Contextual Visual Feature Learning for Zero-Shot Recognition of Human-Object Interactions

Tsung-Wei Ke[1] **Dong-Jin Kim**[2] **Stella X. Yu**[1,3] **Liang Gou**[4] **Liu Ren**[4]

[1]UC Berkeley / ICSI [2]Hanyang University [3]University of Michigan

[4]Bosch Research North America & Bosch Center for Artificial Intelligence

Abstract

Real-world visual recognition of an object involves not only its own semantics but also those surrounding it. Supervised learning of contextual relationships is restrictive and impractical with the combinatorial explosion of possible relationships among a group of objects. Our key insight is to formulate visual context not as a relationship classification problem, but as a representation learning problem, where objects located close in the feature space have similar visual contexts. Such a model is infinitely scalable with respect to the number of objects or their relationships.

We develop a contextual visual feature learning model without any supervision on relationships. We characterize visual context in terms of spatial configuration of semantics between objects and their surrounds, and derive pixel-to-segment learning losses that capture visual similarity, semantic co-occurrences, and structural correlation. Visual context emerges in a completely data-driven fashion, with objects in similar contexts mapped to close points in the feature space.

Most strikingly, when benchmarked on HICO for recognizing human-object interactions, our unsupervised model trained only on MSCOCO significantly outperforms the supervised baseline and approaches the supervised state-of-the-art, both trained specifically on HICO with annotated relationships!

1 Introduction

Real-world visual perception of an object is far more complex than its own semantic categorization. What surrounds the object has a great impact. For example, drivers pay more attention to pedestrians hustling through an intersection than trolling down a sidewalk; A baby holding a *knife* versus a *bottle* would be seen *and* reacted differently by their caregivers. That is, real-world object recognition is not about attaching class labels to individual objects in isolation, as studied in computer vision recognition benchmarks nowadays, but about recognizing objects *along with* their contexts.

Visual context has been conventionally characterized by statistical co-occurrences of patches and objects, although its definition varies with different formulations: It has been modeled as spatially organized image feature (e.g., *scene gits* [1]), co-occurring object semantics [2, 3, 4, 5, 6], instance statistics [7], or co-occurring instance graphs [8].

Higher-level visual tasks naturally require the differentiation of visual contexts. In human-object interaction (HOI) detection [9, 10, 11], action recognition [12, 13], or scene graph generation[14], semantic classes are defined not just based on the collection of objects themselves, but their poses and relationships with each other: A *person* could *push* a *bike*, *ride* a bike, or *lean against* a bike. In image captioning [15, 16] and visual question answering [17, 18], co-occurring statistics is further refined to reflect that interesting events (not *any* events) are more likely to be named. To understand a movie [19], object relationships are extensively reasoned spatio-temporally and semantically.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

One way to learn contextual relationships is from annotations. While annotating the semantic category of objects is time-consuming but not infeasible, annotating visual contexts quickly becomes impractical with an increasing number of object categories. For example, Visual Genome [14] has $33,877$ objects and $42,374$ *pair-wise* object relationships alone. With group-wise or spatio-temporal contexts, the actual number of relationships explodes exponentially. It is not only hard for humans to annotate, but also ineffective for models to predict a large set of contextual relationships.

Our key insight is to approach visual context as a representation learning problem, *not* a classification problem [20, 21]. Instead of predicting discrete relationship categories, we learn to map object instances of similar (*dissimilar*) contexts close (*far*) in some feature space. Where an object instance is located in the feature space indicates the type of visual context it belongs to. Such a representation learning model is infinitely scalable, unconstrained by the total number of objects or relationships.

We demonstrate the above concept by learning contextual visual features for recognizing human-object interactions without using any annotations on such relationships. Existing annotated interactions only consider a restrictive subset of object-pair relationships, e.g., the pairwise relationship of *a person riding a horse* detected on green grass vs. a soccer field carries different perceptual qualities. We therefore use large-scale generic image datasets such as MSCOCO [16] as the training set, although they are only annotated with object instances and their semantic categories [22].

We model visual context in terms of spatial configuration of semantics between objects and their surrounds, and train their feature representations in a contrastive fashion accordingly. We use a convolutional neural network (CNN) to learn a pixel-wise feature mapper that encodes visual information centered at each pixel semantically and spatially. Pixels that are closer in the feature space have not only similar visual appearances in the same semantic category, but also similar spatial arrangements of surrounding semantics. For example, *a person riding a horse on grass*, *a person riding a horse on street*, *a person walking a horse on grass* should form their individual clusters instead of being mixed up in one cluster.

We formulate a pixel-to-segment contrastive learning loss [20] for *contextual visual feature learning*, where pixels are attracted to their positive segments and repelled from their negative segments. The positive and negative segment sets for each pixel are defined based on not only its own instance and semantic information [23], but also its surrounding semantics.

Visual contexts emergent in such contrastively learned features are completely data-driven and more general than supervisedly learned models. Benchmarked on HICO [10] for recognizing human-object interactions for each person instance, our unsupervised model trained only on MSCOCO with annotations of semantics not relationships outperforms the basic supervised relationship classifier and approaches the state-of-the-art supervised model, both specifically trained on HICO relationships! In addition, we show that unsupervised characterization of visual context helps learn more discriminate features that can improve semantic segmentation performance.

2 Related Work

Instance context. Earlier works studied instance contextual relationships mainly to improve object detection. [8] proposed an instance-wise exemplar and 2D spatial graph to model context. [3, 6] and [4, 5] proposed Hand-crafted features and tree-based models to capture context in terms of co-occurring statistics and spatial configurations among objects and their semantics [7]. Recent works have developed graphs [24] or spatial memory [25] to encode context in their deep learning models. We instead capture context implicitly in our learned feature space, and remarkably, we are able to recognize high-level contextual relationships (e.g., human-object interactions) automatically..

Human-object interaction. Since various Human-Object Interaction detection works constructed large-scale labeled image datasets [9, 10, 26, 27], significant progress has been achieved for this problem with different methods such as box transformations [9, 28, 29, 30], two channel interaction [10, 31], mutual contexts of human pose and object [29, 32, 33], Contextual correlation [34], correlation prior of interactions [35], Visual transformer [36, 37, 38, 39], or Graph modeling [40, 41, 42, 43, 44]. Beyond typical human and object appearance features, in order to improve the generalizablilty of the relationship detection, HOI detection works devise various information as input such as human pose [29, 45, 46] or linguistic prior knowledge [47, 40, 48, 49, 50], which require extra human labeling effort to capture such knowledge. More recent works combine

these various cues [29, 46, 51, 52]. Our model does not require additional cues such as human pose or language other than RGB images.

Weakly-supervised [53] and zero-shot relationship detection [54, 55, 56, 49, 51, 57] have been studied to improve data efficiency. However, existing zero-shot learning works require large-scale external data to pre-train linguistic knowledge. Although [35, 58] uses prior knowledge that can be obtained from the target training data itself, their method shows limited generalizability on unseen image domains. Weakly-supervised learning still requires laborious image-level annotations.

In contrast, our unsupervised learning is more general: It requires neither target domain information nor relationship labels when training on source data such as MSCOCO. That is, our unsupervised visual context predictor delivers better zero-shot performance than the supervised counterpart!

3 Contextual Visual Feature Learning without Supervision

We approach contextual relationship recognition as a feature learning problem. We map pixels to points in a feature space, such that object instances are grouped (separated) if they have similar (different) contextual relationships. Our model does not use or output any pre-defined relationship categories; it groups objects according to their own semantics *and* visual contexts. If a relationship label is desired, we retrieve nearest neighbours of a query in the feature space and transfer their labels.

Unlike supervised learning methods that train a model based on annotated (restrictive) relationships, e.g., *a person riding a bike*, our unsupervised relationship learning method trains a model based on the semantic category distribution at surrounding neighboring patches of the centered object instance.

3.1 Our Task: Unsupervised Visual Relationship Learning

Supervised setting. Given an image and a set of detected objects, visual relationship labeling [9, 10, 14] infers the relationship among object instances. Supervised methods [28, 36, 42] can only reason in restricted terms specified by training labels, e.g., between *a pair of objects*. To understand the relationship among *a group of objects*, higher-order information needs to be further extracted.

Unsupervised setting. In a stark contrast to these existing methods, we consider a more general but unsupervised learning setting. We assume no prior knowledge of relationship categories. We train our model on a generic image dataset, given only semantic and instance labels on pixels. Our goal is to infer the relationship of each object instance in a test image. For simplification, we detect object instances using off-the-shelf detectors or ground-truth bounding boxes. For inference, we extract features within an object's bounding box, retrieve their nearest neighbors from a labeled set, and predict relationships by transferring neighbors' labels (see Fig. 1).

Evaluation metric. We evaluate the retrieval performance based on the interpolated average precision (AP) metric [59, 60]. We calculate recall (R) and precision (P) by comparing the query's label to the retrieved ones. AP measures the interpolated area under the PR-curve, and is commonly adopted for instance detection and segmentation tasks [16]. See [60] for more details.

3.2 Our framework: Pixel-to-segment Contrastive Learning

SegSort [20] is an end-to-end feature learning framework that learns pixel-wise features and the corresponding segmentation based on EM-optimization that maximizes the discrimination among image segments from the entire dataset.

Specifically, a CNN ϕ maps image I to pixel-wise features $V = \{\boldsymbol{v}_i\}$, where $\boldsymbol{v}_i = \phi(x_i)$ denotes the unit-length features centered at pixel x_i. When V is fixed, SegSort generates an image segmentation using the spherical K-Means algorithm [61]. The E-step assigns pixels to their nearest segments. The M-step updates segment features $U = \{\boldsymbol{u}_s\}$ as the length-normalized average pixel feature within each segment: $\boldsymbol{u}_s = \frac{\sum_{i \in R_s} \boldsymbol{v}_i}{\| \sum_{i \in R_s} \boldsymbol{v}_i \|}$, where R_s is the area of segment s.

Let $S = \{s\}$ be the set of segments and z_i the segment index of pixel i. The posterior probability of pixel i belonging to segment s is formulated as: $p(z_i = s | \boldsymbol{v}_i, U) = \frac{\exp(\kappa \boldsymbol{u}_s^\top \boldsymbol{v}_i)}{\sum_{t \in S} \exp(\kappa \boldsymbol{u}_t^\top \boldsymbol{v}_i)}$, where κ is the

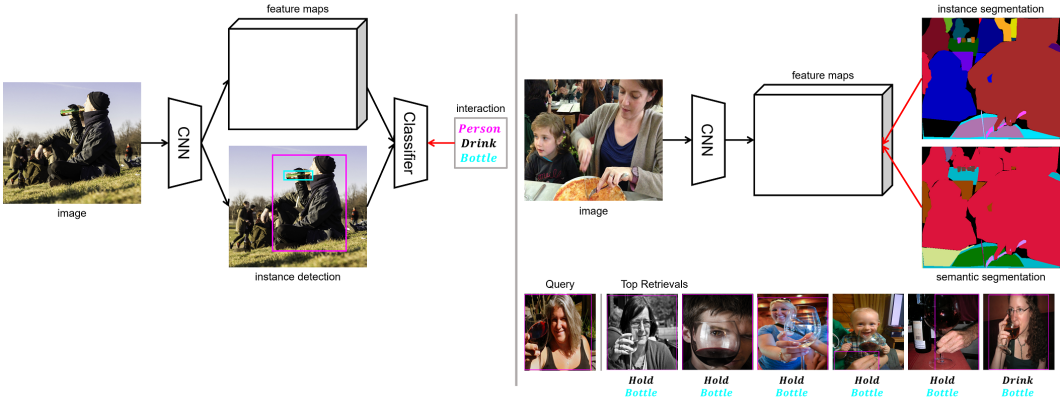


Figure 1: Our framework can discover high-level visual contextual relationships automatically. Take recognition of human-object interactions for example [10], interaction labels are composed of (**person, interaction,** object) triplets. **Left:** Supervised frameworks consider the task as a discrete classification problem. They can only reason in restricted terms specified by training labels, e.g., between *a pair of objects*. **Right:** Our framework tackles the task as a feature learning problem. We learn the feature mappings from semantic and instance labels on pixels. Without any prior knowledge of relationship categories, we predict the interactions of the query person subject by transferring nearest neighbors' labels. Red arrows indicate loss signals.

concentration hyper-parameter. To increase the discrimination among segments, pixel features are optimized to minimize the corresponding negative log-likelihood loss: $-\log p(z_i = s|\boldsymbol{v}_i, U)$.

When ground-truth labels C are provided, SegSort adapts the loss in a soft neighborhood assignment formulation [62] to enhance groupings of same-label segments. The pixel-segment contrastive loss is:

$$L(C) = -\log \sum_{s \in C_i^+} p(z_i = s|\boldsymbol{v}_i, U) = \frac{\sum_{s \in C_i^+} \exp(\kappa \boldsymbol{u}_s^\top \boldsymbol{v}_i)}{\sum_{t \in C_i^+ \cup C_i^-} \exp(\kappa \boldsymbol{u}_t^\top \boldsymbol{v}_i)} \tag{1}$$

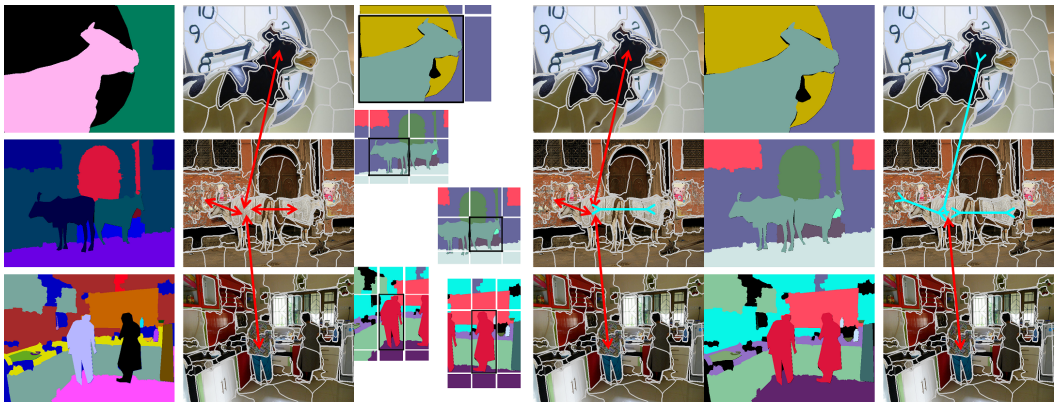
where C defines the positive (negative) set C_i^+ (C_i^-) for pixel i. C_i^+ includes all same-label segments except i's own segment, and C_i^- denotes the set of different-label segments.

3.3 Our Loss: Contextual Visual Feature Learning

The ideal contextual feature mapper should capture not only the visual appearance of the object itself, but also the statistical distribution and spatial organization of the surrounding semantics. We optimize the pixel-wise feature mapper with three pixel-to-segment contrastive losses that encode local-to-global visual contexts: **1)** instance-wise discrimination, **2)** instance-level co-occurring semantic statistics, and **3)** image-level semantic co-occurrences (Fig. 2). We also introduce an additional regularization term, resulting in a total of 4 terms in the loss function.

Instance-wise discrimination. The idea is to push instances away from others such that only visually similar instances stay close in the feature space. Following [23], we contrast pixels with segments based on their instance labels C_O. Positive segments are the ones within pixel i's instance; negative segments include different-instance segments within and other than i's image.

Instance-level co-occurring semantic statistics. The surrounding context should also indicate how to develop feature mappings for each object instance. For example, a bike rider should be distinguished from a motorbike rider. We quantify the surrounding contexts and define contrastive relationship accordingly. Specifically, we calculate the semantic category distribution at the center and eight neighboring patches of the centered object, where the patch size is the same as object's height and width (Fig. 3). Within each patch, we measure the occurrence of each semantic category (including both *things* and *stuff*), resulting in a binary contextual feature vector.



(a) instance-wise discrimination (b) instance-level co-occurring semantic statistics (c) image-level co-occurring semantic statistics

Figure 2: We construct three types of contrastive relationships to encode local-to-global visual contexts in our learned feature mappings. Pixels are attracted to (repelled by) segments: **(a)** of the same (different) instance, **(b)** of similar (distinctive) semantic surrounds, and **(c)** belong to similar (different) image-level scenes. Such global scenes can be approximated by the occurrence of semantic categories in the image. The idea contextual feature mapper should capture both the visual appearance of the object itself, and the statistical distribution and spatial organization of the surrounds.

We define local context pseudo labels C_L based on such second-order statistics. We compute the Hamming distance between the contextual feature vectors of different objects. For pixel i, we define positive segments as the ones belonging to the top-ranked neighbors of i's object; others are negative segments. Positive segments are restricted to have the same semantic category. Our goal is to encourage groupings of object instances embedded in similar contexts.

Image-level co-occurring semantic statistics. We impose a more global regularization at the scene context level. Following [21], we characterize scene context in terms of the occurrences of semantic categories. Images with similar distribution of semantic categories tend to have similar scenes.

We ignore the spatial layout, and measure the occurrence of semantic categories within each image, from which we define global context pseudo labels C_G. For pixel i, its positive set includes all segments from the set of images which share at least one semantic category with i's image. All other segments are considered negative. That is, we desire pixels to be separated by their scene types.

Predictive coding regularization. We additionally impose a predictive coding loss to explicitly enforce structured correlation among pixel features. Intuitively, the feature at one pixel should help predict the feature at other pixels in the image. Following [63], we apply the regularization in a denoising autoencoding manner. We derive a noisy set of pixel features V' by randomly masking out a subset of pixel features from V. The goal is to reconstruct V from V' using multiple encoder and decoder layers. Let ψ be the autoencoder, the loss is: $L_M = \|V - \psi(V')\|_2^2$. See [63] for details.

Total training loss. There are 3 pixel-to-segment contrastive feature loss terms and 1 predictive coding regularization term: $L = \lambda_O L(C_O) + \lambda_L L(C_L) + \lambda_G L(C_G) + \lambda_M L_M$. The two types of losses are complementary to each other: The former enhances feature discrimination without any regard to spatial correlation, whereas the latter enforces structured correlation among pixels within an image, without any regard to instances in different images. We integrate these two aspects in the overall loss to optimize our contextual visual feature.

4 Experiments

We detail our training/testing procedures, and then benchmark our unsupervised visual context model on zero-shot recognition of human-object interactions and additionally semantic segmentation.

Dataset: HICO [10] is a generic human-object interaction dataset. It is labelled with 600 human-object interaction categories w.r.t 80 object categories. Both human and object bounding boxes are provided. The dataset has $38,118$ and $9,658$ images for training and testing.

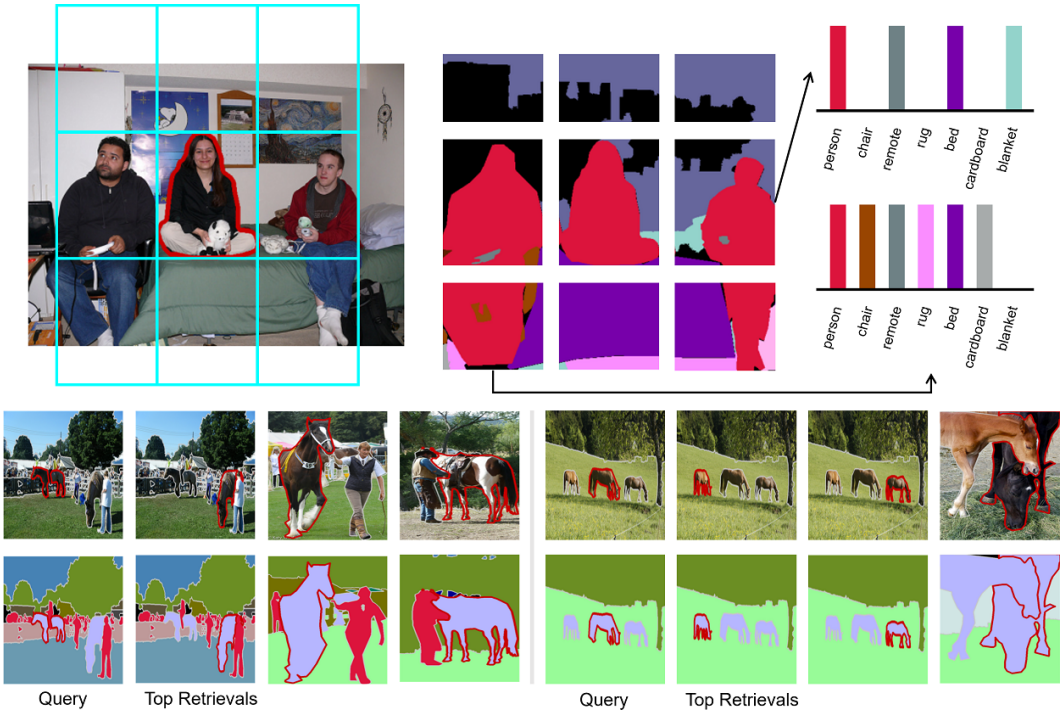


Figure 3: We quantify local visual contexts by calculating the semantic category distribution within each of the nine patches of the centered object (circled by red lines). The patch size is the same as object's height and width. **Top:** For each patch, we measure the occurrence of each semantic category, resulting in a binary contextual feature vector. **Bottom:** Based on such statistical contextual features, we conduct nearest neighbor search for each query object. Top-ranked retrievals have similar semantic surrounds as the query. We contrast pixels according to such second-order statistics to encode visual contexts.

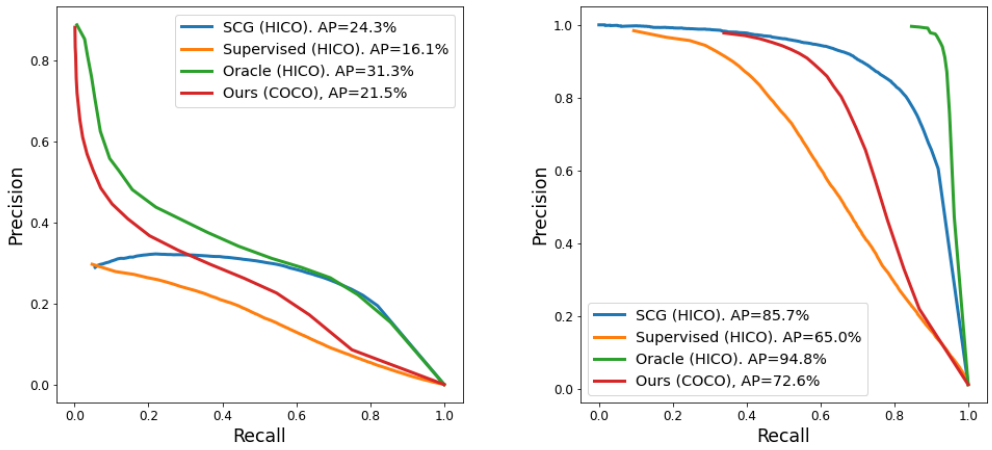
Dataste: MSCOCO [16] is a complex scene parsing dataset with 80 *things* and 91 *stuff* categories. Images have a high variety of visual scenes, such as dining, snow skiing, boat piloting, horse riding *etc*. We adopt *train2017* split (118K images) for training.

Dataset: Cityscapes [64] is a dataset for urban street scene parsing. It contains 19 *things* and *stuff* categories, such as road, pedestrian, and cars *etc*. $5,000$ images are annotated with high-quality pixel labels, which are split into $2,975$, 500 and $1,525$ for training, validation and testing.

Dataset: Pascal VOC 2012 is an object-centric semantic segmentation dataset, labelled with 20 object categories and a background class. Compared to MSCOCO, the image scenes are less complex, with an average of 2.3 objects occur per image (7.3 objects for MSCOCO). We augment the training set with additional images [65], resulted in $10,582$ and $1,449$ for training and validation.

Supervised baselines. We consider two kinds of supervised baseline methods for comparison: 1) Spatially Conditioned Graphs (SCG) [42], and 2) vanilla binary classifier. For SCG, we perform inference using HICO-trained ResNet50-FPN model weights, such that both object detector and interaction classifier are fine-tuned on HICO dataset. For vanilla binary classifier, we adopt exactly the same architecture as our method, but average pool pixel-wise features within each human bounding box. Additional two 1×1 convolutional layers are used as the binary classifier to predict the occurrence of each kind of interaction. Notably, SCG requires pairing a human with an object to classify their interaction, whereas, vanilla binary classify considers each human individually.

Oracle baselines. We consider an oracle baseline method using ground-truth semantics for HOI recognition. On HICO dataset, we compute instance-level co-occurring semantic statistics using ground-truth bounding boxes. We convert bounding boxes into instance and semantic pixel labels. For each human instance, we calculate semantic category distribution at the center and eight neighboring patches. We perform nearest neighbor search using such binary context-induced features to infer HOI



(a) Interaction category (b) Interacting object category

Figure 4: Our unsupervised model approaches the supervised state-of-the-art for recognizing HOIs on HICO. **Left:** Performance evaluated on 600 interaction categories. **Right:** Performance evaluated on 80 interacting object categories. Our framework discovers high-level contextual relationships without any prior knowledge of relationship categories.

for the query human instance. Notably, our framework applies the procedure only during training on MSCOCO, whereas, the oracle baseline infers using HICO ground-truths.

Testing. For inference with our framework on HICO, we average pool and length normalize pixel-wise features within each human bounding box. We use the ground-truth human boxes but not object boxes. For SCG, object boxes are predicted with the object detector. For the oracle baseline and our method, We retrieve 20 nearest neighbors to predict interaction labels. For each query, we count the occurrence of each interaction category, and adjust the threshold from minimum to maximum number of occurrence. For both supervised baselines, we adjust the threshold w.r.t the classification scores. Threshold is applied to decide if the interaction category is detected. We plot the PR-cure and calculate AP performance correspondingly.

For inference of semantic segmentation on Cityscapes and VOC, we follow [20] to predict pixel labels by nearest neighbor search. See [20] for more details.

For all experiments, we dot not use multi-scale but only single-scale images during inference.

Results on HICO: HOI recognition. We present the quantitative results for HOI recognition on HICO dataset. As shown in the left figure of Fig. 4, our method is upperbounded by the oracle baseline, and we achieve 68.7% of the oracle performance ($21.5\% vs. 31.3\%$ AP). Remarkably, our method has never seen any HICO image and label, but still obtains comparable performance w.r.t the SOTA supervised baseline: SCG ($21.5\% vs. 24.3\%$ AP). We report the performance based on the interacting object category not the interaction category in the right figure of Fig. 4. Our method achieves 76.5% and 84.7% performance with respect to the oracle and SCG baseline.

We summarize that training using ground-truth labels does not guarantee good testing performance to distinguish humans with different interactions. Our learned contextual features work as well as supervised classifiers. However, there is still room for improvement for our method to group instances according to co-occurring object semantics more precisely.

Note that HICO dataset annotates the human bounding boxes for each interaction label. Although the same human instance could have multiple interactions, we conduct inference on the human bounding box of each interaction label, individually. We do not filter out duplicated human instances, resulting in noisier predictions and less optimal performance than the ones reported in [42].

Figure 5: High-level contextual semantics emerge from our learned feature mappings. On HICO, we compute the average features within each human bounding box and conduct nearest neighbor retrievals. The ground-truth interaction label are shown in the form of <interaction, object> pair and put below each human instance. Strikingly, we found instances with the similar contextual relationships are close in the learned feature space. N/A denotes 'no interaction' category.

Dataset	Method	mIoU.
Cityscapes	SegSort	69.49
	Our framework	70.38
VOC	SegSort	75.98
	Our framework	77.71

Table 1: Our contextual regularizations improve semantic segmentation.

Results on VOC and Cityscapes: semantic segmentation. We summarize the efficacy of the proposed contextual regularizations for semantic segmentation on VOC and Cityscapes dataset in Table 1. Compared to SegSort [20], which uses only semantic pixel labels, we improve the semantic segmentation performance by 0.89% and 1.73% mIoU on Cityscapes and VOC. We show that our proposed regularizations help recognition in terms of capturing not only pixel itself, but also the surrounding contexts.

Visual results. We present visual results of nearest neighbor retrievals using our learned feature mappings in Fig. 5. Human instances of the similar contextual relationships are grouped.

Summary. We develop a contextual visual feature learning model to tackle recognition of human-object interactions. Without any supervision on relationships, our model approaches the supervised state-of-the-art and is able to discover such high-level contextual relationships automatically. **Limitations:** Our model still requires human-labeled supervision, in terms of semantic and instance labels on pixels. **Potential negative societal impacts:** Our work does not introduce new societal impacts but share them with any other recognition or segmentation works.

References

[1] Oliva, A., Torralba, A.: The role of context in object recognition. Trends in cognitive sciences **11**(12) (2007) 520–527

[2] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.J.: Objects in context. In: ICCV. (2007)

[3] Lee, Y.J., Grauman, K.: Object-graphs for context-aware category discovery. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE (2010) 1–8

[4] Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting hierarchical context on a large database of object categories. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE (2010) 129–136

[5] Choi, M.J., Torralba, A., Willsky, A.S.: A tree-based context model for object recognition. IEEE transactions on pattern analysis and machine intelligence **34**(2) (2011) 240–252

[6] Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 891–898

[7] Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6619–6628

[8] Malisiewicz, T., Efros, A.: Beyond categories: The visual memex model for reasoning about object relationships. In: NIPS. (2009)

[9] Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)

[10] Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 ieee winter conference on applications of computer vision (wacv), IEEE (2018) 381–389

[11] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020)

[12] Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

[13] Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6047–6056

[14] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1) (2017) 32–73

[15] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. (2015) 2641–2649

[16] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755

[17] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. (2015) 2425–2433

[18] Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 6700–6709

[19] Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: European Conference on Computer Vision, Springer (2020) 709–727

[20] Hwang, J.J., Yu, S.X., Shi, J., Collins, M.D., Yang, T.J., Zhang, X., Chen, L.C.: Segsort: Segmentation by discriminative sorting of segments. In: ICCV. (2019)

[21] Ke, T.W., Hwang, J.J., Yu, S.X.: Universal weakly supervised segmentation by pixel-to-segment contrastive learning. ICLR (2021)

[22] Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: CVPR. (2019)

[23] Hwang, J.J., Ke, T.W., Yu, S.X.: Contextual image parsing via panoptic segment sorting. In: Multimedia Understanding with Less Labeling on Multimedia Understanding with Less Labeling. (2021) 27–36

[24] Chen, X., Li, L.J., Fei-Fei, L., Gupta, A.: Iterative visual reasoning beyond convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7239–7248

[25] Chen, X., Gupta, A.: Spatial memory for context reasoning in object detection. In: ICCV. (2017)

[26] Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1017–1025

[27] Zhuang, B., Wu, Q., Shen, C., Reid, I., van den Hengel, A.: Hcvrd: a benchmark for large-scale human-centered visual relationship detection. In: AAAI Conference on Artificial Intelligence (AAAI). (2018)

[28] Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8359–8367

[29] Gupta, T., Schwing, A., Hoiem, D.: No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 9677–9685

[30] Wang, T., Yang, T., Danelljan, M., Khan, F.S., Zhang, X., Sun, J.: Learning human-object interaction detection using interaction points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 4116–4125

[31] Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. arXiv preprint arXiv:1808.10437 (2018)

[32] Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE (2010) 17–24

[33] Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: 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) 482–490

[34] Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J.: Deep contextual attention for human-object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5694–5702

[35] Kim, D.J., Sun, X., Choi, J., Lin, S., Kweon, I.S.: Detecting human-object interactions with action co-occurrence priors. In: European Conference on Computer Vision, Springer (2020) 718–736

[36] Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al.: End-to-end human object interaction detection with hoi transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 11825–11834

[37] Dong, Q., Tu, Z., Liao, H., Zhang, Y., Mahadevan, V., Soatto, S.: Visual relationship detection using part-and-sum transformers with composite queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 3550–3559

[38] Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage hoi detection. Advances in Neural Information Processing Systems **34** (2021)

[39] Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating hoi detection as adaptive set prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 9004–9013

[40] Gao, C., Xu, J., Zou, Y., Huang, J.B.: Drg: Dual relation graph for human-object interaction detection. In: European Conference on Computer Vision (ECCV). (2020)

[41] Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 401–417

[42] Zhang, F.Z., Campbell, D., Gould, S.: Spatially conditioned graphs for detecting human-object interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 13319–13327

[43] Wang, H., Zheng, W.s., Yingbiao, L.: Contextual heterogeneous graph network for human-object interaction detection. In: European Conference on Computer Vision, Springer (2020) 248–264

[44] Ulutan, O., Iftekhar, A., Manjunath, B.S.: Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 13617–13626

[45] Li, Y.L., Liu, X., Lu, H., Wang, S., Liu, J., Li, J., Lu, C.: Detailed 2d-3d joint representation for human-object interaction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020)

[46] Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)

[47] Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. In: AAAI Conference on Artificial Intelligence (AAAI). (2020)

[48] Kato, K., Li, Y., Gupta, A.: Compositional learning for human object interaction. In: European Conference on Computer Vision (ECCV). (2018)

[49] Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 1981–1990

[50] Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)

[51] Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 9469–9478

[52] Xu, B., Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Interact as you intend: Intention-driven human-object interaction detection. IEEE Transactions on Multimedia (2019)

[53] Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. In: IEEE International Conference on Computer Vision (ICCV). (2017)

[54] Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: Proceedings of the European conference on computer vision (ECCV). (2018) 51–67

[55] Hou, Z., Peng, X., Qiao, Y., Tao, D.: Visual compositional learning for human-object interaction detection. In: European Conference on Computer Vision (ECCV). (2020)

[56] Liang, K., Guo, Y., Chang, H., Chen, X.: Visual relationship detection with deep structural ranking. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)

[57] Wang, S., Yap, K.H., Yuan, J., Tan, Y.P.: Discovering human interactions with novel objects via zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 11652–11661

[58] Kim, D.J., Sun, X., Choi, J., Lin, S., Kweon, I.S.: Acp++: Action co-occurrence priors for human-object interaction detection. IEEE Transactions on Image Processing (TIP) **30** (2021) 9150–9163

[59] Salton, G., McGill, M.J.: Introduction to modern information retrieval. mcgraw-hill (1983)

[60] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)

[61] Buchta, C., Kober, M., Feinerer, I., Hornik, K.: Spherical k-means clustering. Journal of Statistical Software (2012)

[62] Goldberger, J., Hinton, G.E., Roweis, S.T., Salakhutdinov, R.R.: Neighbourhood components analysis. In: NIPS. (2005)

[63] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)

[64] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. (2016)

[65] Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 International Conference on Computer Vision, IEEE (2011) 991–998

[66] Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: Upsnet: A unified panoptic segmentation network. In: CVPR. (2019)

[67] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)

[68] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. (2017)

[69] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV. (2017)

[70] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915 (2016)

[71] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee (2009) 248–255

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See our experimental results presented in Sec. 4
   (b) Did you describe the limitations of your work? [Yes] Please see the discussion in Sec. 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see the discussion in Sec. 4.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code and instructions needed to reproduce expereiments will be made public when this paper is made published.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see the Dataset section Sec. 4 and the Architecture and Training section in Sec. A.3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see the Training section in Sec. A.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Please see the Dataset section of Sec. 4.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

We develop a contextual feature learning framework that tackles zero-shot recognition of human-object interactions. Our proposed regularizations enforce encodings of spatial semantic context in the latent features. Visual context emerges in a completely data-driven fashion. Our framework achieves competitive performance against SOTA supervised baselines on HICO dataset. Our proposed losses also deliver performance gain for semantic segmentation tasks on Pascal VOC and Cityscapes. In this supplementary, we include more details as followings:

- We present ablation study on our proposed regularizations in A.1.
- We detail our model architecutres in A.2.
- We detail our training procedure and choice of hyper-parameters in A.3.

$L(C_O)$	L_M	$L(C_G)$	$L(C_L)$	AP
✓	-	-	-	70.6
✓	✓	-	-	71.1
✓	✓	✓	-	71.9
✓	✓	✓	✓	72.6

λ_M	AP
0.0	71.7
0.5	72.3
1.0	72.6
2.0	71.1

λ_G	AP
0.0	71.6
0.5	72.6
1.0	71.6

λ_L	AP
0.0	70.9
0.66	72.6
1.0	71.6

Table 2: Each of our proposed loss regularizations helps feature mappings to capture visual contexts better. We report the performance evaluated by interacting object categories on HICO. **From left to right:** the performance gain resulted from the addition of each loss, and the effects of loss weightings.

A.1 Ablation study.

We summarize the efficacy of each proposed regularization in Table 2. We report the AP performance based on interacting object category. By successively adding loss terms L_M, $L(C_G)$ and $L(C_L)$, we improve the performance by 0.5%, 0.8% and 0.7% AP, compared to the models training with only instance discrimination loss. We also study the weightings for each loss, and adopt the best set of hyper-parameters for training.

A.2 Architecture.

For HOI recognition on HICO, we follow UPSNet [66] to build our model architecture. It consists of a ResNet50 [67] backbone, followed by a FPN [68] layer to generated multi-scale features. The channel dimension of output features are 256. We fuse the multi-scale features using a deformable convolutional [69] layer, resulted in 128-dim unit-length output features. For semantic segmentation on Cityscapes and VOC, we adopt deeplab-v2 [70] model architecture, where ResNet101 is used as the backbone CNN. The output feature dimension is set to 64.

A.3 Trainig.

For all experiments, we fine-tune ResNet50 backbone, which is pre-trained on ImageNet [71] dataset. We use 2 Nvidia V100 cards for training. We set initial $\mathrm{learning\_rate}$ to 0.003, $\mathrm{momentum}$ to 0.9, and $\mathrm{weight\_decay}$ to 0.0001. Following [70], we adopt poly learning rate policy by multiplying base learning rate by $1 - (\frac{iter}{max\_iter})^{0.9}$.

On MSCOCO, we set $\mathrm{crop\_size}$ to 640×640, $\mathrm{batch\_size}$ to 12, training $\mathrm{iterations}$ to $60,000$. We iterate spherical K-Means algorithm for 10 steps to partition an image into 49 segments, which are furthered refined by instance and semantic pixel labels (see [20]). For contrastive losses, we set κ to 12, 16 and 16 for $L(C_O)$, $L(C_L)$ and $L(C_G)$. λ_O, λ_L, λ_G, λ_M are set to 1.0, 0.66, 0.5, and 1.0.

For training on Cityscapes and VOC, we set $\mathrm{crop\_size}$ to 512×512, $\mathrm{batch\_size}$ to 12, training $\mathrm{iterations}$ to $30,000$. We iterate spherical K-Means algorithm for 10 steps to partition an image into 36 segments. Such image oversegmentation is likewise refined by instance and semantic pixel labels. We adopt the same settings for the learning losses.