Diagnosing Errors in Video Relation Detectors

Shuo Chen
Pascal Mettes
Cees G.M. Snoek
{s.chen3,p.s.m.mettes,cgmsnoek}@uva.nl

VIS Lab,
University of Amsterdam,
The Netherlands

Abstract

Video relation detection forms a new and challenging problem in computer vision, where subjects and objects need to be localized spatio-temporally and a predicate label needs to be assigned if and only if there is an interaction between the two. Despite recent progress in video relation detection, overall performance is still marginal and it remains unclear what the key factors are towards solving the problem. Following examples set in the object detection and action localization literature, we perform a deep dive into the error diagnosis of current video relation detection approaches. We introduce a diagnostic tool for analyzing the sources of detection errors. Our tool evaluates and compares current approaches beyond the single scalar metric of mean Average Precision by defining different error types specific to video relation detection, used for false positive analyses. Moreover, we examine different factors of influence on the performance in a false negative analysis, including relation length, number of subject/object/predicate instances, and subject/object size. Finally, we present the effect on video relation performance when considering an oracle fix for each error type. On two video relation benchmarks, we show where current approaches excel and fall short, allowing us to pinpoint the most important future directions in the field. The tool is available at https://github.com/shanshuo/DiagnoseVRD.

1 Introduction

This paper performs an in-depth investigation into the video relation detection task. Video relation detection, introduced by Shang *et al*. [19], requires spatio-temporal localization of object and subject pairs in videos, along with a predicate label that describes their interaction. To tackle this challenging problem, Shang *et al*. [19] first proposed a three-stage approach: split a video into snippets, predict the predicate, and associate the snippets over time. Such a three-stage tactic has since become popular for video relation detection [8, 17, 22, 26, 27]. Among them, Tsai *et al*. [26], Qian *et al*. [17] and Xie *et al*. [27] focus on improving predicate prediction. Tsai *et al*. and Qian *et al*. construct graphs to pass messages between object nodes, while Xie *et al*. utilizes multi-modal features. Alternatively, both Di *et al*. [8] and Su *et al*. [22] shift their attention to a better association process.

Not all works follow a canonical three-stage approach. Cao *et al*. [5], for example, propose a 3D proposal network to learn relational features in an end-to-end manner. Sun *et*

al. [23] and Liu *et al*. [15] rely on a sliding window to generate proposals and recognize predicates within proposals. Chen *et al*. [7] learn interaction primitives to generate interaction proposals [6] and recognize predicates. While video relation results keep progressing, there is still a lot of room for improvement. For example, Xie *et al*. [27], the winner of the Video Relation Detection task from the Video Relation Understanding Challenge 2020, combines a wide variety of multi-modal features for each subject-object tubelet pair to predict the relations with an improved detection performance. Nonetheless, their final mAP (mean Average Precision) is only 9.66% on the VidOR validation set [20]. In short, the task is far from solved. Moreover, it is unclear which factors are most critical for better results. We seek to fill this void.

We take inspiration from error diagnosis in the spatial domain for object detection [9, 10] and in the temporal domain for action detection [2, 16]. These works have previously performed a deep dive into the main sources of errors for their respective tasks, including false positive analysis, false negative analysis, and mAP sensitivity tests for object attributes or action characteristics. The analyses have helped to explain limitations in the field and to provide guidance for the next steps [1, 2, 3, 4, 10, 11, 12, 13, 28, 29]. In a similar spirit, we shine a light on the spatio-temporal domain for video relation detection, where the spatial challenges of object detection and the temporal challenges of action detection need to be simultaneously addressed.

We provide an error diagnosis for video relation detection, which starts with an outline of current benchmarks, evaluation protocols, the algorithms under consideration, and a categorisation of different possible error types. Under this setup, we make the following analytical contributions:

- false positive analysis outlining which types of errors are most common, along with potential cures for each error type, evaluated on two state-of-the-art approaches;
- false negative analysis along with a categorization of the kind of relation characteristics that are most difficult to detect;
- analysis of the different video relation characteristics and their influence on the performance, including relation length, number of subject/object/predicate instances, and spatio-temporal subject and object size;
- oracle analysis to identify which aspects lead to the biggest improvements.

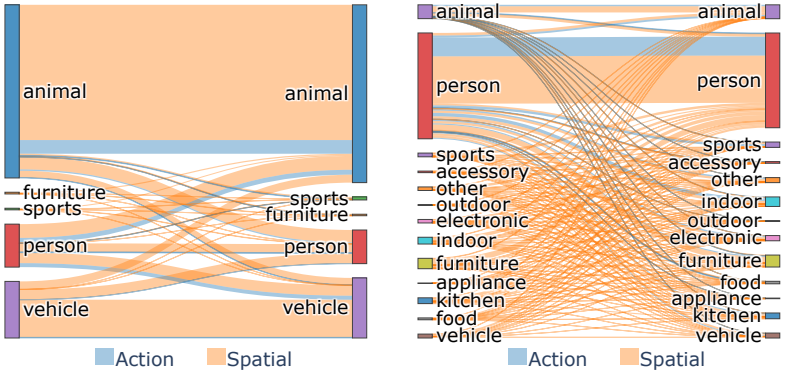
2 Error diagnosis setup

As a starting point of the error diagnosis, we first outline the core characteristics and biases of the current video relation detection datasets, the definitions of different error types, and the methods from the literature under investigation.

2.1 Dataset characterization

We perform our analysis on the two existing datasets in video relation detection, namely ImageNet-VidVRD [19] and VidOR [20].

ImageNet-VidVRD [19] consists of 1,000 videos, created from the ILSVRC2016-VID dataset [18]. There are 35 object categories and 132 predicate categories. The videos are densely annotated with relation triplets in the form of $\langle$*subject-predicate-object*$\rangle$ as well as the corresponding subjects and objects trajectories. Following [19, 26], we use 800 videos



(a) ImageNet-VidVRD. (b) VidOR.

Figure 1: Subject, object, and predictate diagrams on ImageNet-VidVRD and VidOR. On both datasets, knowledge about animals, person, vehicles, and spatial relations will go a long way for video relation detection due to a large bias towards these overarching category types.

for training and the remaining 200 for testing. We analyze the method performance on the 200 test videos.

VidOR [20] contains 10,000 user-generated videos selected from YFCC-100M [24], for a total of about 84 hours. There are 80 object categories and 50 predicate categories. Besides providing annotated relation triplets, the dataset also provides the bounding boxes of objects. The dataset is split into a training set with 7,000 videos, a validation set with 835 videos, and a testing set with 2,165 videos. Since the ground truth of the test set is not available, we use the training set for training and the validation set for testing, following [15, 17, 22, 27]. We report the analysis of method performance on the VidOR validation set.

Prevalent relations. To gain insight into the large number of possible combinations of subjects, objects, and interactions in ImageNet-VidVRD and VidOR, we first categorize all into super categories and investigate patterns among the super categories. For VidOR, the object categories are based on MS-COCO [14] and we, therefore, use its 12 object super categories, along with an *other* category for exceptions. For the predicates, we employ the hierarchy in VidOR that makes a split into *action-based* and *spatial* predicates. In the supplementary materials, we show the prevalent objects and predicates of ImageNet-VidVRD and VidOR. The animals and persons are the dominant subjects and objects, while spatial predicates form the dominant interactions between them. This is not surprising, as spatial relations are common and omnipresent.

Predicate biases. For a given dataset, the number of relations consists of all combinations of subjects, objects, and predicates. Most combinations are however not likely to occur, resulting in a bias towards common and generic $\langle$*subject-predicate-object*$\rangle$ triplets. We find that subject and object labels are highly predictive of predicate labels. Figure 1 shows which subjects and objects are likely to be in interaction and indicates which type of predicate commonly occurs between super categories of subjects and objects. To quantify the bias towards predicate categories for subject-object pairs, we predict the predicate using a naïve Bayes classifier built upon training set statistics between subjects and objects. On ImageNet-VidVRD, the predicate accuracy on the validation set is 14.02% compared to 0.8% for random guessing. On VidOR, the accuracy is 36.11% compared to 2.0% for random guessing. Evidently, there is not only a strong bias towards common predicates but also

Error type	Definition
Classification error	Overlap between discovered and ground truth relation is above 0.5, the relation triplet labels are not identical.
Localization error	Overlap between discovered and ground truth relation is between 0.1 and 0.5, the relation triplets labels are identical.
Confusion error	Overlap between discovered and ground truth relation is between 0.1 and 0.5, the relation triplets are not identical.
Double detection	Overlap between discovered and ground truth relation is above 0.5, the relation triplet are identical, but the ground truth instance has already been detected.
Background error	Overlap between discovered and *any* ground truth relation is lower than 0.1.
Missed ground truth	An undetected ground truth instance not covered by other errors.

Table 1: Categorization of six different types covering all errors that a video relation detector can make. The error types are used for our in-depth false positive analysis.

from subjects and objects to predicates. Empirically, we will investigate whether current video relation detection approaches also mirror this bias.

2.2 Evaluation protocol and error types

In the literature, the mean Average Precision (mAP) is widely used for video relation detection evaluation [15, 17, 19, 21, 22, 23, 25, 27]. Different from conventional Average Precision evaluation for detection [9], the averaging per category is performed over videos, not categories. Let G be the set of ground truth instances for a video such that an instance $g^{(k)}=(\langle s,p,o\rangle^g,(T_s^g,T_o^g))$ consists of a relation triplet label $\langle s,p,o\rangle^g$ with subject and object bounding-box trajectories (T_s^g,T_o^g). Let P be the set of predictions such that a prediction $p^{(i)}=(p_s^{(i)},\langle s,p,o\rangle^p,(T_s^g,T_o^g)$ consists of a relation triplet score $p_s^{(i)}$, a triplet label $\langle s,p,o\rangle^p$, and predicted subject and object trajectories. To match a predicted relation instance $(\langle s,p,o\rangle^p,(T_s^p,T_o^p))$ to a ground truth $(\langle s,p,o\rangle^g,(T_s^g,T_o^g))$, we require:

i their relation triplets to be exactly the same, i.e. $\langle s,p,o\rangle^p=\langle s,p,o\rangle^g$;
ii their bounding-box trajectories overlap s.t. $\text{vIoU}(T_s^p,T_s^g)\geq 0.5$ and $\text{vIoU}(T_o^p,T_o^g)\geq 0.5$, where vIoU refers to the voluminal Intersection over Union;
iii the minimum overlap of the subject trajectory pair and the object trajectory pair $\text{ov}_{pg}=\min(\text{vIoU}(T_s^p,T_s^g),\text{vIoU}(T_o^p,T_o^g))$ is the maximum among those paired with the other unmatched ground truths G, *i.e.*, $\text{ov}_{pg}\geq\text{ov}_{pg'}(g'\in G)$.

While calculating the score, we only consider the top-200 predictions for each video. After we get AP for each video, we finally calculate the mean AP (mAP) over all testing/validation videos. The above criteria make it hard for the ground truth to match the prediction. In this work, we are not only interested in the matches, but also in analyzing the mismatches. In Table 1, we have outlined six possible error types, five False Positives, and one False Negative. We visualize and show qualitative examples of true positives as well as different error types in Figure 2. We will use these error types to investigate common pitfalls in current video relation detection approaches.

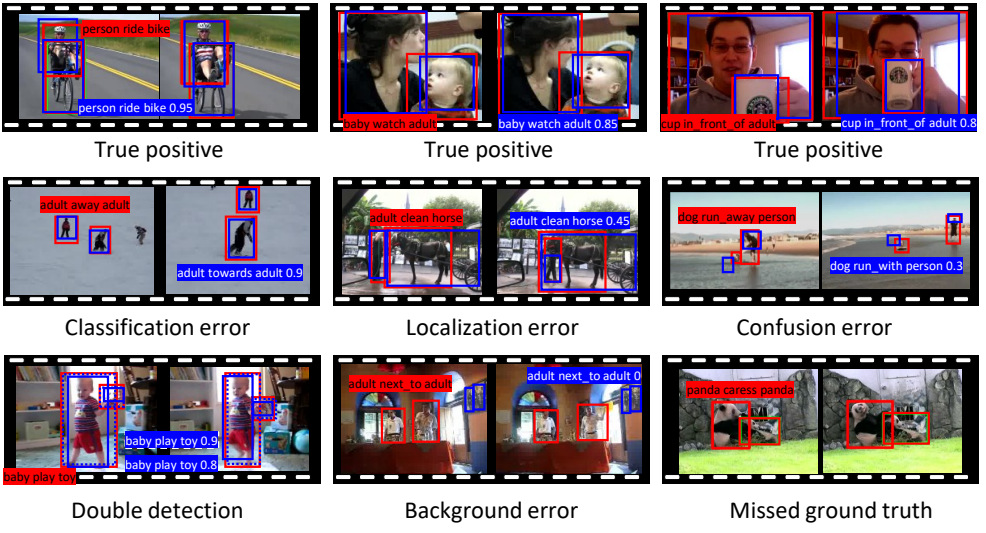


Figure 2: Video relation detection examples of true positives and the six error types from Table 1. Red boxes indicate ground truth and blue boxes specify predictions. The number in the blue box is the vIoU between the detection and the ground truth. The dashed boxes in double detection represent the best mapped prediction to this ground truth. To match a prediction to a ground truth is difficult and many factors could influence the final performance.

2.3 Algorithms under investigation

We exemplify the use of our diagnostic tool by studying two state-of-the-art approaches which have conducted experiments on ImageNet-VidVRD and VidOR. Both methods tackled the problem in a three-stage manner, similar to [19]. However, there are design differences in each stage which are relevant to highlight.

Liu *et al*. [15] avoid the need to split videos into snippets. In a first stage they generate object tubelets for the whole videos. The second stage refines the tubelet-features and finds relevant object pairs using a graph convolutional network. The third stage focuses on predicting the predicates between related pairs. In this manner, interactions can be detected without a need for snippet splitting.

Su *et al*. [22] is based on the three-stage architecture proposed in Shang *et al*. [19]. A video is first split into short snippets and subject/object tubelets are generated per snippet. Then, short-term relations are predicted for each tubelet. In the second stage, spatio-temporal features of each pair of object tubelets are extracted and used to predict short-term relation candidates. In the third stage, they maintain multiple relation hypotheses during the association process to accommodate for inaccurate or missing proposals in the earlier steps.

3 Findings

In this section, we demonstrate the generality and usefulness of our analysis toolbox by exploring what restricts the performance of video relation detection approaches. We first conduct a false positive analysis, composed of the first five error types defined in Table 1

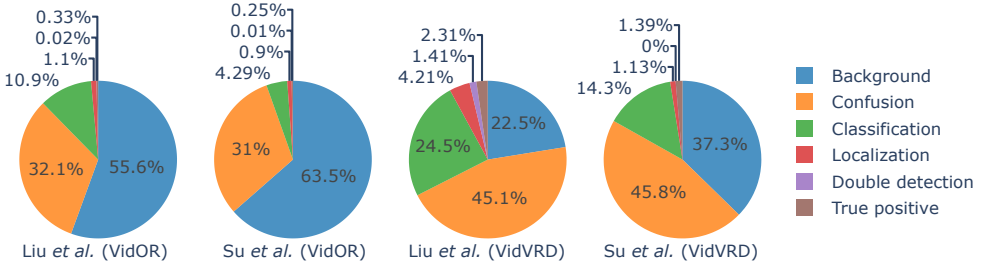


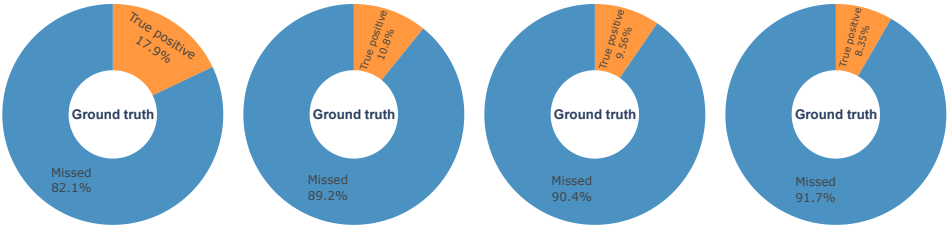
Figure 3: The false positive error breakdown in Liu *et al*. [15] and Su *et al*. [22] on the VidOR and ImageNet-VidVRD datasets. The classification error, which is also one cause of confusion error, as well as background error, should be solved first in future research.

(classification, localization, confusion, double detection, background). Then, we analyze the false negatives, *i.e*., missed ground truth (Miss), along with different relation characteristics that correlate with the false negatives. Finally, we contribute the mAP gain of each error type.

3.1 False positive analysis

The first experiment investigates which error types are prevalent in current approaches. To answer this question, we break down the false positives and present the distribution of errors for Liu *et al*. [15] and Su *et al*. [22] on ImageNet-VidVRD and VidOR in Figure 3. To our surprise we find that in all four cases, the localization error takes only a small part of all false positive in the spatio-temporal detection task. Since in diagnostic papers on well-established detection tasks such as object detection [4, 11] and temporal action detection [2], localization error is important and takes a much larger ratio. Due to the large amount of possible triplet combinations, it is more common to have both low overlapping volumes as well as wrong triplet labels, categorized as confusion error. Next, we see that there is almost no double detection error. When predicting predicates, Liu *et al*. and Su *et al*. keep the top 20 prediction results for each subject-object pair. Thus, the diversity in the predicted detection results make it difficult to map the multiple detections to the same ground truth.

Comparison across methods. From Figure 3 we can observe that the background error ratio is much lower in Liu *et al*. compared to Su *et al*. Liu *et al*. generate less detections where no interesting relations are involved. We attribute this to their proposal generation and filtering stages. Su *et al*.'s split and merge pipeline might be unable to remove bad proposals efficiently. Another observation is that Liu *et al*.'s classification error is much higher than the one of Su *et al*. on ImageNet-VidVRD. Su *et al*.'s multiple hypothesis association enables to connect neighbour segments with low predicate prediction scores. When ranking detection results, the scoring reflects the reliability of forming the corresponding hypothesis video relation, enabling a more robust ranking for those with a lower predicate prediction score. This is beneficial especially for ImageNet-VidVRD with more predicate categories but less training data, resulting in undistinguished classification scores for predicates. Su *et al*. have fewer true positives than Liu *et al*., but higher mAP. This also shows that Su *et al*.'s scoring algorithm outperforms Liu *et al*. In VidOR, with more training data and fewer predicate categories, Su *et al*. have a lower classification error ratio than Liu *et al*., but the gap is not as large as on ImageNet-VidVRD. We conclude that Liu *et al*. and Su *et al*. have their own



(a) Liu *et al*. (VidVRD) (b) Su *et al*. (VidVRD) (c) Liu *et al*. (VidOR) (d) Su *et al*. (VidOR)

Figure 4: The missed ground truth error (false positive) ratio on ground truth in Liu *et al*. [15] and Su *et al*. [22] on ImageNet-VidVRD and VidOR datasets. Both have many ground truths undetected.

advantages for dealing with different error types. Both have in common that the background error and classification error should have higher priority than the other error types to gain the most in performance.

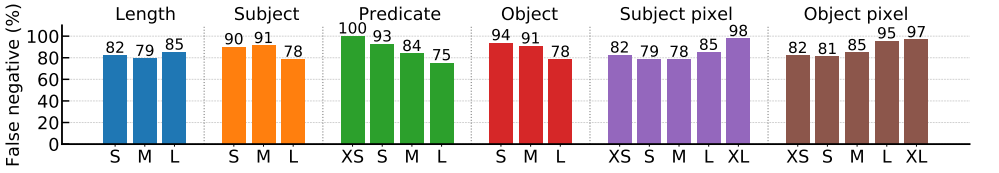
3.2 False negative analysis

So far, we have only considered the types of false positive errors introduced by the detection algorithms. However, false negative errors (missed ground truth) also influence the mAP.

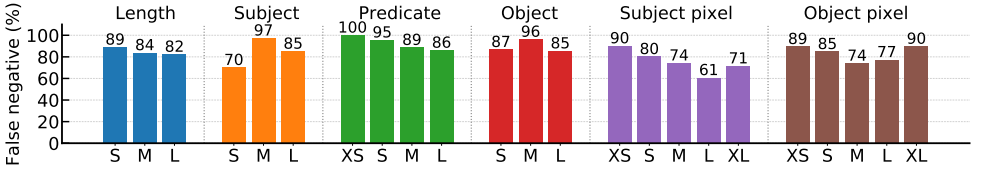
In Figure 4 we present the missed ground truth ratios for Liu *et al*. and Su *et al*. on ImageNet-VidVRD and VidOR. For both ImageNet-VidVRD and VidOR, roughly 90% of the ground truth relation instances remain undetected. VidOR has a higher missed ground truth ratio, highlighting the more complex nature of the dataset. On ImageNet-VidVRD, Liu *et al*. detect more instances than Su *et al*. but attribute them with lower scores, leading to a lower mAP value. This tells us that proposal-based methods can cover more relations, while Su *et al*.'s scoring method helps to better rank detected predictions. It is insightful to study what makes these missed ground truth instances difficult to detect. Towards this end, we group the instances according to six relation characteristics defined below:

- **Length:** we measure relation length by the duration in seconds and create three different length groups: Short (S: (0, 10]), Medium (M: (10, 20]), and Long (L: > 20). Overall, most of the instances are short, both in ImageNet-VidVRD (94.11%) and VidOR (80.06%). The number of medium and long relations are roughly similar.
- **Number of predicate instances**: we count the total number of predicate instances over all videos and create four categories: XS: (0, 10]; S: (10, 100]; M: (100, 1000]; L: (1000, 10000]; XL: (10000, 100000]; XXL: >100000.
- **Number of subject instances**: idem but for subjects.
- **Number of object instances**: idem but for objects.
- **Subject pixel scale**: we take the average of the bounding boxes for the subject trajectories and group the mean bounding box. We define subjects with pixel areas between 0 and 162 as extra small (XS), 162 to 322 as small (S), 322 to 962 as medium (M), 962 to 2882 as (L), and 2882 and above as extra large (XL).
- **Object pixel scale**: idem but for objects.

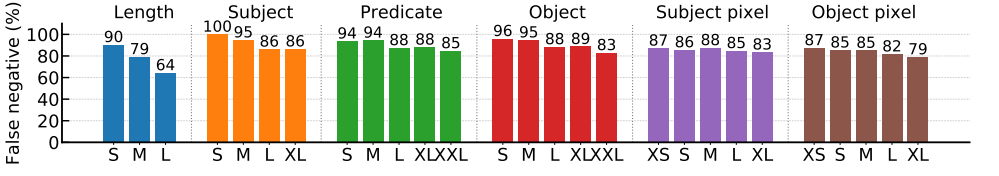
Figure 5 shows the overview of the effect for all relation characteristics for both Liu *et al*. and Su *et al*. on ImageNet-VidVRD and VidOR. We first observe a long-tail issue for the



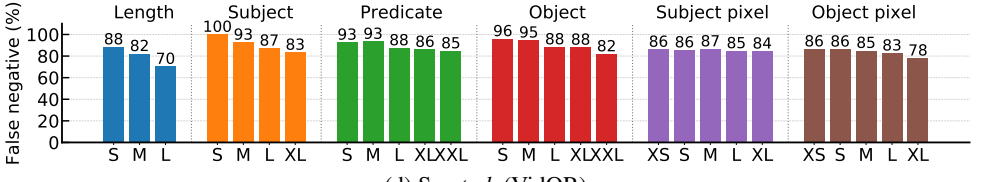
(a) Liu *et al*. (ImageNet-VidVRD).



(b) Su *et al*. (ImageNet-VidVRD).



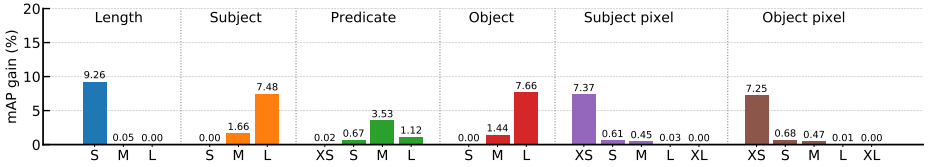
(c) Liu *et al*. (VidOR).



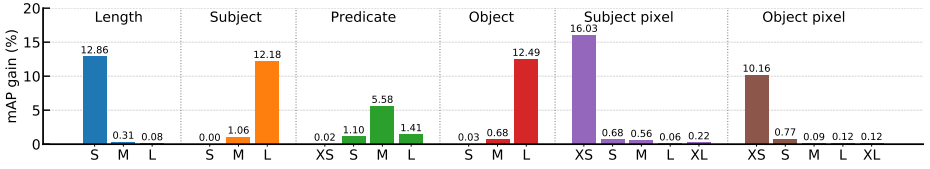
(d) Su *et al*. (VidOR).

Figure 5: Relation characteristics of Liu *et al*. and Su *et al*. on ImageNet-VidVRD and VidOR. Relations with fewer subject/predicate/object instances and smaller subject/object pixel areas are more difficult to detect.

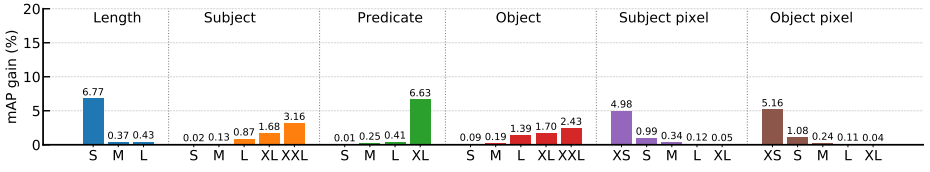
predicates. On ImageNet-VidVRD, both methods completely fail on relation instances for which the predicate category has fewer than 10 samples. This means that datasets with more training samples are essential to this task, or methods should better exploit the few available samples. Another observation is that Su *et al*. have fewer false negatives on long-range relations on ImageNet-VidVRD, even though Liu *et al*. focus on long-range representations in their approach. This may be due to the construction of the ImageNet-VidVRD dataset, which was built through asking annotators to label segment-level visual relation instances in decomposed videos. This annotation procedure results in an abundance of relations that can be recognized without the need for long-range information. VidOR is annotated differently. Given a pair of object tubelets, the annotators are asked to find and temporally localize relations, resulting in more long-lasting relations. The patterns regarding the number of subject and object instances are intuitive in VidOR; the more instances to train on the better. Moreover, subjects and objects with larger size are easier to detect than smaller size. This pattern does, however, not hold for ImageNet-VidVRD, which could be due to the overall dataset size. Since the numbers of 'XL' subpxl and 'XL' objpxl in ImageNet-VidVRD are



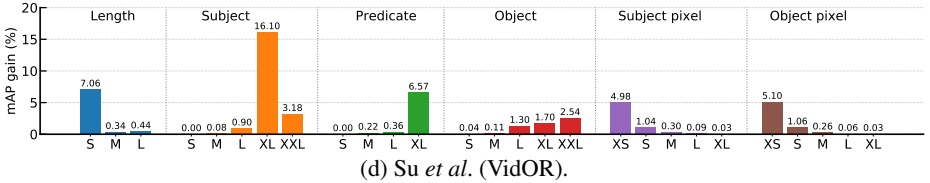
(a) Liu *et al*. (ImageNet-VidVRD).



(b) Su *et al*. (ImageNet-VidVRD).



(c) Liu *et al*. (VidOR).



(d) Su *et al*. (VidOR).

Figure 6: The mAP gain on relation characteristics of Liu *et al*. and Su *et al*. on ImageNet-VidVRD and VidOR. Focusing on detecting relation instances with a short temporal timespan, a large number of instances, and small pixel areas for the subject and object will improve the mAP by the largest margin.

much lower than in VidOR.

To deepen the analysis of each characteristic's effect, we calculate the mAP gain after dropping the missed ground truths under this characteristic. From Figure 6, we observe that not all characteristics contribute equally to gains in mAP. It reveals that to improve the final metric the most, methods should focus on detecting relation instances with a short temporal timespan, a large number of instances, and small pixel areas for the subject and object.

3.3 mAP sensitivity

Where we have so far looked into which errors are most prevalent, we also want to examine to what extent each error type in Table 1 is holding back progress. We do so by quantifying the impact on the mAP for each error type by means of an oracle fix. We show how the mAP changes when each error type would be fixed. Rather than only removing the predictions causing this error [2], we define the following cures for each of the main error types:

- **Classification cure**: Correct the class of the detection (thereby making it a true positive). If this results in a duplicate detection, remove the lower scoring detection.

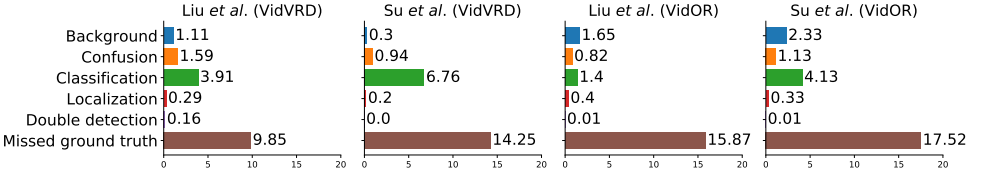


Figure 7: The mAP gain in Liu *et al*. [15] and Su *et al*. [22] on ImageNet-VidVRD and VidOR. Fixing missed ground truth error will maximize the performance improvement.

- **Localization cure**: Set the localization of the detection equal to the ground truth localization (thereby making it a true positive). If this results in a duplicate detection, remove the lower scoring detection.
- **Confusion cure**: Since we cannot be sure of which ground truth the detector was attempting to match to, we remove the false positive detection.
- **Double detection cure**: Remove the duplicate detection with lower score.
- **Background cure**: Remove the background detection.
- **Missed ground truth cure**: Reduce the number of ground truth instances in the mAP calculation by the number of missed ground truth.

Figure 7 shows the error types impact on the mAP. Note that the sum of each error type's mAP gain is not 100%. The reason is due to the property of mAP. If we fix the error types progressively, the final mAP will be 100%. But the later fixed error types will gain more weights than earlier fixed error types. For a meaningful comparison, we fix them separately. In Figure 7, fixing missed ground truth errors will improve the mAP by a large margin, Su *et al*. with 14.25% on ImageNet-VidVRD and 17.52% on VidOR. However, in practice, we cannot simply drop these missed ground truths. The solution is to include more ground truths in the selected top 200 detections of a video. And many detections that could be matched to missed ground truths are not selected due to their low scores. We believe one direction is improving the performance of the predicate prediction module, to give the background proposals low scores and proposals of correct predicate categories high scores. This will also fix the classification errors and background errors to boost the final mAP further.

4 Conclusion

This work performs a series of analyses to understand the challenging problem of video relation detection better. Using two canonical approaches, we first perform false positive analyses and define the different types of errors. Two error types are prevalent across approaches and datasets: confusion with non-matching ground truth relations and detecting relations that are part of the background. We then perform false negative analyses, which show that most ground truth instances are missed entirely. Focusing on detecting relation instances with a short temporal length, a large number of instances, and small pixel areas for the subject and object will improve the mAP the most. Lastly, to create a future outlook, we investigate several cures for common errors and find that the ability to discard background relations provides the shortest path to improve video relation detection performance. Our toolbox is generic and can be employed on top of any video relation detection approach. We make the toolbox and evaluation scripts publicly available to help researchers dissect their video relation detection approaches. Currently our tool only consider the single variant's effect to the final metric, we will investigate a multivariate statistical analysis in the future.

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016.

[2] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *ECCV*, 2018.

[3] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV Workshops*, 2014.

[4] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. TIDE: A General Toolbox for Identifying Object Detection Errors. In *ECCV*, 2020.

[5] Qianwen Cao, Heyan Huang, Xindi Shang, Boran Wang, and Tat-Seng Chua. 3-D Relation Network for visual relation recognition in videos. *Neurocomputing*, 2021.

[6] Shuo Chen, Pascal Mettes, Tao Hu, and Cees GM Snoek. Interactivity proposals for surveillance videos. In *ICMR*, 2020.

[7] Shuo Chen, Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Social fabric: Tubelet compositions for video relation detection. In *ICCV*, 2021.

[8] Donglin Di, Xindi Shang, Weinan Zhang, Xun Yang, and Tat-Seng Chua. Multiple hypothesis video relation detection. In *BigMM*, 2019.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[10] Heng Fan, Fan Yang, Peng Chu, Yuewei Lin, Lin Yuan, and Haibin Ling. Tracklinic: Diagnosis of challenge factors in visual tracking. In *WACV*, 2021.

[11] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, 2012.

[12] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. How good are detection proposals, really? In *BMVC*, 2014.

[13] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *PAMI*, 2015.

[14] 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.

[15] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *CVPR*, 2020.

[16] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. *BMVC*, 2020.

[17] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *ACM MM*, 2019.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

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

[20] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, 2019.

[21] Xindi Shang, Junbin Xiao, Donglin Di, and Tat-Seng Chua. Relation understanding in videos: A grand challenge overview. In *ACM MM*, 2019.

[22] Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, and Tat-Seng Chua. Video relation detection via multiple hypothesis association. In *ACM MM*, 2020.

[23] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. Video visual relation detection via multi-modal feature fusion. In *ACM MM*, 2019.

[24] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.

[25] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019.

[26] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *CVPR*, 2019.

[27] Wentao Xie, Guanghui Ren, and Si Liu. Video relation detection with trajectory-aware multi-modal features. In *ACM MM*, 2020.

[28] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *CVPR*, 2016.

[29] Hongyuan Zhu, Shijian Lu, Jianfei Cai, and Quangqing Lee. Diagnosing State-Of-The-Art Object Proposal Methods. *BMVC*, 2015.