Supplementary: 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

1. 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.

1.1. 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 [2]. 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 [6].

1.2. Deformable Decoder

The decoder in our model is similar to the one proposed in DDETR [1], 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 [1, 6].

1.3. 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).

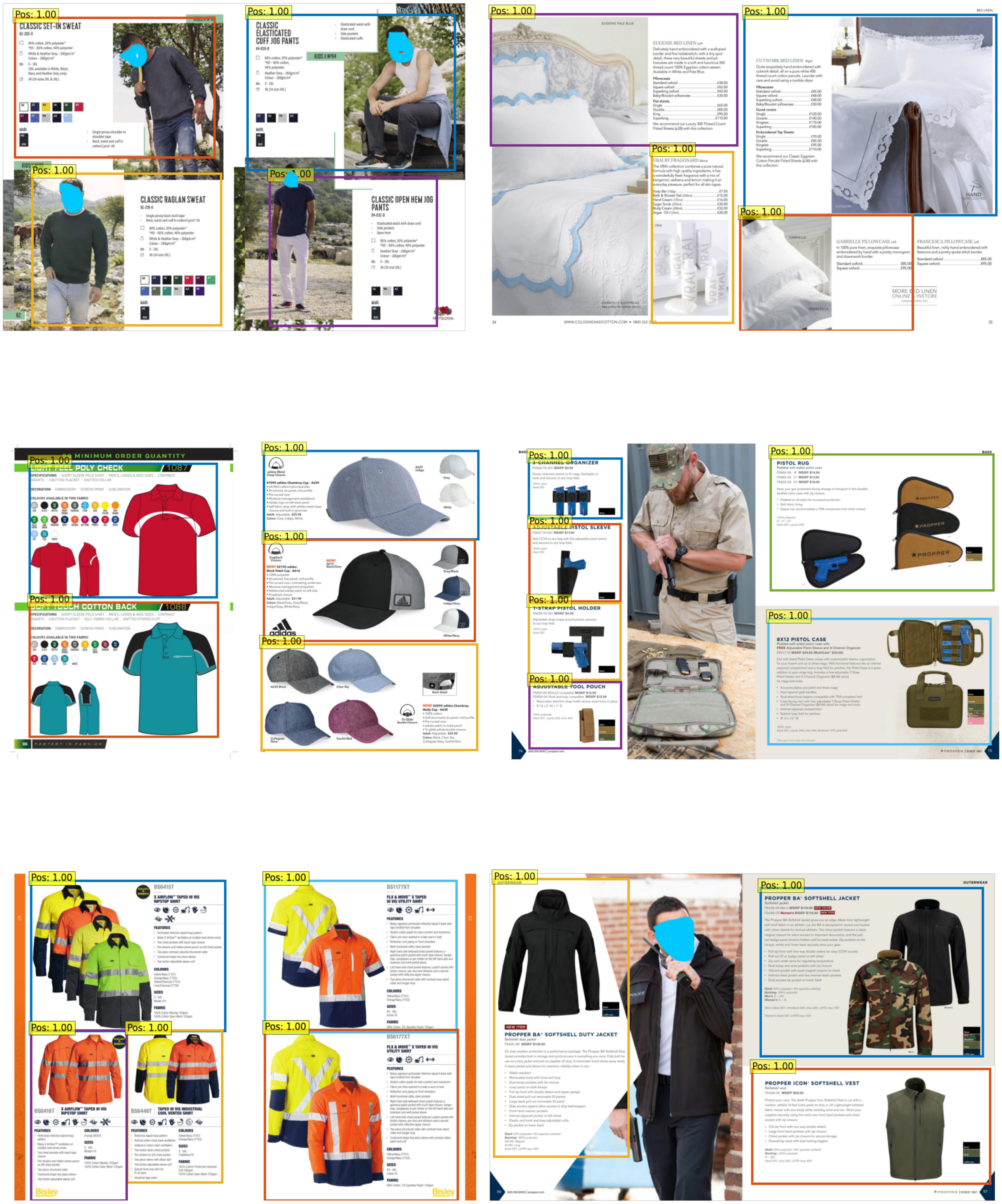

Figure 1. The example of the diverse samples from our annotated dataset. We can observe that samples are unbounded by category and are highly diverse in nature because of the unstructured nature of the catalogs and multi-modal information.

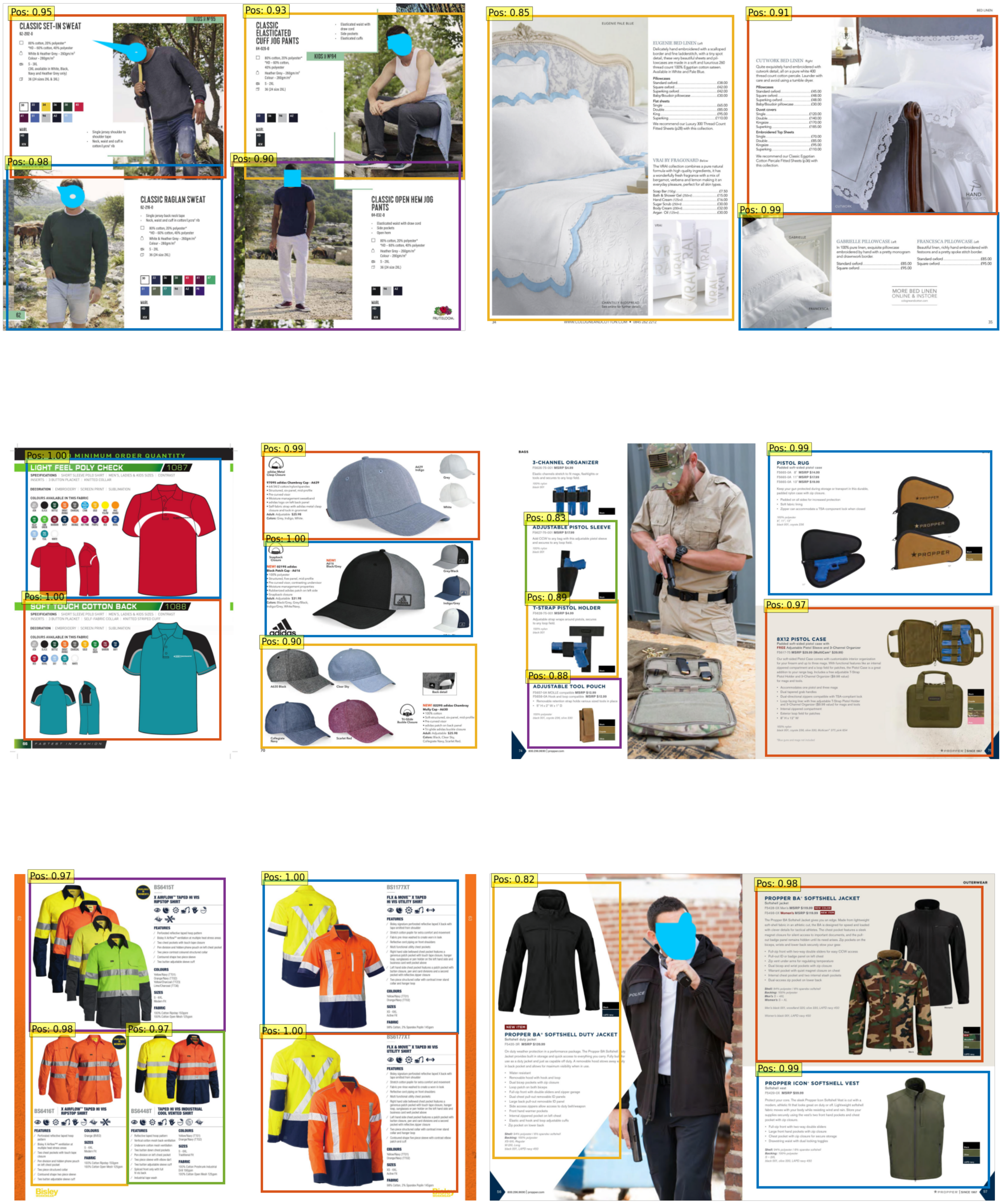

Figure 2. The predicted bounding box for the catalog dataset as compared to the ground truth given in the figure-1. We can observe that on the complex multi-modal catalog the proposed model shows promising results.

2. 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-1 from various catalogs like handbags, bedlinen, 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 [3–5] 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.

2.1. Dataset Diversity

In the Figure-1 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.

2.2. Qualitative Results

The Figure-2 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.

References

[1] 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] 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. 1

[3] 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. 4

[4] 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. 4

[5] 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. 4

[6] 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