Supplemental Materials to " SIM: Semantic-aware Instance Mask Generation for Box-Supervised Instance Segmentation"

Anonymous CVPR submission

Paper ID 532

In this supplemental file, we provide the following materials:

- Details about the online weakly-supervised Copy-Paste operation, and some visualization results (*cf*. Sec3.3-Online Weakly-Supervised Copy-Paste in the main paper);
- Implementation details on Mask2Former [2] (*cf*. Sec4.1-Implementation Details in the main paper);
- Analysis on positive weighting strategy and more visualizations (*cf*. Sec3.2.2-Positive Mask Weighting and Sec4.4-Visualizations of Weights in the main paper);
- More qualitative results (*cf*. Sec4.3-Qualitative Results in the main paper);
- More parameter analyses (*cf*. Sec4.4-Ablation Study in the main paper).

A. Online Weakly-Supervised Copy-Paste

Copy-paste is a simple yet effective way to improve the data efficiency of instance segmentation models. By pasting objects of various categories and scales to different images, copy-paste could achieve solid performance improvements on strong baseline models [3–6]. In addition, Ghiasi *et al*. [6] demonstrated the efficacy of copy-paste under the semi-supervised learning setting. However, Copy-Paste has rarely been explored for weakly-supervised instance segmentation. In this work, we use copy-paste to create new training data for better handling the object occlusions and rare object categories. Following the work [6], we adopt a simple strategy of randomly picking objects and pasting them on the target image, which could provide a considerable boost on top of baselines.

Importance Sampling. The quality of pseudo mask varies significantly across different instances within an image. A poor mask may produce an unconvincing paste, while a good mask could result in a convincing one. Therefore, we adopt an importance sampling strategy to select instances with high-quality masks. Specifically, we calculate the averaged mask score S that reflects the importance of each instance by:

$$S_k = \frac{\sum_i^{N_k} M_{prob}^{k,i} \cdot \mathbb{1}(\hat{M}^{k,i} = 1)}{\sum_i^{N_k} \mathbb{1}(\hat{M}^{k,i} = 1)}, \tag{1}$$

where N_k denotes the number of pixels from the k-th instance map of M_{prob}, and S_k is the score of the k-th instance. We also store S into the memory bank $\mathcal{M}$.

For each training iteration, we randomly sample an image $\{X', Y', B', \hat{M}', S'\}$ from $\mathcal{M}$ and randomly extract a subset of instances according to their mask scores S', so that the instances with higher-quality masks are more likely to be selected.

Examples of online weakly-supervised Copy-Paste are shown in Fig. S1, we cut objects which have significant contrast with the background, and randomly paste them on the target image. In this way, more challenging data with various occlusion patterns can be created, which could effectively improve the model ability to handle the occlusions between objects. In addition, since small objects are the majority (more than 40%) in the COCO dataset, more small objects will be created so that the Copy-Paste strategy will benefit small objects much (please refer to the Tab. 1 in our main paper).

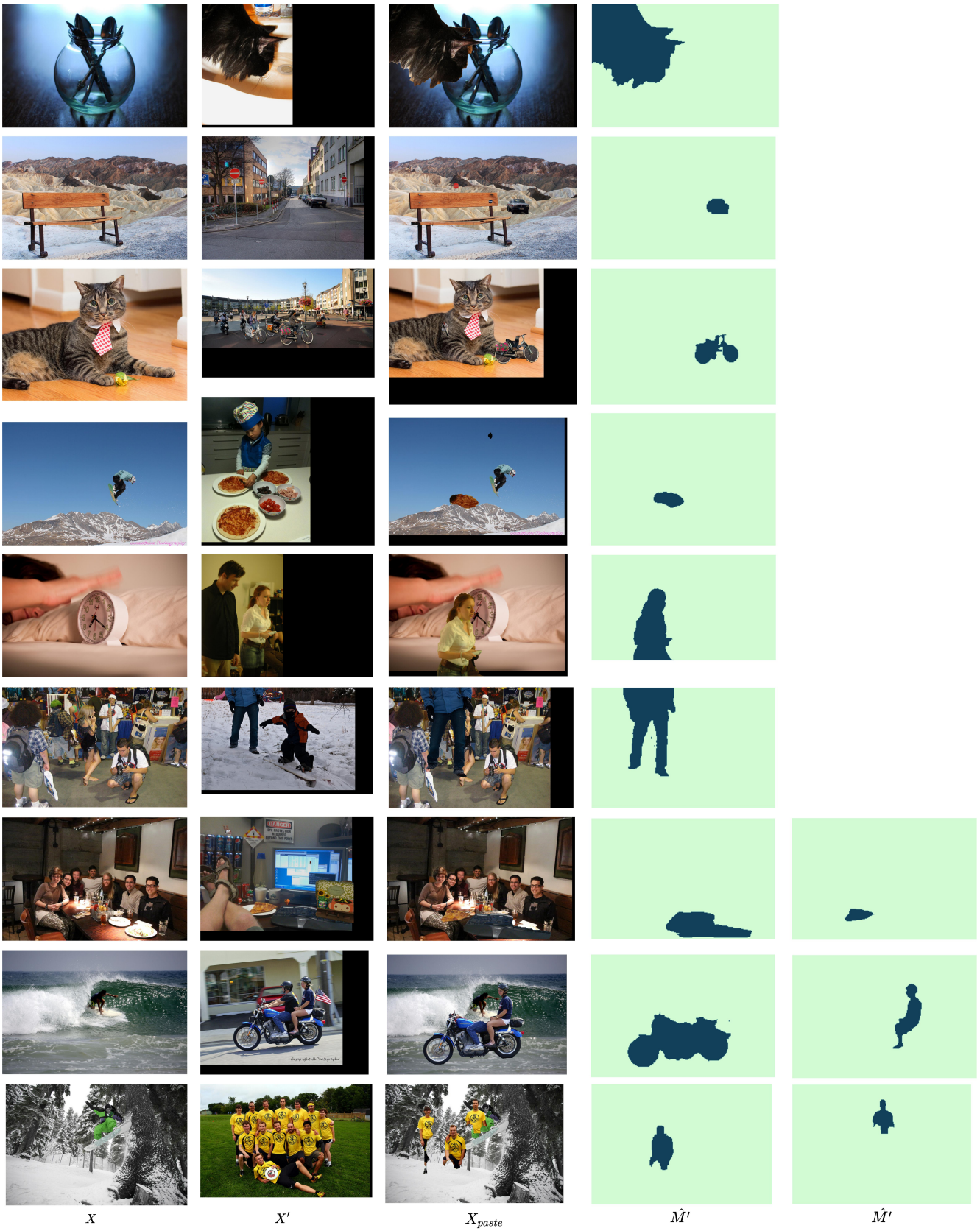

Figure S1. Examples of online weakly-supervised Copy-Paste. We use $\hat{M}'$ to extract instances from X' and paste them onto X, resulting in new training data X_{paste}.

Figure S2. Visualizations of weights for different positive samples.

B. Implementation Details

We adopt CondInst [9] and Mask2Former [2] as our baselines. For Mask2former, we use Detectron2 and follow the commonly used settings for each dataset. More specifically, we use AdamW [7] and the *poly* [1] learning rate schedule with an initial learning rate of 10^{-4}. A learning rate multiplier of 0.1 is applied to CNN backbones. The more advanced multi-scale deformable attention transformer [11] is employed as the default pixel decoder.

C. Analysis on Positive Weights

In Sec. 3.2.2 of our main paper, we use the instance map to rectify the semantic map so that those falsely activated objects and regions will be suppressed while the correct ones are enhanced. Since the quality of masks produced by different positive samples varies significantly, we propose a positive mask weighting strategy to integrate different masks according to their quality. We visualize the weights of different positive samples in Fig. S2. One can observe that large weights tend to locate in the central regions of foreground objects, and the smaller weights tend to locate in ambiguous regions.

Effect of μ. The parameter μ in Eq. 6 of our manuscript controls the relative gaps between different weights. We investigate its effect in Tab. S1. Out method attains the best performance when μ is set to 5.0. When we equally treat different positive samples (*i.e.*, set μ to 0), the performance decreases by 0.8% AP. When we set μ to a larger value, *e.g.*, $\mu = 20$, the final result drops by 0.9% AP. This is because we rely on some certain positive samples while ignoring other useful ones. Actually, different positive samples provide complementary information and they should be fully integrated for producing better results.

μ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0	31.4	52.9	31.6	14.3	33.7	46.9
1	31.6	53.4	32.0	14.5	34.1	47.2
5	32.2	54.0	33.0	15.8	34.5	48.3
10	31.8	53.9	32.3	15.0	34.8	46.8
20	31.3	53.1	31.9	15.1	33.9	46.1

Table S1. Effect of parameter μ.

τ_{low}	τ_{high}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.5	0.5	31.4	53.6	31.9	14.7	34.3	46.8
0.4	0.6	31.9	53.9	32.3	15.4	34.8	47.7
0.3	0.7	32.2	54.0	33.0	15.8	34.5	48.3
0.2	0.8	31.8	53.6	32.8	15.5	34.1	47.2

Table S2. Effect of thresholds τ_{low} and τ_{high}.

temperature(τ)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.1	30.9	52.5	30.9	14.5	33.0	45.0
0.5	31.7	53.3	32.2	15.5	34.1	46.4
1.0	32.2	54.0	33.0	15.8	34.5	48.3
2.0	31.9	54.1	32.1	15.6	34.0	47.7
5.0	31.6	53.8	32.5	15.6	34.2	47.5

Table S3. Effect of temperature τ.

λ_1	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.1	31.9	53.7	32.8	15.4	34.4	47.2
0.3	32.1	54.0	33.1	15.9	34.2	47.9
0.5	32.2	54.0	33.0	15.8	34.5	48.3
0.7	31.4	53.6	32.2	15.2	33.9	46.9

Table S4. Effect of parameter λ_1.

D. Qualitative Results

Fig. S3 shows some qualitative segmentation results of our method and BoxInst [10] on COCO *val* split. On the one hand, our method could better segment foreground instances that heavily tangle with the background or other objects with similar appearances. On the other hand, our method is good at separating overlapping objects of the same semantics and can keep the integrity of objects.

Fig. S4 gives more results of SIM with the ResNet-101-FPN backbone and $3\times$ training schedule. One can see that our method achieves precise predictions around the objects' boundaries.

E. Parameter analysis

Tab. S2 shows the effects of the two thresholds τ_{low} and τ_{high}. When we set $\tau_{low} = \tau_{high} = 0.5$, all pixels provide supervision, which inevitably introduces much noise. When we set $\tau_{low} = 0.2$ and $\tau_{high} = 0.8$, many pixels are neglected and hence limited supervision is provided and the performance is slightly degraded.

Tab. S3 shows the effect of temperature τ. We employ the instance map M_{I} as a weight map to online rectify the semantic map M_{S}, where the temperature τ controls the modulation intensity. When $\tau \to 0$, the modulation intensity increases so that the final pseudo mask $\hat{M}$ will rely more on M_{S}. When $\tau \to \infty$, the modulation intensity decreases so that the final pseudo mask $\hat{M}$ will rely more on M_{I}.

Tab. S4 shows the segmentation results by using different weights λ_1 in pseudo loss. One can see that our method is insensitive to this parameter when $0.1 < \lambda_1 < 0.5$.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1, 3

[3] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 1

[4] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. 1

[5] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–691, 2019. 1

[6] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[8] Tal Remez, Jonathan Huang, and Matthew Brown. Learning to segment via cut-and-paste. In *Proceedings of the European conference on computer vision (ECCV)*, pages 37–52, 2018.

[9] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020. 3

[10] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 4, 6

[11] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 3

Figure S3. Qualitative results of BoxInst [10] (in the red box) and our method (in the blue box) on COCO `val2017`.

Figure S4. Qualitative results of our method on COCO `val2017`.