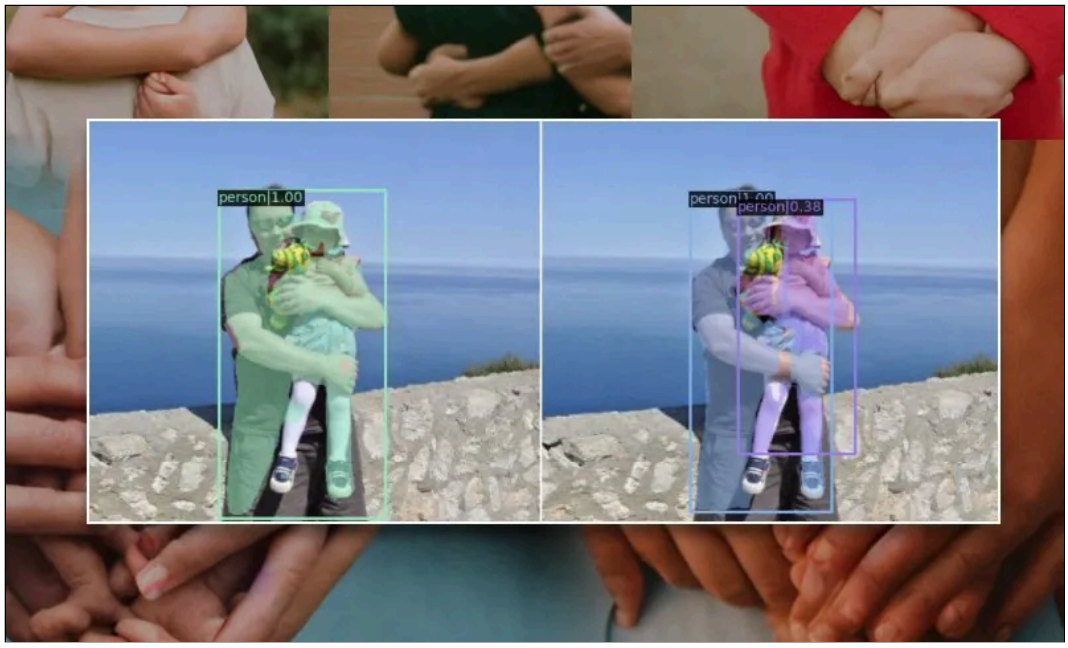
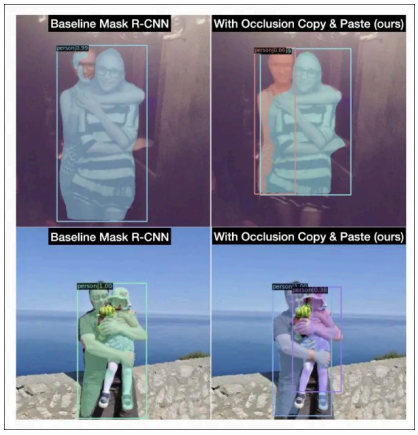
Separating ‘Fused’ Humans in Computer Vision

Originally published at https://www.unite.ai/separating-fused-humans-in-computer-vision/



A new paper from the Hyundai Motor Group Innovation Center at Singapore offers a method for separating ‘fused’ humans in computer vision – those cases where the object recognition framework has found a human that is in some way ‘too close’ to another human (such as ‘hugging’ actions, or ‘standing behind’ poses), and is unable to disentangle the two people represented, confusing them for a single person or entity.



Two become one, but that’s a not a good thing in semantic segmentation. Here we see the paper’s new system achieving state-of-the-art results on individuation of intertwined people in complex and challenging images. Source: https://arxiv.org/pdf/2210.03686.pdf

This is a notable problem that has received a great deal of attention in the research community in recent years. Solving it without the obvious but usually unaffordable expense of hyperscale, human-led custom labeling could eventually enable improvements in human individuation in text-to-image systems such as Stable Diffusion, which frequently ‘melt’ people together where a prompted pose requires multiple persons to be in close proximity to each other.

Embrace the horror – text-to-image models such as DALL-E 2 and Stable Diffusion (both featured above) struggle to represent people in very close proximity to each other.

Though generative models such as DALL-E 2 and Stable Diffusion do not (to the best of anyone's knowledge, in the case of the closed-source DALL-E 2) currently use semantic segmentation or object recognition anyway, these grotesque human portmanteaus could not currently be cured by applying such upstream methods – because the state of the art object recognition libraries and resources are not much better at disentangling people than the CLIP-based workflows of latent diffusion models.

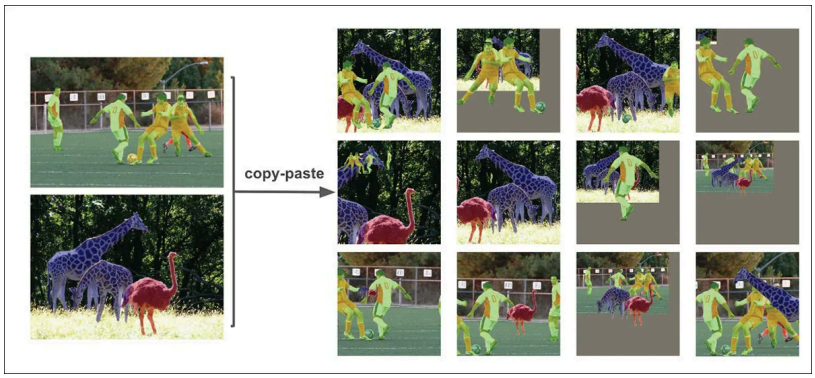
To address this issue, the new paper – titled *Humans need not label more humans: Occlusion Copy & Paste for Occluded Human Instance Segmentation*– adapts and improves a recent 'cut and paste' approach to semi-synthetic data to achieve a new SOTA lead in the task, even against the most challenging source material:

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**

The new Occlusion Copy & Paste methodology currently leads the field even against prior frameworks and approaches that address the challenge in elaborate and more dedicated ways, such as specifically modeling for occlusion.

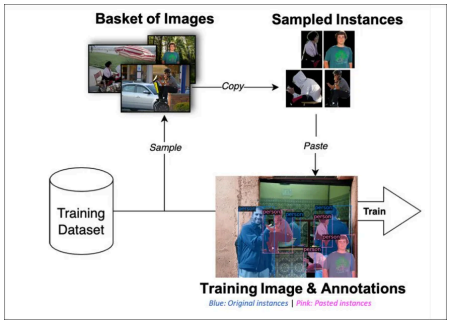
Cut That Out!

The amended method – titled *Occlusion Copy & Paste* – is derived from the 2021 *Simple Copy-Paste* paper, led by Google Research, which suggested that superimposing extracted objects and people among diverse source training images could improve the ability of an image recognition system to discretize each instance found in an image:



From the 2021 Google Research-led paper 'Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation', we see elements from one ph with the objective of training a better image recognition model. Source: https://arxiv.org/pdf/2012.07177.pdf

The new version adds limitations and parameters into this automated and algorithmic 'repasting', analogizing the process into a 'basket' of images full of potential candidates for 'transferring' to other images, based on several key factors.



The conceptual workflow for OC&P.

Controlling the Elements

Those limiting factors include *probability* of a cut and paste occurring, which ensures that the process doesn't just happen all the time, which would achieve a 'saturating' effect that would undermine the data augmentation; the *number of images* that a basket will have at any one time, where a larger number of 'segments' may improve the variety of instances, but increase pre-processing time; and *range*, which determines the number of images that will be pasted into a 'host' image.

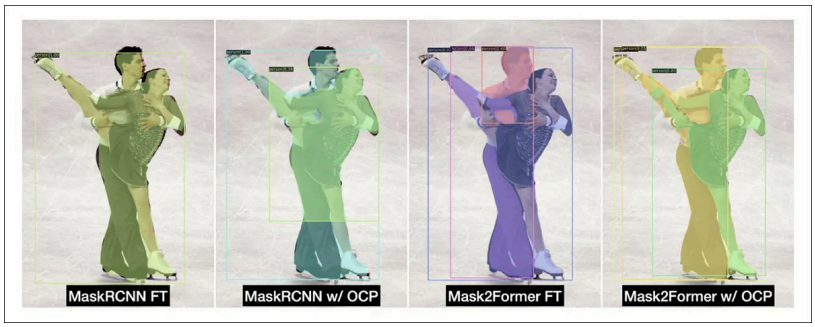
Regarding the latter, the paper notes *'We need enough occlusion to happen, yet not too many as they may over-clutter the image, which may be detrimental to the learning.'*

The other two innovations for OC&P are *targeted pasting* and *augmented instance pasting*.

Targeted pasting ensures that an apposite image lands near an existing instance in the target image. In the previous approach, from the prior work, the new element was only constrained within the boundaries of the image, without any consideration of context.

Though this 'paste in', with targeted pasting, is obvious to the human eye, both OC&P and its predecessor have found that increased visual authenticity is not necessarily important, and could even be a liability (see 'Reality Bites' below).

Augmented instance pasting, on the other hand, ensures that the pasted instances do not demonstrate a 'distinctive look' that may end up classified by the system in some way, which could lead to exclusion or 'special treatment' that may hinder generalization and applicability. Augmented pasting modulates visual factors such as brightness and sharpness, scaling and rotation, and saturation – among other factors.



From the supplementary materials for the new paper: adding OC&P to existing recognition frameworks is fairly trivial, and results in superior individuation of peopl
https://arxiv.org/src/2210.03686v1/anc/OcclusionCopyPaste_Supplementary.pdf

Additionally, OC&P regulates a *minimum size* for any pasted instance. For example, it may be possible to extract an image of one person from a massive crowd scene, that could be pasted into another image – but in such a case, the small number of pixels involved would not likely help recognition. Therefore the system applies a minimum scale based on the ratio of equalized side length for the target image.

Further, OC&P institutes scale-aware pasting, where, in addition to seeking out similar subjects as the paste subject, it takes account of the size of the bounding boxes in the target image. However, this does not lead to composite images that people would consider to be plausible or realistic (see image below), but rather assembles semantically apposite elements near to each other in ways that are helpful during training.

Reality Bites

Both the previous work on which OC&P is based, and the current implementation, place a low premium on authenticity, or the 'photoreality' of any final 'montaged' image. Though it's important that the final assembly not descend entirely into Dadaism (else the real-world deployments of the trained systems could never hope to encounter elements in such scenes as they were trained on), both initiatives have found that a notable increase in 'visual credibility' not only adds to pre-processing time, but that such 'realism enhancements' are likely to actually be counter-productive.

From the new paper's supplementary material: examples of augmented images with 'random blending'. Though these scenes may look hallucinogenic to a person nonetheless have similar subjects thrown together; though the occlusions are fantastical to the human eye, the nature of a potential occlusion can't be known in ac is impossible to train for – therefore, such bizarre 'cut offs' of form are enough to force the trained system to seek out and recognize partial target subjects, without develop elaborate Photoshop-style methodologies to make the scenes more plausible.

Data and Tests

For the testing phase, the system was trained on the *person* class of the MS COCO dataset, featuring 262,465 examples of humans across 64,115 images. However, to obtain better-quality masks than MS COCO has, the images also received LVIS mask annotations.

Released in 2019, LVIS, from Facebook research, is a voluminous dataset for Large Vocabulary Instance Segmentation. Source: https://arxiv.org/pdf/1908.03195.pdf

In order to evaluate how well the augmented system could contend against a large number of occluded human images, the researchers set OC&P against the OCHuman (Occluded Human) benchmark.

Examples from the OCHuman dataset, introduced in support of the Pose2Seg detection project in 2018. This initiative sought to derive improved semantic segmentation of people by using their stance and pose as a semantic delimiter for the pixels representing their bodies. Source: https://github.com/liruilong940607/OCHumanApi

Since the OCHuman benchmark is not exhaustively annotated, the new paper's researchers created a subset of only those examples that were fully labeled, titled OCHuman^{FL}. This reduced the number of *person* instances to 2,240 across 1,113 images for validation, and 1,923 instances across 951 actually images used for testing. Both the original and newly-curated sets were tested, using Mean Average Precision (mAP) as the core metric.

For consistency, the architecture was formed of Mask R-CNN with a ResNet-50 backbone and a feature pyramid network, the latter providing an acceptable compromise between accuracy and training speed.

With the researchers having noted the deleterious effect of upstream ImageNet influence in similar situations, the whole system was trained from scratch on 4 NVIDIA V100 GPUs, for 75 epochs, following the initialization parameters of Facebook's 2021 release Detectron 2.

Results

In addition to the above-mentioned results, the baseline results against MMDetection (and its three associated models) for the tests indicated a clear lead for OC&P in its ability to pick out human beings from convoluted poses.

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**

Besides outperforming PoSeg and Pose2Seg, perhaps one of the paper's most outstanding achievements is that the system can be quite generically applied to existing frameworks, including those which were pitted against it in the trials (see the with/without comparisons in the first results box, near the start of the article).

The paper concludes:

> *'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.'*

Potential for Improving Text-to-Image Synthesis

Lead author Evan Ling observed, in an email to us*, that the chief benefit of OC&P is that it can retain original mask labels and obtain new value from them 'for free' in a novel context – i.e., the images that they have been pasted into.

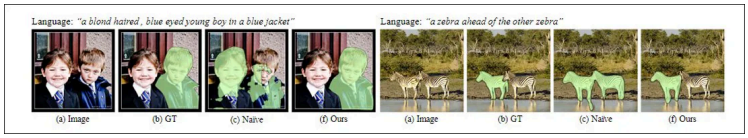
Though the semantic segmentation of humans seems closely related to the difficulty that models such as Stable Diffusion have in individuating people (instead of 'blending them together', as it so often does), any influence that semantic labeling culture might have with the nightmarish human renders that SD and DALL-E 2 often output is very, very far upstream.

The billions of LAION 5B subset images that populate Stable Diffusion's generative power do not contain object-level labels such as bounding boxes and instance masks, even if the CLIP architecture that composes the renders from images and database content may have benefited at some point from such instantiation; rather, the LAION images are labeled for 'free', since their labels were derived from metadata and environmental captions, etc., which were associated with the images when they were scraped from the web into the dataset.

'But that aside,' Ling told us. 'some sort of augmentation similar to our OC&P can be utilised during text-to-image generative model training. But I would think the realism of the augmented training image may possibly become an issue.

'In our work, we show that 'perfect' realism is generally not required for the supervised instance segmentation, but I'm not too sure if the same conclusion can be drawn for text-to-image generative model training (especially when their outputs are expected to be highly realistic). In this case, more work may need to be done in terms of 'perfecting' realism of the augmented images.'

CLIP is already being used as a possible multimodal tool for semantic segmentation, suggesting that improved person recognition and individuation systems such as OC&P could ultimately be developed into in-system filters or classifiers that would arbitrarily reject 'fused' and distorted human representations – a task that is hard to achieve currently with Stable Diffusion, because it has limited ability to understand where it erred (if it had such an ability, it would probably not have made the mistake in the first place).



Just one of a number of projects currently utilizing OpenAI's CLIP framework – the heart of DALL-E 2 and Stable Diffusion – for semantic segmentation. Source: https://openaccess.thecvf.com/content/CVPR2022/papers/Wang_CRIS_CLIP-Driven_Referring_Image_Segmentation_CVPR_2022_paper.pdf

'Another question would be,' Ling suggests. 'will simply feeding these generative models images of occluded humans during training work, without complementary model architecture design to mitigate the issue of "human fusing"? That's probably a question that is hard to answer off-hand. It will definitely be interesting to see how we can imbue some sort of instance-level guidance (via instance-level labels like instance mask) during text-to-image generative model training.'

** 10th October 2022*

First published 10th October 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More