GeometryPaste: Geometry-Based Copy-Paste Data Augmentation for Instance Segmentation

Nicholas Dunn
Department of Mathematics and Computer Science
Columbus State University
Columbus, GA
dunn_nicholas2@students.columbusstate.edu

Anurag Ghosh
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA
anuraggh@andrew.cmu.edu

Christoph Mertz
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA
cmertz@andrew.cmu.edu

***Abstract*—Instance segmentation models require large datasets of annotated images to achieve adequate performance. However, annotated datasets are difficult to build or obtain. There are several large-scale datasets for common objects, but few exist for rare objects. Detecting rare objects has many practical applications, such as autonomous vehicles detecting roadwork objects. Copy-Paste is a data augmentation method for generating images and has been utilized successfully to improve instance segmentation performance. Prior works have studied both random Copy-Paste, where objects are randomly pasted onto images, and pasting objects based on the surrounding visual context. In this paper, we develop GeometryPaste, a method of pasting objects according to the geometry and context of the objects and background images. We build a small dataset of roadwork objects and fine-tune a pre-trained instance segmentation model to evaluate our method. Our results are compared against both baseline and random Copy-Paste APs. The results suggest that GeometryPaste may provide performance improvements over both baseline and random Copy-Paste augmentation for instance segmentation of rare object categories in small datasets.**

I. Introduction

Image segmentation is a problem in computer vision involved with grouping pixels according to different characteristics. Three primary pixel groupings studied are semantic segmentation, instance segmentation, and panoptic segmentation [1]. Semantic segmentation [2] is the task of assigning a class label to each pixel in an image. In contrast, instance segmentation [3] attempts to detect every instance of an object and segment all pixels belonging to each instance. Panoptic segmentation [4] combines semantic and instance segmentation to produce both per-pixel class labels and per-object segments. This paper focuses on instance segmentation of roadwork objects.

Detecting rare object categories has practical applications in many fields, including autonomous vehicles and robots, medical imaging, and manufacturing. Roadwork objects are particularly important for self-driving cars due to the dynamics of the situation. The presence of roadwork objects creates a complex environment in which the typical driving rules no longer apply. For example, the car may need to drive in opposite lanes or disobey speed limits or traffic signs. This situation presents a very challenging navigational problem for self-driving cars, and detecting the presence of roadwork objects is one of the first steps toward mitigating this issue.

Fig. 1: Example image of roadwork objects from the RoadBotics dataset.

To achieve instance segmentation performance capabilities good enough for real-world applications, machine learning models are typically pre-trained on a large, annotated dataset of common objects and then fine-tuned on a dataset representative of the context in which the model will be deployed and containing the categories relevant for detection. Several large-scale datasets exist for common objects, including COCO [5], Mapillary Vistas [6], Cityscapes [7], and ADE20K [8]. However, building large-scale, annotated datasets is a very costly and time-consuming process. For example, over 70k worker hours were utilized to annotate the COCO dataset [5]. Due to these difficulties, few large-scale, annotated datasets exist for rare object categories. To mitigate this issue, methods of creating new images by augmenting existing datasets have been explored.

Copy-Paste [9]–[12] is a data augmentation method that has been utilized successfully to improve instance segmentation performance. Prior works have studied Copy-Paste in a variety of settings and contexts, including Simple Copy-Paste [9], where objects are randomly pasted onto images, pasting objects based on the surrounding visual context [10], and pasting objects in different locations within the same image [11]. In this paper, we develop a method of pasting objects according to the geometry and context of the objects and background images. Because the objects of interest for this work are roadwork objects, we ensure they are pasted on or beside the road and scaled based on known camera parameters

and vanishing point estimation.

The dataset utilized for our study is custom-built from images from different cities from the RoadBotics dataset. A sample image from the dataset is shown in Fig. 1. We use the dataset both with and without augmentation to fine-tune a Mask2Former [13] instance segmentation model with a ResNet-50 backbone [14] pre-trained on COCO to evaluate our method. The metrics for evaluation are AP scores, which are compared against baseline and random Copy-Paste APs. Our results suggest that a geometry-based, context-aware Copy-Paste data augmentation strategy may outperform other methods.

II. Related Work

A. Instance Segmentation

Instance segmentation [3] is a well-known and challenging task in computer vision. The goal of instance segmentation is to detect every instance of an object in an image and segment all the pixels belonging to each instance. Several machine learning architectures, such as Faster R-CNN [15] and Mask R-CNN [16], have been applied to this task. A recent model, Mask2Former [13], utilizes the Transformer [17] architecture to create a unified image segmentation architecture capable of performing semantic, instance, and panoptic segmentation. Several datasets have also been created to assist with instance segmentation, including COCO [5], Mapillary Vistas [6], Cityscapes [7], and ADE20K [8]. Additionally, many works utilize a model that is pre-trained on these datasets to evaluate their method. This paper employs a Mask2Former model with a ResNet-50 backbone pre-trained on COCO and fine-tuned on a custom-built dataset to evaluate our method.

B. Copy-Paste Data Augmentation

Because instance segmentation models require large amounts of annotated data to achieve practical performance capabilities, much effort has gone into augmenting existing datasets by generating new training data using various methods, including Copy-Paste. Copy-Paste is a well-known data augmentation strategy that has been shown to be effective at improving instance segmentation performance. The basic concept of Copy-Paste is to generate new training images by copying relevant objects from images and pasting them into new backgrounds. Prior works have utilized a variety of different methods to perform Copy-Paste.

In Cut, Paste and Learn [12], objects and background images without annotations are collected. Both relevant objects and distractor objects are pasted onto randomly chosen background images. The objects are placed at random locations, and a variety of blending techniques are employed to ensure their object detector model ignores local pixel artifacts to prevent performance degradation. The objects undergo 2D and 3D rotations, random scaling, occlusion, and truncation. In contrast, our method chooses a context-based location, performs geometry-based scaling, and has no rotations or occlusions for pasted objects.

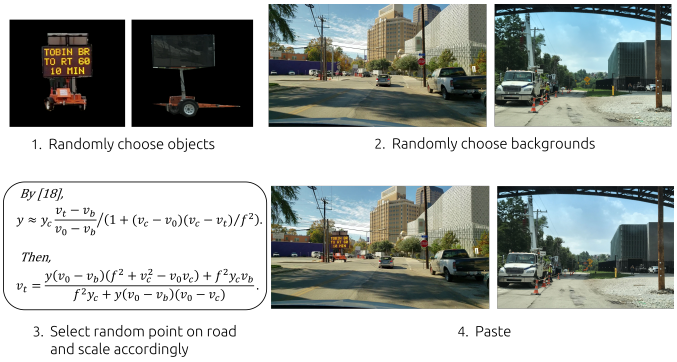

Fig. 2: Overview of GeometryPaste. Objects are copied from their original images and pasted onto new backgrounds using the geometry and context of the objects and background images.

Simple Copy-Paste [9] is another strategy that pastes objects onto a random location. In [9], pairs of images are randomly selected and subjected to random scale jittering and horizontal flipping. A subset of objects from one image is then randomly selected and pasted into the other. Existing annotations are updated, with fully occluded objects removed. Like [9], we utilize background images with existing objects and update their annotations accordingly after pasting. However, we choose a context-based location for pasting and scale the objects being pasted rather than the images.

As with this work, contextual Copy-Paste [10] and InstaBoost [11] choose pasting locations based on the context of the object and background image. However, different methods are used for determining context. In [10], a context model is trained to predict the appropriate context, and [11] moves the object to a nearby location within the same image rather than pasting it into a new image. Our approach maintains context by pasting onto a randomly chosen point on the existing road segmentation in the background image.

III. Methods

In this paper, we propose a method of generating realistic training images with Copy-Paste data augmentation to increase instance segmentation performance on rare object categories. We copy objects from their original images and paste them onto new background images using both the geometry and context of the objects and background images. First, we randomly choose an object and a background image. Then we choose a random point on the existing road segmentation in the background image to paste the object onto. Next, the object is scaled to the appropriate size based on the location of the chosen point. The object is then pasted onto the new image, allowing for partial truncation. Finally, existing object annotations are updated to account for occlusions from the pasted object. Fig. 2 provides a system overview.

A. Object Selection

Although multiple objects can be used with our method, we focus on objects that are underrepresented and difficult for our model to detect based on AP scores. Specifically, we

choose only the TTC Message Board from our dataset to paste into new background images. Because TTC Message Boards have many sizes, configurations, and messages, we select 27 TTC Message Boards with high-quality annotations from our training dataset to ensure diversity in the generated images. Despite being the highest quality, some quality issues exist with the selected TTC Message Boards, including small occlusions from other objects and missing parts due to annotation errors.

B. Background Selection

Background images are chosen from our training dataset and contain existing objects. Because most images contain at most one TTC Message Board, we select background images without TTC Message Boards. Again, due to annotation errors, many images do not contain road segmentations. Since the objects will be pasted onto points selected from the road segmentation, we ensure the chosen images contain a road segmentation. Another constraint is that the image must contain its predicted vanishing point. This constraint is imposed to assist with visual analysis of the quality of the vanishing point prediction, which is used in our scaling function. In total, there are 1,640 candidate background images.

C. Pasting, Truncation, and Occlusions

To maintain appropriate context, objects are pasted onto randomly chosen points from the road segmentation of background images. Truncation occurs when the object is pasted such that the image boundary partially occludes it. Our policy for truncation is the same as in [12]. That is, we ensure at least 25% of the object's bounding box remains in the image. If, after scaling the object, the chosen location results in a truncation of more than 75%, a new location is selected. In addition to truncation, occlusions of existing objects may occur after pasting the object into the new image. Existing object annotations are updated accordingly to account for occlusions for objects that remain at least 25% visible and removed otherwise.

D. Object Scaling

Having the camera parameters for our dataset and ensuring the vanishing line is within the image allows us to scale the object to the appropriate size. Let y be the object height, f the camera focal length, v_c the camera optical center y-coordinate, y_c the camera height, v_0 the y-coordinate of the horizon line, and v_t and v_b the object top and bottom coordinates, respectively. Then by Equation (5) in [18], we have

$$y \approx y_c \frac{v_t - v_b}{v_0 - v_b} / \left(1 + (v_c - v_0)(v_c - v_t)/f^2\right).$$

Therefore,

$$v_t = \frac{y(v_0 - v_b)(f^2 + v_c^2 - v_0 v_c) + f^2 y_c v_b}{f^2 y_c + y(v_0 - v_b)(v_0 - v_c)}.$$

Letting v_b be the y-coordinate of the new location, the new height becomes $v_t - v_b$, and the width is scaled accordingly to maintain the aspect ratio. The y-coordinate of the horizon line v_0 is predicted using NeurVPS [19], a deep neural network vanishing point detector. The object's height y in its original image is obtained using its ground-truth annotation, known camera parameters, and predicted vanishing point with Equation (5) from [18].

TABLE I: Distribution of objects in the baseline dataset.

Category	Train (3435)	Val (492)	Test (981)
Cone	7604	1203	2224
Fence	880	139	411
Drum	1189	150	1141
Barricade	1436	185	566
Barrier	1275	212	401
Work Vehicle	3106	457	870
Vertical Panel	5175	879	1912
Tabular Marker	4269	567	1562
Arrow Board	212	29	105
TTC Message Board	90	14	33
Other Roadwork Objects	191	20	63
Guide Sign	420	59	59
Road	2030	276	697
TTC Sign	2500	320	681
Work Equipment	280	44	41

E. Blending

We apply either no blending or Gaussian blurring to pasted objects. When Gaussian blurring is applied, each image is generated twice: once with no blending and once with Gaussian blurring. The background image remains the same, and the object maintains the same position and scale, with the only difference being the blending strategy.

IV. Experiments

A. Dataset

We fine-tune our pre-trained model on a specialized dataset of images containing roadwork objects from different cities from the RoadBotics dataset. However, many images in our dataset contain missing or low-quality annotations, and there are also inconsistent labeling issues. In constructing our dataset, irrelevant categories and images without annotations are removed, yielding 15 categories of roadwork objects and 4,908 images, from which we create training/validation/testing splits of sizes 70%/10%/20%. Table I shows the distribution of objects in the dataset.

B. Augmented Datasets

We build several augmented datasets according to different strategies for this study: random Copy-Paste, Geometry-Paste, and GeometryPaste blended. In the random Copy-Paste dataset, objects are pasted onto randomly chosen locations and randomly scaled to 0.1 - 2.0 times their original size. The GeometryPaste and GeometryPaste blended datasets are constructed using our method, with the only difference being that GeometryPaste blended contains both blended and non-blended images, whereas GeometryPaste contains only non-blended images. We utilize 27 TTC Message Boards 64 times each, generating 1,728 new training images for the

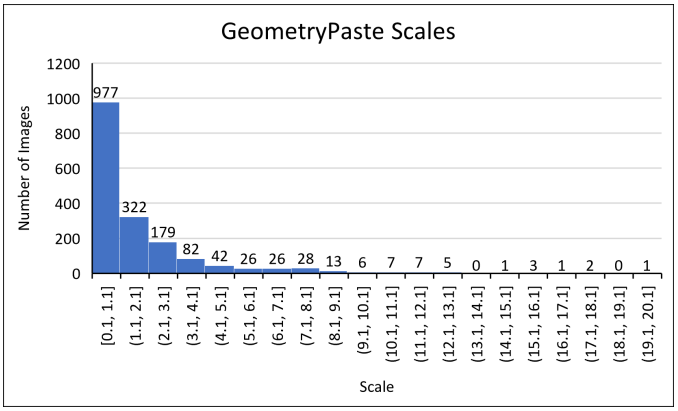

Fig. 3: The object scales generated by GeometryPaste follow a right-skewed distribution, with the majority ranging from 0.1 – 1.1.

random Copy-Paste and GeometryPaste datasets and 3,456 new training images for the GeometryPaste blended dataset.

C. Training and Evaluation

We employ the Mask2Former [13] architecture with a ResNet-50 backbone [14] pre-trained on COCO to evaluate our method. Mask2Former builds on MaskFormer [20] and utilizes the Detectron2 [21] framework. We follow the baseline Mask2Former settings, which include using the AdamW [22] optimizer, an initial learning rate of 0.0001, and a batch size of 16. We fine-tune on each dataset for 45k iterations on 8 Bridges2 [23] GPUs. The baseline training regime utilizes only the original dataset, whereas all training with augmented datasets consists of the augmented and original datasets. We report the overall AP and TTC Message Board AP scores.

V. Results

Fig. 3 shows that our method generates a right-skewed distribution of scales ranging from 0.1 - 20.1, with the vast majority being from 0.1 – 1.1. The overall appearance of the images generated is more realistic than the ones generated with random Copy-Paste. However, some objects are placed off the ground due to low-quality road segmentations. This, along with vanishing point prediction errors, resulted in some objects being scaled incorrectly since our scaling function depends on the coordinates of the road and vanishing point.

Fig. 4 shows each method's overall AP and TTC Message Board AP scores. We see that the highest AP score for the TTC Message board was 35.5, achieved by GeometryPaste with the model trained for 30k iterations. We also see that the highest overall AP score was 31.3, achieved by GeometryPaste with Gaussian blurring with the model trained for 35k iterations. Fig. 5 provides an example where GeometryPaste was the only method that detected the TTC Message Board in the given image.

VI. Discussion

Instance segmentation is a computer vision task with many important applications. However, instance segmentation models may have difficulties detecting rare objects due to a lack

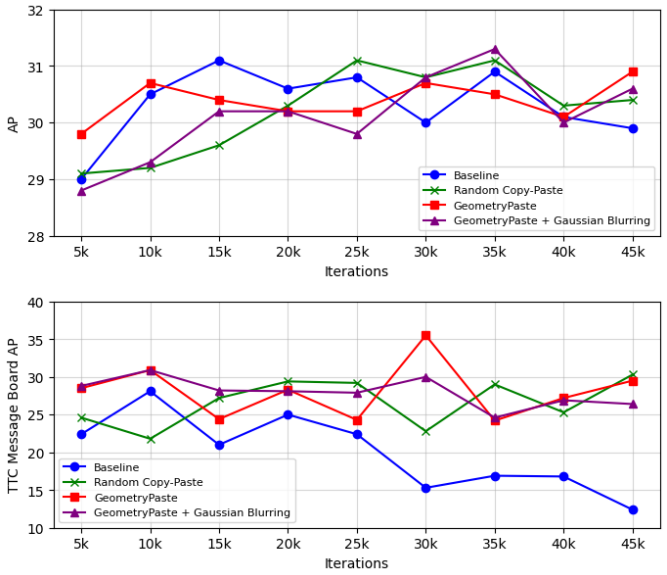

Fig. 4: AP scores for each method. The highest overall AP score was 31.3, achieved by GeometryPaste with Gaussian blurring with the model trained for 35k iterations. The highest TTC Message board AP score was 35.5, achieved by GeometryPaste with the model trained for 30k iterations.

of training data. In this paper, we presented GeometryPaste, a method of generating realistic training images by copying objects from their original images and pasting them onto new backgrounds using both the geometry and context of the objects and background images. We showed that GeometryPaste outperforms other methods in the TTC Message Board AP scores, and that GeometryPaste with Gaussian blurring outperforms other methods in the overall AP scores.

Although we obtained positive results, further work is needed. Our method needs to be verified against additional Copy-Paste strategies. Furthermore, many images in our dataset contain missing or low-quality annotations, and there are also inconsistent labeling issues, which can affect the entire pipeline. Future work may include pasting additional objects and verifying annotation quality, especially in backgrounds used to generate new images, to avoid exacerbating the quality issues.

Acknowledgements

The author would like to thank Dr. Christoph Mertz, Anurag Ghosh, Dr. John Dolan, Rachel Burcin, Maria Ferrato, the entire RISS and CMU community, and the NSF.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.

[2] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation," in *European Conference on Computer Vision (ECCV)*, January 2006. [Online]. Available: https://www.microsoft.com/en-us/research/publication/textonboost-joint-appearance-shape-and-context-modeling-for-mulit-class-object-recognition-and-segmentation/

(a) Ground Truth (b) Baseline

(c) Random Copy-Paste (d) GeometryPaste

(e) GeometryPaste + Blurring

Fig. 5: Visualization of results obtained from the model trained for 30k iterations. Only GeometryPaste detected the TTC Message Board.

[3] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 297–312.

[4] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[6] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[8] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[9] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. 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*, 2021, pp. 2918–2928.

[10] N. Dvornik, J. Mairal, and C. Schmid, "Modeling visual context is key to augmenting object detection datasets," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 364–380.

[11] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu, "Instaboost: Boosting instance segmentation via probability map guided copy-pasting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 682–691.

[12] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1301–1310.

[13] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal of Computer Vision*, vol. 80, pp. 3–15, 2008.

[19] Y. Zhou, H. Qi, J. Huang, and Y. Ma, "Neurvps: Neural vanishing point scanning via conic convolution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 864–17 875, 2021.

[21] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[23] S. T. Brown, P. Buitrago, E. Hanna, S. Sanielevici, R. Scibek, and N. A. Nystrom, "Bridges-2: A platform for rapidly-evolving and data intensive research," in *Practice and Experience in Advanced Research Computing*, ser. PEARC '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3437359.3465593