Appendices

A. Details of datasets

We list the information of each dataset in Table A, including number of categories, data scale of training and validation sets, and metrics.

Table A: Details about the different datasets. mAP, mmAP and MR^{-2} are abbreviations of the mean Average Precision at overlap 0.5, mean mAP over overlap ranging in [0.5, 0.95], and log average Miss Rate over false positives per image ranging in $[10^{-2}, 10^{0}]$.

Dataset	Category	Train	Validation	Metric
Open Images	500	1.7m	40k	mAP
Objects365	365	600k	30k	mmAP
COCO	80	115k	5k	mmAP
Caltech	1	42k	3k	MR^{-2}
CityPersons	1	3k	0.5k	MR^{-2}
VOC	20	16k	5k	mAP
WiderFace	1	13k	3k	mAP
KITTI	3	4k	4k	mAP
LISA	4	8k	2k	mAP
DOTA	15	14k	5k	mAP
Watercolor	6	1k	1k	mAP
Clipart	20	0.5k	0.5k	mAP
Comic	6	1k	1k	mAP
Kitchen	11	5k	2k	mAP
DeepLesions	1	28k	5k	mAP

B. Rules for architecture selection

As mentioned in the Section4.2.1 of paper, there are some prescribed rules for architecture selection. We denote the minimum total depth and maximum total depth as d_{min} and d_{max}. The depth of model is denoted as d and we have $d' = d_{max} - d_{min}$ for simplicity. The pool of rules and respective sampling probability are shown as follow:

- models with $d = d_{min}$, $p = \frac{1}{8}$.
- models of the $d = d_{min} + 0.25d'$, $p = \frac{1}{8}$.
- models of the $d = d_{min} + 0.5d'$, $p = \frac{1}{8}$.
- models of the $d = d_{min} + 0.75d'$, $p = \frac{1}{8}$.
- models of the $d = d_{max}$, $p = \frac{1}{8}$.
- random models, $p = \frac{3}{8}$.

C. Visualization results on each dataset

We visualize the detection results on each dataset, as demonstrated in Figure A.

D. The adapted architectures to each dataset

We list the selected architecture for each downstream task(Table 5 in paper) in Table B. "†": In CityPersons dataset, the default input size is 1024×2048, thus we build our search space of input scale surrounding the default input size with a step of 128 pixels.

Table B: Details about the adapted architectures.

Dataset	Scale	Depth	Width
Open Images	640	[4,6,29,4]	[64,80,160,192,640]
Objects365	720	[3,4,23,3]	[64,64,128,192,640]
COCO	720	[3,4,23,3]	[64,64,128,192,640]
Caltech	880	[2,4,17,2]	[48,48,128,256,640]
CityPersons†	1152	[3,2,4,3]	[64,64,96,192,384]
VOC	640	[3,4,29,4]	[64,64,128,256,512]
WiderFace	880	[4,4,4,2]	[64,64,96,192,384]
KITTI	880	[3,4,6,3]	[48,64,96,192,384]
LISA	720	[2,4,17,3]	[64,64,128,192,512]
DOTA	880	[4,6,4,2]	[32,48,96,192,512]
Watercolor	640	[3,2,29,3]	[48,80,96,256,640]
Clipart	640	[3,6,17,3]	[32,64,128,320,640]
Comic	640	[2,6,17,2]	[48,64,160,320,512]
Kitchen	720	[3,6,23,2]	[48,48,160,192,512]
DeepLesions	720	[4,4,17,2]	[32,80,96,256,512]

E. The selected data for each downstream task

Given few images from downstream tasks as query, we show the relevant data collected by GAIA in Figure B. (Table 6 in paper)

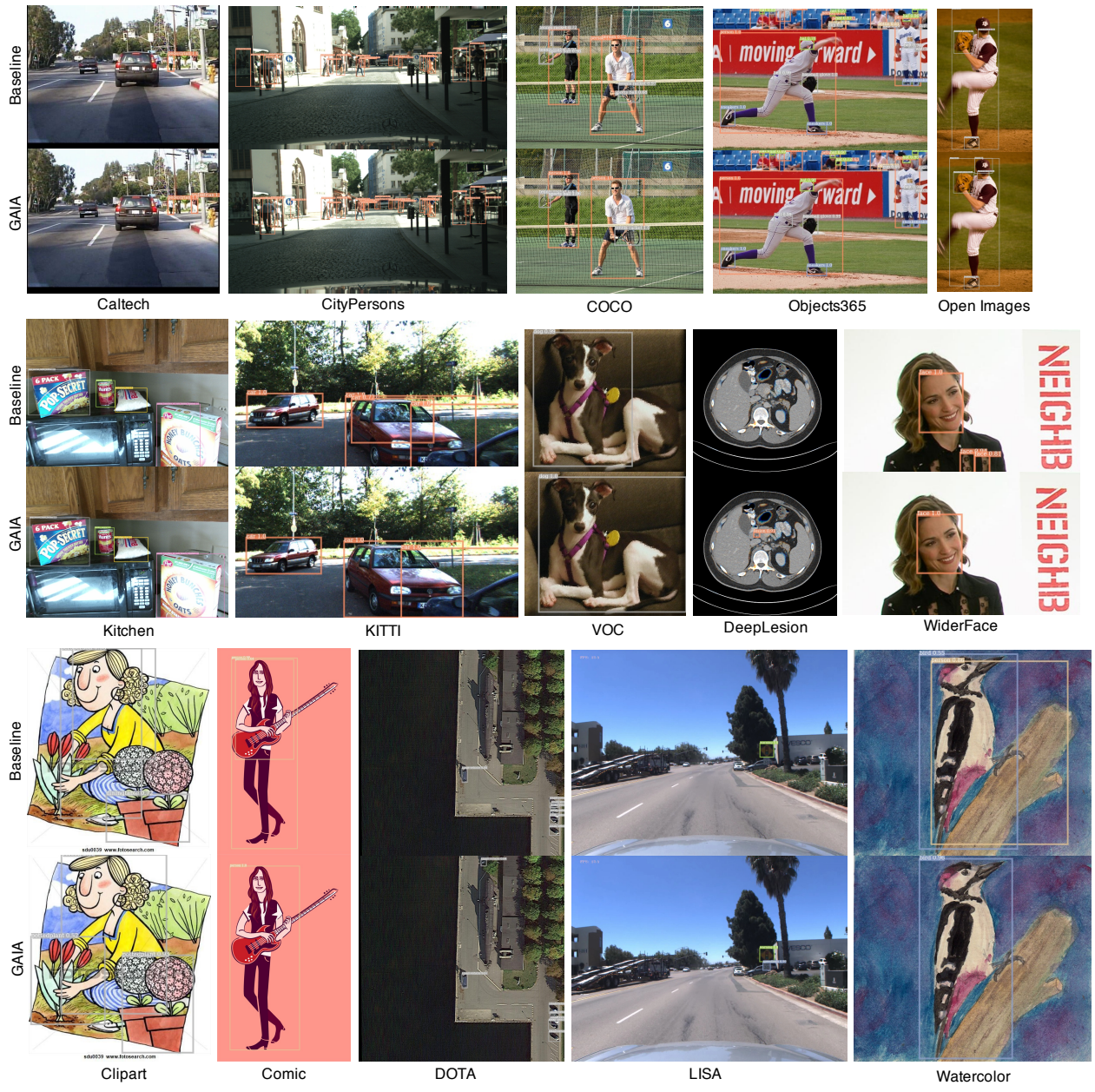
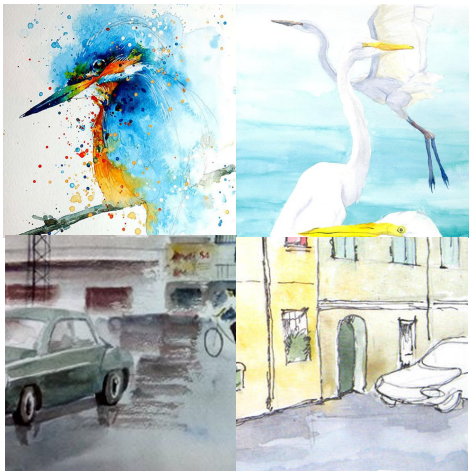
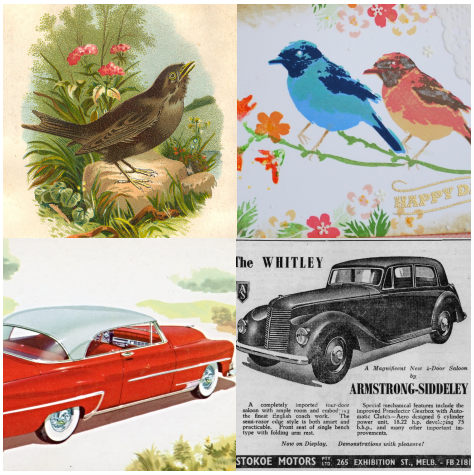


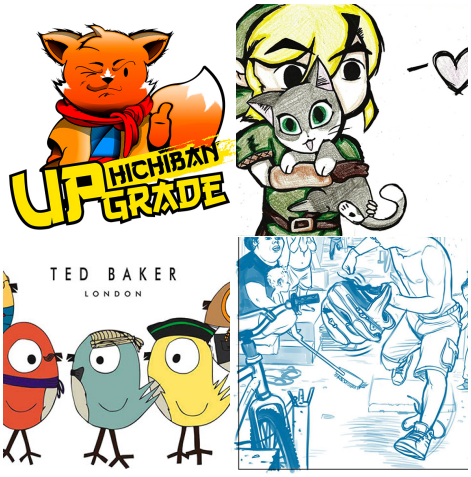
Figure A: Examples of detection results from ImageNet baseline and GAIA on each dataset.



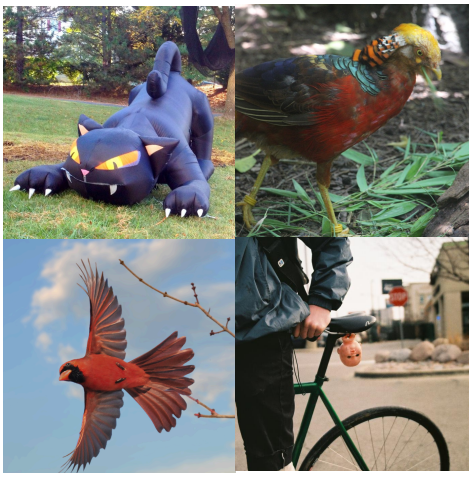
Watercolor



Selected data

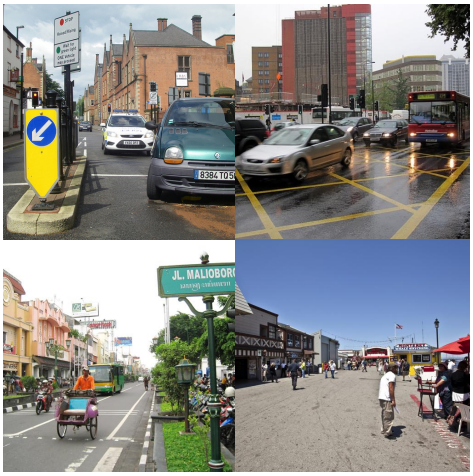


Comic



Selected data

KITTI



Selected data

Figure B: Examples of data selection results. From top to bottom: Watercolor, Comic, and KITTI. From left to right: downstream tasks data and their corresponding images selected from upstream datasets.