v-CLR: View-Consistent Learning for Open-World Instance Segmentation Supplementary Material

Chang-Bin Zhang[1] Jinhong Ni[1] Yujie Zhong[2] Kai Han[1*]
[1]Visual AI Lab, The University of Hong Kong [2]Meituan Inc.
{cbzhang, jhni}@connect.hku.hk jaszhong@hotmail.com kaihanx@hku.hk

1. Experimental Details

1.1. Auxiliary Views

Our learning framework leverages multiple transformed views of the original natural image. Specifically, we apply off-the-shelf models to transform natural images into art-stylized and colorized depth images. For the art-stylized transformation, we utilize the pre-trained StyleFormer [16] model, which is trained on the WikiArt [11] dataset. For each natural image, we randomly select a target style from the WikiArt [11] dataset. For the colorized depth transformation, we employ the off-the-shelf ZoeDepth [1] model, pre-trained on the NYU Depth v2 [12] and KITTI [5] datasets. Additionally, for edge maps used in our ablation study, we apply the off-the-shelf RCF [9] model for edge detection. Examples of natural, art-stylized, and colorized depth images are shown in Fig. 1. Notably, no human annotations are used for generating depth maps or stylized images, ensuring that our method avoids any information leakage.

1.2. Experiments on the CLEVR Dataset

The CLEVR dataset [6] is a synthetic dataset featuring objects characterized by four attributes:

- Size: large, small
- Shape: cube, sphere, cylinder
- Color: gray, red, blue, green, brown, purple, cyan, yellow
- Material: rubber, metal

In this work, we focus on two attributes—color and material—for illustrative simplicity. Specifically, we designate *red metal* objects as the known class, while objects with any other attribute combination are treated as unknown classes. The CLEVR dataset [6] comprises 70,000 training images and 15,000 validation images. We apply vanilla DINO-DETR [18] and train the model under two settings: with and without colorized depth images and stylized images. When using colorized depth images, the model randomly selects either a natural image, a depth map or a stylized image as input, each with equal probability. With 300 denoising queries in DINO-DETR [18], we train the model for 2,000 iterations with a batch size of 8, while retaining the remaining training configurations identical to those of vanilla DINO-DETR [18].

*Corresponding author.

Algorithm 1 Pseudo-code of Parameter Perturbation in a PyTorch-like style

```
# image: input image tensors
# model: the detector
# noise_std: the standard deviation of gaussian noise
def perturbation_forward(image, model, noise_std):
   # adding gaussian noise for each parameter
   for name, param in model.named_parameters():
      param += torch.randn_like(param) * noise_std
   output = model(image)
   return output
```

1.3. Object Proposal Generation

Thanks to large-scale self-supervised learning, neural networks have shown remarkable capabilities in object recognition and localization [10, 15]. Leveraging this advancement, unsupervised instance segmentation [13, 14] has recently achieved significant progress. By benefiting from unsupervised training, these methods exhibit strong instance awareness, making them well-suited for generating object proposals in our work. Throughout this paper, we employ the ImageNet-pretrained Cascade R-CNN [2] from CutLER [14] to infer object proposals from the dataset. For each training image, we apply Non-Maximum Suppression (NMS) with a threshold of 0.7 and select the top-10 proposals based on prediction confidence.

2. Robustness against Parameter Perturbation

Numerous studies [3, 7, 17, 19] have demonstrated that neural networks trained with flatten minima exhibit superior generalization ability, *i.e.,* , the minima of the model should be in wide valleys rather than narrow crevices [3, 7, 17, 19]. In such cases, small perturbations to model parameters should not significantly degrade the performance of a model with strong generalization ability. Consequently, we can assess a model's generalization by introducing random perturbations to its parameters. Specifically, as detailed in Alg. 1, we inject

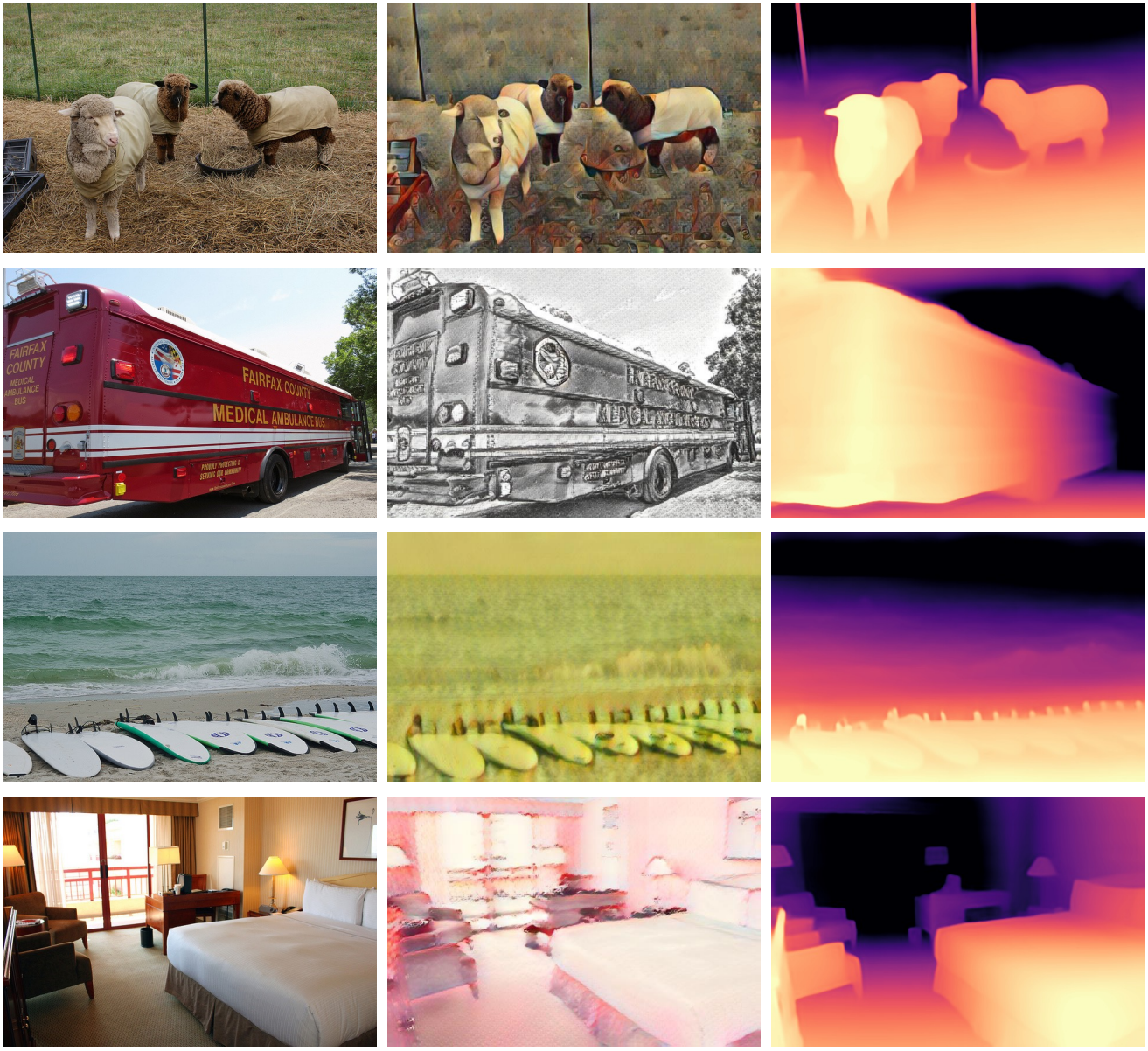


Figure 1. **Visualization of three views** used in our method, natural, art-stylized, and colorized depth images, respectively.

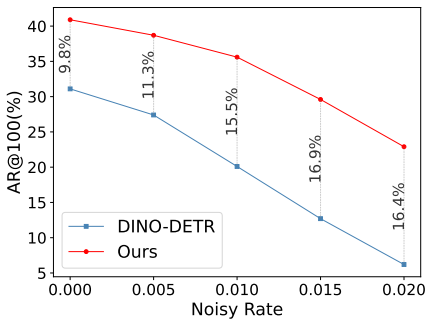


Figure 2. **AR^b_{100} under different noisy rates.** All models are evaluated in the VOC→Non-VOC setting.

Gaussian noise with varying standard deviations into all network parameters and evaluate the resulting performance. All models are trained on VOC classes and evaluated on Non-VOC classes. As illustrated in Fig. 2, we define the noise rate as the standard deviation of the Gaussian noise. With increasing noise rates, both our model and DINO-DETR experience performance degradation. However, at high noise rates, our method consistently outperforms the baseline by a substantial margin, demonstrating greater robustness to parameter perturbations.

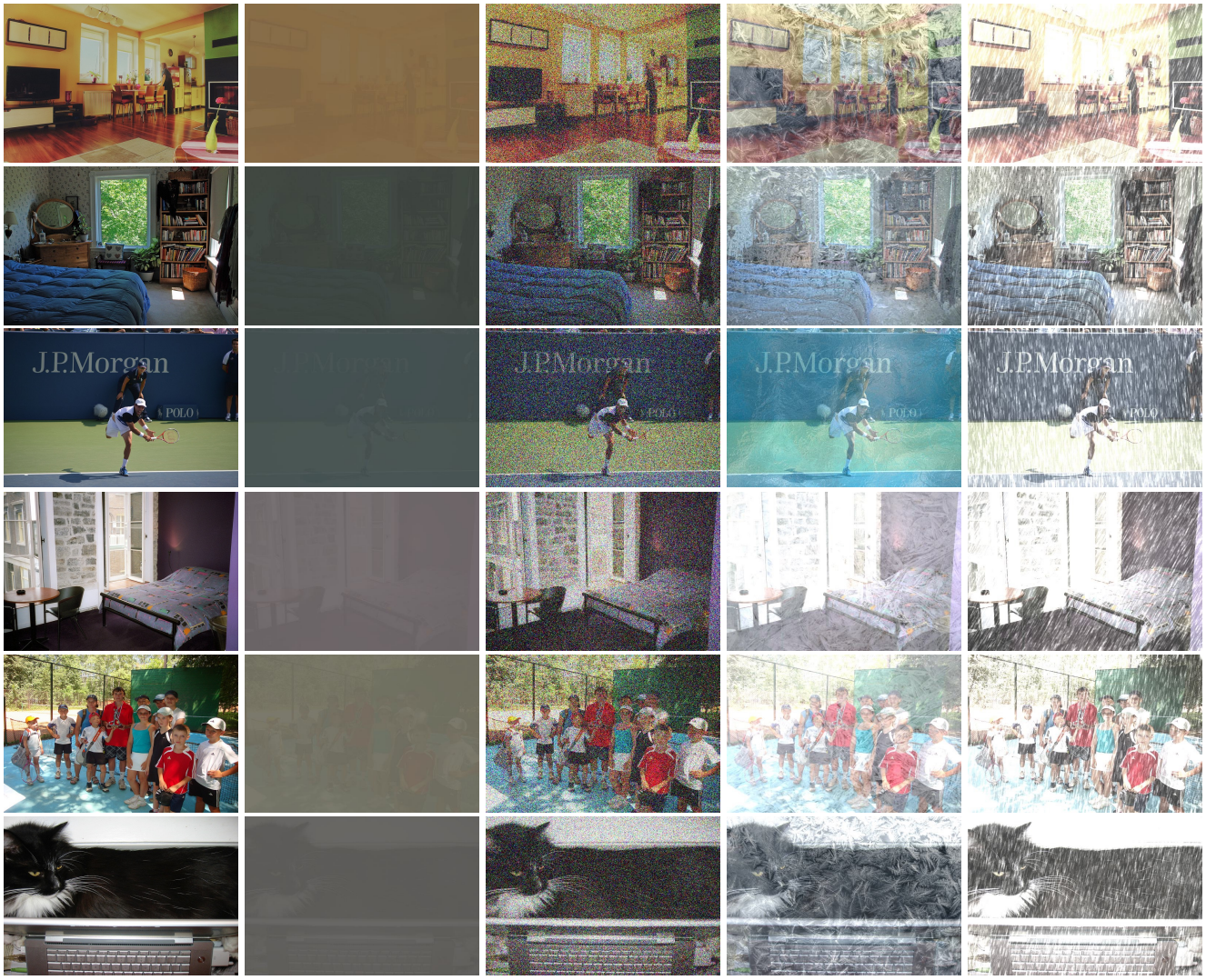
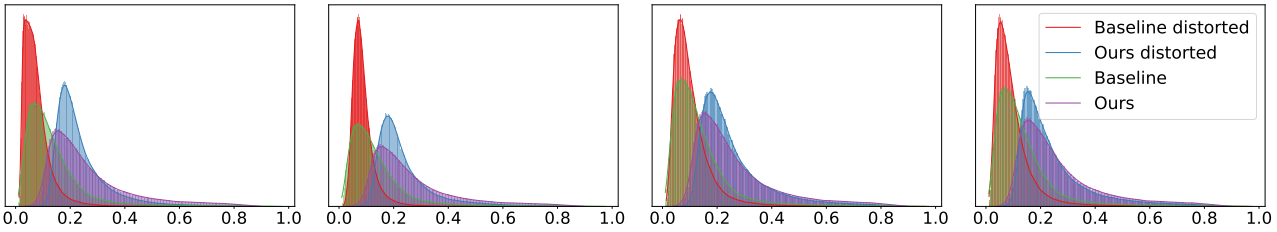


Figure 3. **Examples of distorted images on COCO 2017 [8] validation set.**



(a) Contrast (b) Gaussian Noise (c) Snow (d) Frost

Figure 4. **The distribution of prediction scores from the baseline DINO-DETR [18] and our *v*-CLR under four types of image distortion.** For visualization clarity, we calculate the distribution of *top*-50 prediction scores.

3. Robustness against Image Distortion

To verify the effectiveness of our method under different input perturbations, we evaluate our model under four popular distortions, Contrast, Gaussian Noise, Snow, and Frost. As shown in Fig. 3, we generate validation images with these distortions and evaluate the model's performance on them. We examine the robustness of our *v*-CLR approach against different types of image distortions. In Fig. 4, we plot the distribution of prediction scores for both the baseline

Method	AR^b_1	AR^b_{10}	AR^b_{100}
SiameseDETR [4]	12.4	23.0	30.7
v-CLR (**ours**)	**16.8**	**42.7**	**60.2**

Table 1. **Comparison with Siamese DETR [4].** For a fair comparison, all experiments are conducted on the Non-VOC→VOC setting with Deformable-DETR [20].

DINO-DETR [18] and our method, with and without image distortions. Our method (purple) consistently yields higher prediction scores than the baseline (green) on undistorted images. For distorted images, the distribution of the distorted baseline (red) exhibits a heavier right tail compared to the undistorted baseline (green), indicating that distortions reduce DINO-DETR's prediction confidence. In contrast, our method demonstrates greater robustness to image distortions, as the distributions of prediction scores for distorted and undistorted images show similar right-tail behavior. Surprisingly, the prediction score distribution for our method on distorted images exhibits even lower variance and a slightly higher mean than on undistorted images. This further suggests that image distortions have minimal impact on our model's prediction confidence.

4. Comparison with SiameseDETR

We further compare our method with Siamese DETR [4], a recent self-supervised DETR-like object detector. Siamese DETR employs two augmented views to enforce instance-level consistency. Although it also utilizes transformations, its motivation differs substantially from ours, and the transformations in Siamese DETR do not specifically address texture bias. We evaluate both methods in the Non-VOC→VOC setting, as shown in Tab. 1, ensuring a fair comparison since VOC classes are unknown to both models. Experimental results reveal that our method surpasses Siamese DETR by a significant margin across all evaluation metrics, underscoring the effectiveness of our proposed framework.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1

[3] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019. 1

[4] Ze-Sen Chen, Gengshi Huang, Wei Li, Jianing Teng, Kun Wang, Jing Shao, Chen Change Loy, and Lu Sheng. Siamese detr. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 4

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 2013. 1

[6] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[7] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 3

[9] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[10] 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. 1

[11] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 1

[12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. Comput. Vis.*, 2012. 1

[13] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[14] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1

[15] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 1

[16] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Int. Conf. Comput. Vis.*, 2021. 1

[17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021. 1

[18] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1, 3, 4

[19] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1

[20] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4