Humans need not *label* more humans: Occlusion Copy & Paste for Occluded Human Instance Segmentation

Evan Ling
evan.ling@hmgics.com
Dezhao Huang
dezhao.huang@hmgics.com
Minhoe Hur
minhoe.hur@hmgics.com

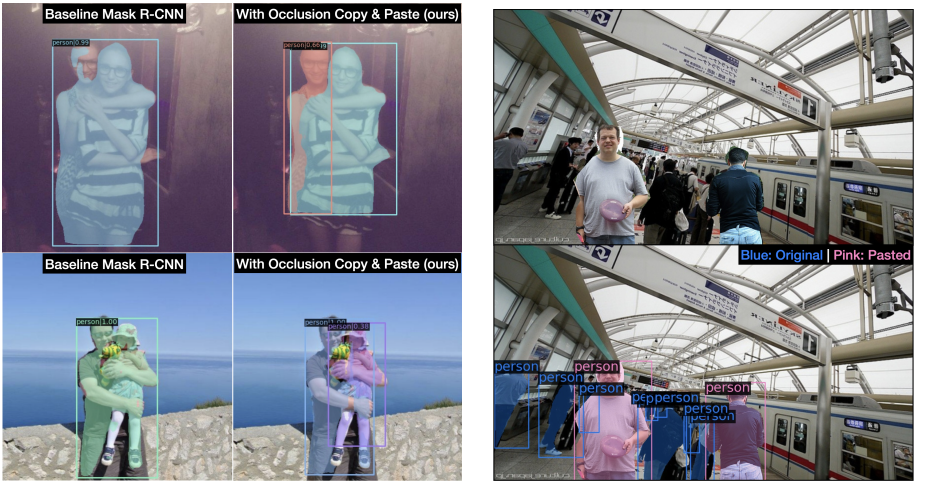
AIR Center
Hyundai Motor Group Innovation Center in Singapore
Singapore

Abstract

Modern object detection and instance segmentation networks stumble when picking out humans in crowded or highly occluded scenes. Yet, these are often scenarios where we require our detectors to work well. Many works have approached this problem with model-centric improvements. While they have been shown to work to some extent, these supervised methods still need sufficient relevant examples (*i.e.* occluded humans) during training for the improvements to be maximised. In our work, we propose a simple yet effective data-centric approach, Occlusion Copy & Paste, to introduce occluded examples to models during training — we tailor the general copy & paste augmentation approach to tackle the difficult problem of same-class occlusion. It improves instance segmentation performance on occluded scenarios for "free" just by leveraging on existing large-scale datasets, without additional data or manual labelling needed. In a principled study, we show whether various proposed add-ons to the copy & paste augmentation indeed contribute to better performance. Our Occlusion Copy & Paste augmentation is easily interoperable with any models: by simply applying it to a recent generic instance segmentation model without explicit model architectural design to tackle occlusion, we achieve state-of-the-art instance segmentation performance on the very challenging *OCHuman* dataset. *Source code is available at* https://github.com/levan92/occlusion-copy-paste.

1 Introduction

There is nothing more interesting to humans, than humans. Object-level visual recognition tasks like object detection and segmentation of the *person* class have overarching applications from pedestrian detection for autonomous driving, to worker safety monitoring in factories and construction sites, to retail analytics in the commercial space. The fact that humans are everywhere necessitates that we need to get very good at picking out humans. However, current state-of-the-art (SOTA) object detection and instance segmentation models



(a) (b)

Figure 1: (a) Instance segmentation predictions of Mask R-CNN on occluded human cases versus one trained with our Occlusion Copy & Paste (base images from *OCHuman* [34]); (b) Example of image augmented with our Occlusion Copy & Paste approach (base image from MS COCO dataset [24])

often fail when people are occluded and in crowded scenes [22, 30, 34]. The issue is exacerbated in instance segmentation when we need to delineate multiple neighboring instances of the same class occluding each other. The current model performance on such cases is illustrated on the left of Figure 1(a). Our work looks at this harder problem of same-class occlusion in instance segmentation, and more specifically, of the *person* class, the most common class of instances in the MS COCO dataset [24]. Moreover, as discussed above, there are ubiquitous applications for recognizing humans.

Methods to tackle most problems in deep learning can be categorized into model-centric or data-centric approaches — most of the time, people found most success using a combination of both. In the space of occluded instance recognition, good model-centric approaches have been proposed to tackle occlusions [22, 23, 30, 34]. However, most of these are supervised, inherently assuming availability of relevant labelled data to learn from, *i.e.*, labelled images with instance occlusions. While it has been very promising in the self-/semi-supervised space [5, 21, 29], the current state of deep learning is still very much label-hungry. Nothing improves a model more than simply having many relevant labelled data.

It is commonplace for detection & segmentation models to be trained (or pre-trained) on open-source large-scale datasets[10, 24]. However, such datasets comprise of common scenes from daily life and difficult scenes with occlusions are few and far between. This possibly contributes to the poor performance of modern models on occluded scenes. A naïve way to go about this is to collect and label a sufficiently large dataset of occluded people. Unfortunately, segmentation tasks are notoriously hard to label for. It took workers an average of 79.2 seconds per instance to annotate the segmentation masks in the MS COCO dataset [24] — a total of 2,500,000 instances took 55,000 man-hours. The effort will be prohibitively large to collect and annotate a large-scale dataset of occluded humans, given that the labelling effort would be even harder due to occluded & intertwined instances.

In light of these limitations, we propose the novel use of instance-level copy & paste as a form of augmentation to directly induce occlusion during training. Figure 1(b) shows

an example training image after applying our augmentation. With our proposed Occlusion Copy & Paste, this straightforward approach significantly improves instance segmentation performance on highly occluded human scenarios (example results on right of Figure 1(a)), all whilst training on existing datasets without additional data collection or labelling. For the first time, to the best of our knowledge, copy & paste is used as an augmentation to directly tackle the issue of occluded instance segmentation. We studied the necessity of realism in augmented images and found that most of the time, it is actually counter-productive. We also experimented and explored the space of using copy & paste to refine its effectiveness as an augmentation technique to tackle occluded instance segmentation. Our approach is complementary to, and should be coupled with model-centric approaches in tackling occlusions to achieve the best performance.

To summarise, our key contributions are as follows: **(1)** We propose the novel use of instance-level copy & paste augmentation to tackle the problem of occluded person instance segmentation; **(2)** We conducted a principled study on the efficacy of our various add-ons that tailors our copy & paste augmentation for occluded instance segmentation and importantly show that sometimes, variety is favoured over realism; **(3)** By simply applying our Occlusion Copy & Paste with a recent instance segmentation model [6], without any explicit model architectural designs to tackle occlusions, we achieve SOTA instance segmentation performance on the very challenging *OCHuman* benchmark [34]; **(4)** We also clarify various labelling and evaluation fairness issues around *OCHuman* and propose a fully-labelled subset for future works to benchmark upon: $OCHuman^{FL}$.

2 Related Work

Object Detection & Instance Segmentation. Object detection and instance segmentation are closely linked in concept and research works. Modern object detection is anchored upon the seminal R-CNN work [14] in 2014 and its succeeding Fast & Faster R-CNN [13, 28] which formed the core framework for two-stage object detection for nearly the decade to come. Two-stage models first propose regions of interest, then refine bounding boxes and classify in the second stage. For instance segmentation, Mask R-CNN [19] was a natural extension of Faster R-CNN [28], extending a mask head parallel to the existing box and classification heads. This completely shifted how instance segmentation is approached, bringing forth an "instance-first" strategy, which proved to be highly effective despite being conceptually simple. Most recently, Mask2Former [6] was introduced as a universal model architecture for segmentation tasks, which broke the SOTA in not just instance but across semantic and panoptic segmentation as well. Using a DETR [3] style transformer-based decoder with masked attention and learnable object queries allowed for efficient training while exploiting the pair-wise self-attention interaction in the transformer layers. The goal of our work is to prove the efficacy of copy & paste augmentation in tackling the issue of occlusions in instance segmentation. Therefore, we chose Mask R-CNN [19] as our baseline model during experimentation due to its simplicity and efficiency. We eventually extend our approach to Mask2Former [6], demonstrating SOTA results on occluded human benchmarks.

Tackling Occlusions. Occlusion is not a new problem — many works over the years have attempted to tackle it from a model-centric perspective, from occlusion in object detection [30, 33] to segmentation like OCFusion [23] and BCNet [22]. More relevant to our work, [34] introduced a new benchmark for occluded person instance segmentation, the Occluded

Human (*OCHuman*) dataset. This challenging benchmark is especially useful, plugging the gap on instance segmentation datasets containing heavily occluded humans. In the same work, they propose to tackle occlusion via Pose2Seg, where instead of bounding boxes, human poses are used as proposals to instance segmentation and used for feature alignment. Following, PoSeg [35] extends it by integrating top-down and bottom-up cues, predicting human pose and instance segmentation mask in a multi-task manner. Similarly, they benchmark their approach on the *OCHuman* dataset. Most works have tackled occlusion with supervised model-centric approaches. Complementing that, our work approaches the occlusion issue from a data-centric direction by providing relevant labelled data.

Copy & Paste as Augmentation. Image augmentation regularizes training by introducing greater variety of training inputs. More complex forms of augmentation bring in multiple images, like mixup [32], CutMix [31] or mosaic augmentation of Yolov4 [2]. These ideas run along the same vein as the concept of copy & pasting instances onto another image. This copy & paste concept is simple and over the years, several works have tried to formalize the idea, proposing variations for different uses. One of the earlier works [9] used it for instance detection: from a separate bank of segmented instances, they pasted instances randomly onto backgrounds, framed as a synthetic data approach instead of as an augmentation technique. They particularly found that their model was over-fitting to pasting artefacts at the edges and implements random blending for the model to look beyond these artefacts and blending styles. A subsequent work by [8] argues that random pasting does not extend to object detection, proposing an additional contextual model to determine if an instance is of suitable class for pasting into an image. Then, InstaBoost [11] eliminates the extra contextual model, making use of original instances within an image and jitter them around their locality instead.

Most recently, [17] primes Simple Copy-Paste (SCP) for improving instance segmentation performance. This work surprisingly rebuts the various claims from the previous works [8, 9]—they claim that blending is not beneficial and that pasting with context is not needed, random pasting works just as well. Similarly in our experiments, we show that such realism enhancers are actually counter-productive most of the time. SCP views copy & paste as a way to generally increase diversity of data points during training, simply taking a pair of images and directly overlaying instances from one image over the other without modifications. In contrast, we use copy & paste as a means to directly induce occlusions and at the same time retain "free" ground-truth labels via Occlusion Copy & Paste. We introduce more direction and stochasticity in pasting—copy & paste is used as image augmentation to directly tackle the issue of occluded persons instance segmentation.

3 Approach

The concept behind copy & paste is straightforward: we start from our Basic Copy & Paste pipeline, then introduce some add-ons and designs that improves efficiency, stochasticity and enhance realism. We show in experiments later that realism enhancements are generally counter-productive. The eventual Occlusion Copy & Paste does not contain most of the realism enhancements, but do contain the add-ons that improve efficiency and stochasticity.

3.1 Basic Copy & Paste

Figure 2 illustrates our Basic Copy & Paste pipeline for each data iteration during training: For every training image, a dynamic basket of images is sampled from the dataset. Then,

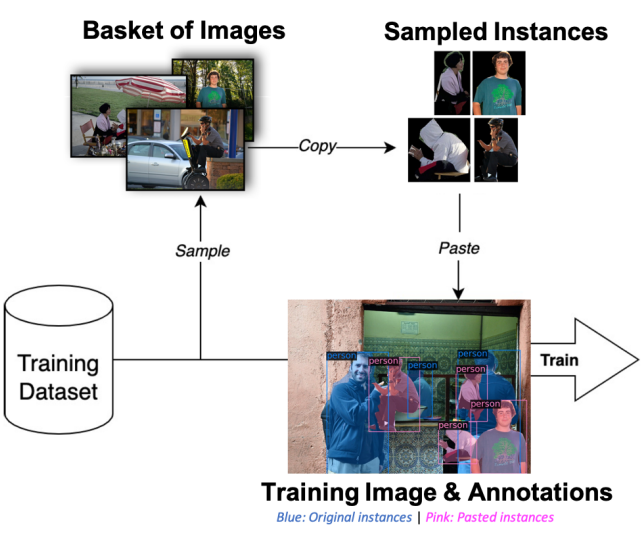


Figure 2: Overview of our Basic Copy & Paste augmentation pipeline.

a random subset of instances from this basket is pasted onto the current image at random locations. Sub-regions of existing instances that overlapped with pasted instances may no longer be visible, so the associated ground-truth mask regions are removed. Some instance may be too occluded to meaningfully enforce predictions: if remaining visible regions is too small, we will remove that instance entirely. Instances are pasted in sequentially, which facilitates the possibility of occlusions amongst pasted instances as well. This runs in real-time in each training iteration, giving rise to novel views & occlusion scenarios in every step, the combinatorial sum of all instances and images provides relevant yet huge variety of training examples to the task of occluded instance segmentation.

The augmentation pipeline has the following hyper-parameters introduced:

- $\mathbf{P}_{CP}$: Overall probability for Copy & Paste augmentation, sampled every iteration, ensures it does not happen all the time, as it is considered a strong augmentation.
- $\mathbf{N}_{basket}$: Number of images to sample into a basket. Larger number may improve variety of instances to be pasted, but increases pre-processing time. On the other hand, smaller $\mathbf{N}_{basket}$ may preserve more context information where people of a certain appearance may tend to appear together in the same image (*e.g.*, a skiing scene).
- $\mathbf{R}_{paste}$: Range of number of sampled instances to be pasted — the actual number of pasted instances each iteration is uniformly sampled within this range. An optimal number is chosen, demonstrated in section 4.4.1 later on. We need enough occlusion to happen, yet not too many as they may over-clutter the image, which may be detrimental to the learning. $\mathbf{N}_{basket}$ will need to increase as $\mathbf{R}_{paste}$ increases such that there are enough instances in the basket to sample from.

Already, in order to address occlusions, our basic implementation differs from SCP [17]: our instances are stochastically selected from a dynamic basket and pasted at random locations, while SCP rigidly takes instances from one image and directly overlays onto another at fixed original positions.

3.2 Occlusion Copy & Paste

Various improvements are then made on top of our Basic Copy & Paste pipeline to improve efficiency and introduce more variety in augmentation, further tailoring it to tackle occlusion. *More details on these add-ons can be found in the supplementary material.*

Targeted Pasting. Targeted pasting works to increase efficiency of the copy & paste augmentation to more directly induce occlusion. In target pasting, the pasting location of each copied instance is randomly chosen within the locality of a random existing instance in the image (previously, each instance is randomly pasted within the image bounds).

Augmented Instance Pasting. Augmented instance pasting works to introduce more variety in pasting instances during training. Augmentation on instances consist of a mixture of random color jittering (saturation, contrast, brightness and sharpness) and geometric jittering (scaling and rotation).

Figure 3: Example of targeted pasting and augmented instance pasting. Pasted instances sampled within image to show contrast after instance augmentation

3.3 Realism Enhancements

As with all kinds of synthetic approaches, a natural question to ask is how realistic are the synthesized images for our model to learn from and generalize to real-world testing. An image's realism depends on various factors: In synthesis via graphics rendering, photo realism is the main factor people care about. However, the very advantage of the copy & paste approach is we get around photo realism by simply re-using parts of other real images. On the other hand, semantic realism is relatively more lacking, as pasted instances often look out of place due to the random pasting. We briefly describe some realism enhancements implemented for experimentation. It is not proven that we need *perfect* realism to train a good model — we will show in later sections that most of these realism enhancements may actually be counter-productive. Furthermore, in single class setting like this, there is an even lesser need to learn inter-class contextual dependencies (classes that naturally occur more often with each other). *Please refer to supplementary for more visualised examples.*

Minimum size filter. We may end up with pasted instances that are too small in the augmented image. Sometimes, tiny instances are annotated and can be predicted by model because of contextual information, for *e.g.*, if an individual person is picked out from the audience stand of a baseball game and pasted elsewhere, it may not be reasonable for any model nor humans to detect the instance if it is too small. We exclude tiny instances for pasting thresholded on the ratio of equalized side length ($\sqrt{area}$) of instance to target image.

Scale-aware Pasting. To improve semantic realism, in scale-aware pasting, we obtain the size distribution of the bounding boxes of original instances within the destination image, and subsequently sample a scale from this distribution for each instance to be pasted in. This way, pasted instances will look more natural and aligned with context in the scene (for *e.g.*, wide shot or close-up shot of a scene, *see supplementary*).

Better quality masks. We experiment if better quality masks for copy & paste lead to better training. We tap on the *LVIS* [16] dataset, which uses MS COCO [24] images but provides better quality mask labels. However, as *LVIS* is labelled in a federated style, it

contains less labelled instances in less images: 15,482 person instances in 2,225 images, compared to 262,465 in 64,115 images in COCO.

Blending. In [9], blending aims to remove boundary artifacts after pasting. Similarly, we implemented Gaussian blurring that operates on the pasting masks. The size and standard deviation of the Gaussian kernel controls how much "blend" we introduce. These hyper-parameters can be randomized at every paste during training (known as "random blending").

4 Experiments

4.1 Dataset and Evaluation Protocol

Datasets. In all experiments, we train on the *person* class subset of the MS COCO dataset [24], with 262,465 human instances in 64,115 images. Our copy & paste augmentation taps on the same training set. Only for the experiment for better quality masks (as described in section 3.3), we additionally utilize *LVIS* [16] mask annotations. To evaluate performance of instance segmentation under highly occluded scenarios, we test on the *OCHuman* [34] validation (4,291 instances in 2,500 images) & test set (3,819 instances in 2,231 images).

Fully Labelled OCHuman. It is found that the *OCHuman* benchmark is not exhaustively annotated: there are some human instances that do not have mask labels. This led to reproducibility and evaluation fairness issues which are *further discussed in our supplementary material*. For fairer evaluation, we provide a subset of *OCHuman* which only contains images that are fully labelled (especially important for reproduction of Pose2Seg results in Table 4). This subset, $OCHuman^{FL}$ has 2,240 instances in 1,113 images (in val) and 1,923 instances in 951 images (in test). We report our main results on both sets: *OCHuman* and $OCHuman^{FL}$. We hope $OCHuman^{FL}$ can re-anchor a new starting point for future instance segmentation works to benchmark upon for fair evaluation and comparison.

Evaluation. We report the segmentation mean average precision (*mAP*), standard evaluation protocol for instance segmentation [24], equivalent to *AP* in a one-class setting.

4.2 Implementation details

For all experiments (less section 4.5), we use the same model for consistency: Mask R-CNN [19] with ResNet-50 [18] backbone & Feature Pyramid Network [25], considered a strong baseline for instance segmentation and strikes a good balance between accuracy & training speed / GPU memory. [12, 20] have shown that when training with strong augmentations, initializing with pre-trained ImageNet [7] weights might actually hurt performance. We notice the same from our initial experiments, and therefore we train our whole model from scratch: convolutional weights are initialized with Kaiming initialization [17] and batch normalization trainable parameters are initialized to 1. We train on single machines with 4 NVIDIA V100 GPUs each, with Synchronized Batch Normalization [27]. We mostly follow training hyper-parameters in [1]: an initial learning rate of 0.0125 at a batch size of 8 (scaled according to the linear scaling rule [15]), momentum of 0.9 and weight decay of 4×10^{-5}. We train for 75 epochs (longer due to training from scratch) with learning rate decay by a factor of 0.1 on the 66^{th} & 72^{nd} epochs. We use MMDetection [4] to perform our experiments, and the pre-trained models which we compare against are also from MMDetection.

4.3 Baseline results

Firstly, we show on Table 1 that just training with our Basic Copy & Paste augmentation, significant improvements over the pre-trained model & baseline vanilla trained model are evident. With the same training settings as the baseline vanilla training, the model trained additionally with Basic Copy & Paste outperforms across the board. Additionally, AP_{person} on COCO (mini 5k) validation set is also reported here to show that in getting better at picking out occluded people, we do not sacrifice on the accuracy in normal cases.

Training Approach	OCHuman		OCHuman^{FL}		COCO
	AP^{val}	AP^{test}	AP^{val}	AP^{test}	AP^{val}_{person}
Pre-trained from [4]	14.9	14.9	24.5	24.9	47.5
Baseline vanilla training	16.5	16.6	27.0	27.4	48.7
+ Basic Copy & Paste (ours)	**18.6**	**17.8**	**29.3**	**28.5**	**49.2**

Table 1: Baseline comparisons with Basic Copy & Paste, tested on *OCHuman*

Further improvements are made after the ablation study in section 4.4 and the *final results* are reported in section 4.5.

4.4 Ablation study

In this section, we step through experiments in our ablation study to further tailor our eventual Occlusion Copy & Paste for the best performance.

4.4.1 Number of pasted instances

We study how number of pasted instances affect the performance of the trained model. The results are plotted out on Figure 4. When there are too little pasted instances, there is little chance of occlusion occurring which contributes to poorer performance. On the other hand, when there are too many pasted instances, performance drops possibly due to over-cluttering affecting learning. Optimal R_{paste} is $[4,6]$ for Basic Copy & Paste.

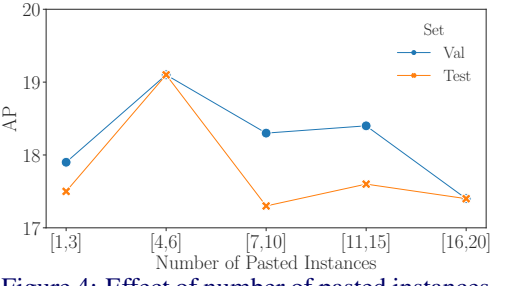


Figure 4: Effect of number of pasted instances on *AP*, tested on *OCHuman*

copy & paste add-ons	AP^{val}		AP^{test}	
Basic Copy & Paste	18.6		17.8	
Minimum Pasting Size	**18.8**		**18.5**	
Minimum Pasting Size + Scale Aware	18.2	-0.6	18.2	-0.3
Minimum Pasting Size + Better Quality Mask	18.4	-0.4	18.0	-0.5
Minimum Pasting Size + Blend (Fixed)	18.6	-0.2	17.9	-0.6
Minimum Pasting Size + Blend (Random)	**19.0**	+0.2	18.4	-0.1

Table 2: Ablation on realism enhancers, tested on *OCHuman*

4.4.2 Realism enhancement experiments

Next, we experimented on the effects of the various realism enhancements described in section 3.3. As seen on Table 2, imposing minimum pasting size led to a slight improvement — not pasting tiny instances which are not meaningful for the model to predict and learn from, does help the model to learn slightly better. However, beyond that, we see that in general,

the rest of the realism enhancements actually caused performance to drop. The argument against realism enhancements is the trade-off with the scope of augmentation variability. Implementing realism means imposing a more restrictive scope on how varied pasting can be: Scale-aware pasting restricts variability on sizes of pasted instances. Using better quality masks may actually help, but in this case, using *LVIS* for better quality masks means fewer human instances to copy from. Variability of pasted instances is more important than quality of instances. Random blending does perform slightly better on the validation set, but the slight improvement does not justify the increase of about 20% in training time due to significant computation in blending operations. Our findings that realism enhancements are generally counter-productive corroborates with [12]. Moving on, we only preserve the minimum pasting size control for Occlusion Copy & Paste.

4.4.3 Efficacy of targeted pasting & instance augmentation

One potential downside of Basic Copy & Paste was that training time increases with a greater number of pasted instance due to the sequential pasting of instances. This downside is minimized with targeted pasting (described in section 3.2). As seen on Table 3, targeted pasting improves the efficiency of pasting such that we achieve the very good performance at low R_{paste} of $[1,3]$. In fact, at higher pasting number with targeted pasting, performance improvements are not as obvious and even starts to drop, possibly due to over-cluttering around a local region. Targeted pasting allows us to achieve high performance with less pasted instances and consequentially, shorter training duration. We also observe that augmented pasting (described in section 3.2) further pushes the performance on *OCHuman*. Once again, this demonstrates that more variability in augmentation helps with performance.

copy & paste add-ons	AP^{val}		AP^{test}	
Basic C&P, $R_{paste} = [1,10]$	18.6		17.8	
+ Targeted	18.6	+0.0	18.2	+0.4
Basic C&P, $R_{paste} = [1,3]$	17.9		17.5	
+ Targeted	19.1	+1.2	18.0	+0.5
+ Targeted & Augm. Paste	19.2	+1.3	18.4	+0.9
+ Targeted, Augm. Paste & Min. Size	**19.5**	**+1.6**	**18.6**	**+1.1**

Table 3: Ablation on Targeted & Augmented Paste, tested on *OCHuman*

Our eventual Occlusion Copy & Paste (OC&P) is made out of the Basic Copy & Paste with minimum pasting size control imposed, targeted pasting at a R_{paste} of $[1,3]$ and augmented instance pasting.

4.5 Pushing the performance & SOTA

On Table 4, we show that OC&P is easily interoperable with any model: besides Mask R-CNN (previously in sections 4.3 and 4.4), we trained OC&P with Pose2Seg [34] and showed a significant $\sim$10% improvement on $OCHuman^{FL}$. Finally, we applied OC&P on Mask2Former (M2F) [6], one of vastly different model architecture (see section 2), and push the SOTA performance on instance segmentation task on the *OCHuman* benchmark. We use M2F with Swin-S [26] backbone, follow their original training schemes and fine-tune on our COCO training set — this includes Large Scale Jittering (LSJ) augmentation, which is a strong augmentation first introduced in [12], which does random image resizing at a larger range of $[0.1,2.0]$, followed by fixed size cropping. This demonstrates that OC&P is also

additive to other strong augmentation methods. *More details on our M2F experiments can be found in supplementary*. As seen on the last row of Table 4, our eventual model trained with OC&P outperforms PoSeg [35], the current SOTA on *OCHuman*. The performance on $OCHuman^{FL}$ doubles that of Pose2Seg [34] as well. The model architectures of Pose2Seg & PoSeg were specially designed to tackle occlusions, and ExPoSeg even utilises a high-performance external pose estimation model. M2F is a generic instance segmentation model, but is able to outperform both models just by training with OC&P — this shows the potential of such data-centric approaches. For completeness of comparison, we also include results from our training runs with Simple Copy-Paste [12] (details in section 2) instead of OC&P and show that OC&P still outperforms the simpler augmentation method.

Model	External Pose Model	Modelled for Occlusion	OCHuman AP^{val}	OCHuman AP^{test}	$OCHuman^{FL}$ AP^{val}	$OCHuman^{FL}$ AP^{test}
Pose2Seg§ [34]	✓	✓	-	-	22.8+	22.9+
+ *Occlusion C&P (ours)*			-	-	25.3+	25.1+
Mask R-CNN§ [19]	✗	✗	14.9	14.9	24.5	24.9
Mask R-CNN†			16.5	16.6	27.0	27.4
+ *Occlusion C&P (ours)*			19.5	18.6	30.6	29.9
PoSeg (JoPoSeg) [35]	✗	✓	25.8⋆	26.4⋆	-	-
PoSeg (ExPoSeg)	✓		26.4⋆	26.8⋆	-	-
Mask2Former§ [6]	✗	✗	25.9	25.4	43.2	44.7
Mask2Former†			26.7	26.3	45.2	46.4
+ Simple Copy-Paste [12]			28.0	27.7	48.9	50.2
+ Occlusion C&P (ours)			**28.9**	**28.3**	**49.3**	**50.6**

Table 4: OC&P improves *AP* across the board and achieves SOTA performance on *OCHuman*. | §: pre-trained models | †: models from our baseline vanilla training

⋆ Directly referenced from paper as no code is published

+ An exhaustively-labelled $OCHuman^{FL}$ is important for fair evaluation in Pose2Seg. Evaluating on $OCHuman^{FL}$ allow us to closely reproduce results reported in [34].

5 Conclusion

All in all, we propose a novel use of copy & paste augmentation to tackle the difficult problem of same-class occlusion in instance segmentation. From Basic Copy & Paste, we experimented with various add-ons: we found realism enhancements are mostly counter-productive, but targeted pasting & augmented pasting improves performance through increased efficiency and variability in augmentation. The eventual Occlusion Copy & Paste augmentation takes these elements and we show that it is interoperable with SOTA instance segmentation models, significantly improving performance on occluded scenarios for "free", without any additional data or labels. Even without explicit architectural design to tackle occlusions, we outperform the SOTA on *OCHuman* by simply applying our Occlusion Copy & Paste on a generic SOTA instance segmentation model. This demonstrates the potential of data-centric approaches. A key benefit of our approach is that it is easily applied with any models or other model-centric improvements. Given the speed at which the deep learning field moves, it is to everyone's advantage to have approaches that are highly interoperable with every other aspect of training. We leave as future work to integrate this with model-centric improvements to effectively solve occluded person instance segmentation.

References

[1] Vaibhav Aggarwal, Yuxin Wu, Wan-Yen Lo, and Ross Girshick. New, improved detectron2 mask r-cnn baselines, Jun 2021. URL https://ai.facebook.com/blog/advancing-computer-vision-research-with-new-detectron2-mask-

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[11] 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.

[12] 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.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[15] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] 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.

[20] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[22] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4019–4028, 2021.

[23] Justin Lazarow, Kwonjoon Lee, Kunyu Shi, and Zhuowen Tu. Learning instance occlusion for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10720–10729, 2020.

[24] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022.

[27] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6181–6189, 2018.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[29] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[30] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018.

[31] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[32] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[33] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.

[34] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019.

[35] Desen Zhou and Qian He. Poseg: Pose-aware refinement network for human instance segmentation. *IEEE Access*, 8:15007–15016, 2020.