Supplementary Materials for 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

1. Further Overview for MISC210K

In this paper, we build a large scale dataset for multi-instance semantic correspondence task called MISC210K based on COCO dataset [9]. Our MISC210K contains 218,179 image pairs composed of 4,812 images from 34 categories. We choose images that contain at most 4 instances for each category to limit the number of instances since too many instances will reduce the quality of data annotation. For each image pair, instance level (instance masks) and fine-grained level (key-point of instances) annotations are provided. In Figure 3∼9, we provide an overview of our dataset. The following part will thoroughly introduce four main stages to construct our dataset namely: candidate category selection, candidate images filtering, design of important keypoints, and annotation workflow.

1.1. Candidate Classes Selection

To select object categories that are suitable for learning instance-level correspondence, we summarized the characteristics of each selected class and show them in Table 1∼4. We finally selected 34 object categories according to the actual difficulty of different categories and evaluated the annotation cost for each class of instances. The 34 classes have commonalities such as good consistency of instances within the category, moderately challenging scenarios, and clear potential key points that can be intuitively described and comprehended.

1.2. Candidate Images Filtering

We have described the main pipeline for collecting candidate images from 34 object categories in COCO [9]. Here we will provide a detailed introduction to our rule-based image filtering mechanism. We first filter out images that contain over 4 object instances from the same category to make the correspondence learning problem become tractable to be solved. Besides, we listed 4 kinds of images that need to be removed: 1) tiny instances; 2) abnormal morphology; 3) partially invisible instances; 4) heavy occlusion. As shown in Figure 1, each problem can bring a great burden to annotation, post-processing and usage for further works.

1.3. Design of Important Keypoints

Here we provide an overview of the keypoint identification system designed for all the 34 object categories. As shown in Figure 2, our keypoint identification system relies on 3D models to find keypoints in six perspectives. We select unique feature points by combining skeletons, contours, and appearances. We also labeled some key points that need to use relative position information for direct description (such as the left side of the neck for bear, cat, and dog). This brings a new challenge as location-association reasoning for multi-instance semantic correspondence methods.

1.4. Annotation Workflow

Following the work [14] we tend to further reduce the workload of manual annotation. As a result, following the work [15], we introduced the human-machine collaborative annotation mechanism. Here we will further introduce this pipeline in five aspects namely: 1) the role of human; 2) data grouping; 3) automatic annotation task selection; 4) quality control; 5) data flow and interaction procedure.

Role of human: In the previous data labeling work [7, 9, 11, 18], crowdsourcing workers are usually introduced for heavy labeling work. Especially for the dataset with more fine-grained tasks such as segmentation [6, 9], detection [5, 9] as well as pose estimation [1], the completely manual labeling method significantly increases the data set construction cycle and labeling cost. As a result, we redesigned the role of human in the labeling process. Instead

∗: Contribution Equally
†: Corresponding Authors

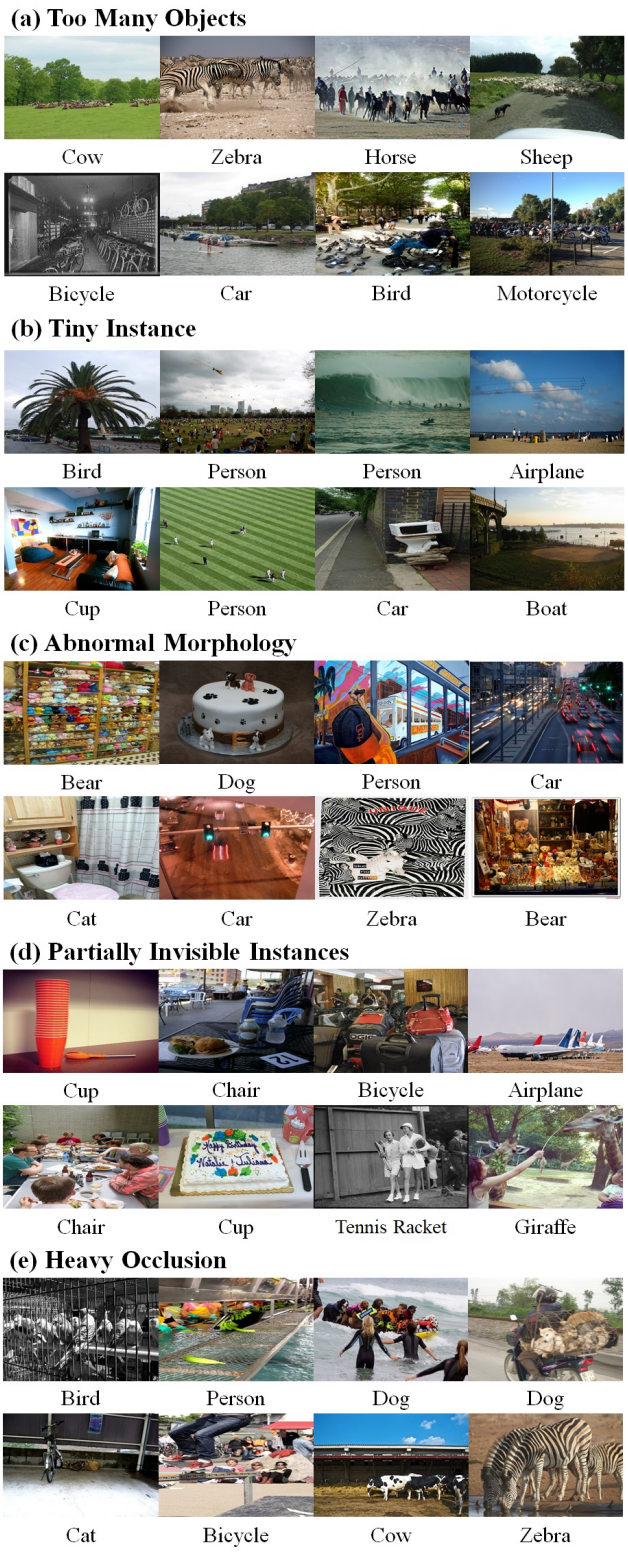

Figure 1. Illustration of images that are ignored during our annotation procedure.

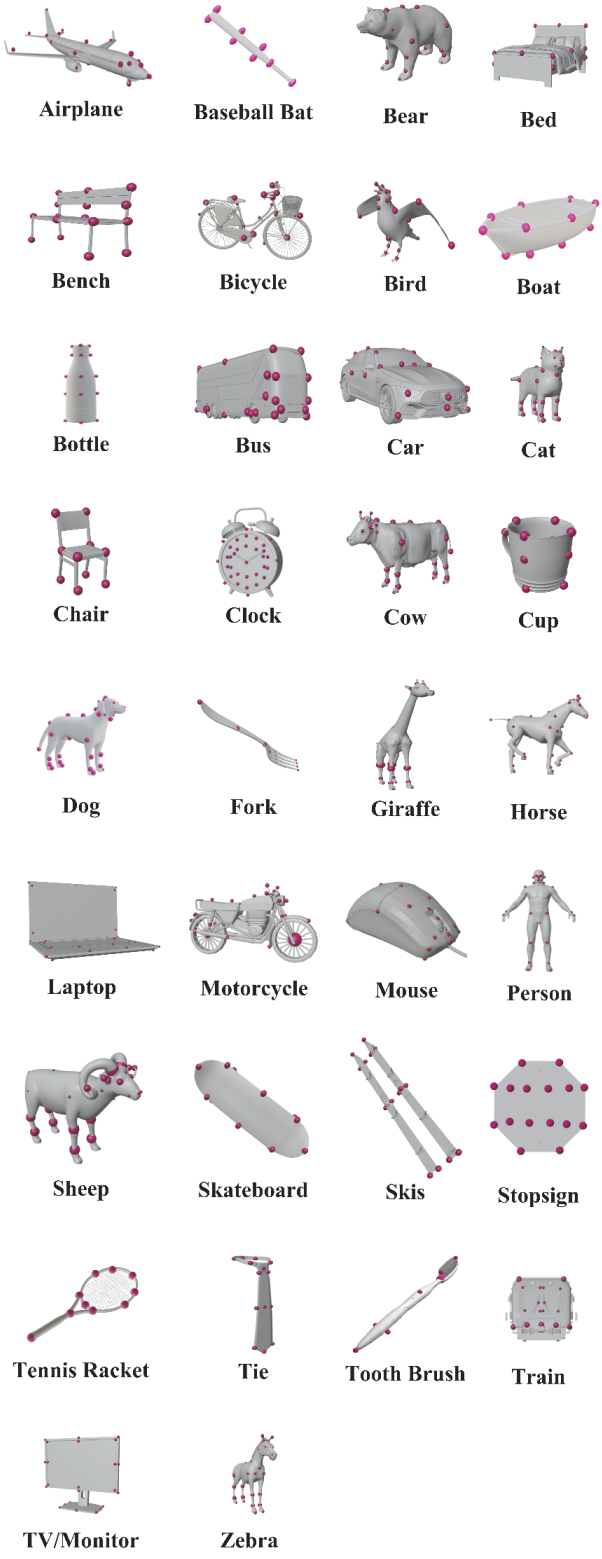

Figure 2. An overview for our designed key-point systems for 34 categories.

of labeling the data directly, labelers are required to review and modify the labeling output of the automated annotation system. This pipeline greatly reduced the workload of crowdsourcing workers, shortened the labeling cycle, and reduced labeling costs.

Data grouping: After data selection, we acquired over 300 images for each category. We manually annotated 40% of the raw images for each category and used them to train the annotation system. While the preserved 60% raw images are divided evenly into task packages for automatic labeling. The manually revision and feedback is also carried out in units of task packages.

Automatic annotation task selection: Since our multi-instance semantic correspondence is a difficult task, current method can not achieve an ideal performance. As a result, we can not design automatic labeling systems directly based on this task. Therefore, we put eyes on an easier task called 2D pose estimation [2, 3]. This task aims at estimating key-points of an image instance by instance. To fit the annotation task we divided an image with multiple instances into several smaller image clipping blocks which contains only one instance according to the instance level mask provided by COCO [9]. After annotation, an automatic data integration script was used. This improves the efficiency of automatic labeling.

Quality control: For a task package, after the automatic annotation, we used the format conversion script to package the system output into JSON files and feed them into the reviewer platform. Our reviewer platform can visualize the key points from JSON directly on the raw images one by one and use the corresponding keypoint description as a prompt for reviewers. The platform can also load 3D keypoint instructions to provide intuitive guidance for reviewers. Three choices are provided to reviewers as accept, manual revision, and discard. Accepted data will be directly used to conduct our final dataset while discarded data will be removed from raw images. For the sample with manual revision, the platform will record the revision results and be used to retrain the automatic annotation system. Through such a manual correction mechanism, we can guarantee the quality of the dataset.

Data flow and interaction procedure: We used the JSON [12] files as the container of annotation data. Here we unify the data format as that of LabelMe [13]. To maintain this protocol, a data format script was designed. Besides, our generated JSON file can directly be read and reversed by the widely used Labelme [13] toolkit which makes our labeling information can be easily gotten by other researchers for further research.

1.5. Agreement

- The MISC210K dataset is available to **non-commer -cial research purposes** only.
- All images of the MISC210K dataset are obtained from the Microsoft COCO dataset [9] which are not property of Academy of Engineering & Technology, Fudan University. Our group is not responsible for the content nor the meaning of these images.
- You agree **not to** reproduce, duplicate, copy, sell, trade, resell or exploit for any commercial purposes, any portion of the images and any portion of derived data including but not limited to annotations, and cropped image parts.
- You agree **not to** further copy, publish or distribute any portion of the MISC210K dataset. Except, for internal use at a single site within the same organization it is allowed to make copies of our dataset.
- Our group reserves the right to terminate your access to the MISC210K dataset at any time.

2. More Results of Benchmark Performance

On top of the MISC210K, we systematically evaluate three kinds of baseline architectures following semantic matching baselines [4, 17] and investigate the joint instance segmentation and semantic correspondence learning tasks based on our DPCL framework. According to these results, we observe that existing methods (MMNet [17] and CATs [4]) fail to perform well in distinguishing key-points in instances of challenging classes such as 'fork' and 'skis' which usually contain serious occlusions. In addition, on basis of another ablation study, selection of joint learning task plays an important role in training procedure of DPCL.

2.1. Basis of Evaluation Metrics Selection

Here we extend the PCK metric as mPCK instead of directly using the mAP (pose) metric in pose estimation [8]. This is because the mAP (pose) metric only defines the instance level unit of positive and negative samples with object keypoint similarity (OKS) [16] calculation. However, in semantic correspondence task, we have to evaluate the result point-by-point to objectively evaluate model performance. While for instance co-segmentation, although original instance segmentation head can be used, models do not need to judge multiple instance categories, but only need to judge the foreground and background. Moreover, we pay more attention to the contour accuracy to ensure the accuracy of matching key point grouping. As a result, we just evaluate the segmentation performance with averaged instance IOU instead of mAP (mask) metric [10].

2.2. Finer Grained Baseline Evaluation

Table 5 shows evaluation results of three multi-instance semantic matching baseline for 34 classes as well as overall average $mPCK@\alpha$ with 7 granularity of α. We also provide instance-level evaluation for co-segmentation on Table

5. According to above results of three kinds of baseline architectures, we figure out two conclusions: 1) MMNet [17] shows the worst results, and DPCL achieves the best performance. We consider that MMNet [17] fails to estimate the number of key-point, but our DPCL can model both the number and location of key-point at the same time. 2) Our designed DPCL architecture can further improve the performance because the instance segmentation co-training plays an important role in multi-instance semantic matching and provides additional information for prediction of key-point. In different classes of our built MISC210K, $mPCK@\alpha$ of different methods demonstrate that most methods perform better on clock and perform well on sheep and cow, but are mostly confused in fork, laptop, and tie due to occlusion and interference.

2.3. The Selection of Joint Learning Task in DPCL

Table 6 shows evaluation results of DPCL trained with four different joint learning tasks for 34 classes as well as overall average $mPCK@\alpha$ with 7 granularity of α. We also provide instance-level evaluation for co-segmentation on Table 6. The result shows the instance segmentation is much better than other co-training tasks. We attribute this to instance segmentation requires models to have stronger ability to distinguish similar instances, which is critical to multi-instance semantic correspondence task for the reason that correctly separated instances help models to estimate the number of key-point and to differentiate semantic information under heavy occlusion and truncation.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 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. 3, 16

[5] 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

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[7] 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

[8] 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. 3

[9] 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. 1, 3

[10] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S Ecker. One-shot instance segmentation. *arXiv preprint arXiv:1811.11507*, 2018. 3

[11] 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. 1

[12] Felipe Pezoa, Juan L Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. Foundations of json schema. In *Proceedings of the 25th International Conference on World Wide Web*, pages 263–273, 2016. 3

[13] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008. 3

[14] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20922–20931, 2022. 1

[15] 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. 1

[16] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 3

[17] 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. 3, 4, 16

[18] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

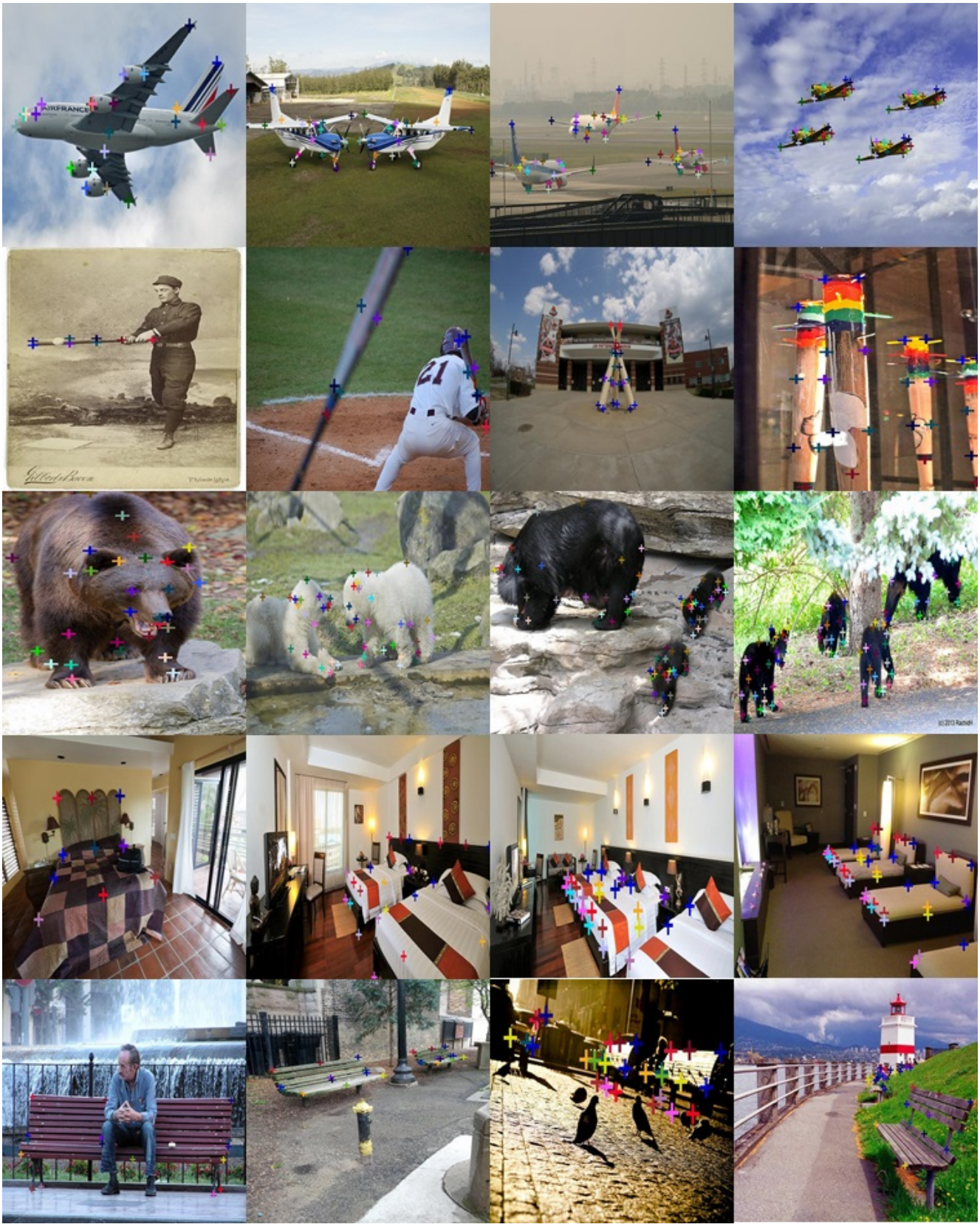

Figure 3. Overview for our dataset (Part 1 of 7). From the left to right are the samples containing 1-4 instances in an image for different categories.

Figure 4. Overview for our dataset (Part 2 of 7). From the left to right are the samples containing 1-4 instances in an image for different categories.

Figure 5. Overview for our dataset (Part 3 of 7). From the left to right are the samples containing 1-4 instances in an image for different categories.

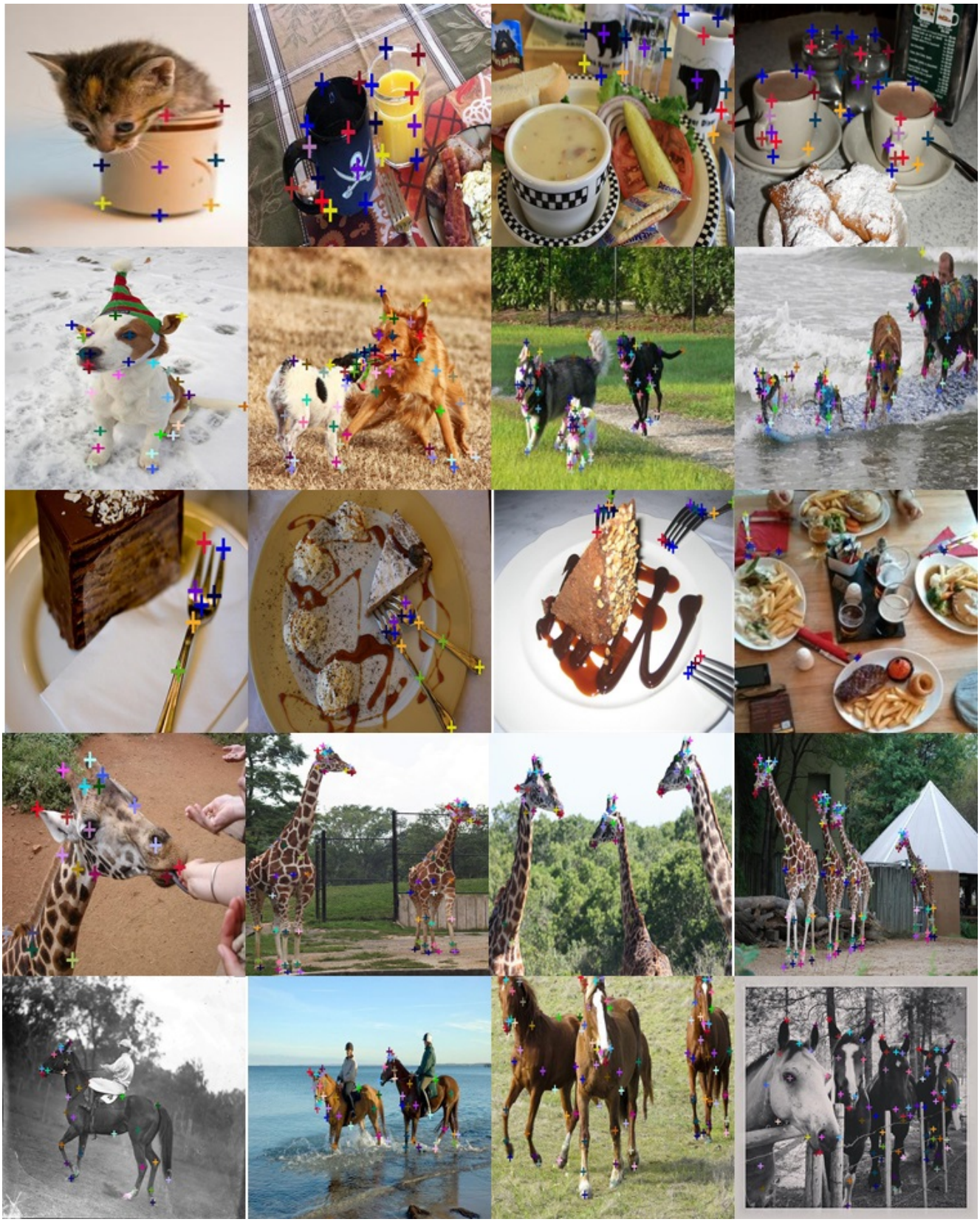

Figure 6. Overview for our dataset (Part 4 of 7). From the left to right are the samples containing 1-4 instances in an image for different categories.

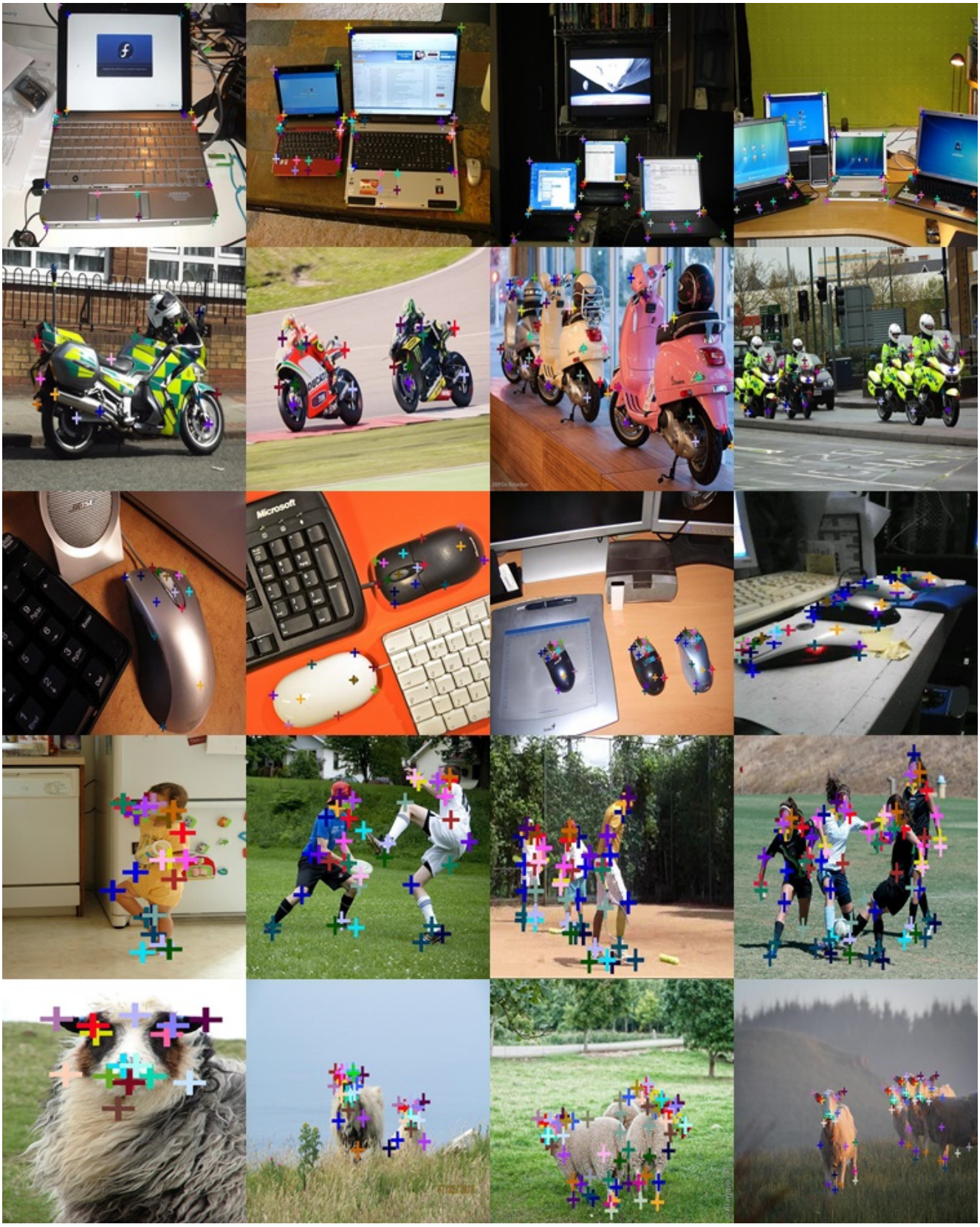

Figure 7. Overview for our dataset (Part 5 of 7). From the left to right are the samples containing 1-4 instances in an image for different categories.

Figure 8. Overview for our dataset (Part 6 of 7). From the left to right are the samples containing 1-4 instances in an image for different categories.

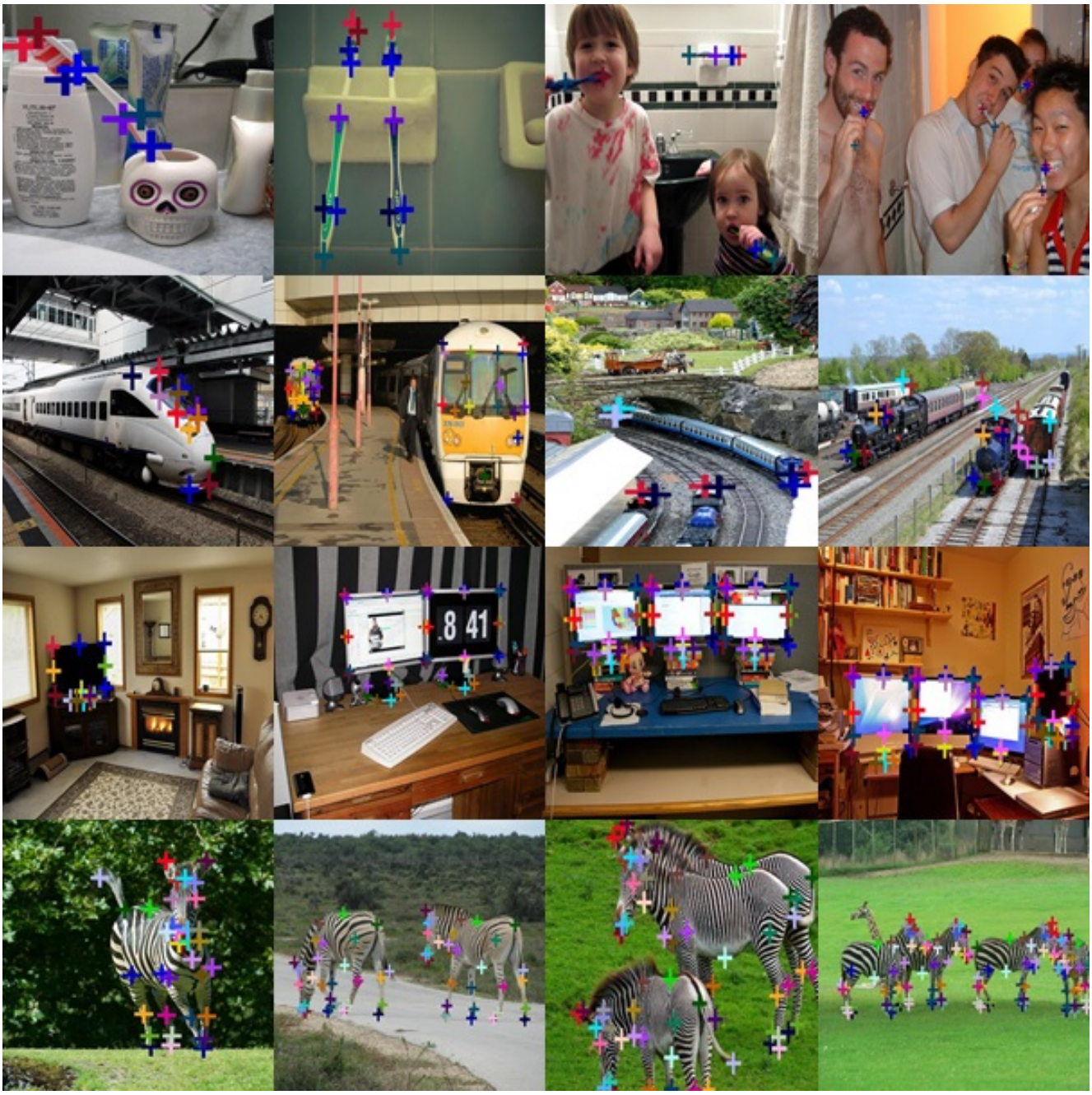

Figure 9. Overview for our dataset (Part 7 of 7). From the left to right are the samples containing 1-4 instances in an image for different categories.

Category	Description
airplane	A clear skeleton and easy to define keypoints. The keypoints have a clear positional relationship with each other. The scenes in which there are large spatial changes and serious differences in perspective.
baseball bat	Clearly definable keypoints exist, and there are clear positional relationships between the keypoints. Small morphological differences exist.
bear	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to label intensively, and the morphology is highly variable, usually for obvious non-rigid transformation.
bed	The keypoints are clear, and there is almost no difference in morphological appearance. It locates in a simple environment where can be directly categorized as a 3D coordinate space transformation and perspective transformation.
bench	The keypoints are clear, between which the positional relationship are clear.
bicycle	It contains a clear skeleton and the keypoints are easy to define. The keypoints have a clear positional relationship with each other. The scenes in which there are large spatial changes and serious differences in perspective.
bird	The presence of obvious keypoints such as the face and wings, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation.
boat	A clear skeleton and easy to define keypoints. The keypoints have a clear positional relationship with each other. The scenes in which there are large spatial changes and serious differences in perspective.
bottle	The presence of key points to be labeled is regular. Objects are usually small and contains challenging issues such as occlusion.
bus	A clear skeleton and easy to define keypoints. The keypoints have a clear positional relationship with each other. The scenes in which there are large spatial changes and serious differences in perspective.
car	A clear skeleton and easy to define keypoints. The keypoints have a clear positional relationship with each other. The scenes in which there are large spatial changes and serious differences in perspective.
cat	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation
chair	The keypoints are clear, and there is almost no difference in morphological appearance. It locates in a simple environment where can be directly categorized as a 3D coordinate space transformation and perspective transformation.
clock	The keypoints are clear, and there is almost no difference in morphological appearance. It locates in a simple environment where can be directly categorized as a 3D coordinate space transformation and perspective transformation.
cow	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation
cup	The presence of key points to be labeled is regular. Objects are usually small and contains challenging issues such as occlusion.
dog	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation.

Table 1. Characteristics of accepted candidate classes (Part 1 of 2).

Category	Description
fork	The presence of key points to be labeled is regular. Objects are usually small and contains challenging issues such as occlusion.
giraffe	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation.
horse	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation.
laptop	Presence of obvious key points such as contours, but existence of complex environment in which the object is located.
motorcycle	A clear skeleton and easy to define keypoints. The keypoints have a clear positional relationship with each other. The scenes in which there are large spatial changes and serious differences in perspective.
mouse	Presence of obvious key points such as contours, but existence of complex environment in which the object is located.
person	There is a clear preamble study discussing how to define the key point system. And the morphology is highly variable, usually for obvious non-rigid transformation.
sheep	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation.
skateboard	There are obvious keypoints, and the shape of the object is usually fixed. However, the scene usually has occlusion problems.
skis	There are obvious keypoints, and the shape of the object is usually fixed. However, the scene usually has occlusion problems.
stop sign	The keypoints are clear, and there is almost no difference in morphological appearance. It locates in a simple environment where can be directly categorized as a 3D coordinate space transformation and perspective transformation.
tennis racket	There are obvious keypoints, and the shape of the object is usually fixed. However, the scene usually has occlusion problems.
tie	Clearly definable keypoints exist, and there are clear positional relationships between the keypoints. Small morphological differences exist.
toothbrush	Key points are clear and have almost no difference in morphological appearance. Object is situated in an environment simply classified as a 3D coordinate space transformation.
train	A clear skeleton and easy to define keypoints. The keypoints have a clear positional relationship with each other. The scenes in which there are large spatial changes and serious differences in perspective.
tv	The keypoints are clear, and there is almost no difference in morphological appearance. It locates in a simple environment where can be directly categorized as a 3D coordinate space transformation and perspective transformation.
zebra	The presence of obvious keypoints such as the face and limbs, but the body and other parts are difficult to mark intensively, and the morphology is highly variable, usually for obvious non-rigid transformation.

Table 2. Characteristics of accepted candidate classes (Part 2 of 2).

Category	Description
apple	Large variation in morphology, difficult to define suitable key points.
backpack	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
banana	Large variation in morphology, difficult to define suitable key points.
baseball glove	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
book	Large variation in morphology and heavy overlapping among instances, difficult to define suitable key points.
bowl	Instances are usually almost invisible, with severe morphological changes.
broccoli	Large variation in morphology, difficult to define suitable key points.
cake	Large variation in morphology, difficult to define suitable key points.
carrot	Large variation in morphology, difficult to define suitable key points.
cell phone	The large variation in the morphology of objects in the same category makes it difficult to define suitable key points.
couch	Often occluded by other objects, lack of samples from multi-instance scenes.
dining table	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
donut	Large variation in morphology, difficult to define suitable key points.
elephant	Presence of large low-texture areas. The instance itself is too large and the environment is complex, so there is often occlusion between multiple instances and occlusion from the environment.
fire hydrant	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
frisbee	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
hair drier	Insufficient number of image samples. The large variation in the morphology of objects in the same category makes it difficult to define suitable key points.
handbag	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
hot dog	Large variation in morphology,difficult to define suitable key points.
keyboard	Key points of detail in the perspective are usually hard to annotate.
kite	The large variation in the morphology of objects in the same category makes it difficult to define suitable key points.
knife	Lack of obvious description of potential keypoints. In most scenes, only the handle of the knife appears, and it is difficult for even the annotator to confirm the category of objects.
microwave	Large morphological differences in the same category. There are few multi-instance scenes to be labeled.

Table 3. Characteristics of abandoned candidate classes (Part 1 of 2).

Category	Description
orange	Frequent congestion between instances, presence of large low-texture areas.
oven	Large morphological differences in the same category. There are few multi-instance scenes to be labeled.
parking meter	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
pizza	Large variation in morphology, difficult to define suitable key points.
potted plant	Large variation in morphology, difficult to define suitable key points.
refrigerator	Large morphological differences in the same category. Few images from multi-instance scenes can be labeled.
remote	Objects are often too small and heavily occluded or blurred. Key points of detail in the perspective are usually hard to annotate.
sandwich	Large variation in morphology, difficult to define suitable key points.
scissors	Objects are often too small and heavily occluded or blurred. Key points of detail in the perspective are usually hard to annotate.
sink	Large morphological differences in the same category. There are few multi-instance scenes to be labeled.
snowboard	Heavy occlusion occurs. Presence of large low-texture areas.
spoon	In most scenes, only the handle of the spoon appears, and hard to confirm the object category
sports ball	Lack of obvious description of potential keypoints. The scenes are usually blurry with problems as tiny instances and occlusion.
suitcase	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
surfboard	Heavy occlusion occurs. Presence of large low-texture areas.
teddy bear	The overall sample size is too small.
toaster	Large morphological differences in the same category. There are few multi-instance scenes to be labeled.
toilet	Large variation in morphology, difficult to define suitable key points.
traffic light	Objects are often too small and heavily occluded or blurred. Key points of detail in the perspective are usually hard to annotate.
truck	A clear skeleton and easy to define keypoints. The keypoints have a clear positional relationship with each other. However, the morphological differences in truck are too great to find the intersection of keypoints.
umbrella	Presence of large low-texture areas. The morphology of objects in the same category varies greatly, making it difficult to define suitable keypoints.
vase	The large variation in the morphology of objects in the same category makes it difficult to define suitable key points.
wine glass	Transparent objects make it difficult to determine the range and shape of the object

Table 4. Characteristics of abandoned candidate classes (Part 2 of 2).

Method	α	airplane	base-ballbat	bear	bed	bench	bicycle	bird	boat	bottle	bus	car	cat	chair	clock	cow	cup	dog	all
MMNet [17]	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
	0.20	51.65	56.50	65.45	34.86	46.52	64.49	68.60	56.13	44.00	41.19	22.91	58.73	42.11	72.52	63.38	49.91	66.98	55.11
	0.25	62.86	69.69	77.22	49.27	59.51	76.30	79.40	66.44	52.04	56.09	34.63	68.77	53.65	78.38	74.01	61.69	79.11	66.93
	0.30	70.33	79.13	83.97	60.77	69.50	83.50	86.04	73.87	60.05	67.23	45.43	75.39	61.88	81.96	80.39	69.67	85.92	75.14
	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
	0.20	48.25	35.02	51.32	32.20	30.64	47.14	48.13	41.16	31.82	49.81	34.21	46.16	32.96	54.61	49.29	30.82	45.17	45.04
	0.25	56.50	44.55	59.45	40.93	39.66	55.40	56.06	48.90	39.17	58.76	41.28	54.09	41.57	61.35	57.92	38.18	53.49	53.12
	0.30	62.52	53.01	65.12	48.61	46.15	62.03	62.39	55.99	45.15	65.26	46.94	60.29	48.39	66.00	63.78	44.44	59.90	59.36
	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
	0.20	46.89	44.91	63.19	26.10	40.24	58.94	63.35	51.45	44.22	39.23	34.74	61.00	32.90	55.91	57.32	46.23	56.92	47.43
	0.25	56.14	57.31	70.71	37.91	49.18	65.85	69.85	57.52	50.20	49.86	42.80	68.37	43.34	62.21	65.16	56.76	66.01	56.54
	0.30	63.05	63.28	75.80	48.72	56.05	71.15	76.02	65.05	54.87	58.89	48.85	73.38	53.70	67.76	70.83	66.32	73.65	63.82
	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
	IoU	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 [17]	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
	0.20	37.32	48.15	56.19	49.56	58.35	52.30	49.09	73.01	55.88	62.07	60.95	71.63	65.89	51.00	63.45	45.63	52.39	55.11
	0.25	48.68	61.67	70.04	62.17	69.31	68.95	62.47	83.10	67.65	75.22	73.20	84.66	75.84	62.67	74.25	59.44	67.75	66.93
	0.30	58.41	71.77	79.50	72.25	77.12	80.17	72.42	88.49	76.09	83.28	80.49	90.37	82.49	70.48	81.15	70.56	78.58	75.14
	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
	0.20	29.34	56.96	59.79	37.49	55.38	34.08	52.45	60.57	42.16	52.38	54.79	42.01	56.06	28.11	52.46	46.61	58.86	45.04
	0.25	36.32	64.73	67.49	47.12	64.80	40.55	60.91	67.71	51.02	59.97	61.12	50.39	63.75	36.01	60.83	55.50	66.94	53.12
	0.30	43.33	70.61	73.21	54.40	70.34	45.53	67.23	72.74	58.41	65.61	66.04	57.57	68.76	42.69	67.06	62.72	72.44	59.36
	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
	0.20	37.18	51.49	49.58	33.89	57.77	37.39	51.46	64.11	47.46	35.33	36.65	56.80	48.53	44.74	50.63	32.98	53.99	47.43
	0.25	44.28	60.48	58.92	45.47	67.22	47.27	61.65	71.27	60.33	45.57	45.54	63.85	59.84	51.20	60.93	45.19	62.92	56.54
	0.30	50.11	66.51	67.44	55.08	73.20	56.37	68.75	77.44	68.51	55.78	53.84	70.64	66.44	58.14	68.05	55.25	69.75	63.82
	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
	IoU	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 5. Evaluation for MMNet [17], 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 (IOU) for DPCL.

Method	α	airplane	base-ballbat	bear	bed	bench	bicycle	bird	boat	bottle	bus	car	cat	chair	clock	cow	cup	dog	all
DPCL-SC	0.05	6.48	2.89	9.14	2.32	1.71	6.81	12.01	8.07	4.07	3.48	2.08	6.72	1.04	7.49	11.20	4.83	8.57	5.76
	0.10	18.29	15.01	23.33	6.44	11.02	18.03	28.93	21.42	12.60	9.86	8.94	18.52	9.20	19.38	25.25	12.33	21.48	15.84
	0.15	31.07	29.34	34.76	12.83	18.49	30.70	39.14	31.52	17.97	19.99	13.62	28.76	17.71	30.16	36.22	21.80	34.33	25.90
	0.20	42.11	40.73	44.44	21.92	28.86	41.48	48.72	41.62	28.50	30.03	22.52	39.28	23.54	40.41	44.63	33.65	45.13	36.17
	0.25	52.95	53.58	52.43	30.83	38.52	50.60	56.18	46.92	40.46	37.86	30.07	49.18	30.97	49.18	51.93	43.55	54.23	46.08
	0.30	60.59	60.14	58.15	41.29	48.46	61.47	62.72	56.01	47.16	47.03	36.54	55.44	39.25	57.72	59.21	50.67	63.02	54.38
	1.00	98.66	99.21	99.63	97.37	99.26	99.64	99.62	99.55	99.43	99.19	97.79	98.30	98.93	99.70	99.93	99.14	99.19	99.21
DPCL-CL	0.05	7.63	5.19	8.72	0.73	3.18	5.74	11.07	7.96	6.28	3.70	2.18	6.69	2.60	13.03	9.75	3.08	7.33	6.19
	0.10	18.40	17.96	21.45	5.15	11.81	17.13	26.90	20.73	13.19	10.12	9.16	19.58	12.32	27.63	24.94	16.77	21.88	17.12
	0.15	28.72	29.88	33.70	12.77	22.61	30.37	40.70	34.54	20.33	19.65	15.80	30.99	21.65	39.87	38.19	25.99	35.10	27.85
	0.20	41.18	39.82	46.45	23.59	32.06	42.13	51.74	44.29	30.43	31.85	23.05	40.61	29.55	49.74	48.52	38.64	43.97	38.26
	0.25	51.01	54.14	55.66	34.79	41.02	55.20	59.61	54.60	39.56	41.81	30.53	51.45	37.92	59.03	57.43	46.72	53.22	48.34
	0.30	59.02	62.63	63.95	44.56	50.59	65.26	66.97	63.80	46.39	50.32	37.64	61.45	47.36	64.36	65.71	54.06	61.38	57.16
	1.00	99.43	99.75	99.35	99.49	97.85	98.27	99.36	100.00	99.70	97.50	98.67	98.97	99.80	100.00	99.35	99.11	99.45	99.34
DPCL-DET	0.05	8.10	6.39	15.76	2.03	2.90	11.76	21.24	10.21	9.68	5.16	5.03	12.46	4.94	12.40	14.46	8.50	11.84	9.21
	0.10	20.01	20.78	32.96	6.22	11.86	27.58	40.61	23.39	21.05	13.60	14.28	33.57	15.99	29.75	32.38	20.25	29.48	22.21
	0.15	33.99	36.69	47.94	15.01	22.13	44.12	51.52	39.39	25.57	24.12	20.64	49.18	24.47	41.55	45.94	28.61	43.19	33.72
	0.20	45.47	46.49	59.56	27.26	35.46	57.50	60.73	52.45	32.58	36.90	29.06	58.58	33.68	50.18	54.86	39.65	54.94	44.49
	0.25	55.16	59.23	67.36	39.75	46.99	65.57	69.17	59.65	41.95	47.20	36.56	68.04	42.75	59.38	62.46	49.27	64.71	54.80
	0.30	62.80	69.00	73.32	50.50	53.99	72.99	75.81	68.03	49.45	55.81	42.11	73.72	53.14	68.73	68.83	59.74	73.26	63.29
	1.00	99.20	99.69	99.25	99.50	99.94	99.87	99.97	99.82	99.40	97.56	99.48	99.41	99.59	98.98	98.92	99.78	98.65	98.99
DPCL-IS (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
	0.20	46.89	44.91	63.19	26.10	40.24	58.94	63.35	51.45	44.22	39.23	34.74	61.00	32.90	55.91	57.32	46.23	56.92	47.43
	0.25	56.14	57.31	70.71	37.91	49.18	65.85	69.85	57.52	50.20	49.86	42.80	68.37	43.34	62.21	65.16	56.76	66.01	56.54
	0.30	63.05	63.28	75.80	48.72	56.05	71.15	76.02	65.05	54.87	58.89	48.85	73.38	53.70	67.76	70.83	66.32	73.65	63.82
	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
	IoU	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
DPCL-SC	0.05	2.89	4.36	5.55	3.05	5.63	5.95	5.40	12.51	4.35	4.94	4.09	5.07	10.18	6.82	6.33	2.37	5.99	5.76
	0.10	9.43	13.07	17.27	9.30	16.48	13.64	18.49	27.14	12.94	19.05	12.42	14.01	20.86	15.86	15.42	6.49	17.01	15.84
	0.15	14.44	21.81	29.19	16.88	29.31	19.70	33.36	37.53	25.18	30.71	21.45	24.04	31.80	21.13	26.45	17.02	28.67	25.90
	0.20	21.89	31.34	39.16	26.09	42.06	27.27	45.54	49.65	36.82	41.08	30.14	36.56	39.24	28.57	39.14	30.27	42.72	36.17
	0.25	31.23	40.19	48.41	38.27	52.35	34.79	56.65	59.95	47.49	48.97	38.26	51.88	51.15	38.00	51.29	42.79	56.19	46.08
	0.30	38.01	48.51	57.34	49.81	60.23	42.78	66.23	67.40	53.63	57.87	44.91	61.08	59.88	47.16	58.49	54.98	64.51	54.38
	1.00	98.84	98.82	99.94	98.34	100.00	99.14	99.61	99.92	99.84	100.00	99.46	98.40	98.68	98.45	99.38	99.28	99.76	99.21
DPCL-CL	0.05	3.16	5.35	5.90	5.36	5.65	7.33	6.75	12.27	4.44	3.23	7.27	7.23	6.19	7.64	4.55	2.42	6.39	6.19
	0.10	9.63	16.23	18.06	15.07	19.70	15.25	17.76	28.91	12.89	17.17	17.77	17.52	15.43	18.21	14.99	9.12	19.94	17.12
	0.15	16.90	26.64	32.00	24.47	37.43	21.00	31.46	40.17	22.74	27.19	26.29	25.73	27.42	26.53	27.52	18.00	32.50	27.85
	0.20	24.82	37.77	43.07	34.84	53.06	23.90	44.91	49.87	32.27	37.37	34.67	36.11	37.04	34.53	37.67	31.50	43.65	38.26
	0.25	34.78	48.76	52.66	45.23	64.32	29.43	56.70	59.99	42.27	51.90	44.10	47.20	45.34	41.84	50.24	44.27	54.36	48.34
	0.30	46.50	58.44	61.79	56.08	72.36	34.87	67.06	67.47	50.10	61.86	52.85	57.44	56.46	51.06	60.37	55.81	62.17	57.16
	1.00	99.34	99.41	99.98	98.92	99.33	99.62	99.66	100.00	99.21	99.61	99.82	100.00	99.46	98.58	98.64	99.47	99.83	99.34
DPCL-DET	0.05	3.85	11.22	10.10	4.00	7.79	9.29	9.19	18.29	6.20	2.99	5.06	9.09	12.49	17.77	8.61	2.13	9.71	9.21
	0.10	10.46	28.05	23.36	14.60	24.36	17.65	23.04	39.61	18.06	11.30	11.85	20.41	21.59	36.66	24.72	8.63	23.94	22.21
	0.15	19.21	37.43	37.41	24.87	39.89	28.57	36.81	53.73	31.58	17.66	22.69	35.63	34.16	41.43	36.46	18.45	36.16	33.72
	0.20	29.45	47.64	46.73	38.24	53.06	37.67	49.35	62.28	41.53	27.48	34.71	48.91	44.18	50.03	45.50	32.28	48.90	44.49
	0.25	43.65	57.78	57.38	50.88	65.03	48.30	58.76	71.46	54.92	38.65	44.64	62.16	56.61	57.18	57.04	45.55	58.48	54.80
	0.30	52.53	65.65	66.22	60.19	74.01	55.80	66.72	79.14	66.19	48.80	54.00	71.43	64.13	62.86	66.10	58.77	65.50	63.29
	1.00	98.98	99.62	99.79	98.62	99.48	97.61	99.26	98.45	98.75	97.96	98.57	99.34	97.03	98.45	97.06	99.83	99.02	98.99
DPCL-IS (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
	0.20	37.18	51.49	49.58	33.89	57.77	37.39	51.46	64.11	47.46	35.33	36.65	56.80	48.53	44.74	50.63	32.98	53.99	47.43
	0.25	44.28	60.48	58.92	45.47	67.22	47.27	61.65	71.27	60.33	45.57	45.54	63.85	59.84	51.20	60.93	45.19	62.92	56.54
	0.30	50.11	66.51	67.44	55.08	73.20	56.37	68.75	77.44	68.51	55.78	53.84	70.64	66.44	58.14	68.05	55.25	69.75	63.82
	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
	IoU	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 6. Evaluation for DPCL trained with only semantic correspondence task (DPCL-SC), DPCL co-trained with image classification task (DPCL-CL), DPCL co-trained with object detection task (DPCL-DET) and DPCL co-trained with instance segmentation task (DPCL-IS). We provide mPCK result on 34 classes in MISC210K with different α metrics. We also provide instance-level evaluation (IOU) for DPCL-IS, which is identical to DPCL in our main paper and Table 5.