Beyond Appearance: a Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks

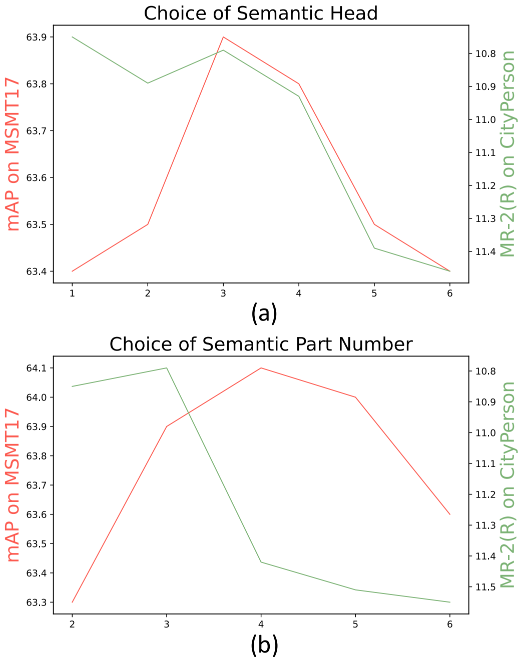
Weihua Chen, Xianzhe Xu, Jian Jia, Hao luo, Yaohua Wang, Fan Wang, Rong Jin, Xiuyu Sun[†]
Alibaba Group
{kugang.cwh, xianzhe.xxz, jj359864, michuan.lh, xiachen.wyh, fan.w, jinrong.jr, xiuyu.sxy}@alibaba-inc.com

In this supplementary material, we first provide experiments on the influence of the semantic head size and the clustered part number in SOLIDER. Then the experiments of λ selection for downstream tasks are present. At last, a pseudo code is given to clarify the whole training process.

1. Choice of Semantic Head Size

This experiment is provided to explore the influence of the size of semantic head. Specifically, we tried semantic head with different number of blocks, ranging 1 to 6 blocks. Each block includes a fully connected layer, a batch norm layer and a relu. We conduct this experiment on two representative downstream tasks, *i.e.* person re-identification and pedestrian detection. Person re-identification is an image-level prediction task, which prefers appearance information for identification purpose. Pedestrian detection is a dense prediction task that focuses on learning global semantic representations from human foreground.

The results are shown in Fig.1 (a). We can see the performance of pedestrian detection is steadily declined during the increase of the semantic head block number (green curve).[1] As we known, the semantic head is used for semantic supervision. A small semantic head would hold limited power on learning a semantic space by itself, which forces the feature maps from the backbone to contain more semantic information to fit the optimization of the semantic supervision. As a result, the representation would preserve more semantic information and provide a better performance on downstream tasks. Hence, a lightweight head would contribute better to the final performance of pedestrian detection.



(a)

(b)

Figure 1. The influence of "the semantic head size" and "the clustered part number" in SOLIDER on person re-identification and pedestrian detection tasks. (a) Results of different block numbers; (b) Results of different clustered part numbers. Best viewed in color.

From another aspect, we can observe that with the increasing of the block number, the performance of person re-identification first increased and then decreased (red curve). It is because the gap between the recognition task and semantic supervision. The semantic supervision would weak the discriminative ability among parts from different images with the same semantic meaning. A too small semantic head leads the backbone to a too strong semantic supervision and harm the performance on recognition tasks, *i.e.* person re-identification. This conclusion is consistent with the design of the lightweight decoder in MAE [1] and SimMIM [2]. However, if the semantic head is too large, the semantic supervision would be useless to the backbone,

[†] Corresponding Author

[1] We invert the axis of the performance ($\text{MR}^{-}2$(R/HO)) of pedestrian detection, to make its tendency consistent with other tasks, *i.e.*, upper is better.

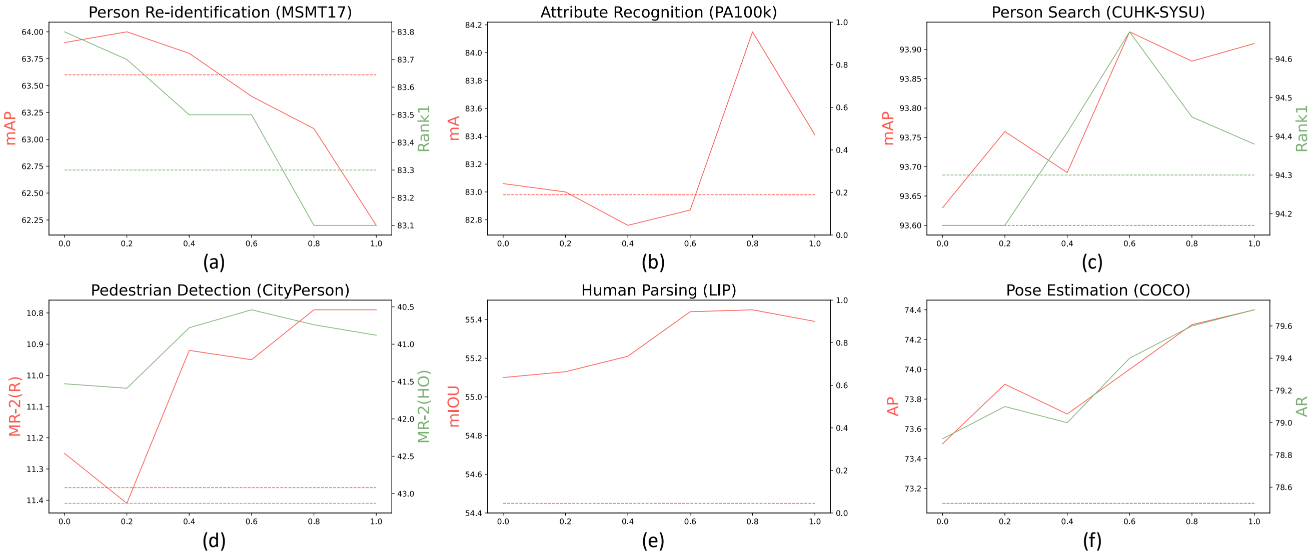


Figure 2. Selections of different λ on six downstream tasks. The dash line means the performance from the original DINO model.

which is also not good for the recognition task.

Base on the observation from two curves, we set the block number to 3 in our experiments for a good balance.

2. Choice of Clustered Semantic Part Number

This number N determines how many semantic regions are discovered from every image. We perform a quantitative ablation study to find the most suitable N for semantic supervision. As shown in Fig.1 (b), with the increase of N, the performance of person re-identification is improved (red curve), while the performance of pedestrian detection drops dramatically (green curve). It shows that the small N benefit to the pedestrian detection task, but is harmful to the person re-identification task. Based on the observation, we provide an analysis that a small N would keep the representation on the global human body in foreground, which benefits the body-level tasks, such as pedestrian detection, but is harmful to those tasks caring about the details in the body. On the contrary, a large N may find more semantic local details that help to build more discriminative representations, but its semantic information is too fragmented to be used for body-level tasks. Meanwhile, a too large N would cause more clustering noises and produce too many detailed fragments, which is also harmful for both tasks. To balance all the tasks in downstream, we choose $N = 3$ for our other experiments.

3. Ablation Study of B&F Clustering and MIM

Background&Foreground (B&F) Clustering and Masked Image Modeling (MIM) are two modifications to improve the Semantic Supervision. We conduct an extra ablation study to verify their effectiveness. The results are in Table. 1. It can be found that both of them benefit to the final performance.

4. Selection λ for Downstream Tasks

The value λ is used to control the ratio of semantic information and appearance information in the output representation. To find the best λ for each task, we conduct experiments of different λ on all downstream tasks. The results are shown in Fig.2. It can be seen that for person re-identification task (a), the performance gets worse with the increase of λ. While for attribute recognition (b), pedestrian detection (d), human parsing (e) and pose estimation (f), the best performance is achieved at a large λ. Unlike person re-identification, these four tasks have a strong dependence on the semantic information. So we set λ to 0.0-0.2 for person re-identification, 0.8-1.0 for attribute recognition, pedestrian detection, human parsing, and pose estimation. As person search consists of person re-identification and human detection, both of which have strong influence on its final performance. From Fig. 2 (c), we select a moderate value, *i.e.*, 0.4-0.6.

5. SOLIDER Training Process

To further clarify the whole training process of the proposed SOLIDER framework, we provide the pseudo code for training, as shown in Algorithm 1.

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable

Table 1. B&F implies Background and Foreground Clustering, while MIM means Masked Image Modeling.

+ B&F	+ MIM	MSMT17↑	CUHK-SYSU↑	CityPerson↓
		60.9/81.1	93.2/93.6	11.4/43.0
✓		61.4/81.9	93.5/94.0	11.3/42.2
✓	✓	61.6/82.2	93.6/94.1	11.1/41.7

Algorithm 1 SOLIDER Training Process

INPUT: The dataset X; The predefined semantic part number N; The SOLIDER model including a teacher backbone F_t and a student backbone F_s; Max epoch number for the training process Ep_{max}; The Batchsize B.

OUTPUT: The trained SOLIDER Model.

```
for epoch=1 to $Ep_{max}$ do
    for $i$=1 to $B$ do
        Randomly sample a binary $\lambda_i$ from {0,1} for Image $x_i$.
        Send Image $x_i$ and $\lambda = \lambda_i$ to the teacher backbone and obtain the feature maps $F_t(x_i, \lambda_i)$, which would be used in the dino loss;
        # Generating Semantic Labels:
        Send Image $x_i$ and $\lambda = 1$ to the teacher backbone and obtain the feature maps $F_t(x_i, 1)$;
        Cluster all the tokens in $F_t(x_i, 1)$ into two categories based on their magnitudes;
        Mark the category with a larger average magnitude as foreground and get the foreground binary mask $M$;
        Cluster the foreground tokens $F_t(x_i, 1)$[M==1] into $N$ semantic parts;
        Get the label $y$ for every token in $F_t(x_i, 1)$, and the label of background tokens is set to 0;
        # Contrastive Supervision:
        Use the label $y$ to randomly mask out a semantic part in image $x_i$ and obtain the masked image $\tilde{x}_i$;
        Send Image $x_i$ and $\tilde{x}_i$ with $\lambda = \lambda_i$ to the student backbone, and obtain $F_s(x_i, \lambda_i)$ and $F_s(\tilde{x}_i, \lambda_i)$;
        Compute the dino loss $L_{dino}(F_t(x_i, \lambda_i), F_s(x_i, \lambda_i))$;
        # Semantic Supervision:
        Compute the (masked) semantic classification loss $L_{sm}(F_s(x_i, \lambda_i), y)$ and $L_{sm}(F_s(\tilde{x}_i, \lambda_i), y)$;
        Compute the total loss $L = \alpha L_{dino} + \lambda_i(1 - \alpha)L_{sm}$;
    end for
    Update the SOLIDER model with the total loss;
end for
```

vision learners. In *CVPR*, 2022. 1

[2] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 1