CACP: Context-Aware Copy-Paste to Enrich Image Content for Data Augmentation

Qiushi Guo[1] Shaoxiang Wang[2] Chun-Peng Chang[2] Jason Rambach[2*]
[1]CSRD, Chengdu, China
[2]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
guoqiushi910@gmail.com
{shaoxiang.wang,chun-peng.chang,Jason.Rambach}@dfki.de

Abstract

Data augmentation is a widely used technique in deep learning, encompassing both pixel-level and object-level manipulations of images. Among these techniques, Copy-Paste stands out as a simple yet effective method. However, current Copy-Paste approaches either overlook the contextual relevance between source and target images, leading to inconsistencies in the generated outputs, or heavily depend on manual annotations, which limits their scalability for large-scale automated image generation. To address these limitations, we propose a context-aware approach that integrates Bidirectional Latent Information Propagation (BLIP) for extracting content from source images. By aligning the extracted content with category information, our method ensures coherent integration of target objects through the use of the Segment Anything Model (SAM) and YOLO. This approach eliminates the need for manual annotation, offering an automated and user-friendly solution. Experimental evaluations across various datasets and tasks demonstrate the effectiveness of our method in enhancing data diversity and generating high-quality pseudo-images for a wide range of computer vision applications.

1. Introduction

Deep Learning-based approaches have become the major paradigm in many computer vision tasks, ranging from classification to segmentation. These approaches outperform traditional ones in terms of accuracy and generalization. However, the bottleneck of supervised deep learning is the quality and quantity of the training dataset. To obtain a dataset, a large volume of images needs to be annotated, which is labour-intensive and time-consuming. For segmentation task, annotating a single image is estimated to take up to 1.5 hours[7]. How to generate high-quality, highly realistic datasets has become an important research question in recent years. Previous

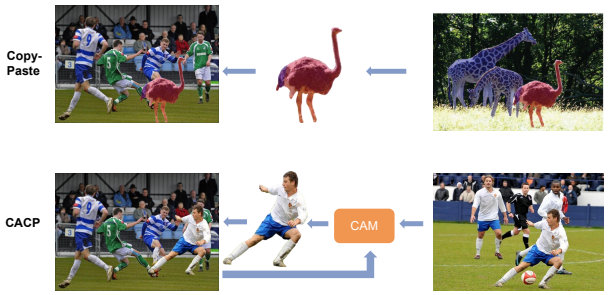

Figure 1. Comparison between the Copy-Paste method (first row) and CACP (second row). The former overlooks the contextual relevance between the base and target images, leading to disharmony. Our approach leverage the semantic information using CAM(Context-Aware-Module) to alleviate this issue.

data augmentation methods increase the diversity of images by applying operations such as flipping, rotating, and adding blur and noise. However, these techniques fail to enhance the content of images at the object level. To address this issue, the Copy-Paste method [14] was proposed. The idea is straightforward and intuitive: objects from target images are pasted onto source images at random positions, resulting in images with enriched content.

However, existing Copy-Paste methods have several drawbacks:

- **Context Neglect**: Methods often neglect the contextual relationship between the copied objects and the target images. For example, a penguin is unlikely to appear in a desert, and a giraffe is improbable on a soccer field. Such contextually incompatible images reduce the practical significance of the augmented dataset.

*Corresponding author

- **Dependency on Masks**: The original Copy-Paste approach depends on publicly available image-mask pairs to generate new images, which limits its applicability and requires additional effort to extend its use. This process is not adaptable to scenarios where masks are unavailable.

To address the aforementioned issues, we propose a novel approach named Context-Aware Copy-Paste (CACP), leveraging large language models (LLMs) and vision foundation modles. Our method integrates several NLP-based models to ensure contextual relevance between the source and target images. The main procedure is as follows: A vision-language model is used to generate captions for the source images (the images to be augmented). Object365, a dataset containing 365 distinct classes, serves as the target image set. For a given source image, a similarity score is computed between its caption and the category names in Object365 using a semantic similarity model based on a transformer architecture. The category with the highest similarity score is selected, and an image from this category is randomly chosen as the target image. An object detection model is then employed to identify objects in the target image. The bounding box of the detected object is processed by a segmentation model to obtain a pixel-level mask. Finally, the object, guided by the mask, is pasted onto the source image.

Our approach can be applied to several computer vision tasks, including classification, object detection, and segmentation. Without requiring extra manual annotation, the target gallery can be easily extended to adapt to specific tasks. Our contributions can be summarized as follows:

- We propose a data augmentation mechanism called Context-Aware Copy-Paste (CACP), which semantically bridges the source and target images. Additionally, this approach is easily extendable to custom tasks without requiring extra annotation.
- We demonstrate that robust segmentation results can be achieved by combining object detector and class activation mapping as prompt generators.
- The experiments results demonstrate that our method outperforms the original Copy-Paste technique and enhances model performance.

2. Related Work

Copy-Paste for Data Augmentation

Copy-Paste has been widely used in semi-supervised scenarios due to its efficiency [15, 16]. Dvornik et al.[12] first proposed the copy-paste approach for object detection tasks based on visual context, which significantly boosted performance on the VOC07 [13] dataset. However, they only used VOC2012[13] as their target gallery, making it challenging to apply the method to other specific scenarios. Additionally, they used a CNN classifier to obtain context information(describing an image only by a word), which is less effective compared to our BLIP-based approach(describing an image by a sentence). Golnaz et al.[14] was the first to propose the Copy-Paste data augmentation method for instance segmentation. They claim that simply pasting objects randomly provides substantial gains over baselines. Although their approach is easy to implement, the random pasting generates images that lack the grounding of real images, as the distribution of object co-occurrences is ignored. Zhao et al. [33] proposed X-Paste, which leverages zero-shot recognition models like CLIP to make the approach scalable. X-Paste demonstrated impressive improvements over CenterNet2. Viktor Olsson el al.[24] introduced ClassMix, which generates augmentations by mixing unlabeled samples based on the network's predictions to respect object boundaries. Inspired by their work, we combined vision-language models with copy-paste to generate augmented images efficiently.

Vision Language Models and Zero-shot Segmentation

Vision-Language Pre-training (VLP) has recently made significant breakthroughs. The zero-shot capability and image-text alignment make it an ideal support for the copy-paste pipeline.

CLIP [26] is a neural network trained on a variety of (image-text) pairs. It can be instructed in natural language to predict the most relevant text snippet for a given image, without direct task-specific optimization. MaskCLIP[11]incorporates a newly proposed masked self-distillation into contrastive language-image pretraining, it distills representation from a full image to the representation predicted from a masked image. BLIP [22] is a new VLP framework that flexibly transfers to both vision-language understanding and generation tasks. Utilizing noisy web data, BLIP achieves state-of-the-art results on several benchmarks and performs exceptionally well on zero-shot tasks.

SAM is a promptable instance segmentation model trained on the largest segmentation dataset to date [20]. It can generalize to new, unseen distributions and tasks, with competitive or even superior performance compared to prior fully supervised results. However, SAM's performance depends heavily on the quality of prompts; insufficient prompts can lead to unstable or unintended segmentation results. SAM is widely used to guide data augmentation. For example, [8] introduced SAMAug, a novel visual point augmentation method for SAM that enhances interactive image segmentation performance. The above approach could be improved by introducing semantic tool to bridge the source and target objects.

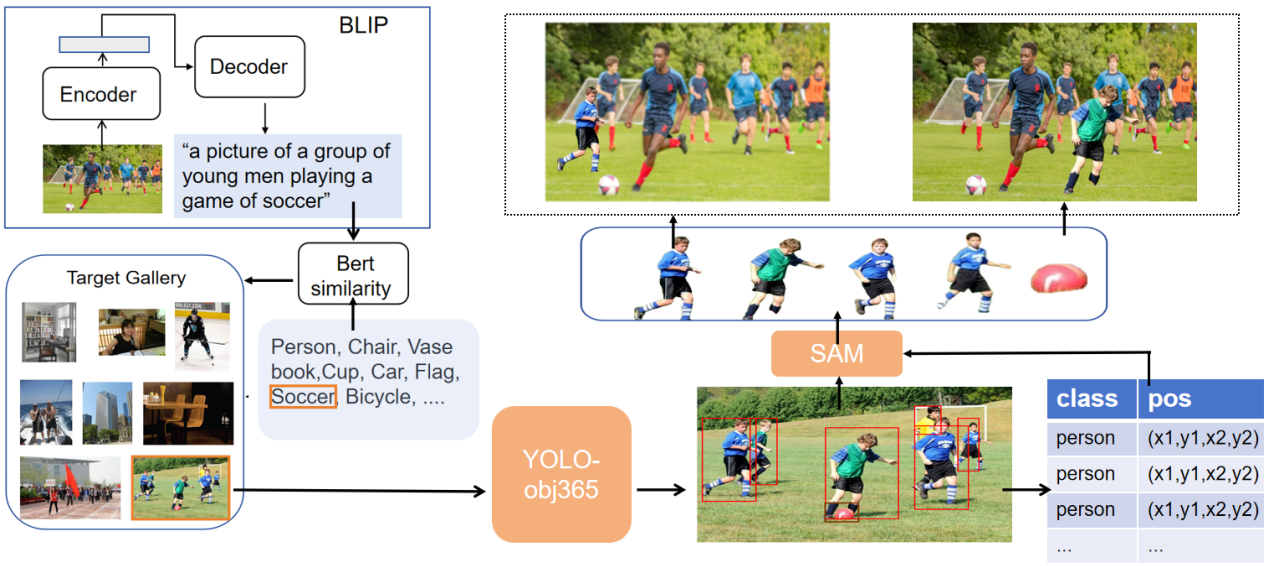

Figure 2. Our method's pipeline involves leveraging BLIP and BERT to select the best-matched target image from a gallery. Subsequently, the corresponding mask is obtained using YOLO and SAM. A single base-target pair can generate multiple augmented images based on user preferences.

3. Motivation

Although copy-paste data augmentation methods have significantly improved computer vision tasks[14], traditional crop-paste pipelines have two notable limitations. First, the crop-paste method is challenging to scale effectively. Second, the semantic gap between the source image and the target image remains unaddressed.

3.1. Scalability and Extendability

Previous copy-paste methods heavily relied on image-mask pairs to perform operations. However, preparing masks for images is costly and time-consuming, making it challenging to apply these methods in a generic manner. To address this issue, previous copy-paste approaches have resorted to using public datasets with pixel-level annotations such as VOC2007 and CamVid[2, 13].

However, these datasets are limited in the number of categories they cover, thereby restricting the diversity of content in generated images. Tab. 1 provides a listing of properties of several public segmentation datasets (ADE20K [34], COCO [23], VOC2007 [13]). These datasets often cannot meet the specific requirements of scenarios. For instance, in a foreign object detection task[29], foreign objects may span hundreds of categories, making it impractical to rely on public datasets or manual annotation to prepare the dataset. An entirely automated copy-paste pipeline is needed to generate large quantities of high-quality images.

dataset	images	classes	resolution
VOC2007[13]	9963	20	-
CamVid[1]	701	32	480*360
CityScape[6]	20000	30	2024*1048
coco[23]	330k	80	640*480
ADE20K[34]	25574	150	1650*2220

Table 1. Summary of properties of common public segmentation datasets

3.2. Content Discrepancy

Previous work has often overlooked the issue of content relevance in the data augmentation process. Irrelevant objects are frequently pasted onto source images, providing minimal training benefit. As depicted in Fig. 4 , images generated by traditional copy-paste methods contribute less to person-related computer vision tasks, as evidenced by corresponding Grad-CAM results. In addition, models trained on such images failed to learn scene context, i.e. which classes are likely to co-exist in an image.

We propose that incorporating highly relevant objects into source images at appropriate positions and scales can enrich the image content and expedite the training process. Experimental results validate our hypothesis: traditionally pasted elements often fail to activate appropriately. In contrast, our method successfully triggers activations, as demonstrated by GradCam visualization techniques Fig. 3, enhancing the content relevant to the desired classes.

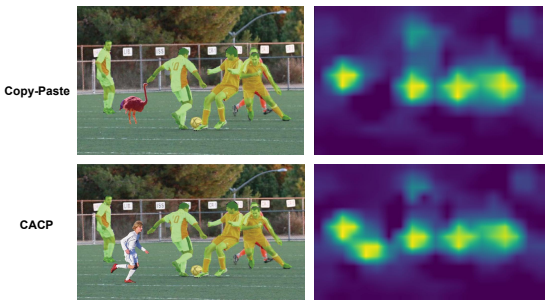

Figure 3. GradCam comparison between the Copy-Paste(top row) and our context-aware copy paste(bottom row). CACP contributes more in person related vision tasks compared to copy-paste.

4. Method

Our CACP approach can be split into the following parts: Gallery Preparation, Context-Awareness Module and Copy-Paste. In the Gallery preparation part, target images are selected to provide object-level content enhancement; In the context aware stage, source image and target image are bridged using a BLIP and BERT-based similarity measurement tool. Once the preferred target image category is determined, object of interests in target image will be cropped and pasted onto the source image considering the size and position. An object detector model and SAM will be leveraged to make the crop-paste process fully automatic. The details of each procedure will be described in follow paragraphs.

4.1. Problem Formulation

Given a source image set $\mathcal{D}_S$ and a target image I_T, our task is to find the most relevant image $I_S \in \mathcal{D}_S$ and most relevant object $obj \in \mathcal{O}$, where $\mathcal{O}$ is the collection of objects in I_S. Specifically, I_S can be obtained as follow:

$$I_s = \arg\min_{I_i} \ \phi(I_i, I_T) \tag{1}$$

where $\phi(\cdot)$ is the function to measure the semantic similarity between two images. Once I_S is determined, obj and corresponding mask M can be inferred as follow:

$$obj, M = \psi(I_S) \tag{2}$$

where $\psi(\cdot)$ are deep learning models, which take images as input and output coordinates(detection task) or labels of each pixel(segmentation task).

$$I_{syn} = I_S \otimes M + I_T \otimes (1 - M) \tag{3}$$

where I_{syn} is the generated image, $\otimes$ is pixel-wise multiplication.

4.2. Data preparation

To enhance the diversity of our dataset, a substantial collection of images is essential for our gallery. In this study, we utilize Object365 [29] as our image gallery. Object365 encompasses 365 classes, featuring over 2 million images and 30 million bounding boxes. These images are characterized by high resolution and quality annotations. In contrast, COCO offers only 80 classes. Object365 significantly expands the range of target objects available for augmentation.

Additionally, we propose an alternative method to leverage images without bounding box annotations, thereby enhancing the applicability of our approach. Specifically, we assume that all images in the galleries are presented without bounding boxes or masks, and each image is labeled solely with its category name. This approach enables users to extend custom categories and adapt them to specific scenarios.

4.3. Context Awareness Module

Image Captioning

To establish semantic coherence between the source and target images, it is crucial to recognize the contents of both images beforehand. Rather than solely detecting objects within the images, we employ a state-of-the-art Visual-Language pre-training (VLM) model as the content extractor. In this role, we utilize BLIP[22]. Compared to object-detection methods, BLIP generates smooth and natural descriptions of input images, rather than isolated words. Furthermore, while object-level approaches may struggle to provide meaningful information when encountering unseen objects, BLIP consistently offers general information applicable to common scenarios.

Target object matching

In the last step we obtain the caption of the source image,namely $C(I_S)$. Due to the large amount of the target image gallery, as a trade-off, we take the class name as the caption of the target image, annotated as $C(I_{T_i}(i = 1, 2, ..., n))$,where n is the total number of categories. To determine the correlation between the $C(I_S)$ and $C(I_{T_i})$, Bert-embedding[10] is leveraged as our measurement tool. In Table 2 we present examples of samples using Bert-embedding to calculate the similarity in our work compared to traditional approaches. From the Tab. 2 we can find that Bert-based distance metric is preferred.

4.4. Copy-Paste

Once the category with the highest similarity score is determined, we randomly select an image from this class as the target image. The SAM is then employed as the pixel-level mask extractor. SAM is a single-shot

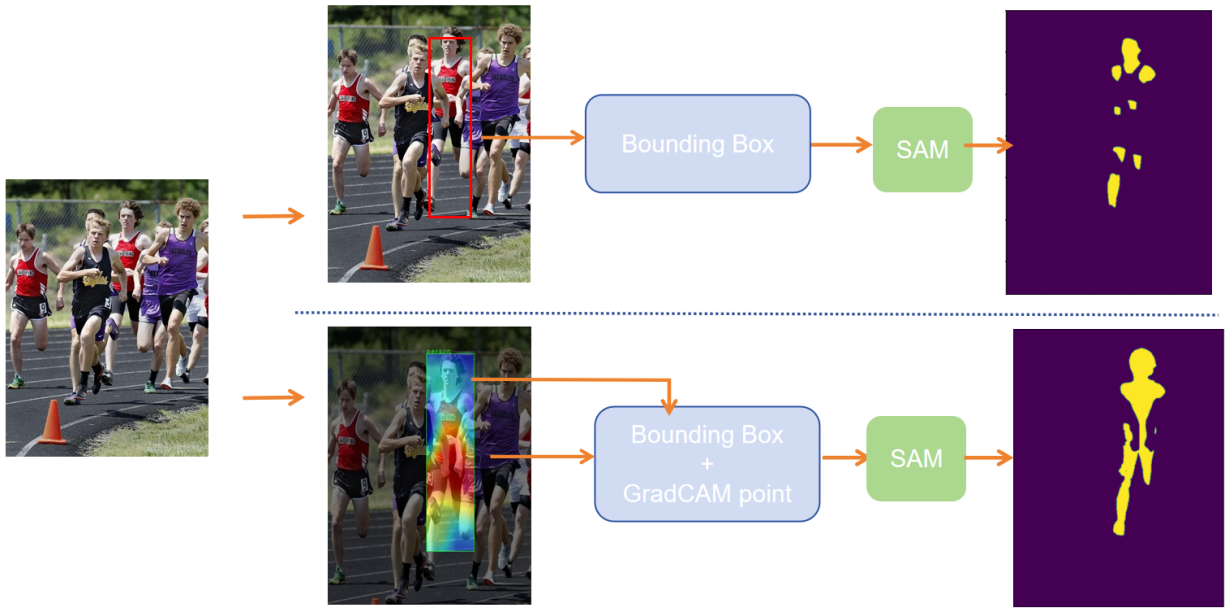

Figure 4. Comparison of SAM segmentation results using different prompts: Single bounding box prompts (upper row) tend to produce incomplete masks, while combining bounding boxes with Grad-CAM points generates more accurate and robust masks.

Image caption	category	Bert
"Two teams are playing	soccer	0.94
football games"	pig	0.41
"A boy is dancing	person	0.89
with a girl in the garden"	goose	0.46
"A boy is standing	flower	0.51
near a red car"	truck	0.89

Table 2. Comparison between BERT-similarity and Cosine-similarity.

segmentation model capable of segmenting any object based on prompts, such as bounding boxes or multiple points indicating the intended objects.

4.4.1. Prompt generation

To obtain the prompts for segmentation, we utilize YOLO-365, a model trained on the Object365 dataset, to detect objects within the target images. For instance, if the YOLO-365 model detects a 'dog' within a target image categorized under dogs, the bounding box of the dog is then forwarded to the SAM model along with the target image. SAM subsequently generates the corresponding segmentation mask for the detected object. Finally, guided by this generated mask, the pixels representing the dog are pasted onto the source images.

In our experimental trials, we observed that feeding pure bounding boxes into the Segment Anything Model (SAM) often results in unintended or incomplete masks for the corresponding target objects. To mitigate this issue, we propose an approach based on Grad-CAM [28] to achieve more accurate segmentation masks.

By inputting the target image into the Grad-CAM module, the resulting heatmap provides valuable positional information about the target. We then use this heatmap to sample points, which are combined with the bounding box as prompts for the SAM model. This hybrid approach, illustrated in Figure 4, improves the accuracy of segmentation results compared to using bounding boxes alone.

4.4.2. Scale and Position

To enhance the realism and harmony of the generated image, rescaling and rendering techniques are implemented. The cropped objects are rescaled according to a ratio interval based on our statistical analysis of the Object365 dataset. We traverse the images in Objects365, and record ratios of image pair, namely $obj1 - obj2 - ratio$. Before pasting the object onto the source images, we extract object pair names and obtain the $ratio_{max}^{obj1,obj2}$ and $ratio_{min}^{obj1,obj2}$ from record.

5. Experiments

5.1. Configuration

The experiments are conducted in the pytorch platform. GPU is RTX 3090ti with 24GB memory. For the

Methods	CamVid						
	Car	Pedestrian	Building	Road	Sky	Tree	↑
U-Net[27]	0.776	0.447	0.764	0.872	0.872	0.834	
U-Net+CACP	0.789(+0.013)	0.481(+0.034)	0.783+0.019	0.893(+0.021)	0.875(+0.003)	0.841(+0.007)	+0.016
FPN[19]	0.792	0.432	0.783	0.897	0.884	0.867	
FPN+CACP	0.813(+0.011)	0.479(+0.047)	0.797(+0.014)	0.903(+0.006)	0.885(+0.001)	0.881(+0.014)	+0.015
PSPNet[32]	0.788	0.445	0.792	0.886	0.875	0.843	
PSPNet+CACP	0.803(+0.015)	0.487(+0.013)	0.799(+0.007)	0.901(+0.015)	0.873(+0.002)	0.862(+0.019)	+0.012
DeepLabV3[4]	0.792	0.461	0.803	0.908	0.891	0.875	
DeepLabV3+CACP	0.811(+0.019)	0.493(+0.028)	0.812(+0.009)	0.922(+0.014)	0.887(-0.004)	0.89(+0.019)	+0.011
DeepLabv3plus[5]	0.803	0.471	0.817	0.927	0.907	0.871	
DeepLabv3plus+CACP	0.817(+0.014)	0.497(+0.026)	0.826(+0.009)	0.933(+0.006)	0.912(+0.005)	0.883(+0.012)	+0.012
PAN[21]	0.794	0.501	0.806	0.931	0.883	0.868	
PAN+CACP	0.812(+0.008)	0.513(+0.012)	0.821(+0.015)	0.937(+0.006)	0.891(+0.008)	0.891(+0.023)	+0.012

Table 3. CACP provides robust gains across popular segmentation architectures in CamVid except **Sky** in DeeplabV3.

classification task, the batch size is set to 16 and the loss function is cross entropy. Adam[18] is used as the optimizer with learning rate of 0.001 over 50 epochs. For the segmentation task, batch size is set to 8, loss function is dice loss,epochs are set to 20. For the detection task, we set batch size as 8, and epochs are set to 50. For segmentation task and object detection task we use the segmentation pytorch model and yolov5-s[30], respectively. Our segmentation models are implemented by Segmentation-Models-Pytorch(SMP) library[17].

5.2. Metric

Precision is selected as the metric for classification task, which is calculated as given below:

$$Accuracy = \frac{Number\_correct\_predictions}{Total\_number\_of\_Predictions} \quad (4)$$

For the segmentation task, the mean Intersection over Union (mIoU) is used to evaluate the model's performance, computed as follows:

$$IoU = \frac{TP}{TP + FP + FN} \quad (5)$$

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} IoU_c \quad (6)$$

To evaluate the performance of the object detection model, we use (mean Average Precision)mAP as our metric. Here we set the threshold as 50%, which means IoU over 0.5 will be considered as correct detection.

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \quad (7)$$

5.3. Evaluation Datasets

To comprehensively assess the viability of our proposed approach, we conducted experiments across two distinct datasets: Cat-Dog classification[25], and CityPersons[31]. These datasets are representative of key tasks in the computer vision field: classification, segmentation, and object detection, respectively.

Methods	**Cat-Dog**	**CityPersons**	
	(Acc)	(mIoU)	(mAP)
B	0.927	0.914	0.557
B+CP	0.941	0.897	0.561
B+aug	0.957	0.903	0.567
B+aug+cp	0.962	0.911	0.571
B+CACP	0.969	0.929	0.577
B+CACP+aug	**0.974**	**0.938**	**0.591**

Table 4. Results between copy-paste(CP) and context-aware Copy- Paste(CACP) in classification, segmentation and detection tasks.

The Cat-Dog dataset contains 25,000 images, each labeled as either a cat or a dog. The CityPersons[31] dataset is a subset of Cityscapes, focusing solely on person annotations. It includes 2,975 images for training and 500 images for validation. The Cambridge driving Labeled Video Database(CamVid)[2] is the first collection of videos with object class semantic labels, complete with metadata. The dataset contains over 700 images with pixel-level annotation. The annotation of images cover 32 class labels from urban and non-urban driving scenes.

5.4. Results

5.4.1. Across Initialization

To validate the robustness of CACP across different initialization, we conduct experiments on CamVid based on two different initialization configurations, namely ImageNet[9] pretrained and normal initialization. As illustrated in **Fig .**5, the results with CACP outperform the

one without CACP in both configurations.

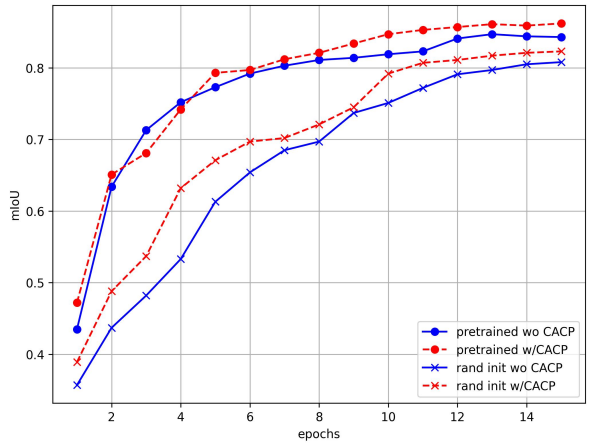

Figure 5. CACP provides gains that are robust to training configurations. We train DeepLabv3 on CamVid for varying number of epochs. The CACP is helpful under with and without pretraining configurations.

5.4.2. Across Tasks

We conduct experiments on different computer vision tasks to validate the usage scenarios of CACP, namely image classification tasks on Cat-Dog, image segmentation and object detection on CityPersons. As illustrated in **Table** 4, the items in the column Methods indicate different combinations. **B** indicates base model. In the classification task this is a Resnet-50 and in segmentation it is DeepLabv3[4], while in object detection it is YOLOv5-s[30]. **CP** and **CACP** indicate random crop-paste and context-aware copy-paste, respectively. **aug** is a combination of traditional data augmentation techniques, including flip, Color jittering, random noise, which is provided by lib Albumentation[3]. All augmentation techniques have the ability to boost the model. CACP outperforms CP and aug across all three tasks.

5.4.3. Across Architectures

As illustrated in **Table** 3, to validate the effectiveness of our method across different architectures, we conduct experiments on CamVid[1] dataset using popular encoder-decoder architectures, namely U-net[27], FPN[19], PSPNet[32], DeepLabv3[4], DeepLabV3+[5], and PAN[21]. The experiment is conducted with or without our CACP augmentation operation. To observe the effect of different categories, we select six classes: [car,pedestrian,building,tree,sky,road]. It can be noticed that almost all classes results are improved with CACP augmentation (except Sky in DeepLabV3 configuration).

5.4.4. Across Partition

To validate the effect of the number of augmented images, we conduct experiments on CamVid as illustrated in **Table** 5. **1/n** indicates 1/n of training images have been augmented using CACP. It is important to note that the total number of training image is fixed, only the ratio of augmented to non-augmented varies. The result illustrates that the performance increases with the rise of partition from 1/8 to 1/2, the trend is stable in all 4 experiments. The increase is saturated when partition is over 1/2.

5.4.5. Speed Up Convergence

We have noticed that CACP contributes to speed up the training process. We trained DeepLabv3 on Camvid with 20 epochs. As illustrated in **Fig.** 6, the CACP augmented one converges rapidly compared to the wo-CACP one. The loss is stable around epoch 15, while the wo-CACP is still not fully converged after epoch 19.

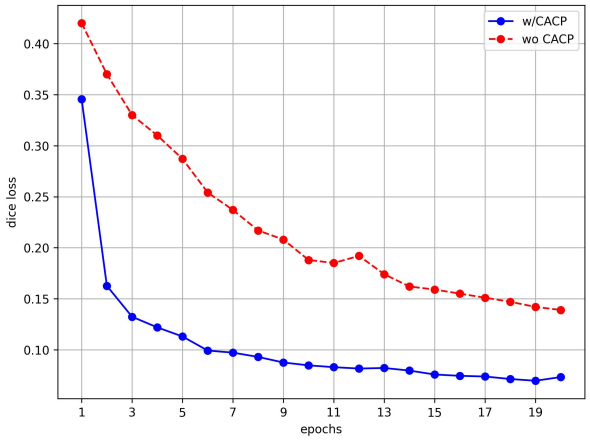

Figure 6. During training, the Dice loss with the CACP configuration converges faster compared to the process without the CACP configuration.

5.4.6. Effect of GradCAM Prompt

To improve the robustness and stability of the output masks of SAM, we propose the GradCAM-guided approach to generate points+bounding box prompts. During our trials, we notice that the number of selected points in GradCAM affects the final segmentation mask. To determine the best point numbers, we conduct the following experiment on CamVid. **Rand** indicates the point is sampled randomly inside bounding box. CAM(n) means n points are extracted in GradCAM map with high value. **Table** 6 indicates that extra prompts can improve the accuracy of SAM; CAM-based point prompts are better than random points. The preferred number of points is around 3 to 5.

Methods	DeepLabv3			
	1/8	1/4	1/2	Full
PSPNet	0.872	0.891	0.887	0.893
U-NET	0.851	0.873	0.879	0.877
PAN	0.862	0.871	0.877	0.874
DeepLabv3 plus	0.883	0.903	0.901	0.907

Table 5. Results between copy-paste(CP) and context-aware Copy- Paste(CACP) in classification, segmentation and detection tasks.

bbox	+rand(1)	+CAM(1)	+CAM(3)	+CAM(5)
0.734	0.841	0.927	0.934	0.933

Table 6. mIoU across different prompts.The performance improves with the number of CAM-generated point prompts and stabilizes when the number of point prompts exceeds three.

6. Discussion

In this paper, we propose a context-aware copy-paste (CACP) data augmentation approach, designed as a versatile plug-and-play module for various computer vision tasks, eliminating the need for additional manual annotation. CACP is particularly effective for custom segmentation in semi-supervised learning, offering a time-efficient and scalable solution that allows users to tailor their target gallery to specific task requirements.

Future research directions include the following:

- Integration with diffusion models to generate synthetic augmented images with well-crafted prompts, particularly for privacy-sensitive objects, thereby mitigating privacy concerns.
- Adaptation to downstream industrial applications, such as obstacle detection and pedestrian detection.
- Enhancement through advanced image composition techniques, including object placement, image blending, image harmonization, and shadow generation, to improve the realism and consistency of augmented images

References

[1] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern recognition letters*, 30(2):88–97, 2009. 3, 7

[2] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern recognition letters*, 30(2):88–97, 2009. 3, 6

[3] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 7

[4] 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. 6, 7

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6, 7

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[8] Haixing Dai, Chong Ma, Zhengliang Liu, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, Dajiang Zhu, Wei Liu, et al. Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*, 2023. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[11] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023. 2

[12] 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. 2

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2, 3

[14] 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, 2, 3

[15] Qiushi Guo. A universal railway obstacle detection system based on semi-supervised segmentation and optical flow. *arXiv preprint arXiv:2406.18908*, 2024. 2

[16] Qiushi Guo, Yifan Chen, Yihang Yao, Tengteng Zhang, and Jin Ma. A real-time chinese food auto billing system based on instance segmentation. In *2023 IEEE Region 10 Symposium (TENSYMP)*, pages 1–5. IEEE, 2023. 2

[17] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019. 6

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 6, 7

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[21] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 6, 7

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 4

[23] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[24] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1369–1378, 2021. 2

[25] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 6, 7

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 5

[29] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 3, 4

[30] Ultralytics. YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com, 2021. 6, 7

[31] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 6

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 2881–2890, 2017. 6, 7

[33] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *International Conference on Machine Learning*, pages 42098–42109. PMLR, 2023. 2

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 3