Skew-Robust Human-Object Interactions in Videos

Apoorva Agarwal
IIT Bombay
apoorva@cse.iitb.ac.in

Rishabh Dabral
IIT Bombay
rdabral@cse.iitb.ac.in

Arjun Jain
UAVIO Labs
arjunjain@gmail.com

Ganesh Ramakrishnan
IIT Bombay
ganesh@cse.iitb.ac.in

Abstract

Humans are, arguably, one of the most important regions of interest in a visual analysis pipeline. Detecting how the human interacts with the surrounding environment, thus, becomes an important problem and has several potential use-cases. While this has been adequately addressed in the literature in the image setting, there exist very few methods addressing the case for in-the-wild videos. The problem is further exacerbated by the high degree of label skew. To this end, we propose SERVO-HOI *, a robust end-to-end framework for recognizing human-object interactions from a video, particularly in high label-skew settings. The network contextualizes multiple image representations and is trained to explicitly handle dataset skew. We propose and analyse methods to address the long-tail distribution of the labels and show improvements on the tail-labels.* SERVO-HOI *outperforms the state-of-the-art by a significant margin (21.1% vs 17.6% mAP) on the large-scale, in-the-wild VidHOI dataset while particularly demonstrating solid improvements in the tail-classes (19.9% vs 17.3% mAP).*

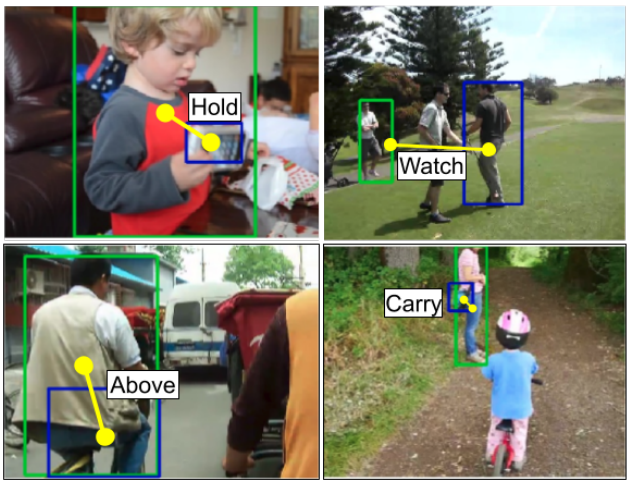


Figure 1: A collection of our results on the VidHOI dataset. We estimate the interaction predicates between the human in green box and the object in the blue, thereby producing a <human, predicate, object> triplet. E.g. in the first image, <boy, holds, toy>.

1. Introduction

Recognizing how humans interact with specific objects and/or persons in the surrounding environment (*e.g.*, a person *holding a* cup, a person *watching* a kid, *etc.* as can be seen in Figure 1) is a key problem and crucial part of Scene Understanding. Availability of such information can be crucial in several real-world downstream applications such as unmanned grocery stores, robotics, surgery monitoring, *etc.*

In this work, we present an approach towards analyzing and detecting interactions between the humans and the objects in an image/video. Human-Object Interaction detection in images has been a well-studied problem in recent years [30, 22, 15, 32, 35, 13, 27, 6]. While there has been a significant amount of published research for the image domain, (thanks to the presence of in-the-wild and challenging datasets such as V-COCO [10], HICO-DET [2] and HOI-A [17]), the same is not true for videos. This can be partly attributed to the unavailability of good, in-the-wild datasets with CAD-120 [14] being the only majorly used dataset for video HOI for several years. However, the CAD-120 dataset has been captured in a highly controlled indoor environment with a limited number of objects and a limited range of interactions restricted to a single person. Further, most of the works trained on CAD-120 are critically hinged upon hand-crafted features that exploit the RGB-D nature of the dataset. Such works are not amenable to generalization owing to unavaiability of depth inputs. However, this has changed with the introduction of the vidHOI dataset [3]. The vidHOI dataset is a subset of the larger VidOR dataset [24] that has been released for the generic vi-

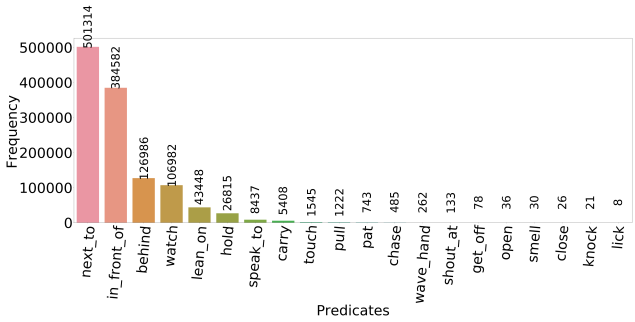


Figure 2: Predicate Label Distribution: x-axis corresponds to predicate categories and y-axis maps them to their frequency in the train set. We only show 20 out of 50 classes for readability. Notice the extreme skew in the label distribution.

sual relationship detection task on videos and is a challenging, in-the-wild, and well-annotated dataset. It is further made difficult, inherently, by the multi-label nature of the annotations. This is fairly realistic; a human and an object may interact in more than one ways at the same time. For example, a person may be *pointing_at* a television screen while also being placed *next_to* the television. Finally, the dataset suffers from a high degree of class-imbalance and long-tailed nature (*c.f.*, Figure 2). For example, the top two classes – *next_to* and *in_front_of* – constitute 63.55% of total labels and the bottom 10 labels constitute only 0.04% of the labels. All these constraints establish VidHOI as a fairly challenging dataset, especially in comparison to the other datasets [14, 10, 2].

With the aforementioned constraints and in mind, we propose SERVO-HOI [1] – a novel and robust method for inferring human-object interactions in videos. We place our method on three footings. Firstly, we propose an end-to-end pipeline that infers multiple human-object interactions in the video (*c.f.* Section 3.1). Our pipeline takes into account human pose cues as well as factors in the positional priors. While human pose, in general, is a good and crucial cue in recognizing the relationship, it becomes extremely noisy if the persons in the scene are heavily occluded or truncated to the effect that including them could become counterproductive. We avoid over-committing to such poses by using a softer representation of heatmaps [28].

Secondly, we address nuances in the multi-label and long-tailed nature of the dataset. As a first step toward addressing the dataset's long-tail distribution and high skew (*c.f.*, Figure 2), we formulate a class-weighted training objective with two variants for determining the weights. The first variant tunes the class weights by performing grid-search over the validation dataset and while yielding the best performance, is computationally intensive. As a more efficient alternative, we adapt the propensity-weighted CE loss [12] whose core idea is to increase the weights of the rare classes while also not drastically underwhelming the abundant classes (*c.f.* Section 3.4). Since the distribution is quite heavily long-tailed, it becomes imperative to demonstrate improvements in rare categories. The propensity-weighted CE loss achieves performance on rare categories comparable to the weight-tuned CE loss while being significantly faster. Additionally, we also present the focal loss formulation [18] as an equally efficient alternative, which we find to be even more robust to noisy detection signals, especially on the tail labels. We also factor-in the muti-label nature of the problem using a simple yet effective threshold tuning mechanism [1](*c.f.* Section 3.5). The result is a pipeline that improves the state-of-the-art by a significant margin on multiple protocols. We achieve a mean Average Precision (mAP) score of 21.2% compared to 17.6% on the challenging VidHOI dataset [3]. Note, that this is a 20% improvement and a significant improvement on VidHOI tasks. On another evaluation mode (detection), we improved the mAP by 50%, achieving 4.8% compared to 3.2% of the best method.

As a third contribution, we identify and discuss issues with the existing evaluation protocol and propose a solution that is consistent with the existing evaluation setups. For this improved protocol, we provide benchmark results and also evaluate existing methods on it.

In summary, we introduce a novel, state-of-the-art method to estimate in-the-wild human-object interactions in videos by exploiting spatial and postural cues and incorporating multi-label attributes while also addressing the high degree of dataset skew.

2. Related Works

We discuss the related works from three vantage points - methods exploring human-object interactions in videos (Sec 2.1), methods using pose for performing HOI (Sec 2.2) and visual relationship detection methods (Sec 2.3) that form a generalization of the HOI problem.

2.1. HOI in Videos

Human-Object Interactions, akin to Scene Graphs, is inherently well suited for graph-like formulations. Indeed, there have been several works using such a formulation [22, 26, 9, 33, 23]. The Graph Parsing Neural Network (GPNN) [22] takes into account the structural knowledge and utilizes message-passing in a deep neural network setup for learning and inference while offering a scalable and generic HOI representation applicable to both static and dynamic settings. Zhang *et. al.* [35] demonstrate a spatio-temporal recurrent neural network (STRNN) which jointly integrates spatial and temporal information in RNN, as well as learns discriminative features. However, these methods

[1] which can be expanded as **S**k**E**w **R**obust **V**ide**O**

have been trained and tested on CAD-120 and overly rely on hand-crafted ground-truth features that the dataset provides. The issue of over-reliance on hand-crafted features has been addressed by recent approaches. The authors in [26] propose a generalizable, multi-level model, , for identifying Human-Object Interactions from videos where video-based HOI estimation is performed on learnt visual features. In addition, lends itself naturally to static and image-based settings.

Likewise, Chiou *et. al.* [3] propose ST-HOI, a Spatial-Temporal baseline for Human-Object Interaction detection in videos (ST-HOI), which predicts HOI with instance-wise spatial-temporal features based on trajectories. This also happens to be the method closest to our work. However, we propose several improvements over ST-HOI with our tail-aware loss formulation and a different network architecture.

2.2. Pose for HOI

Since humans are the main subjects of an HOI task, it is natural to derive cues from the physical properties of the human. In this regard, several works have attempted to use human body pose [5, 4, 19] as a signal for interaction recognition. Work in [30] propose a multi-level relation reasoning for HOI detection which utilizes human pose to capture global configuration and for extracting detailed local appearance cues uses attention. The network has a modularized architecture to predict HOI, giving an output which is interpretable based on relation affinity and part attention. In [17], the authors propose a single-stage, real-time solution, that tackles the task of HOI detection as a point detection and matching problem on HICO-Det and HOI-A benchmark. While the work of [33] proposes Interactive-Graph (in-Graph), a graph-based interactive reasoning model for inferring HOIs. They also propose inGraph-Net comprising of in-Graphs for detecting HOI and this network is free from the need of costly human pose annotations. While authors of [16] propose a Pose-based Modular Network (PMN) that explores the "absolute" and "relative spatial" pose features to improve the detection of HOI. Most of these methods differ in the way the human poses are integrated in the network. In our method, too, we use human postural cues and analyse when poses help and when they hurt the overall task.

2.3. VRD methods

Visual Relationship Detection (VRD) is a generalized version of the Human-Object Interaction (HOI) detection problem. As a result, several HOI related works draw inspiration from the VRD literature. The work of [25] proposes a Video Visual Relation Detection (VidVRD) task with an aim to explore relationships between objects in videos. This is done by detecting the visual relations in videos through object tracklet proposal, relation prediction and greedy relational association. For video relation detection, the authors in [23] make use of 1) graph convolution network (VRD-GCN) which predicts objects and their dynamic relationships, and 2) an online association method with a siamese network for relation instances association. [29] creates a Conditional Random Field (CRF) on a fully-connected spatio-temporal graph that makes use of statistical dependency of spatial and temporal structure of object relationships in videos. Additionally, gated energy function parametrization learns adaptive relations conditioned on visual observations. In [20], the authors propose a pipeline that first performs relationship detection using a graph-convolutions and then classifies the kind of interaction in an end-to-end manner. In [34], the authors propose to address the dataset skew in VRD datasets by learning to embed the visual features in the same space as the textual embedding of the object/relationship classes, which helps in generalizing even across unseen classes.

Our method, while not directly borrowing from the VRD literature, is inspired by [34] wherein we demonstrate improvements in tail-labels while using a rather simple network architecture.

3. Method

We now discuss the proposed pipeline in detail. Given a video stream $\mathcal{I} = \{I_1, I_2, \ldots, I_T\}$ of T frames, the goal is to output all the <human, predicate, object> triplets in the video. For this, the proposed network regresses human scores s_h and object category scores s_o for all the humans and objects in the scene and computes predicate scores s_{ho} for each proposed human-object pair. Note, that there can also be human-human pairs; in which case, one human is considered to be a major and the other to be a minor subject. The final triplet score s_{hoi} can then be computed as $s_{hoi} = s_h * s_o * s_{ho}$.

We discuss the network architecture, loss functions and plug-in classifiers for multi-label classification, in subsequent sections.

3.1. SERVO-HOI Pipeline

The proposed pipeline is depicted in Figure 3. The input video frames are first featurized by sequentially passing them through a pretrained ResNeXt-101 feature extractor, resulting in $f(\mathcal{I})$, a $T \times C \times d_h \times d_w$ dimensional feature vector . Given the object trajectory, $\mathcal{B}^o = \{B_1^o, B_2^o, \ldots, B_N^o\}$, with B_i^o being the bounding box at frame i, human bounding boxes $\mathcal{B}^h$ and union boxes, $\mathcal{B}^{ho}$, we extract RoI features from the ResNeXt features. These trajectories, $\mathcal{B}^h$, $\mathcal{B}^o$ and $\mathcal{B}^{ho}$ can either be fetched from the ground-truth annotations, or be estimated by detection/tracking depending on the mode of evaluation. We next perform RoIAlign [11] on the ResNeXt-101 features to obtain features $\Omega_h, \Omega_o, \Omega_{ho}$ for the human, object and union

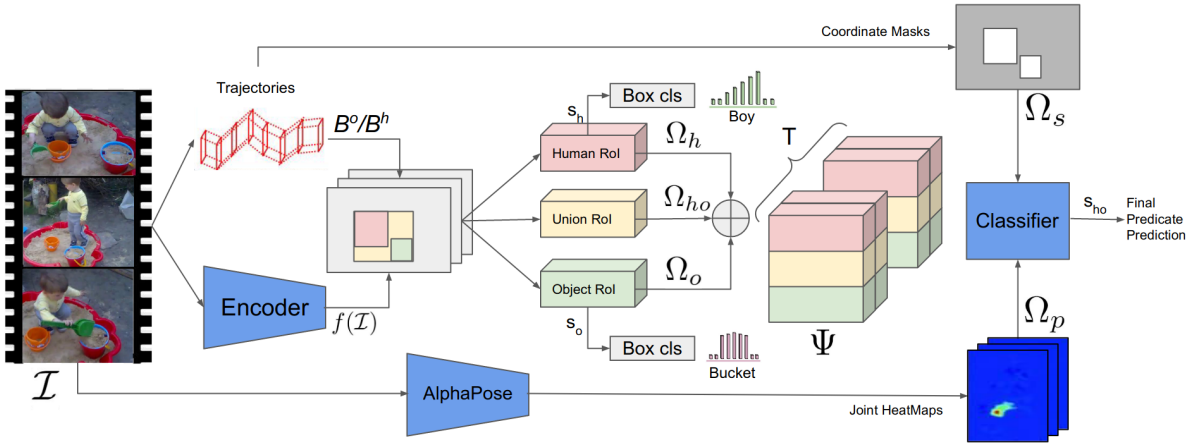


Figure 3: Overall pipeline of the proposed SERVO-HOI architecture. Given an input video segment with T frames and bounding box coordinates of the humans and objects in every frame, we (a) first extract the visual features using ResNext-101 and the RoI trajectories of humans and objects. Next, (b) the per-ROI features are extracted using RoI align and (c) aggregated together along with the pose and spatial features before being finally put through a 3D-convolutional classifier that classifies the interaction.

RoIs, respectively, with $\Omega_h, \Omega_o, \Omega_{ho} \in \mathcal{R}^{T\times C\times d_h\times d_w}$. Union RoIs corresponds to the area including both, the human as well as the object. These volume features are then simply concatenated along the channel dimension, giving us $\Psi = \Omega_h \bigoplus \Omega_o \bigoplus \Omega_{ho}$. Ψ is finally subjected to a 3D-convolution based classifier that provides the final per-predicate-class scores. It is worth noting that we do not commit to temporal pooling in the network. We believe early temporal pooling, as done in [3], leads to information loss at an early stage. Instead, we defer the time-dimension contextualization until to the 3D CNN classifier that produces the final class logits given the entire temporal context.

3.2. Incorporating Postural Cues

Human body poses could be crucial cues for recognizing the type of interaction between the human and the surroundings. While several approaches have been proposed for incorporating this information in the network, the best way to do so is still an open question. Since our pipeline already involves 3D convolutions, we found the heatmap representation [28] to be naturally amenable to our pipeline.

To this end, we first estimate the body keypoints of all the persons in the video using an off-the-shelf pose estimator [8]. Next, we construct the per-joint heatmap for each person, thereby giving us a $T \times K \times H_p \times W_p$ dimensional set of maps per person, where K is the number of body joints. The heatmaps are constructed by placing a unit gaussian around the joint's location in the human bounding box. We bring the H_p, W_p down to $d_h \times d_w$ by a series of 1×1 convolutions to get $\Omega_p \in \mathcal{R}^{T\times K\times d_h\times d_w}$. Ω_p can then be concatenated with the rest of the features to give $\Psi = \Omega_h \bigoplus \Omega_o \bigoplus \Omega_{ho} \bigoplus \Omega_p$.

3.3. Incorporating Spatial Cues

In addition to postural cues, the relative spatial ordering between the humans and the objects also provides important signals for recognizing the interactions. For example, the 2D location of the object with respect to the box-center of the human can help the model differentiate between *standing_on* and *next_to*. Prior works lilke [17] have also demonstrated that this spatial relationship can be further granularized by relating the 2D location of the objects with respect to the body joints. Such spatial information, when viewed in conjunction with postural cues, adds very little compute overhead, while providing high-quality features to the classifier. Specifically, we extract the spatial features by constructing the person-relative coordinate masks $\mathcal{I}^m \in \mathcal{R}^{T\times K\times H\times W\times 2}$ for each object in the clip. We construct K coordinate-masks for each object - one for each body keypoint. Let $l = (i, j)$ be a pixel inside the object bounding box. Considering the keypoint, $k = x, y$, we populate all the pixel positions (i, j) inside the object bounding box with the relative coordinates $(x - i, y - j)$. The relative positions outside the object box are populated with 0. These masks are processed with convolutional layers to get Ω_s, which are then concatenated with visual and postural features to produce $\Psi = \Omega_h \bigoplus \Omega_o \bigoplus \Omega_{ho} \bigoplus \Omega_p \bigoplus \Omega_s$.

3.4. Loss Functions & Sensitization to Label Skew

As discussed earlier, the VidHOI dataset - like many other visual relationship detection datasets - suffers from a high degree of dataset skew. This problem is partly inherent to the label choices, since some labels such as *next_to* and *in_front_of* are significantly more generic compared to, say, *lick*. While the prior works on video-HOI have left this

aspect unattended, we explicitly address this by experimenting with three varieties of losses designed for this purpose - weight-tuned cross entropy, propensity-weighted CE, and focal loss. Given per-predicate score $s_{ho}(i,l)$ for label l and for the i^{th} example, we compute the one-vs-all probability as $p_l^i = sigmoid\,(s_{ho}(i,l))$.

Weight-tuned CE Loss: Consider the cross entropy loss $\mathcal{L}_l$ for label l specified as

$$\mathcal{L}_l = -\sum_i [p_l^i log(p_l^i) + (1-p_l^i) log(1-p_l^i)]$$

Our first two losses are specific instances of a generic label-weighted binary cross entropy loss wherein w_l is the weight for label l

$$\mathcal{L} = \sum_l w_l \mathcal{L}_l \tag{1}$$

To begin with, we perform a grid search on the possible weight candidates (*c.f* Section 3.6) to tune the weights w_l. Naturally, this grid search is expensive and a key motivator to informed weighting methods as discussed next.

Propensity-weighted CE Loss: We observe (*c.f.*, supplementary) that the dataset has label noise such as owing to missing labels. On inspecting the output of SERVO-HOI (vanilla) on validation data, we realize that the tail labels are often missing and frequently correspond to labels with high probabilities of the mis-classified examples. To address this propensity of tail labels to go missing, we also experiment with propensity-driven cross entropy (CE) loss weighting [12] which has demonstrated improvements over other weighting methods in the literature. The propensity score ρ_l for each label, l, can be defined as:

$$\rho_l = \frac{1}{1 + C \exp -A \log(N_l + B)}$$

where N_l is the frequency of label l, A and B are hyper-parameters and $C = (\log N - 1)(B+1)^A$. Inverting the propensity score ρ_l gives us the label weight w_l, which can be plugged into eq. (1)

$$w_l = 1/\rho_l = 1 + C(N_l + B)^{-A}$$

Focal Loss: The focal loss [18] dampens the loss for well-predicted instances, thereby increasing the contribution of the poorly-predicted instances. Given hyper-parameter γ, the focal loss $\mathcal{L}_f$ can be defined as:

$$\mathcal{L}_f = -\sum_l \sum_i \left(1 - p_l^i\right)^\gamma log(p_l^i)$$

This yields us significant gains and helps avoid overfitting the data as can be seen later in Figure 5.

3.5. Optimizing for performance measures

Visual relationship detection is inherently a multi-label classification task. Two objects (or a human and an object) can be related in more than one ways. Therefore, we need a mechanism to retrieve multiple positive predictions for each human-object pair such that a Key Performance Indicator (KPI) such as mean averaged precision or F1 measure is maximized. Plug-in classifiers [21, 1] achieve this by learning to predict *Class Probability Estimate* (CPE) scores p_l^i where the final classifier is of the form $(p_l^i - \eta)$. Here, η is a threshold that is tuned to maximize one of classification accuracy, class-wise precision or a complex multivariate performance measure such as F-measure, mAP *etc.* The same CPE model can be reused to target several performance measures by simply changing the threshold tuning step. We use these tuned η values to evaluate under the Protocol 2 proposed in Section 4.

3.6. Implementation Details

We use AlphaPose [8] to extract the human pose features which produces $K = 17$ body joints; we ignore the ear and eye keypoints. The network is trained for 10 epochs using the Adam optimizer using a base learning rate of 1e-3. We follow a multi-step learning rate schedule, with learning rate dropping by a factor of 10 after the 3^{rd} and 8^{th} epochs. Training takes approximately 30 hours on a single 1080Ti GPU. In order to tune the weights for weight-tuned CE, we divide the labels into abundant (top 5 most frequent) and rest. We then experiment with weights (0.01, 0.1, 0.2, 0.5) for the abundant labels and weights (1.0, 2.0, 10.0) for the rest. We found (0.1, 1.0) to be the most optimal combination of weights.

4. Experimental Setup

We train and evaluate the proposed pipeline for the task of video Human-object interaction detection on the VidHOI dataset [3]. The VidHOI dataset has been crafted out of the larger and challenging VidOR [25] dataset. It uses only those annotations and images in VidOR which have at least one human as the subject. Note, that it allows for human-human interaction triplets as well. Overall, there are 50 predicate labels (such as *next_to, in_front_of, watch, behind*, *etc.*) and 78 object labels including the person label. This results in 557 unique <human, predicate, object> triplets. Of these triplets, 315 are considered **Rare** and appear in less than 25 instances whereas the remaining 242 triplets are considered as **Non-Rare**. Our training data consists of 6366 videos (approx. $6.5M$ frames) while the validation set is made up of 756 videos (approx. $700K$ frames). The reader is referred to [3] for more statistics of the dataset.

The dataset is evaluated under two modes, *viz.*, *Oracle* and *Detections*. In the *Oracle* mode, the network is trained

Table 1: Comparison of our approach against baselines and the state-of-the-art in *Oracle* mode. Here, 2D Model [30] and 3D Model [30] are baselines from an image-based HOI method. We use the *Protocol 1* for mAP estimation in this table.

Method	Full	Non-Rare	Rare
2D Model [30] Baseline	14.1	22.9	11.3
3D Model [30] Baseline	14.4	23.0	12.6
LIGHTEN [26]	13.4	-	-
ST-HOI [3] (vanilla)	17.3	26.9	16.3
ST-HOI [3] + Pose	17.6	27.2	17.3
SERVO-HOI (vanilla)	19.5	28.8	19.5
SERVO-HOI (weight-tuned)	20.2	28.7	18.9
SERVO-HOI (Propensity)	19.8	28.9	**19.9**
SERVO-HOI (Focal Loss)	**21.1**	**29.2**	19.5

and evaluated with ground-truth bounding box trajectories and human/object categories. In the *Detection* mode, the trajectories and class scores of humans and objects are retrieved from an off-the-shelf object detector. The recently reported work [3] on this dataset employs mean Average Precision (mAP) as the evaluation metric. A proposed positive is considered as true positive if (a) the IoU between the ground-truth and predicted bounding boxes is higher than 0.5, (b) the predicted category (including human) is correct, and (c) the predicted predicate/relationship is correct.

We evaluate our pipeline under two evaluation protocols. The first protocol, as used in [3], computes mAP by considering the top-100 proposals as positive proposals. We call this *Protocol 1*. As it assumes the top-100 proposals to be positive, the CPE-threshold based model described in Sec 3.5 cannot be evaluated using this protocol.

Therefore, we also propose *Protocol 2* wherein we drop the consideration of top-100 proposals and instead require the method to submit the class-specific confidence thresholds for the computation of positive proposals. In the absence of per-class thresholds, the methods are evaluated at 0.5 and 0.2 threshold values for all classes. We tabulate the results with this *Protocol 2*, in Table 3. We also evaluate the previous works with this protocol (with and w/o threshold-tuning) and observe that mAP scores reduce significantly when evaluated on this protocol. This drop in performance can be attributed to the fact that this protocol is relatively stricter when it comes to choosing the positive proposals.

4.1. Results

We tabulate the results on *Oracle* mode and *Protocol 1* of VidHOI dataset in Table 1. Even without bells and whistles, our vanilla model already improves upon the state-of-the-art performance. Note that this is a significant jump in performance, brought about by context inclusion and a rather simplistic network design. Further training with skew-robust losses naturally improves the performance. We observe that

Table 2: Comparison of our approach with baselines and the state-of-the-art on *Detection* mode. We use the *Protocol 1* for mAP estimation in this table. As expected, using focal loss significantly improves the detection performance, particularly for rare classes.

Method	Full	Non-Rare	Rare
2D Model [30] Baseline	2.6	4.7	1.7
3D Model [30] Baseline	2.6	4.9	1.9
ST-HOI [3] (vanilla)	3.0	5.5	2.0
ST-HOI [3] + Pose	3.2	6.1	2.0
SERVO-HOI (vanilla)	4.0	6.4	3.2
SERVO-HOI (weight-tuned)	4.2	6.4	2.9
SERVO-HOI (Propensity)	4.4	6.6	3.2
SERVO-HOI (Focal Loss)	**4.8**	**6.8**	**4.1**

Table 3: Comparison of methods on Protocol 2. In the first column, we present mAP results by a default $\eta = 0.5$ thresholding for each 1-vs-rest classifier. In the second column, the mAP results are obtained by optimizing the threshold values η for each class, based on the mAP scores on validation set (*c.f* Section 3.5).

Method	mAP ($\eta = 0.5$)	mAP (η optimized)
ST-HOI [3] (vanilla)	5.5	12.4
SERVO-HOI (weight-tuned)	9.9	14.6
SERVO-HOI (Propensity)	13.3	15.3
SERVO-HOI (Focal Loss)	7.7	16.8

such losses like focal loss and propensity-weighted CE outperform manually weight-tuned cross-entropy loss. Similar trend could be observed when we evaluate SERVO-HOI in *Detection* mode. In Table 2, we tabulate the performances on the *Detection* mode using *Protocol 1*. Being a more difficult setting, we observe natural worsening of performance. Yet, our method performs better than prior works by a respectable margin (4.8% vs 3.2%). It is worth noting that Focal loss and Propensity-weighted loss perform well under such noisy setting, especially for rare classes.

Protocol 2: Finally, we provide the evaluation results on *Protocol 2* in Table 3. Note that these numbers are not comparable with those from *Protocol 1*. As expected, optimizing the threshold η (*c.f* Section 3.5) based on mAP scores on the validation set (column 2) yields consistent performance gains over the default threshold of $\eta = 0.5$ (column 1). Also all skew sensitized losses yield performances better than than the vanilla ST-HOI.

Dataset Skew: Addressing the dataset skew is important, and this becomes evident when we train the vanilla model with weight-tuned binary cross-entropy loss. This is also evident from the improved performance on the rare classes. In order to visually display this, we plot the confusion matrix of the three modes - vanilla (*i.e* with non-weighted

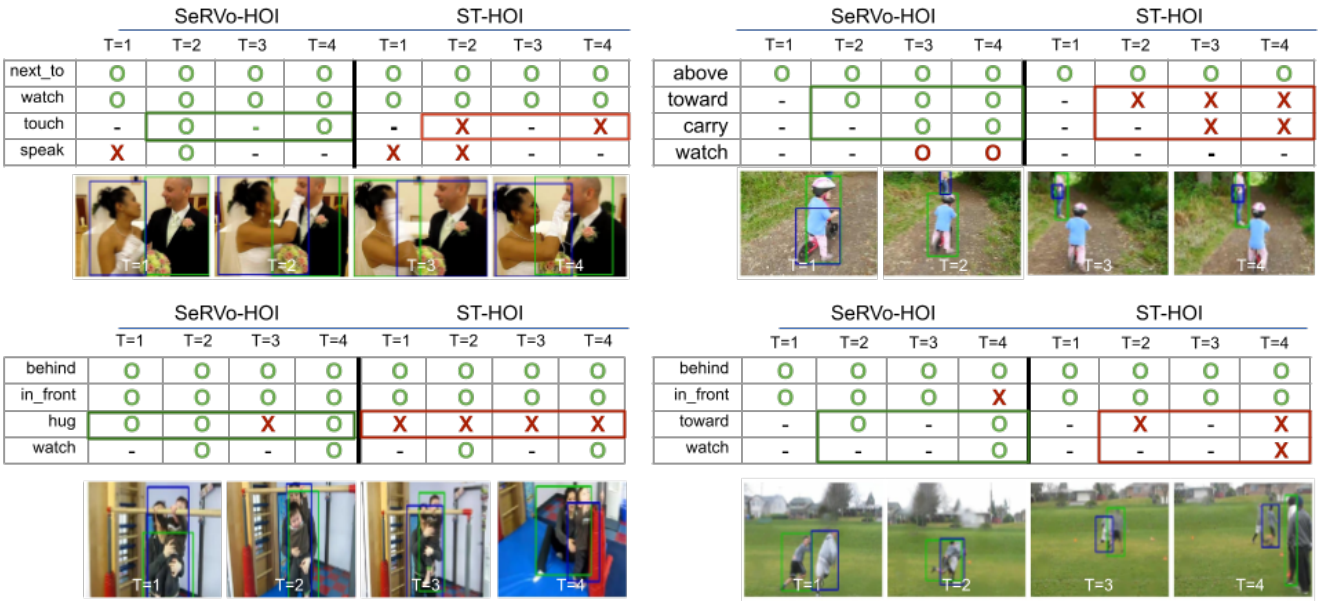


Figure 4: Qualitative performance of our method compared with ST-HOI. For each video, we show keyframes and their corresponding results on top 4 classes. While ST-HOI misses on crucial and rare labels like *carry*, *touch*, *hug*, SERVO-HOI succeeds in identifying those interactions. Also worth noting is the relative improvement in temporal predicates like *toward*.

Table 4: Ablation analysis of the multiple design choices for the network. We confirm that adding union features help but also observe that addition of human pose features does not improve performance. Further discussed in Sec 4.2.

Method	mAP
SERVO-HOI (vanilla)	19.50
SERVO-HOI w/o Ω_{ho}	19.49
SERVO-HOI w/ RNN Classifier	20.34
SERVO-HOI w/ Ω_{ho} w/ 3D Conv Classifier	21.10
SERVO-HOI + Pose	20.71
SERVO-HOI + Spatial	20.76
SERVO-HOI + Spatial + Pose	20.54

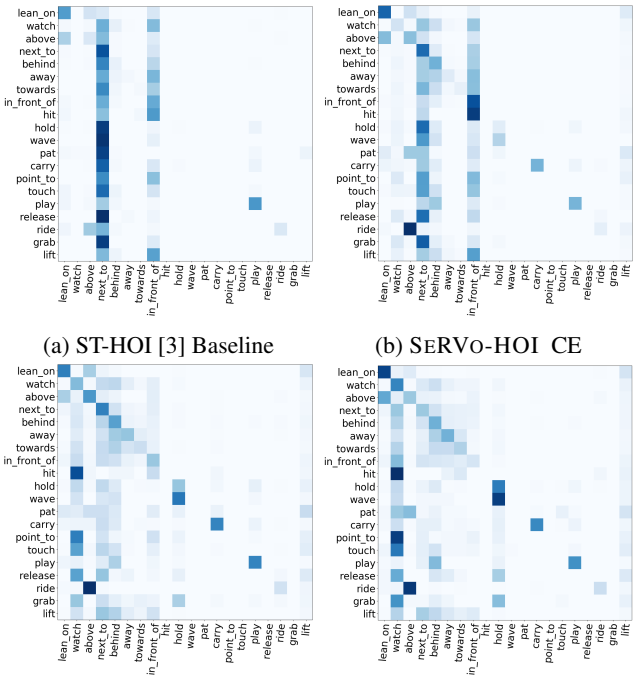
cross-entropy loss), weight-tuned cross-entropy loss and the focal loss - as well as the state-of-the-art in Figure 5. The vanilla SERVO-HOI model as well as ST-HOI suffer immensely from dataset skew, with most predictions degenerating to *next_to* and *in_front_of* classes. This picture changes when we introduce weight-tuned binary cross-entropy and focal losses. While still far from perfect, the confusion is scattered and much less concentrated on the two most dominant classes.

Qualitative Results: We provide a qualitative comparison of our method with the state-of-the-art in Figure 4. As is evident, SERVO-HOI performs well at predicting rare classes whereas ST-HOI predictions are overwhelmingly restricted to non-rare classes. We provide additional results in the supplementary material.

4.2. Ablation Studies

Effect of Pose and Spatial Features: Human pose information is well known to be a key determinant in human-object interaction. However, this can also be a double-edged sword. The VidHOI dataset consists of extremely occluded and truncated persons as shown in Figure 6. Plainly estimating human poses and feeding them into the network can be detrimental in such cases as it leads to noisy signals that further cascade to the HOI network. We note the occurence of this in Table 4, wherein the performance of SERVO-HOI suffers mildly. However, we do not observe such degradation when tested on CAD-120 with a similar network (*c.f.* Supplementary Material). In this regard, CAD-120 is a convenient dataset with full-person visibility, thereby being more amenable to postural cues. Finally, While spatial features indeed help with the performance, adding them with pose features degrades the results as can be seen in the table.

Focal loss vs weight-tuned BCE vs Propensity BCE: It is clear from Tables 1 and 2 that addressing dataset skew improves the performance in general and on the rare classes in particular. However, we are still left with the question: which amongst focal loss, propensity-weighted CE and weight-tuned CE is the most effective? It turns out that we observe similar performances with both, focal loss and propensity-weighted loss. While weight-tuned CE performs at par with other losses with ground-truth boxes using *Protocol* 1, we note that focal loss outperforms all the others in the noisy detection setting of Table 2 and *Protocol* 2 in Table 3. In either case, all the variants improve perfor-



(a) ST-HOI [3] Baseline (b) SERVO-HOI CE

(c) SERVO-HOI weight-tuned (d) SERVO-HOI Focal Loss

Figure 5: Confusion Matrix plots of multiple training variants. While ST-HOI and unweighted Cross-Entropy trained models lead to degenerated confusion across the two most abundant classes, using weight-tuned Cross-Entropy and Focal Loss noticeably disperses the confusion and produces stronger diagonals. Note, that we only plot for the top-1 predictions of 20 out of 50 classes for readability.

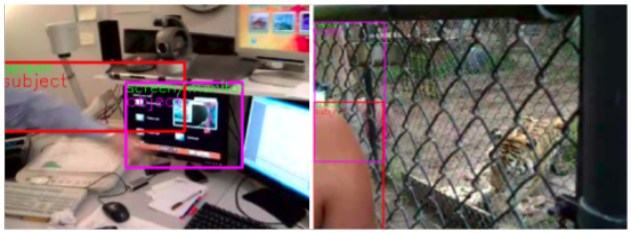


Figure 6: The heavy amount of occlusion and truncation of the human body parts in VidHOI makes it difficult to extract supervision signals from the body.

mance on rare classes compared to the baseline models. It is also worth noting that weight-tuned CE, while performing well in Table 1, has been extensively tuned for multiple values. This leads to multiple trial-and-errors to find the optimal weight configuration. Propensity weighing, on the other hand, requires minimal tuning and is a more efficient choice in this regard .

Network Design: We experiment with several architectural design choices before settling on the proposed one. Specifically, we attempted learning temporal relationships with an RNN-like sequence model as an alternative. We also experimented with graph convolutions to model the inter-object relationships in the spatial context. However, we found the SERVO-HOI to be the most optimal in terms of performance and simplicity. We discuss these network designs and their performances in the Supplementary Material.

5. Limitations and Discussion

Video based Human-Object Interaction recognition has very recent origins. That said, the proposed work leaves several research questions open for exploration. Most importantly, ours is only a *recognition* method; we assume that the start and end frames of an interaction are known. There is a strong case for also performing interaction *detection* wherein the method is required to also estimate the start and end frames of the interaction along with the type of interaction. Contemporary works [7] in video-based Visual Relationship Detection have attempted to solve this problem and similar approaches need to be considered for Human-Object Interaction domain as well.

Further, this work (and ST-HOI) does not address the sequential nature of interactions between humans and objects. For example, 'A person *holding* an object', is likely to *release/place* the object in a subsequent clip. Encoding these sequential semantics can be vital in constructing a smooth human-centric scene graph of the videos. This argument also extends to spatial relationships between the objects. *E.g.*, a person *riding* an object (an animal, in this case) is very likely to be placed *above* the object/animal. We leave modeling such relationships as a future endeavour.

Finally, although we manage to suppress the ill-effects of long tailed label distribution, more needs to be done to sufficiently address the problem. Future works in this domain would necessitate approaches such as [7, 34] that propose creative solutions to the extreme-classification problem at hand. Closely related are recent works [31] that explore zero shot HOI detection in images. We believe there is a strong case for proposing methods that perform few-shot HOI detection in videos.

6. Conclusion

In this paper, we proposed a novel pipeline that identifies human-object interactions in videos. We achieve state-of-the-art results by carefully crafting a network that accounts for the spatial and postural cues of the human body. In addition to this, we address the problem of dataset-skew and demonstrate improved performance on rare classes. Finally, we discuss issues with the existing evaluation protocols and propose solutions to avoid them. In future, we intend to further work on the long-tail label distribution problem in the context of HOI as also propose a pipeline for holistic HOI detection and recognition.

References

[1] Apoorv Aggarwal, Sandip Ghoshal, Ankith M. S. Shetty, Suhit Sinha, Ganesh Ramakrishnan, Purushottam Kar, and Prateek Jain. Scalable optimization of multivariate performance measures in multi-instance multi-label learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017*, 2017.

[2] Y.W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[3] Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. St-hoi: A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval*, 2021.

[4] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *2019 International Conference on 3D Vision (3DV)*, 2019.

[5] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, 2018.

[6] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021.

[7] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *ICCV*, 2021.

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *International Conference on Computer Vision (ICCV)*, 2017.

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollar, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. In *arXiv preprint arXiv:1505.04474*. 2015.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[12] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *ACM SIGKDD*, 2016.

[13] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021.

[14] H.S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. In *The International Journal of Robotics Research*, 2013.

[15] H.S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *TPAMI*, 2016.

[16] Zhijun Liang, Junfa Liu, Yisheng Guan, and Juan Rojas. Pose-based modular network for human-object interaction detection. *arXiv preprint arXiv:2008.02042*, 2020.

[17] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[20] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[21] Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures. In *Proceedings of NIPS*, 2014.

[22] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[23] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[24] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. ACM, 2019.

[25] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM MM*, 2017.

[26] Sai Praneeth Reddy Sunkesula, Rishabh Dabral, and Ganesh Ramakrishnan. Lighten: Learning interactions with graph and hierarchical temporal networks for hoi in videos. In *ACM Multimedia*, 2020.

[27] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, June 2021.

[28] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NeurIPS*, 2014.

[29] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[30] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019.

[31] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with novel objects via zero-shot learning. In *CVPR*, 2020.

[32] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019.

[33] Dongming Yang and Yuexian Zou. A graph-based interactive reasoning for human-object interaction detection. *arXiv preprint arXiv:2007.06925*, 2020.

[34] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *AAAI*, 2019.

[35] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, 2018.