MISC210K: A Large-Scale Dataset for Multi-Instance Semantic Correspondence

Yixuan Sun[1,*], Yiwen Huang[2,*], Haijing Guo[2], Yuzhou Zhao[2], Runmin Wu[3],
Yizhou Yu[3], Weifeng Ge[2,†], Wenqiang Zhang[1,2,†]
[1]Academy of Engineering & Technology, Fudan University, Shanghai, China
[2]School of Computer Science, Fudan University, Shanghai, China
[3]The University of Hong Kong, Hong Kong, China
{wfge, wqzhang}@fudan.edu.cn

Abstract

Semantic correspondence have built up a new way for object recognition. However current single-object matching schema can be hard for discovering commonalities for a category and far from the real-world recognition tasks. To fill this gap, we design the multi-instance semantic correspondence task which aims at constructing the correspondence between multiple objects in an image pair. To support this task, we build a multi-instance semantic correspondence (MISC) dataset from COCO Detection 2017 task called MISC210K. We construct our dataset as three steps: (1) category selection and data cleaning; (2) keypoint design based on 3D models and object description rules; (3) human-machine collaborative annotation. Following these steps, we select 34 classes of objects with 4,812 challenging images annotated via a well designed semi-automatic workflow, and finally acquire 218,179 image pairs with instance masks and instance-level keypoint pairs annotated. We design a dual-path collaborative learning pipeline to train instance-level co-segmentation task and fine-grained level correspondence task together. Benchmark evaluation and further ablation results with detailed analysis are provided with three future directions proposed. Our project is available on https://github.com/YXSUNMADMAX/MISC210K.

1. Introduction

Building dense visual correspondences is a sub-task of image matching, which aims at finding semantic associations of salient parts and feature points of objects or scenes [4,6,34,49]. This task has established a new way for understanding commonalities among objects in a more fine-grained manner and has been widely used for various computer vision tasks, including few shot learning [11,21,48], multi-object tracking [24], and image editing [10, 15, 32]. To learn general semantic correspondence, several popular datasets, such as Caltech-101 [14], FG3DCar [37], PF-WILLOW [8], PF-PASCAL [8], and SPair-71k [27], have been proposed by researchers to train machine learning models. These datasets were designed to capture large intra-class variations in color, scale, orientation, illumination and non-rigid deformation. However, although these datasets provide rich annotations, they are still far from real-world applications because each object category is only allowed to have at most one instance in each image. Moreover, for most object recognition tasks and applications, multiple objects of the same category often appear at the same time. Existing datasets only focus on one-to-one matching without considering multi-instance scenes, and thus cannot be used as simulations of real-world applications.

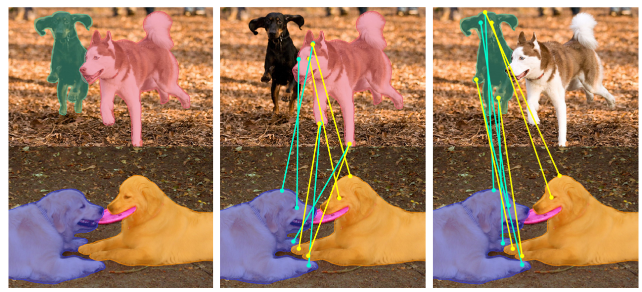


Figure 1. An instructive overview for our multi-instance semantic correspondence task. In this task, models are required to determine the corresponding relationships among multiple instances, where instance masks are introduced for grouping matching key-points.

In this paper, we aim to reduce the gap between one-to-one matching and many-to-many matching by building a new multi-instance semantic correspondence dataset. Following PF-PASCAL [8] and SPair-71k [27], we label keypoints on objects to construct the dataset. There are several key challenges during data labeling. First, how to choose the collection of raw images that contain multiple objects in natural scenes? Second, how to choose candi-

*: Contribution Equally
†: Corresponding Authors

date object categories that have rich keypoints with discriminative semantics? Third, how to ensure label quality while maintaining high annotation efficiency? Last, how to design an evaluation protocol for multi-instance semantic correspondence? While addressing the above issues, we construct a new multi-instance semantic correspondence dataset, called MISC210K, which collects 34 different object classes from COCO 2017 detection challenge [20] and contains 218,179 image pairs with large variations in viewpoint, scale, occlusion, and truncation. Compared with popular PF-PASCAL [8] and SPair-71k [27], MISC210K has much more annotated instances, covers a boarder range of object categories and presents a large number of many-to-many matching cases. Besides, we also design a new protocol to evaluate many-to-many semantic matching algorithms. All these characteristics make MISC210K appealing to the relevant research community.

We summarize main characteristics of MISC210K as follows. First, MISC210K provides annotations for many-to-many matching. Unlike previous datasets [8, 27] which only exploit one-to-one matching, we find out all semantic correspondences among multiple objects (up to 4) across image pairs as shown in Figure 1. Second, MISC210K has more complicated annotations. The number of keypoints in SPair-71k varies from 3 to 30 across categories. In contrast, we well design more keypoints to highlight object contours, skeletal joints, and other distinctive feature points that can characterize objects in detail. Third, MISC210K has a larger scale in comparison to existing datasets. It contains 218,179 image pairs across 34 object categories, which is three times larger than the previous largest dataset, SPair-71k. Fourth, intra-class variations in MISC210K are more challenging. In addition to variations considered in Spair-71k [26], we also introduce more challenging variations, such as mutual occlusion of multiple objects and perspective distortions in complex scenes.

To investigate whether MISC210K can help learn general correspondences across multiple object instances, we evaluate previous state-of-the-art methods, MMNet [49], CATs [4] on MSIC210K. We also propose a dual-path multi-task learning pipeline to solve the complicated multi-instance semantic correspondence problem. For both tasks of correspondence and instance co-segmentation, we designed multi-instance PCK (mPCK) and mIOU (instance) from works [8,25,49]. According to the results, we identify new challenges in this task: (1) extracting discriminative features plays a precursory role to find out commonalities across multiple objects; (2) the uncertainty in the number of matching keypoints makes the matching process more difficult; (3) multiple object instances bring occlusion, interlacing, and other challenging issues. These observations indicate that multi-instance semantic correspondence is a challenging problem deserving further investigation.

This paper is organized as follows. We first describe the MISC210K dataset, its collection process, and statistics. Then we introduce a generic framework for multi-instance semantic correspondence, which enables neural networks to associate salient feature points of object instances across different images. The proposed dual-path collaborative learning (DPCL) pipeline outperforms the transfer of previous one-to-one semantic correspondence algorithms. We further analyze the characteristics of MISC210K and discuss key issues in multi-instance semantic correspondence.

2. Related Work

2.1. Semantic Correspondence Dataset

Caltech-101 [14] provides binary mask annotations of objects of interest for 1515 pairs of images to conduct rough matching. PF-WILLOW [8] and PF-PASCAL [8] provide keypoint annotations for semantic points for evaluating semantic correspondence algorithms. But these two datasets only contain 900 and 1,300 image pairs respectively, which are insufficient for training large semantic correspondence models. Later, Min *et al.* [27] proposed a large-scale semantic correspondence dataset, SPair-71k, which contains 70,958 image pairs with diverse intra-class variations. This dataset soon becomes popular in the research community and leads to breakthrough algorithms, including HPF [26], CATs [4], and MMNet [49]. Considering real-world applications, understanding complex scenes with multiple instances has become an important part of object recognition tasks. Work [17] firstly transferred PASCAL 3D+ dataset for semantic correspondence among multiple instances. Nevertheless, its original design for 3D pose estimation task results in lack of non-rigid object classes and skeleton-centric annotations, which is far from the request for real-world multi-instance semantic correspondence task. To fill this gap, we proposed MISC210K dataset targeting this task. As shown in Table 1, MISC210K contains 34 well chosen categories and over 210K well annotated samples for multiple instances in each image pair.

2.2. Semantic Correspondence Models

Methods for semantic correspondence can be roughly categorized into several groups: handcrafted feature based methods [2, 5, 23, 35, 38], learnable feature based methods [18, 19, 28, 39], graph matching and optimization based methods [22,41,46,47], methods focusing on geometry displacement [3, 8, 9, 13, 40], and etc. Hand crafted features, such as SIFT [23], HOG [37] and DAISY [38], design robust feature descriptors with low level statistics. In NC-Net [33], DualRC-Net [19] and GOCor [39], high level semantic features of CNNs are used to build dense correspondences. SCOT [22] and DeepEMD [47] formulate the semantic correspondence as an optimal transport prob-

Dataset (Year)	Samples (classes)	Source Dataset	Annotations	Match Object
Caltech 101 [14] (2006)	1,515 (101)	Original	Instance Mask	Single
CUB [42] (2010)	∼120,000 (200)	Original	Part Locations (15), Binary-Attributes (312), Bounding Box	Single
Animal-parts [30] (2016)	∼7,000 (100)	ImageNet	Key-point (1∼6)	Single
PF-WILLOW [8] (2017)	900 (5)	PASCAL VOC, Caltech-256	Key-point (10)	Single
PF-PASCAL [8] (2017)	1,300 (20)	PASCAL VOC	Key-point (4∼17), Bounding Box	Single
SPair-71k [26] (2019)	70,958 (18)	PASCAL3D+, PASCAL VOC	Key-point (3∼30), Bounding Box Instance Mask	Single
DISCOBOX [17] (2021)	36,292 (12)	PASCAL 3D+	Key-point (1∼12)	Multiple
MISC210K (2022, ours)	218,179 (34)	MS COCO	Key-point (5∼52), Bounding Box, Instance Mask, Text Description	Multiple

Table 1. Comparison of previous semantic correspondence datasets and our proposed MISC210K. The result shows our MISC210K a large-scale dataset with careful annotation on different granularities designed for the multi-instance semantic correspondence task.

lem and give closed-form solutions. PCA-GM [41] focuses on solving a general quadratic assignment programming (QAP) problem. Besides, PHM [3, 8] and SCNet [9] develop the probabilistic Hough matching in a Bayesian probabilistic framework to model the geometry displacement of objects. These methods have achieved impressive performance on single-instance semantic correspondence tasks. However, expanding semantic correspondence task to multiple instance scenes is vital for applications. Although previous works such as work [17] tried to define this task on PASCAL 3D+ [44], large-scale annotated data in high quality is still lacking. In addition, a series of protocols for multi-instance semantic correspondence (such as train/test set division, evaluation metrics, etc.) requires further definition. Hence, supported by our proposed MISC210K, we designed the whole protocol of training and evaluation based on a dual-path collaborative learning pipeline.

3. The MISC210K Dataset

To advance research on semantic correspondence towards more challenging instance-level correspondence, we build the MISC210K dataset based on the procedures shown in Figure 2. The MISC210K provides instance-level masks along with dense keypoint annotation for each instance.

3.1. Task Definition and Design Protocols

Our MISC210K dataset is composed of image pairs. In each image pair, images are named as the source and target images. Each image pair is designated an object category of interest and exists one or more object instances in that category for each image. We define the task of multi-instance semantic correspondence as follows. Suppose the mask and dense keypoint locations of every instance in the category of interest in the source image of an image pair are given.

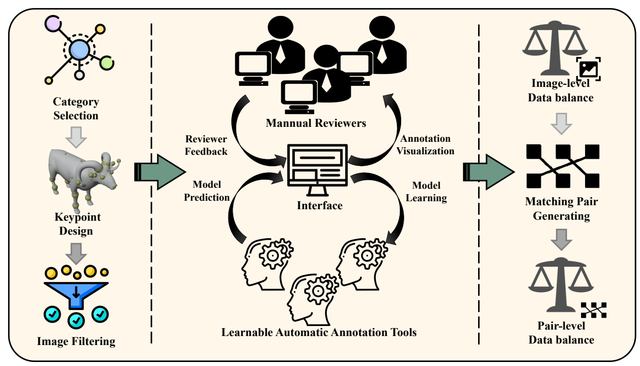


Figure 2. An overview of dataset construction pipeline contains raw data collection, annotation and post processing.

The goal is to obtain, for all the keypoints of each instance in the source image, their corresponding keypoints on every instance in the category of interest in the target image. Note that this definition does not preclude the existence of object instances in categories other than the designated one in the image pairs. This task involves more complicated scenarios than usual and should be tackled more carefully by observing the following design protocols.

Dense keypoint Distribution. In the previous literature, the number of keypoints to match is often smaller than ten for an image pair, and only those most significant ones are chosen. This setting suits the common understanding of 'semantic points', however, a limited number of keypoints hinder the process of truly understanding every part of an object, which is the ultimate goal of semantic correspondence. Hereby, MISC210K aims to provide dense keypoint annotation for every object to make it possible to explore and learn dense semantic correspondences.

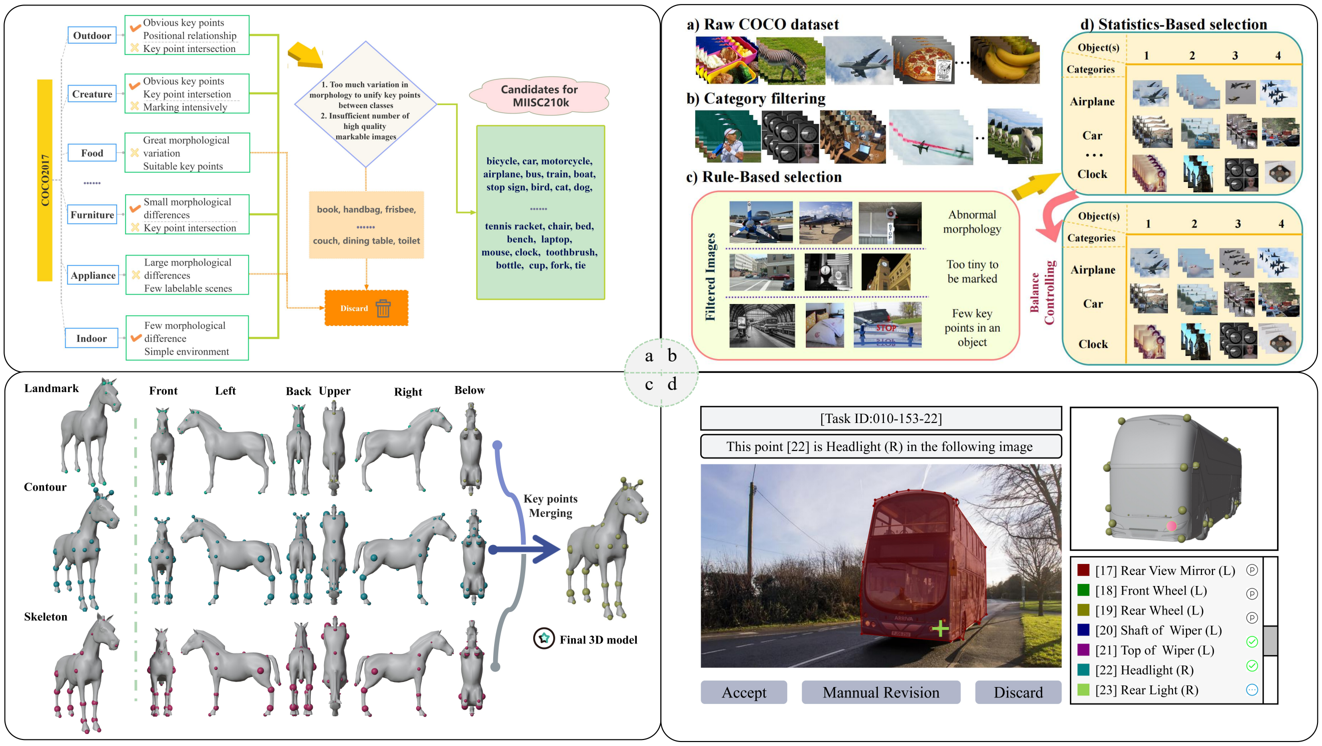


Figure 3. Detailed illustration of MISC210K construction procedure. For each sub-figure, (a) our workflow to select candidate categories from COCO [20]; (b) pipeline for raw image collection; (c) mechanism for object keypoint design; (d) the reviewer platform for annotation.

3D Models as Category Prototypes. Dense keypoint annotation demands a more visually comprehensive and consistent way to define keypoints for all instances of a category. The previous semantic matching literature does not need to tackle this issue because the number of keypoints is small so that their information can be easily communicated without ambiguity. To this end, we employ one 3D model per category as the 3D category prototype. We define a uniform set of keypoints for each category over its 3D prototype to clearly and unambiguously convey their spatial layout and semantics. And this set of keypoints is applied to all instances of the same category.

Multiple Relatively Integral Instances. In everyday life, images of multiple instances in the same category are quite common. Establishing keypoint correspondences among multiple instances simultaneously is key to understanding the layout of image content. However, as a keypoint matching task, each instance in an image must reveal a sufficient number of important parts. Images with too many instances are unsuited for this task as the average space for each instance is severely limited thus making keypoints occluded or indistinguishable. In MISC210K, the number of instances in the same category in each image is at most 4 to reduce the chances of ambiguous keypoints.

3.2. Dataset Construction

Raw Image Collection: Images in MISC210K are collected from the large-scale object detection dataset, Microsoft COCO [20]. COCO gathers images of everyday scenes containing objects in their natural context, thus it includes many images that contain multiple instances in the same category, which suits our instance-level semantic matching task. To ensure that images have the above multi-instance property and are of good quality for instance and dense keypoint annotation, we conduct category filtering and image selection. Certain categories in COCO are removed because they do not have a well-defined 3D prototype (e.g. no shared model for pizza and beef instance for 'food'). After this step, we keep 34 categories as in Figure 3(a). Among images in these categories, we manually remove those with poor-quality or incorrect instance masks. Images with overly small or incomplete instances are also removed because they are unsuited for annotation. We finally choose over 300 images for each category while maintaining a balanced distribution of per-image object counts within each category. The overall distributions of numbers of object classes and instances per image are globally balanced via statistical selection as shown in Figure 3(b).

Matching Pair Annotation: For data annotation, two steps are designed namely category keypoint system definition

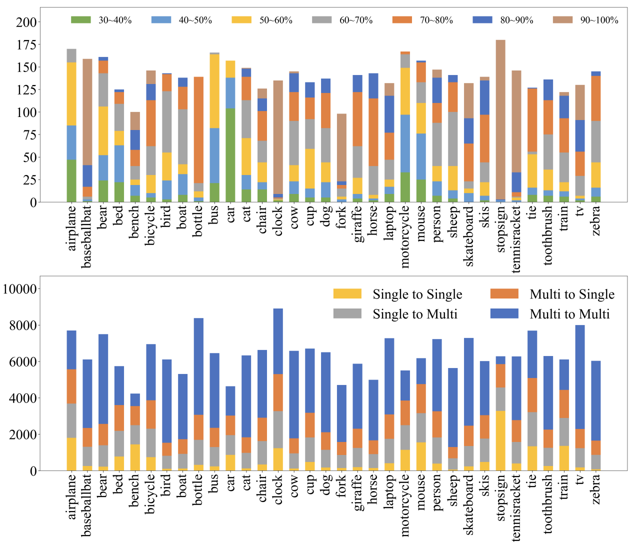


Figure 4. Statistic results for MISC210K. The above graph counts the images based on ratio of annotated points to the entire point set defined for classes. The bottom one shows the distribution of matching instances for an image pair per class.

and human-machine collaborative annotation. Two steps are introduced as follows:

1) Category keypoint system definition: One standard 3D model is chosen as the prototype for each category so that keypoints marked on the model can clearly and visually convey their associated semantics. For keypoint selection, we use three keypoint generation schemes that focus on the skeleton, contour, and appearance, respectively. The skeleton scheme generates candidate keypoints that are skeletal joints on the 3D model (e.g. knee joints in animal-like categories). The contour scheme generates candidate keypoints lying on one of the model contours (e.g. head top point) corresponding to a set of viewpoints. The appearance scheme finds points with unique local appearance and semantics as candidate keypoints (e.g. eyes and nose). To compare the quality of candidate keypoints generated by different schemes, we evaluate them using six distinct perspective views of the 3D model (i.e. upper, below, left, right, front, and back). As shown in Figure 3(c), we follow a voting procedure where three scores between 0 to 1 are used to evaluate the viability of candidate keypoints. The three scores are salience, completeness, and uniformity. Salience evaluates how easily a set of points can be located. Completeness reflects how thoroughly a set of points depict the model shape in a specific view. And uniformity describes how evenly a set of points are spatially distributed within a specific view. Five annotators are asked to grade three sets of candidate keypoints respectively generated by the three schemes within each of the aforementioned six perspective views, and the three resulting scores for each set of candidate keypoints are averaged. The three scores have different weights during averaging, 0.5 for salience, 0.3 for completeness, and 0.2 for uniformity. For each view, the sets of candidate keypoints receiving an average score higher than 0.5 are chosen. keypoints chosen for all six views are merged to form the set of final keypoints for their corresponding category.

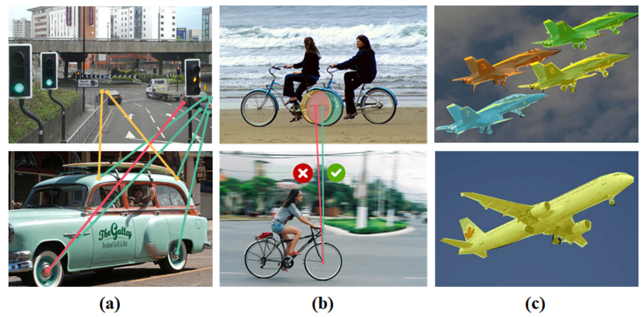
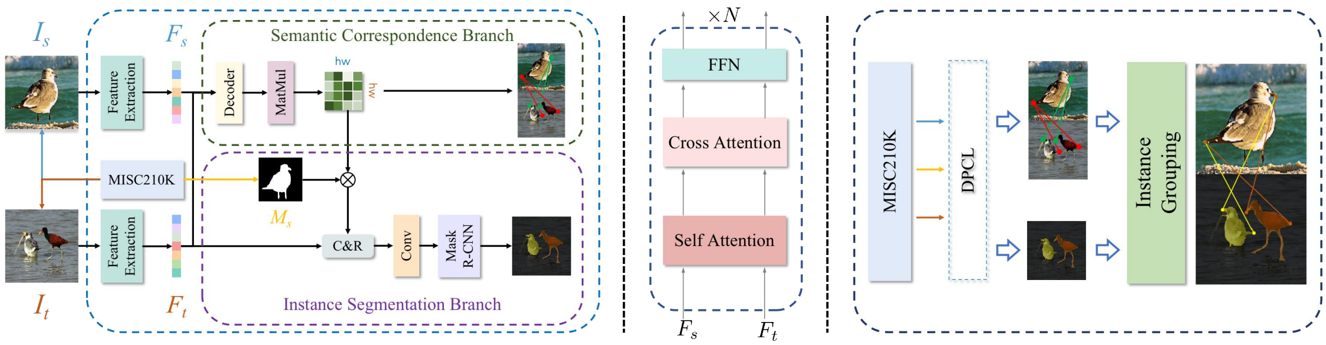


Figure 5. Challenging examples in MISC210K dataset. Images from left to right present (a) severe scale variation, (b) occlusion and (c) shape inconsistency, respectively.

2) Human-machine collaborative annotation: Inspired by the work [43], we introduce an automatic annotation tool and construct a human-machine collaborative semi-automatic annotation pipeline. First, we ask the annotators to manually label 40% randomly chosen images in our dataset and use these annotated images to fine-tune an automatic labeling module (ALM). The trained ALM is utilized to label 30% of the remaining images. With the usage of the platform in Figure 3(d), reviewers are asked to accept, discard or slightly drag each automatically annotated keypoint to the desired location. Such human feedback is then used to retrain the ALM. The retrained model is used to annotate the remaining data followed by human review. In this way, human annotation has switched to reviewing, which greatly reduces human workload.

Dataset Statistics: Our final dataset includes 34 categories and a total of 4,812 images with 1 to 4 instances within each image. In each category, 30,000 initial image pairs are generated. We filter these image pairs to assure that at least 30% of keypoints are shared between the source and target images within each pair. This leads to 218,179 image pairs in our dataset. In addition to keypoint locations, textual descriptions of keypoint semantics are also provided. Furthermore, we associate each object instance with its mask annotation in COCO [20]. Detailed image-level (top) and pair-level (bottom) statistics are shown in the histograms in Figure 4. The constructed MISC210K dataset contains over 210K pairs from 34 object categories and covers most possible real-world object matching scenarios, such as occlusion, overlap, and significant scale differences (examples are given in Figure 5). It greatly expands the application scenarios and supports training for complex deep neural networks. Devising a unified solution for these matching



(a) Dual-Path Collaborative Learning Pipeline **(b) Decoder** **(c) MISC Inference Pipeline**

Figure 6. Illustration of the overall structure. **(a)** demonstrates our collaborative training pipeline, containing multi-instance semantic correspondence branch as well as instance segmentation branch. **(b)** shows the implementation of the decoder. **(c)** illustrates the implementation of inference procedure with our DPCL.

scenarios represents a great challenge. User agreement is required before accessing to our MISC210K dataset.

4. Benchmark Performance

In this section, we propose a dual-path collaborative learning (DPCL) pipeline which performs co-training for instance co-segmentation and multi-instance semantic correspondence to achieve instance-level key-point matching. Our DPCL pipeline, evaluation protocol and performance are given below. We also conduct ablation experiments to illustrate some characteristics of our dataset.

4.1. Dual-Path Collaborative Learning

Pipeline overview. The pipeline is composed of a backbone for feature extraction, a transformer decoder, two prediction branches for multi-instance semantic correspondence and instance co-segmentation respectively, as shown in Figure 6. Specifically, we use iBOT [51] as the feature extraction backbone and 6 cascaded transformer blocks identical to those in [12] as the decoder. Each transformer block contains one self-attention layer, one cross-attention layer and one FFN block. Given an image pair $(\mathbf{I}_s, \mathbf{I}_t)$, the source $(\mathbf{I}_s)$ and target $(\mathbf{I}_t)$ images are fed into the backbone to extract feature maps $\mathbf{F}_s, \mathbf{F}_t$. These feature maps are passed through their respective decoders and then used to calculate a matrix multiplication between them to produce a 4D cost volume, which contains the similarity scores for all possible pixel pairs across the two decoded images. The cost volume is directly used for key-point matching and also concatenated with $\mathbf{F}_t$ for the segmentation head in Mask R-CNN.

Design of two branches. For the task of multi-instance semantic correspondence, we follow previous work on crowd localization [1, 36] and bottom-up pose estimation [16], and cascade a Sigmoid function, non-maximum suppression [29] and static thresholding to obtain final predictions. However, precise key-point matching results alone cannot solve multi-instance semantic correspondence entirely because of the inability to confirm whether multiple matching key-points in the target image actually belong to the same instance. Therefore we use an instance segmentation branch to predict instance masks used for grouping same-instance matching key-points in the target image.

4.2. Experiment Setup

Evaluation protocol. Although DISCOBOX [17] has inherented mAP metric from multi-instance pose estimation for this task, the definition of positive and negative examples in units of instances cannot well reflect the matching effect of key points. Hence, we extend the previous PCK into $mPCK@\alpha$ (multi-instance PCK as eq 1) for this task.

$$mPCK(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_\alpha^i}{TP_\alpha^i + FP_\alpha^i}, \tag{1}$$

in this metric, we consider the difference in the number of predicted key points and ground truth and matching effect between a pair of keypoints. For given N groups of predicted and ground-truth keypoints, we firstly use Kuhn-Munkres Algorithm $\texttt{KMA}(P_{gt}, P_{pred}, \alpha)$, to find out matches for ground-truth P_{gt} and predicted points P_{pred}. When predicted results fall into the circle of radius $\alpha \times d$ (d is the longer side of an image, thus α stands for precision), we consider (P_{gt}, P_{pred}) as a correct match. We defined the amount of unmatched P_{gt} as FN_α, unmatched P_{pred} as FP_α and number of matched pairs as TP_α. We use averaged IOU metric for each instance (mIOU in short) to evaluate performance on multi-instance co-segmentation task.

4.3. Implementation Details

We conduct all experiments on PyTorch-GPU [31] using NVidia RTX 3090 GPUs. All input images are resized to 256×256 and the resolution of cost volumes are scaled to 32×32 with cost volume upscale algorithm in [49]. During

Method	α	airplane	base-ballbat	bear	bed	bench	bicycle	bird	boat	bottle	bus	car	cat	chair	clock	cow	cup	dog	all
MMNet [49]	0.05	5.26	5.57	6.38	2.12	5.70	6.72	7.47	6.56	6.30	3.06	1.26	6.86	4.09	11.53	5.68	5.04	6.61	5.68
	0.10	20.38	21.81	25.95	9.86	18.03	26.80	27.98	23.64	20.50	11.85	5.96	24.03	15.10	39.78	23.92	18.79	26.70	21.58
	0.15	37.38	40.60	48.14	20.79	32.30	47.89	51.37	41.46	34.26	25.70	13.26	43.39	28.74	61.37	45.92	35.50	49.48	39.66
	1.00	99.76	99.94	99.98	99.39	98.63	99.71	99.92	100.00	99.99	99.73	99.89	99.57	99.91	99.96	99.85	99.47	99.69	99.76
CATs [4]	0.05	10.95	4.68	10.50	4.64	3.95	9.18	10.76	8.58	5.43	11.12	8.27	10.41	4.53	15.88	12.27	4.86	8.62	10.00
	0.10	25.55	14.47	27.09	13.02	12.40	23.68	24.11	20.86	14.55	27.13	19.87	24.44	13.28	33.29	27.83	15.07	22.79	23.88
	0.15	38.27	25.04	40.56	22.18	21.16	36.94	36.63	31.64	23.27	39.69	27.50	36.46	23.88	46.00	39.05	23.50	34.93	35.45
	1.00	88.02	91.85	88.41	84.60	83.53	88.90	89.56	88.96	91.70	90.74	83.21	89.17	87.97	91.41	90.80	87.53	88.60	89.15
DPCL	0.05	9.81	9.23	17.43	2.06	6.03	14.74	22.01	11.98	12.24	5.08	6.63	15.64	4.30	17.91	16.57	8.66	13.14	11.32
	0.10	22.96	23.50	36.83	7.37	16.24	32.90	43.04	27.93	23.47	14.37	15.93	38.96	13.62	37.37	35.74	20.24	31.45	25.21
	0.15	35.97	35.17	51.52	16.60	23.94	47.66	54.38	40.76	33.46	26.75	25.34	52.18	24.19	48.47	49.03	34.22	44.78	37.01
	1.00	93.90	94.04	97.29	95.25	94.67	92.79	96.77	96.29	96.23	96.32	92.06	95.37	94.17	97.27	94.65	95.46	93.60	95.07
	mIoU	21.39	1.74	44.36	27.81	32.69	24.92	24.59	21.37	4.42	52.27	16.17	33.03	3.94	15.57	37.50	0.93	30.17	22.80

Method	α	fork	giraffe	horse	laptop	motor-cycle	mouse	person	sheep	skate-board	skis	stop-sign	tennis-racket	tie	tooth-brush	train	tv	zebra	all
MMNet [49]	0.05	4.19	3.76	5.39	5.20	5.22	4.26	3.80	7.78	5.96	5.54	7.21	7.21	5.68	4.85	7.59	4.64	4.35	5.68
	0.10	13.63	16.24	20.18	19.33	21.38	16.91	16.11	30.81	22.68	23.15	25.53	28.96	24.08	18.50	26.54	17.16	17.11	21.58
	0.15	24.93	32.49	38.71	34.95	42.19	33.02	32.80	55.63	41.13	44.39	44.90	51.55	47.83	35.18	46.41	31.58	34.30	39.66
	1.00	99.71	99.82	99.97	99.87	98.34	99.78	99.99	99.73	99.81	99.89	99.44	99.91	100.00	99.86	99.77	99.96	99.72	99.76
CATs [4]	0.05	5.34	15.85	14.82	5.34	11.43	8.19	13.22	17.82	7.22	11.47	18.19	8.56	15.73	4.94	11.66	10.42	14.93	10.00
	0.10	13.81	34.07	35.04	15.94	29.47	17.91	28.90	38.07	19.67	29.50	34.89	20.57	34.57	12.58	28.52	24.37	34.35	23.88
	0.15	21.97	47.42	49.38	27.41	44.62	26.78	41.59	51.80	31.79	42.09	46.48	32.19	47.00	19.92	41.49	36.20	48.57	35.45
	1.00	88.63	92.24	91.64	87.54	90.34	82.90	89.58	92.71	87.76	89.40	86.95	92.45	89.30	87.22	88.21	91.96	91.76	89.15
DPCL	0.05	11.66	14.77	11.68	4.18	11.41	8.74	8.99	21.38	10.24	6.15	6.93	15.13	17.30	16.26	10.10	2.48	13.46	11.32
	0.10	21.93	31.38	25.94	13.92	29.86	15.99	23.73	42.47	25.12	16.14	15.39	34.37	27.93	27.82	28.21	8.19	29.22	25.21
	0.15	28.92	41.82	40.18	22.90	45.19	27.00	38.62	54.26	37.44	24.24	26.69	48.97	38.57	35.24	41.54	20.10	43.36	37.01
	1.00	96.29	94.66	95.18	94.40	91.28	93.06	95.77	96.21	94.82	94.87	93.02	96.97	94.53	96.15	93.58	95.61	96.62	95.07
	mIoU	0.14	29.03	38.05	38.07	41.13	4.92	11.11	33.48	6.81	0.00	41.77	6.06	2.25	10.60	49.22	29.13	40.48	22.80

Table 2. Evaluation for MMNet [49], CATs [4] and our proposed dual-path collaborative learning pipeline (DPCL). We provide mPCK result on 34 classes in MISC210K with different α metrics. We also provide instance-level evaluation (mIOU) for DPCL.

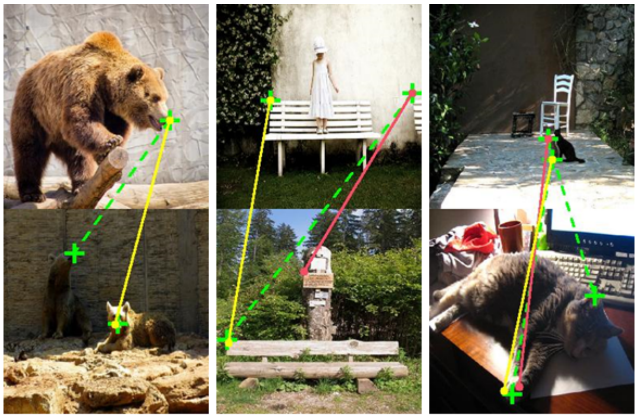
training, after getting cost volume from different baseline methods, we follow work [49] to extract predicted similarity heatmap, generate ground-truth heatmap and optimize model's parameters with binary cross entropy loss. Also, an SGD optimizer is used, where learning rate is set to 0.0005 with momentum of 0.9. For both training and evaluation, the batch-size is set to 16 for all experiments. To generate keypoints from predicted similarity heatmap, we set NMS radius as 5 and threshold as 0.7 to select valid predictions.

4.4. Baseline Evaluation

We provide a comprehensive evaluation of our pipeline on both tasks. For multi-instance semantic correspondence task, we transfer our multi-instance matching head to MMNet [49] and CATs [4] and compare results with them.

We provide evaluation results for 34 categories with three mostly used α for one-to-one semantic correspondence (0.05, 0.1, 0.15). In which $mPCK@\alpha = 0.05$ is the most strict matching criterion. According to Table 2, we find that our method outperforms the baseline in most α granularity which shows that instance-level matching can help build up fine-grained correspondence. Comparing performance across object classes, we found that vehicles and furniture perform worse than animals. We attribute this to the indoor scenes often appear occlusion and interference from similar objects of different categories while multi-instance scenes in vehicles are usually overlapping with others and have severe perspective changes. Besides, we also found that although α is set to 1.0, the $mPCK$ for DPCL can not achieve 100%, which indicates that the methods for multi-instance matching should not only determine the location of matching keypoint, but also estimate the number of matching pairs. Therefore, our multi-instance matching becomes a more challenging task than previous single-instance matching. We also provide instance-level evaluation for co-segmentation on Table 2.

Visualization evaluation: We also provide visualization results for our proposed pipeline and two baselines in Figure 7, where three main challenges are revealed. 1) In multi-instance semantic matching, the model will generally encounter the situation of redundant or missing prediction for key points. We believe this is due to the fact that the current feature representation and alignment scheme cannot guarantee consistency for the same semantic regions. 2) Even if the matching target is kept unchanged, the apparent difference of multiple source instances will obviously affect the matching result. We attribute this to the fact that the current model features can not well achieve fine-grained information consistency, and the extraction of common information between different instances with the same semantics in an image will be a new direction. 3) When the model deals with scale variation between instances, its effect is



(a) (b) (c)

Figure 7. Three main reasons lead to failure cases in MISC210K: (a) Missing Key-point; (b) Inconsistent Prediction; (c) Perspective Distortion. Yellow/Red solid line means correct/incorrect pairs predicted by our DPCL. Green dotted line indicates ground-truth correspondence not be predicted by our DPCL.

significantly reduced. This can be explained as the excessive downsampling for extracting deep semantics limits the resolution of the cost volume. As a result, achieving the matching at the same resolution as the original image will effectively improve the model performance.

Ablation study: Here we hold on two experiments to prove the effectiveness of our dual path learning mechanism and the design of MISC210K dataset. For dual path design, we do ablation on both matching header and co-segmentation header. For both tasks when removing the supervision of the other task, performance occurs an obvious decline (8.8% for multi-instance semantic matching and 2.6% for instance co-segmentation). We believe that the fine-grained commonality enables the model to focus on the matching task and the collaborative segmentation enables the model to focus on the global consistency of the objects that are commonly needed in both tasks. However, whether there is a better task for co-training with multi-instance semantic matching and the characteristics of these tasks are waiting to be discussed. Furthermore, we tried to investigate whether our dataset can help models to handle the semantic correspondence task better. Here we first pretrained CATs and MMNet on MISC210K and then fine-tuned them on SPair-71K. The performance has improved by 1.3% and 2.2%. We attribute this to the designed more difficult samples and more complex tasks in our dataset.

5. Future Direction

Unseen Key-point Discovery. With the usage of fine-grained keypoint annotation in MISC210K, we can propose a series of subtasks so as to investigate the generalizability of models. For example, we propose a new task named key-point discovery, where annotations about certain keypoints are removed in training set, but during testing, the model is required to discover all keypoint including hidden ones.

Matching Closed-loop Constraint. In the multi-instance semantic correspondence task, more than one key-points representing the same semantic information in both source image and target image are available. Therefore, multi-instance semantic correspondence models can be trained with a closed-loop constraint among separated instances from an image pair. Specifically, we introduce the instances form our MISC210K as $\mathbf{I}_s$, $\mathbf{I}_{s+1}$, ..., etc. Given a certain point $\mathbf{P}_s$ in $\mathbf{I}_s$, corresponding point $\mathbf{P}'_s$ in the same instance is calculated from a matching chain $\mathbf{I}_s \rightarrow \mathbf{I}_{s+1} \rightarrow \ldots \rightarrow \mathbf{I}_s$. As a consequence, the offset between $\mathbf{P}_s$ and $\mathbf{P}'_s$ can be an effective constraint to evaluate performance of methods.

Correspondence based Recognition Tasks. Our dataset can also be used for few-shot segmentation [11, 25], object detection [7], medical image processing [50] and pose estimation [45] with the multi-grained annotation. Our dataset also provides a validation platform for the methods aiming at improving the performance of a specific task with joint use of multiple granularity labels. Besides, MISC210K can also be used for applied research such as multi-object tracking, where template matching is a major component [24].

6. Conclusion

In this paper, we proposed a new multi-instance semantic matching task with a large-scale dataset (MISC210K). Compared with existing semantic matching datasets, our MISC210K has many distinctive characteristics: 1) The first defined multi-instance semantic correspondence task; 2) Evidence-based fine-grained keypoint design; 3) Human-machine collaborative annotation with closed-loop constraint and quality control; 4) Object category diversity for the robustness of semantic matching methods. For handling the problem of building up instance-to-instance correspondence as well as co-segmentation masks, we proposed the dual-path collaborative learning pipeline which proved that this schema for learning two tasks synchronously is beneficial to both sides of correspondence and segmentation. These results present some important challenges and uncover critical messages for advancing the area of semantic matching and multi-object recognition in the future.

7. Acknowledgment

This work was supported by National Natural Science Foundation of China (No.62072112, No.62106051), National Key R&D Program of China (No.2020AAA 0108301), Scientific and Technological Innovation Action Plan of Shanghai Science and Technology Committee (No.22511101502, 22511102202), the Shanghai Pujiang Program Nos.21PJ1400600.

References

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 872–881, 2021. 6

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 2

[3] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 3

[4] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021. 1, 2, 7

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 2

[6] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[7] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020. 8

[8] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1711–1725, 2017. 1, 2, 3

[9] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1831–1840, 2017. 2, 3

[10] Nasir Hayat, Hazem Lashen, and Farah E. Shamout. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath, editors, *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 461–477. PMLR, 06–07 Aug 2021. 1

[11] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 1, 8

[12] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 6

[13] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2016. 2

[14] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2307–2314, 2013. 1, 2, 3

[15] Kazuma Kobayashi, Ryuichiro Hataya, Yusuke Kurose, Mototaka Miyake, Masamichi Takahashi, Akiko Nakagawa, Tatsuya Harada, and Ryuji Hamamoto. Decomposing normal and abnormal features of medical images for content-based image retrieval of glioma imaging. *Medical Image Analysis*, 74:102227, 2021. 1

[16] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018. 6

[17] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021. 2, 3, 6

[18] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10196–10205, 2020. 2

[19] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33:17346–17357, 2020. 2

[20] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4, 5

[21] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crcnet: Few-shot segmentation with cross-reference and region–global conditional networks. *International Journal of Computer Vision*, pages 1–18, 2022. 1

[22] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020. 2

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[24] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022. 1, 8

[25] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S Ecker. One-shot instance segmentation. *arXiv preprint arXiv:1811.11507*, 2018. 2, 8

[26] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3395–3404, 2019. 2, 3

[27] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 1, 2

[28] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020. 2

[29] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006. 6

[30] David Novotny, Diane Larlus, and Andrea Vedaldi. I have seen enough: Transferring parts across categories. 2016. 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[32] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13481, 2022. 1

[33] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Proc. NIPS*, 2018. 2

[34] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Ncnet: Neighbourhood consensus networks for estimating image correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[35] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[36] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R Venkatesh Babu. Locate, size, and count: accurately resolving people in dense crowds via detection. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2739–2751, 2020. 6

[37] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4246–4255, 2016. 1, 2

[38] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009. 2

[39] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. GOCor: Bringing globally optimized correspondence volumes into your neural network. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020. 2

[40] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 2

[41] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3056–3065, 2019. 2, 3

[42] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 3

[43] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 2022. 5

[44] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 3

[45] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 398–416. Springer, 2022. 8

[46] Xu Yang, Zhi-Yong Liu, and Hong Qiao. A continuation method for graph matching based feature correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1809–1822, 2020. 2

[47] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[48] Shan Zhang, Dawei Luo, Lei Wang, and Piotr Koniusz. Few-shot object detection by second-order pooling. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[49] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3354–3364, 2021. 1, 2, 6, 7

[50] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022. 8

[51] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022. 6