positive, negative, and no-object categories. The PN has N predictions, which is significantly larger than the number of actual objects present in a single image (cite reference).

B. Results and Discussion

In this section we are showing the diversity of the dataset samples and some qualitative results. The samples shown in the Figure-7 from various catalogs like handbags, bed-linen, suits, jacket etc. As we observe that these samples are highly complex and the spatial alignment of the image and text such that model can predict both into a same bounding box is challenging. This problem is completely different compared to recent multimodal object detection [16, 21, 22] where they focus to image search in the unimodal domain given the another mode of information. Also, they does not require to preserve the spatial information on the text since no text is there in their images.

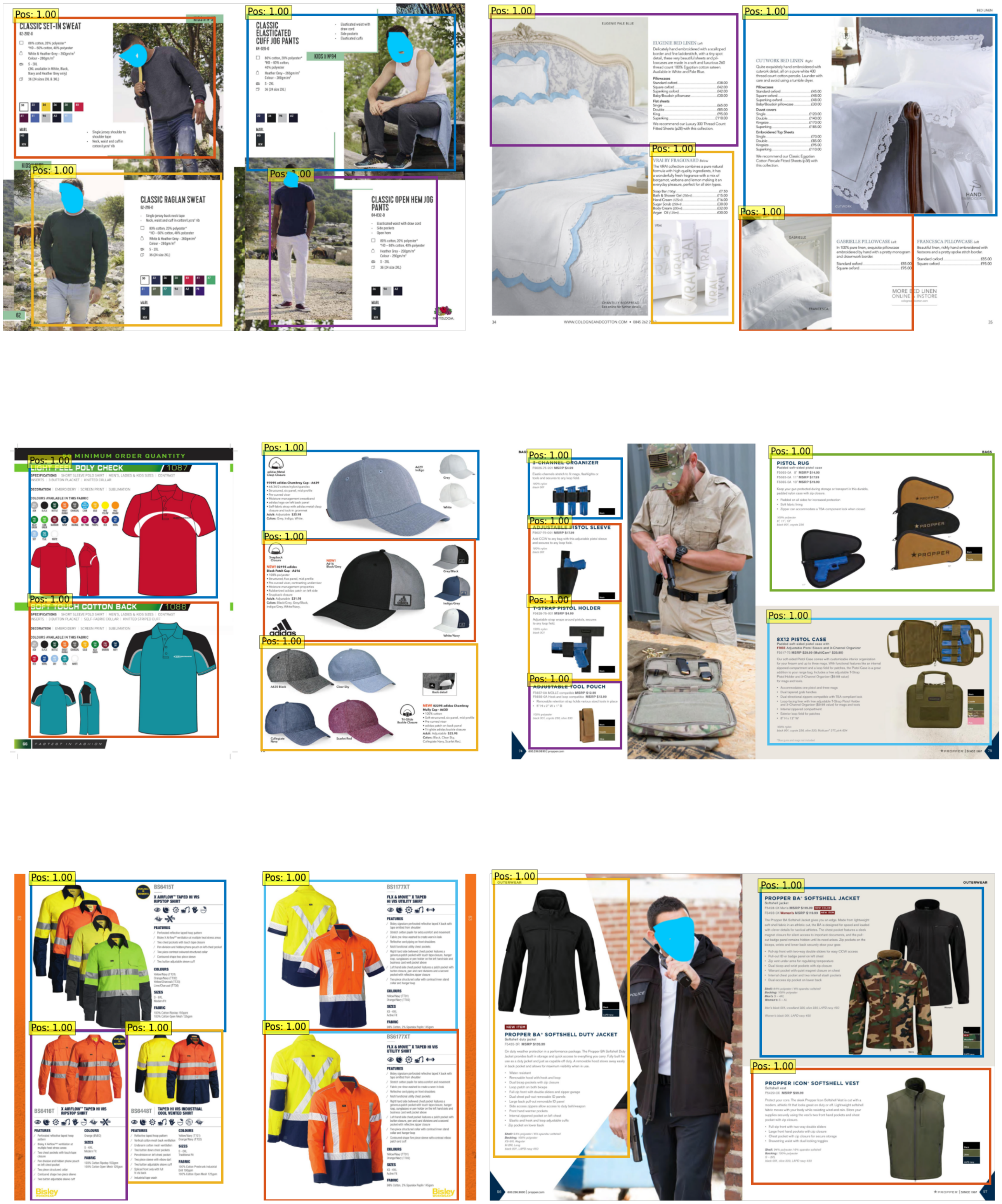

Figure 7. The example of the diverse samples from our annotated dataset. We can observe that samples are unbounded by category and are highly diverse in nature because of the unstructured nature of the catalogs and multi-modal information.

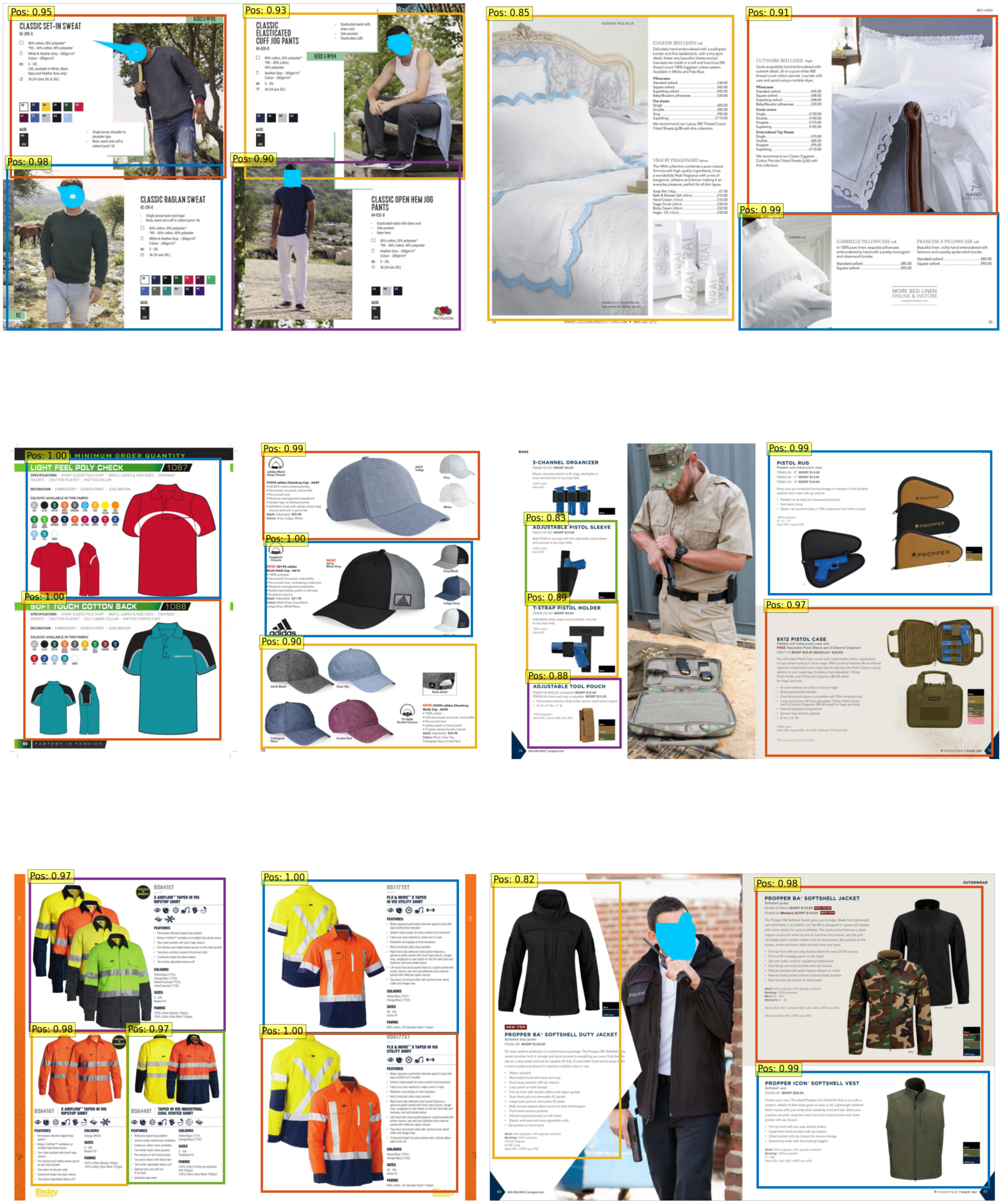

Figure 8. The predicted bounding box for the catalog dataset as compared to the ground truth given in the figure-7. We can observe that on the complex multi-modal catalog the proposed model shows promising results.

B.1. Dataset Diversity

In the Figure-7 we have shown the few samples from the catalog dataset. We can observe that samples are extremely diverse in nature and the multi-modal interaction between the product image and descriptions makes the problem even more harder. The samples of the same class (e.g. electronics) shows high diversity, also the same product across the different sellers are extremely diverse. Because of the high inter-class and intra-class diversity and multiple product interaction model gets confused and predict the incorrect bounding boxes.

B.2. Qualitative Results

The Figure-8 shows the qualitative results predicted by the proposed model. We can observe that model shows the promising result as compared to the ground truth and most of the time model predict the bounding box correctly. In the image [1,1] ([row, col]) image is incorrectly predicted by model. The image contains four product while model predict three, this is because of the complex nature of the image and the description. The model confuses the image and corresponding description and predicts the wrong bounding box for the product.