Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching

Yang Liu
Zhejiang University, China
yangliu9610@zju.edu.cn

Muzhi Zhu
Zhejiang University, China
zhumuzhi@zju.edu.cn

Hengtao Li
Zhejiang University, China
liht@zju.edu.cn

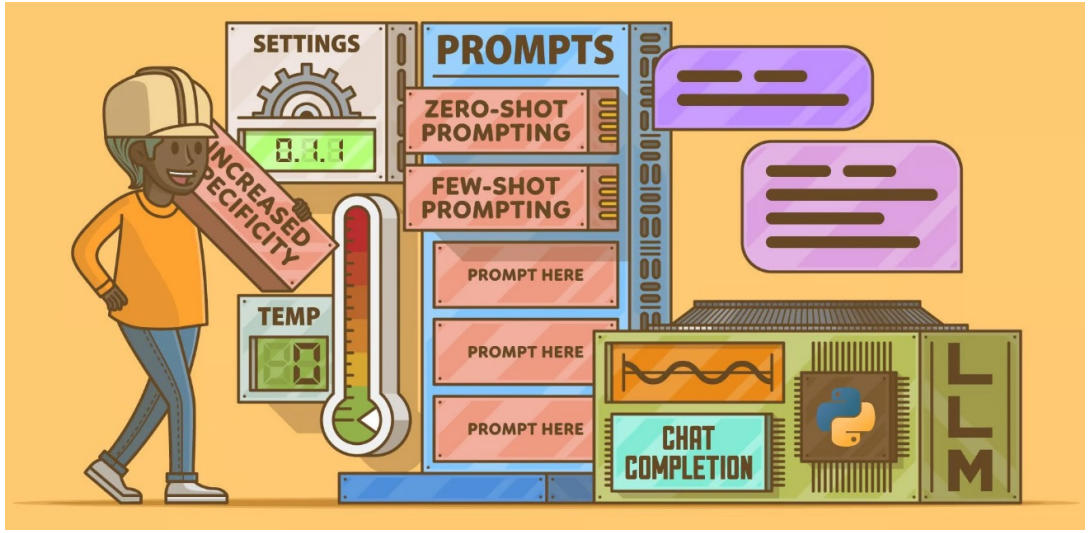
Hao Chen
Zhejiang University, China
haochen.cad@zju.edu.cn

Xinlong Wang
Beijing Academy of Artificial Intelligence
wangxinlong@baai.ac.cn

Chunhua Shen
Zhejiang University, China
chunhuashen@zju.edu.cn

Motivation

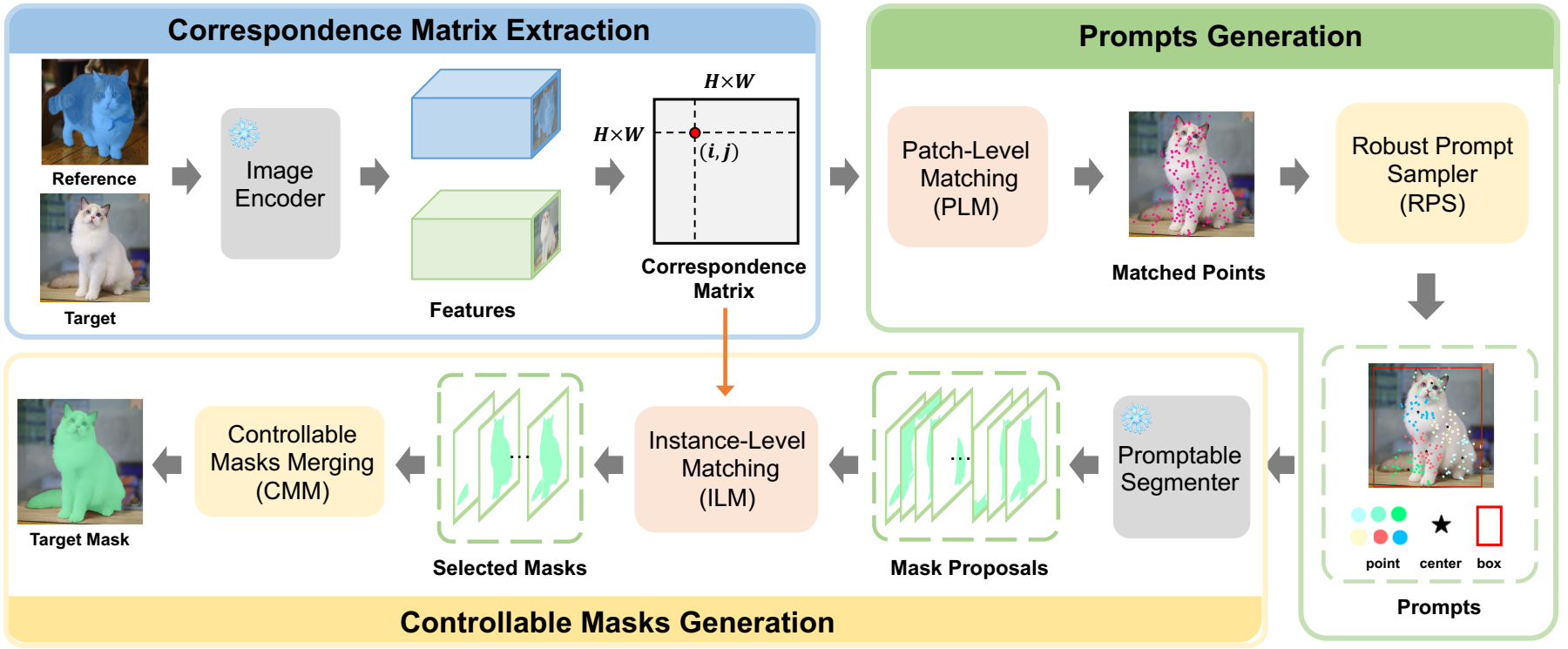
Large language models (LLMs) show powerful zero-shot and few-shot generalization and solve various language tasks well.



How to use vision foundation models (VFMs) to perform various perception tasks in a training-free way?

This work aims to find a new visual research paradigm: investigating the utilization of VFMs for effectively addressing a wide range of perception tasks, e.g., semantic segmentation, part segmentation, and video object segmentation, without training.

Method



- Correspondence Matrix Extraction
- Prompts Generation
- Controllable Masks Generation

Correspondence Matrix Extraction

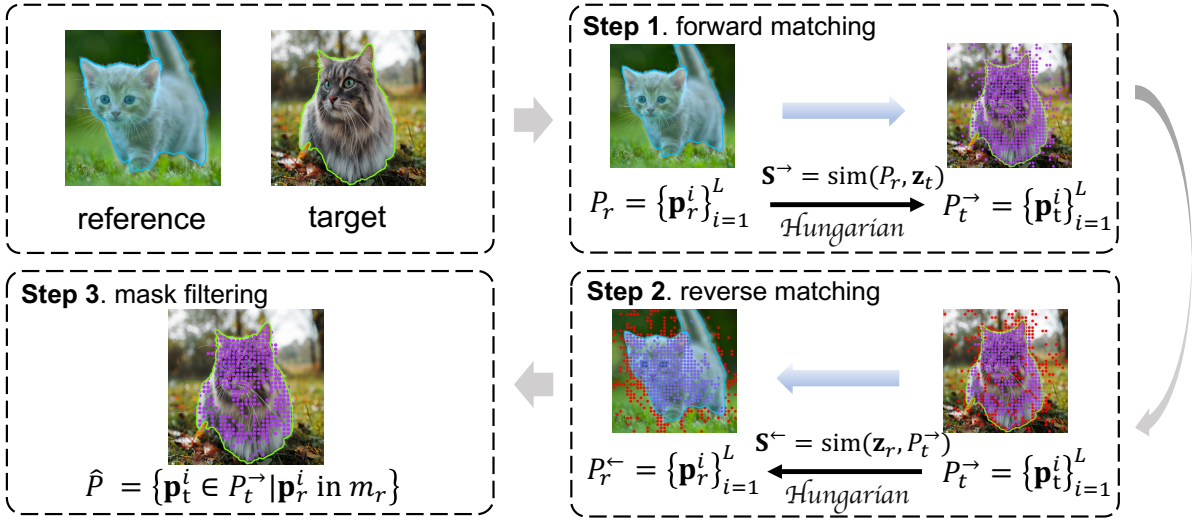
Given inputs $\mathbf{x}_r$and $\mathbf{x}_t$, the image encoder outputs patch-level features $\mathbf{z}_r, \mathbf{z}_t \in \mathbb{R}^{H\times W\times C}$. We define the correspondence matrix $\mathbf{S} \in \mathbb{R}^{HW\times HW}$ as follows:

$$(\mathbf{S})_{ij} = \frac{\mathbf{z}_r^i \cdot \mathbf{z}_r^j}{\left\|\mathbf{z}_r^i \cdot\right\| \cdot \left\|\mathbf{z}_r^j\right\|}$$

The above formulation can be denoted as $\mathbf{S} = \mathrm{sim}(\mathbf{z}_r, \mathbf{z}_t)$.

Prompts Generation

Patch-Level Matching



Robust Prompt Sampler

Cluster $\hat{P}$ based on locations into K clusters $\hat{P}_k$. Three types of subsets are sampled as prompts:

- Part-level prompts are sampled within each cluster $P^p \subset \hat{P}_k$.
- Instance-level prompts are sampled within all matched points $P^p \subset \hat{P}$.
- Global prompts are sampled within the set of cluster centers $P^g \subset$ C to encourage coverage, where $C = \{c_1, c_2, \ldots, c_k\}$ are the cluster centers.

Controllable Masks Generation

Three effective metrics to select high-quality masks :

- *emd:* structural distance between dense semantic features inside the masks to determine mask relevance.
- $purity = \frac{Num(\hat{P}_{mp})}{Area(mp)}$: higher degree of $purity$ promotes the selection of part-level masks.
- $coverage = \frac{Num(\hat{P}_{mp})}{Num(\hat{P})}$: higher degree of $coverage$ promotes the selection of instance-level masks.

The false-positive mask fragments can be filtered by:

$$score = \alpha \cdot (1 - emd) + \beta \cdot purity \cdot coverage^{\lambda}$$

Experiments

Methods	Venue	COCO-20^i one-shot	COCO-20^i few-shot	FSS-1000 one-shot	FSS-1000 few-shot	LVIS-92^i one-shot	LVIS-92^i few-shot
specialist model							
HSNet (Min et al., 2021)	ICCV'21	41.2	49.5	86.5	88.5	17.4	22.9
VAT (Hong et al., 2022)	ECCV'22	41.3	47.9	90.3	90.8	18.5	22.7
FPTrans (Zhang et al., 2022a)	NeurIPS'22	47.0	58.9	-	-	-	-
MSANet (Iqbal et al., 2022)	arXiv'22	51.1	56.8	-	-	-	-
generalist model							
Painter (Wang et al., 2023a)	CVPR'23	33.1	32.6	61.7	62.3	10.5	10.9
SegGPT (Wang et al., 2023b)	ICCV'23	56.1	67.9	85.6	89.3	18.6	25.4
PerSAM†‡ (Zhang et al., 2023)	arXiv'23	23.0	-	71.2	-	11.5	-
PerSAM-F‡	arXiv'23	23.5	-	75.6	-	12.3	-
Matcher†‡	this work	**52.7**	**60.7**	**87.0**	**89.6**	**33.0**	**40.0**

Table 1: Results of few-shot semantic segmentation on COCO-20^i, FSS-1000, and LVIS-92^i. Gray indicates the model is trained by in-domain datasets. † indicates the training-free method. ‡ indicates the method using SAM. Note that the training data of SegGPT includes COCO.

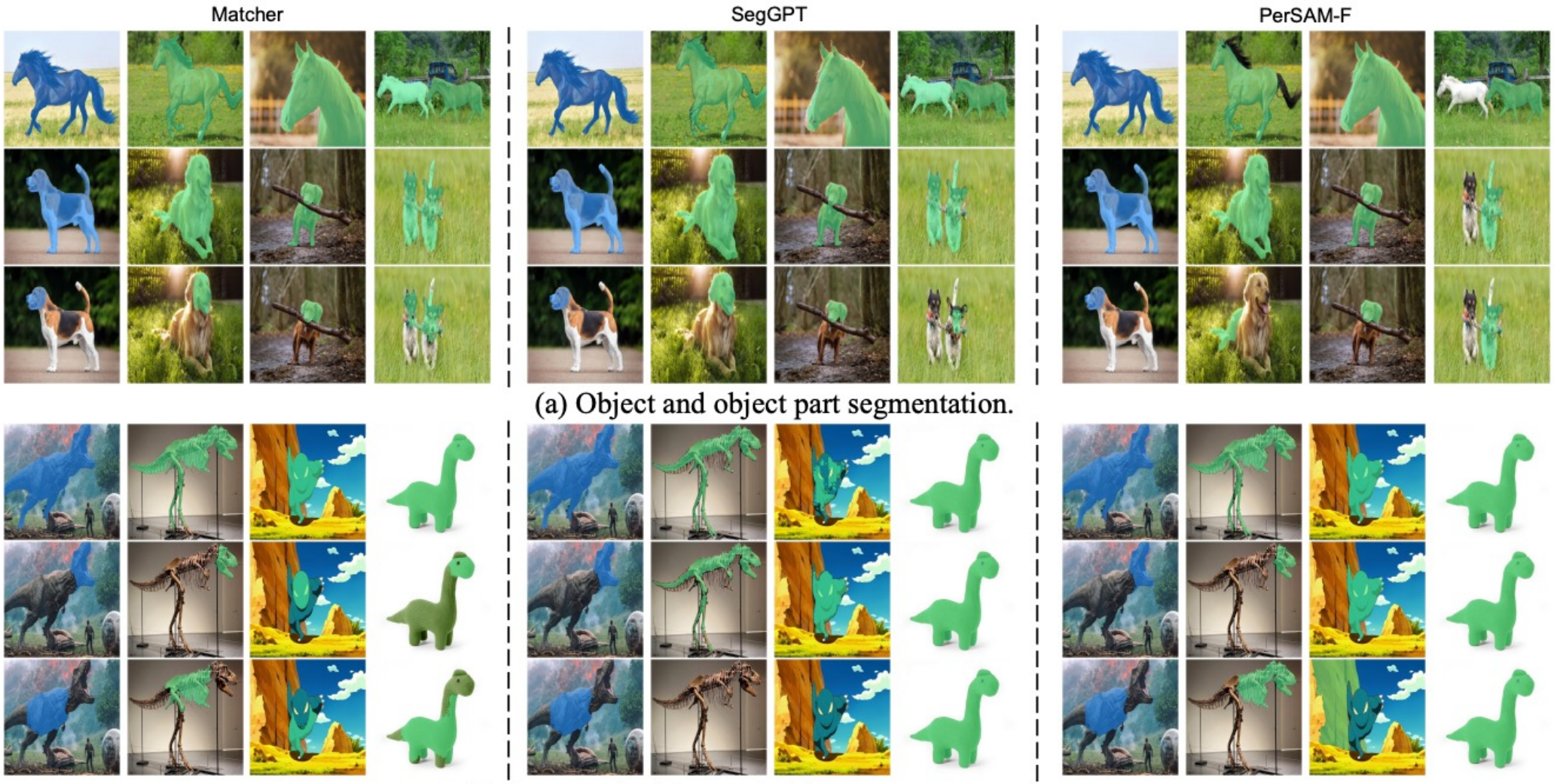
Methods	Venue	PASCAL-Part animals	indoor	person	vehicles	mean	PACO-Part F0	F1	F2	F3	mean
HSNet (Min et al., 2021)	ICCV'21	21.2	53.0	20.2	35.1	32.4	20.8	21.3	25.5	22.6	22.6
VAT (Hong et al., 2022)	ECCV'22	21.5	55.9	20.7	36.1	33.6	22.0	22.9	26.0	23.1	23.5
Painter (Wang et al., 2023a)	CVPR'23	20.2	49.5	17.6	34.4	30.4	13.7	12.5	15.0	15.1	14.1
SegGPT (Wang et al., 2023b)	ICCV'23	22.8	50.9	31.3	38.0	35.8	13.9	12.6	14.8	12.7	13.5
PerSAM†‡ (Zhang et al., 2023)	arXiv'23	19.9	51.8	18.6	32.0	30.1	19.4	20.5	23.8	21.2	21.2
Matcher†‡	this work	**37.1**	**56.3**	**32.4**	**45.7**	**42.9**	**32.7**	**35.6**	**36.5**	**34.1**	**34.7**

Table 2: Results of one-shot part segmentation on PASCAL-Part and PACO-Part. † indicates the training-free method. ‡ indicates the method using SAM.

Methods	Venue	DAVIS 2017 val *J&F*	*J*	*F*	DAVIS 2016 val *J&F*	*J*	*F*
with video data							
AGAME (Johnander et al., 2019)	CVPR'19	70.0	67.2	72.7	-	-	-
AGSS (Lin et al., 2019)	ICCV'19	67.4	64.9	69.9	-	-	-
AFB-URR (Liang et al., 2020)	NeurIPS'20	74.6	73.0	76.1	-	-	-
AOT (Yang et al., 2021)	NeurIPS'21	85.4	82.4	88.4	92.0	90.7	93.3
SWEM (Lin et al., 2022)	CVPR'22	84.3	81.2	87.4	91.3	89.9	92.6
XMem (Cheng & Schwing, 2022)	ECCV'22	87.7	84.0	91.4	92.0	90.7	93.2
without video data							
Painter (Wang et al., 2023a)	CVPR'23	34.6	28.5	40.8	70.3	69.6	70.9
SegGPT (Wang et al., 2023b)	ICCV'23	75.6	72.5	78.6	83.7	83.6	83.8
PerSAM†‡ (Zhang et al., 2023)	arXiv'23	60.3	56.6	63.9	-	-	-
PerSAM-F‡	arXiv'23	71.9	69.0	74.8	-	-	-
Matcher†‡	this work	**79.5**	**76.5**	**82.6**	**86.1**	**85.2**	**86.7**

Table 3: Results of video object segmentation on DAVIS 2017 val, and DAVIS 2016 val. Gray indicates the model is trained on target datasets with video data. † indicates the training-free method. ‡ indicates the method using SAM.

Visualization

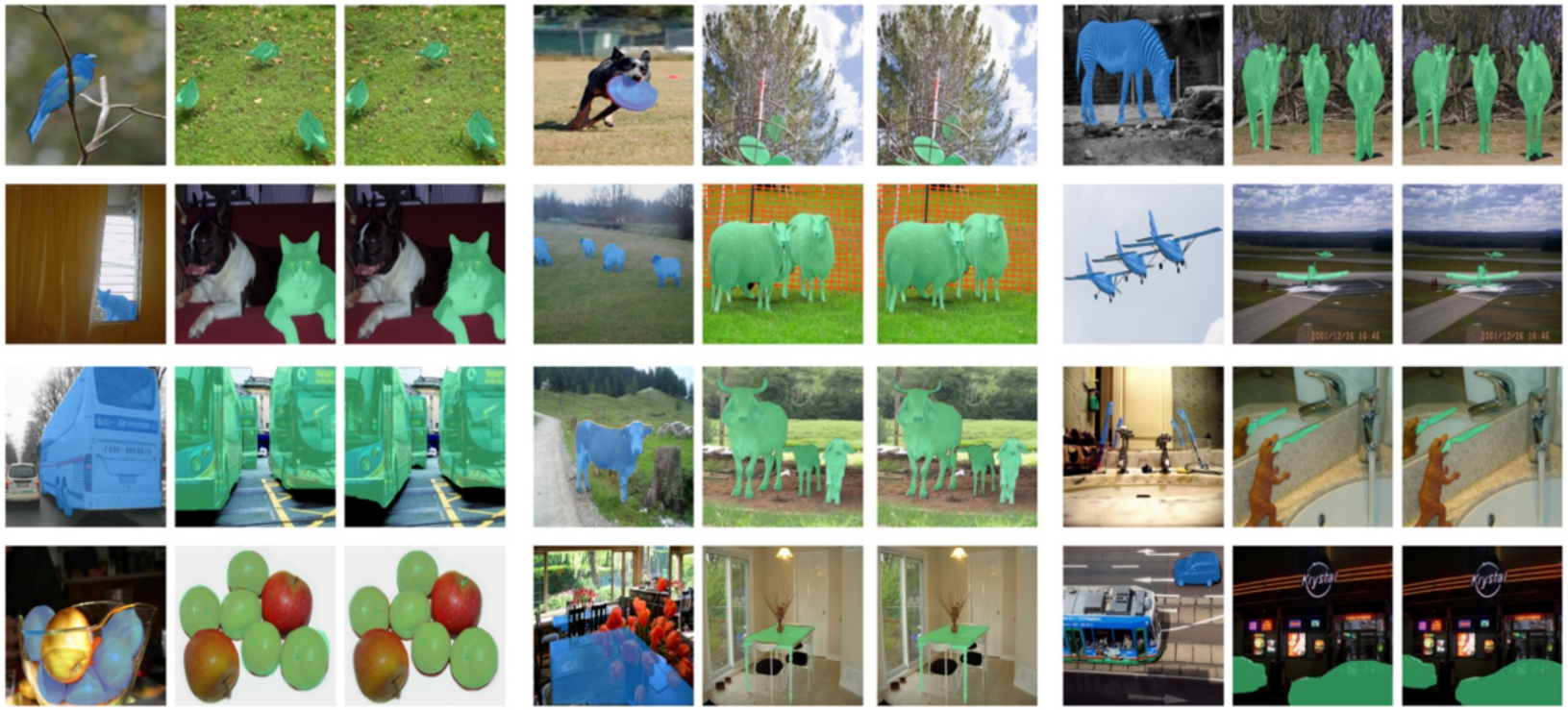


(a) Object and object part segmentation.

(b) Cross-style object and object part segmentation.

Visualization

One-Shot Semantic Segmentation



One-Shot Object Part Segmentation

Thanks for listening.