PaintSeg: Training-free Segmentation via Painting

Supplementary Materials

A More Comparison with Mask-RCNN

Mask RCNN PaintSeg Mask RCNN PaintSeg

Figure A: **Comparison with Mask RCNN with objects beyond 80 COCO categories.**

We present more results compared with supervised Mask RCNN [5]. As shown in Fig. A, we compare box-prompted segmentation with Mask RCNN on objects beyond 80 COCO categories. In the shown examples, we observe that Mask RCNN has difficulty segmenting the correct shape of the object. Instead, PaintSeg provides more accurate object segmentation. As Mask RCNN is only trained on 80 COCO object categories, there is still a substantial gap between the seen and the unseen. In contrast, PaintSeg is a solution that does not require training, which makes it more general and capable of handling new object categories.

B More Ablation Experiments

In this section, we provide additional ablation studies to illustrate the design choices of PaintSeg.

N	1	2	3	4	5	6
IoU	78.8	79.2	79.6	80.1	80.6	80.8

Table A: **Ablation study on the painted image number N for each step**.

B.1 Sampling Number for Each Step

We average N painted images in each step to obtain the final mask prediction due to the randomness of the generative painting model. We present an ablation study to illustrate the impact of the number of painted images in each step. As shown in Table A, we report the performance on the ECSSD [7] dataset with coarse mask prompt from TokenCut [9]. We notice that the performance gradually improved with more painted images averaged in each step. As there is no significant difference in performance between five or six painted images used, we set the number of painted images to five in the PaintSeg process.

B.2 Image Projector

We conduct an ablation study on image projector $\mathcal{E}$ as illustrated in Table B. We compare the widely used DINO [3] VIT-S/8 and the latest DINO [6] VIS-S/14. The results demonstrate that DINO with a

Preprint. Under review.

DINO [3] VIT-S/8	DINO-V2 [6] VIS-S/14
80.6	80.0

Table B: **Ablation study on image projector $\mathcal{E}$ used in AMCP**.

small patch size achieves better performance. It follows that we consider a smaller patch size since PaintSeg requires fine-grained visual information. A larger patch size will blur the object boundary, resulting in a performance drop.

C More Potential Application

In this section, we discuss more potential applications of PaintSeg beyond prompt-guided object segmentation.

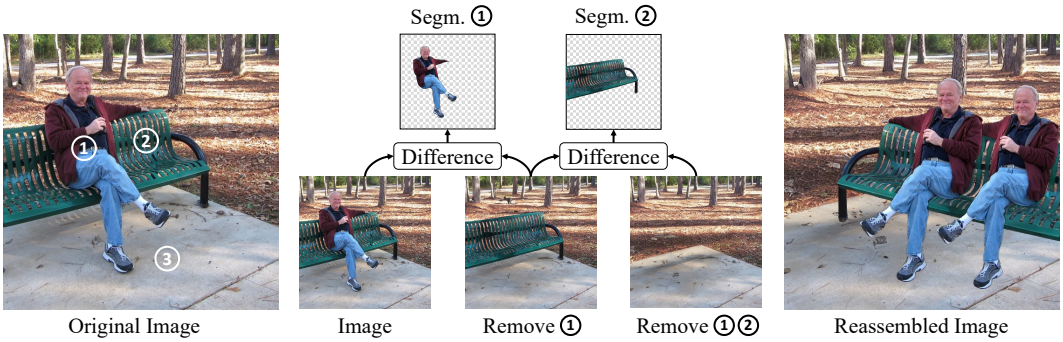

Figure B: **Potential application in image edition and amodal segmentation.** PaintSeg can step-by-step remove objects in the image by using the painted image in I-step. With the segmented object and painted image without objects, we can freely assemble them into a new image. Further, PaintSeg supports amodal segmentation, with the painting capability enabling segmentation of the occluded areas.

C.1 Image Edition

In the I-step of AMCP, the painted image will remove the target object while keeping all other contents in the image. In this way, with the segmented objects and an image without target objects, we can reassemble them into a new image as shown in Fig. B.

C.2 Amodal Segmentation

As shown in Fig. B, PaintSeg can layer-by-layer segment objects. By using the painted image in I-step as the input to the next iteration, PaintSeg can attach the amodal capability. We notice that the bench is occluded by the men in Fig. B. With the PaintSeg, the full shape of the bench can be segmented.

D More Discussion about PaintSeg

In PaintSeg, we introduce a latent variable I_{paint} which is characterized by an off-the-shelf generative model $p(I_{paint}|I \circ M)$ conditioned on an image I and a mask M. $\circ$ represents Hadamard product. In our method, we leverage the AMCP process to estimate and convert the latent variable I_{paint} into mask prediction M with alternating I-step and O-step. Mathematically, both I-step and O-step can be formulated as an expectation-maximization-like process.

- **Expectation**: We introduce a latent variable I_{paint} in the proposed PaintSeg which is modeled by an off-the-shelf generative painting model $p(I_{paint}|I \circ M)$. We assume the generative model will pick the most likely outcome I_{paint} given I and M for every step.

- **Maximization**: After obtaining the latent variable I_{paint}, we define a contrastive potential Φ and utilize clustering to binarize the mask. Mathematically, the contrasting and clustering processes maximize a posteriori probability $p(M|I_{paint}, I) = e^{-\frac{1}{\|M\|_0}\Phi(I_{paint}, I, M)}$.

Although we term I-step and O-step separately, they can be formulated as the same EM process. PaintSeg advances the predicted mask to the ground truth by iteratively conducting the EM process in each step.

E Difference with Previous Segmentation Approaches

In this section, we discuss the major differences between the proposed PaintSeg and previous object segmentation methods as follows.

Discriminative *v.s.* Generative + Discriminative . Conventional object segmentation is a discriminative task that leverages a neural network θ to model the conditional probability of the object mask M given the image I as condition $p_\theta(M|I)$. In PaintSeg, we have mask, paint, and contrast operations in each step. Specifically, in paint operation, we enroll a generative model to estimate painted image I_{paint} with mask M and image I as conditions. After that, the mask can be obtained by comparing the generated image with the original one with a contrastive potential Φ. As discussed in Section D, the paint operation is a generative process to estimate latent variable $p(I_{paint}|I \circ M)$ and the contrast operation is a discriminative process to obtain a mask prediction based on $p(M|I_{paint}, I)$. PaintSeg achieves training-free by constructing a bridge to generative painting models which permits object shape consistency and background content consistency.

Pixel *v.s.* Pixel difference. Conventional object segmentation leverages a network to project an image to the feature space and then binarize (cluster) each pixel into foreground or background classes. Differently, instead of directly clustering over the input image, PaintSeg utilizes the difference between the painted and original image, as a proxy, to leverage the object shape prior and background consistency. The contrastive scheme is rooted in the decomposable nature of images and paves a way to incorporate generated images to segment objects.

Training *v.s.* Training-free. Conventional object segmentation approaches train the neural network to segment objects requiring time-consuming and expensive data labeling. Some unsupervised segmentation methods [1, 4, 2, 8] find a segment from a generative model while they typically require training a network on top of the generative model. Instead, our method is a training-free unsupervised method that learns to segment objects from a generative painting model. We consider the PaintSeg provides a way to bridge the generative model and segmentation which may inspire future research.

F Failure Case Analysis

We analyze the failure case here. As shown in Fig. C, we visualize a failure case when using a point as the prompt. We notice the adjacent car is segmented as a false positive, which is due to the semantic and visual similarity between the target and false positive cars. Despite our method is capable of handling multiple objects with point prompt (right of Fig. C), crowded scenarios can make it difficult to segment the accurate object boundary. However, the issue can be overcome through box prompt.

Figure C: Illustration of failure case.

G More Visualization

In this section, we demonstrate more visualization of PaintSeg. We show more qualitative results with box prompt in Fig. D, with point prompt in Fig. E and with coarse mask prompt in Figs. F and G.

Figure D: **More visualization of PaintSeg with box prompt on COCO MVal.**

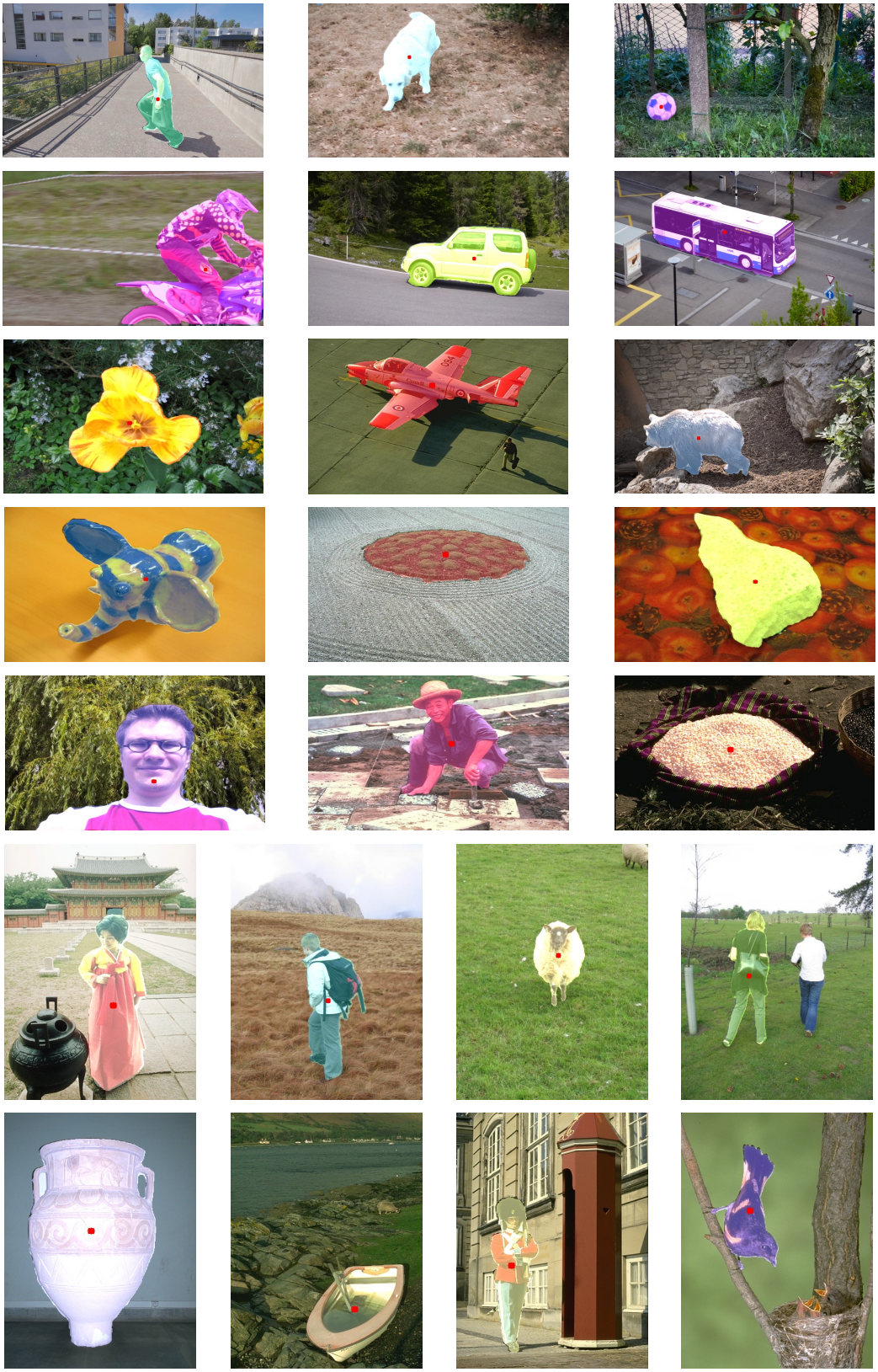

Figure E: **More visualization of PaintSeg with point prompt. The point prompt is illustrated by the red point on the image on DAVIS and Berkeley and GrabCut.**

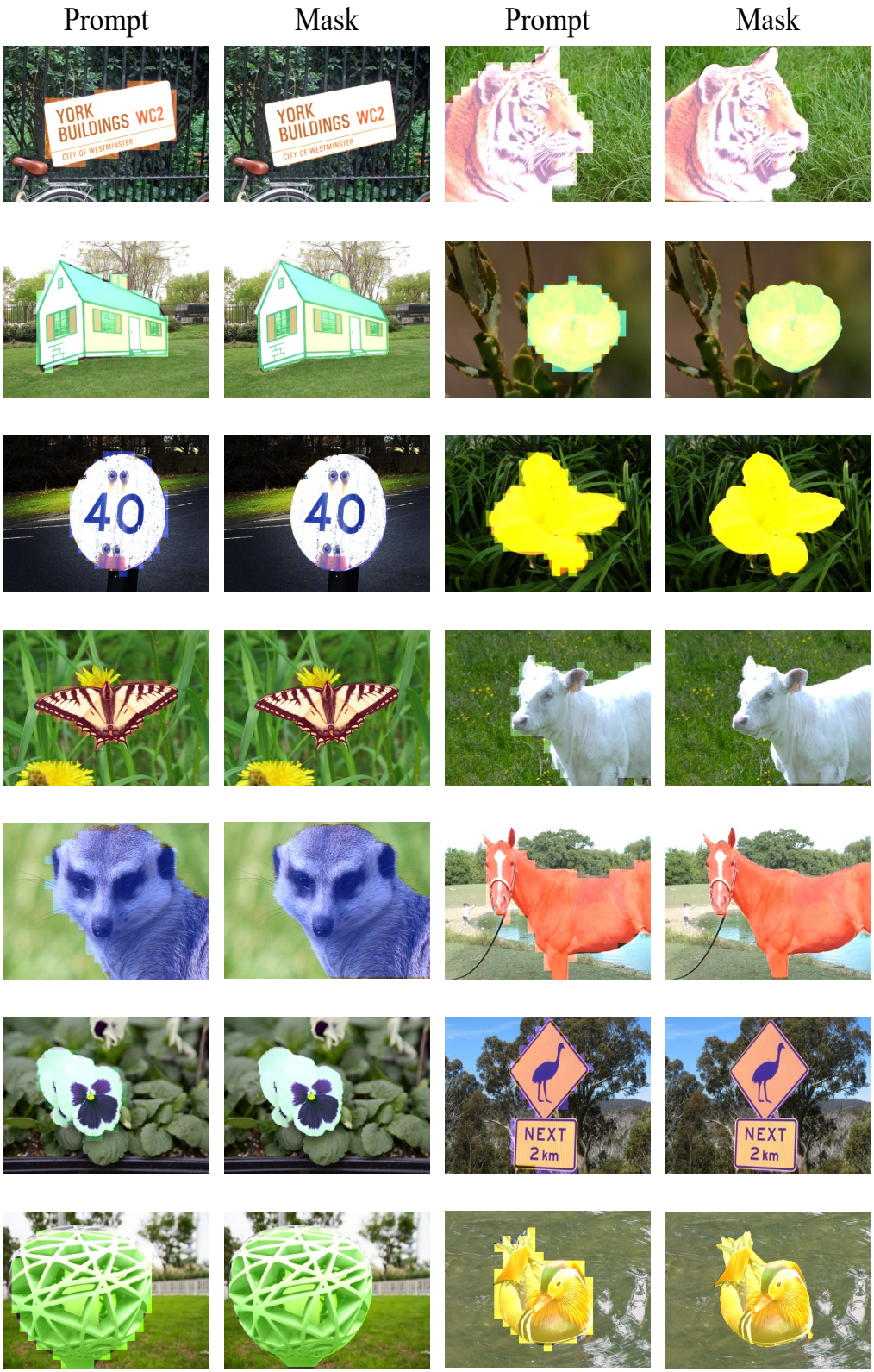

Figure F: **More visualization of PaintSeg with coarse mask prompt on ECSSD.**

Figure G: **More visualization of PaintSeg with coarse mask prompt on ECSSD.**

References

[1] Yaniv Benny and Lior Wolf. Onegan: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 514–530. Springer, 2020.

[2] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[4] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. *Advances in neural information processing systems*, 32, 2019.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[7] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended CSSD. *IEEE TPAMI*, 2016.

[8] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. 2021.

[9] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *CVPR*, 2022.