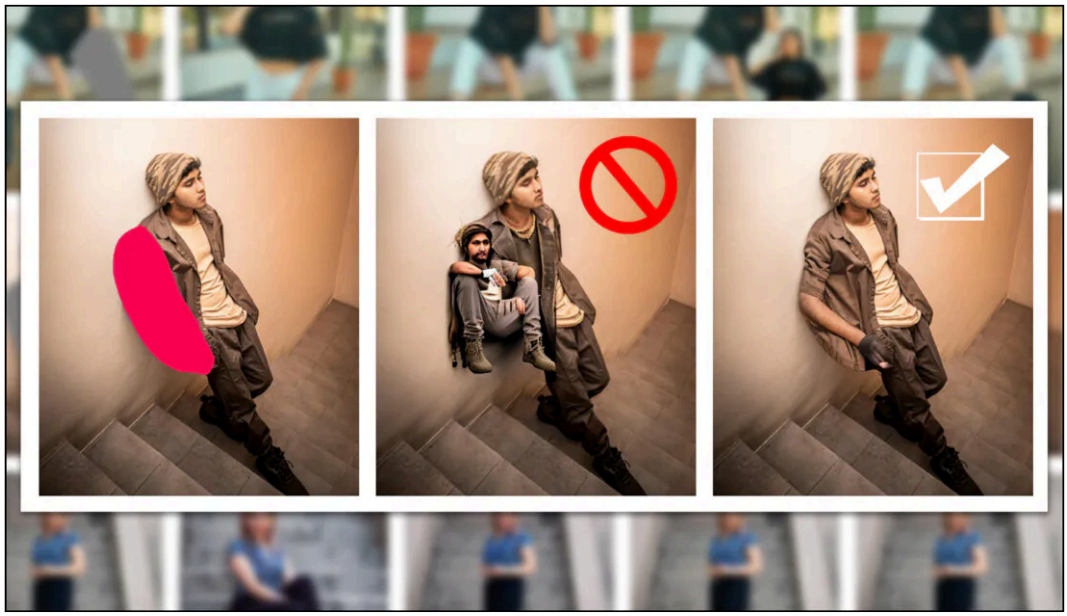
Restoring and Editing Human Images With AI

Originally published at https://www.unite.ai/restoring-and-editing-human-images-with-ai/



A new collaboration between University of California Merced and Adobe ADBE 5.73% offers an advance on the state-of-the-art in *human image completion* – the much-studied task of 'de-obscuring' occluded or hidden parts of images of people, for purposes such as virtual try-on, animation and photo-editing.

Besides repairing damaged images or changing them at a user's whim, human image completion systems such as CompleteMe can impose novel clothing (via an adjunct reference image, as in the middle column in these two examples) into existing images. These examples are from the extensive supplementary PDF for the new paper. Source: https://liagm.github.io/CompleteMe/pdf/supp.pdf

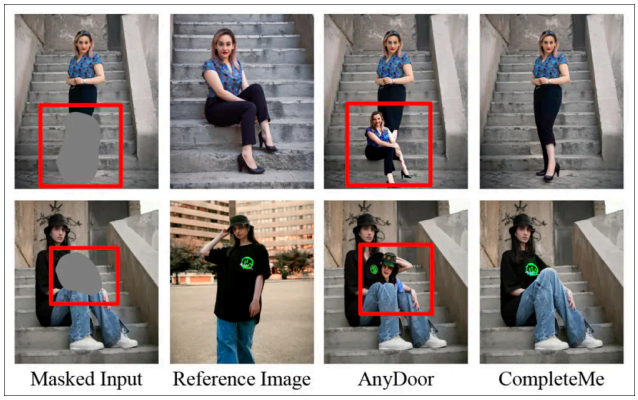
The new approach, titled *CompleteMe: Reference-based Human Image Completion*, uses supplementary input images to 'suggest' to the system what content should replace the hidden or missing section of the human depiction (hence the applicability to fashion-based try-on frameworks):

The CompleteMe system can conform reference content to the obscured or occluded part of a human image.

The new system uses a dual U-Net architecture and a *Region-Focused Attention* (RFA) block that marshals resources to the pertinent area of the image restoration instance.

The researchers also offer a new and challenging benchmark system designed to evaluate reference-based completion tasks (since CompleteMe is part of an existing and ongoing research strand in computer vision, albeit one that has had no benchmark schema until now).

In tests, and in a well-scaled user study, the new method came out ahead in most metrics, and ahead overall. In certain cases, rival methods were utterly foxed by the reference-based approach:

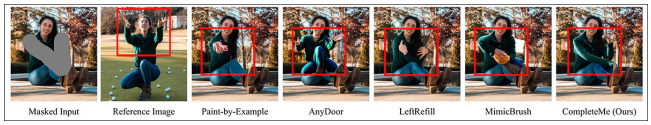


From the supplementary material: the AnyDoor method has particular difficulty deciding how to interpret a reference image.

The paper states:

> *'Extensive experiments on our benchmark demonstrate that CompleteMe outperforms state-of-the-art methods, both reference-based and non-reference-based, in terms of quantitative metrics, qualitative results and user studies.*
>
> *'Particularly in challenging scenarios involving complex poses, intricate clothing patterns, and distinctive accessories, our model consistently achieves superior visual fidelity and semantic coherence.'*

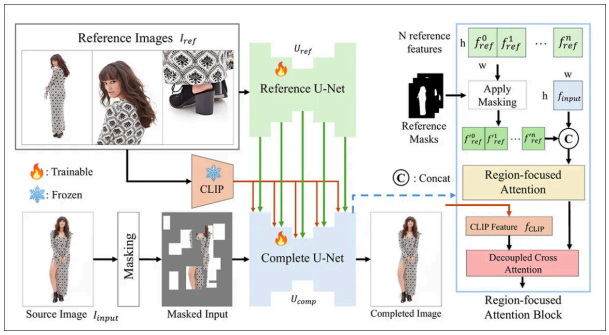
Sadly, the project's GitHub presence contains no code, nor promises any, and the initiative, which also has a modest project page, seems framed as a proprietary architecture.



Further example of the new system's subjective performance against prior methods. More details later in the article.

Method

The CompleteMe framework is underpinned by a Reference U-Net, which handles the integration of the ancillary material into the process, and a cohesive U-Net, which accommodates a wider range of processes for obtaining the final result, as illustrated in the conceptual schema below:



The conceptual schema for CompleteMe. Source: https://arxiv.org/pdf/2504.20042

The system first encodes the masked input image into a latent representation. At the same time, the Reference U-Net processes multiple reference images – each showing different body regions – to extract detailed spatial features.

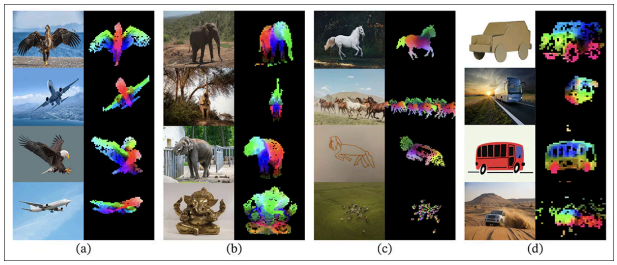
These features pass through a Region-focused Attention block embedded in the 'complete' U-Net, where they are selectively masked using corresponding region masks, ensuring the model attends only to relevant areas in the reference images.

The masked features are then integrated with global CLIP-derived semantic features through decoupled cross-attention, allowing the model to reconstruct missing content with both fine detail and semantic coherence.

To enhance realism and robustness, the input masking process combines random grid-based occlusions with human body shape masks, each applied with equal probability, increasing the complexity of the missing regions that the model must complete.

For Reference Only

Previous methods for reference-based image inpainting typically relied on *semantic-level* encoders. Projects of this kind include CLIP itself, and DINOv2, both of which extract global features from reference images, but often lose the fine spatial details needed for accurate identity preservation.



From the release paper for the older DINOV2 approach, which is included in comparison tests in the new study: The colored overlays show the first three principal components from Principal Component Analysis (PCA), applied to image patches within each column, highlighting how DINOv2 groups similar object parts together across varied images. Despite differences in pose, style, or rendering, corresponding regions (like wings, limbs, or wheels) are consistently matched, illustrating the model's ability to learn part-based structure without supervision. Source: https://arxiv.org/pdf/2304.07193

CompleteMe addresses this aspect through a specialized Reference U-Net initialized from Stable Diffusion 1.5, but operating without the diffusion noise step*.

Each reference image, covering different body regions, is encoded into detailed latent features through this U-Net. Global semantic features are also extracted separately using CLIP, and both sets of features are cached for efficient use during attention-based integration. Thus, the system can accommodate multiple reference inputs flexibly, while preserving fine-grained appearance information.

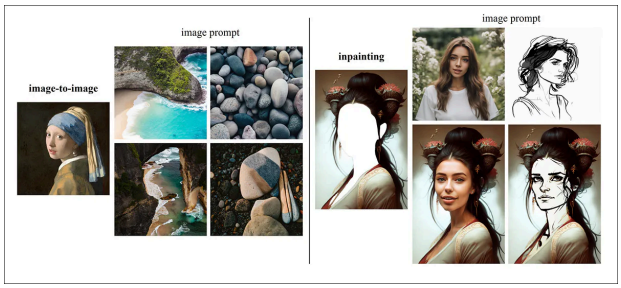
Orchestration

The cohesive U-Net manages the final stages of the completion process. Adapted from the inpainting variant of Stable Diffusion 1.5, it takes as input the masked source image in latent form, alongside detailed spatial features drawn from the reference images and global semantic features extracted by the CLIP encoder.

These various inputs are brought together through the RFA block, which plays a critical role in steering the model's focus toward the most relevant areas of the reference material.

Before entering the attention mechanism, the reference features are explicitly masked to remove unrelated regions and then concatenated with the latent representation of the source image, ensuring that attention is directed as precisely as possible.

To enhance this integration, CompleteMe incorporates a decoupled cross-attention mechanism adapted from the IP-Adapter framework:



IP-Adapter, part of which is incorporated into CompleteMe, is one of the most successful and often-leveraged projects from the last three tumultuous years of development in latent diffusion model architectures. Source: https://ip-adapter.github.io/

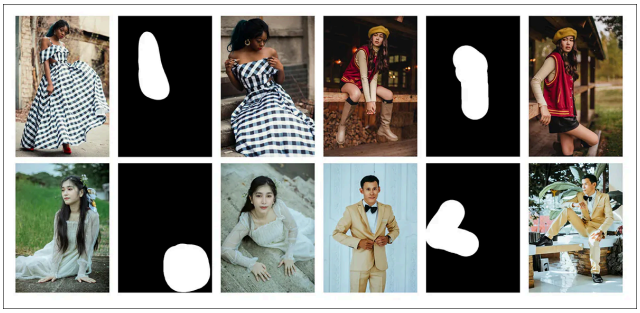
This allows the model to process spatially detailed visual features and broader semantic context through separate attention streams, which are later combined, resulting in a coherent reconstruction that, the authors contend, preserves both identity and fine-grained detail.

Benchmarking

In the absence of an apposite dataset for reference-based human completion, the researchers have proposed their own. The (unnamed) benchmark was constructed by curating select image pairs from the WPose dataset devised for Adobe Research's 2023 UniHuman project.

Examples of poses from the Adobe Research 2023 UniHuman project. Source: https://github.com/adobe-research/UniHuman?tab=readme-ov-file#data-prep

The researchers manually drew source masks to indicate the inpainting areas, ultimately obtaining 417 tripartite image groups constituting a source image, mask, and reference image.



Two examples of groups derived initially from the reference WPose dataset, and curated extensively by the researchers of the new paper.

The authors used the LLaVA Large Language Model (LLM) to generate text prompts describing the source images.

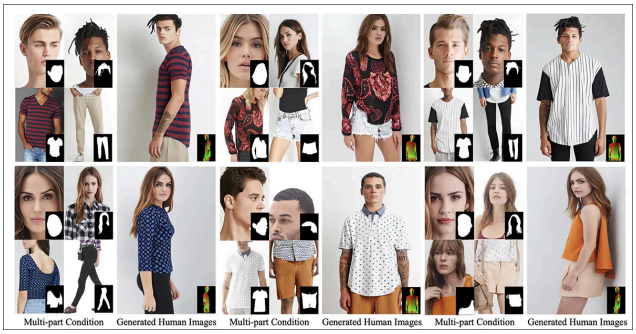
Metrics used were more extensive than usual; besides the usual Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS, in this case for evaluating masked regions), the researchers used DINO for similarity scores; DreamSim for generation result evaluation; and CLIP.

Data and Tests

To test the work, the authors utilized both the default Stable Diffusion V1.5 model and the 1.5 inpainting model. The system's image encoder used the CLIP Vision model, together with projection layers – modest neural networks that reshape or align the CLIP outputs to match the internal feature dimensions used by the model.

Training took place for 30,000 iterations over eight NVIDIA A100† GPUs, supervised by Mean Squared Error (MSE) loss, at a batch size of 64 and a learning rate of 2×10^{-5}. Various elements were randomly dropped throughout training, to prevent the system overfitting on the data.

The dataset was modified from the Parts to Whole dataset, itself based on the DeepFashion-MultiModal dataset.



Examples from the Parts to Whole dataset, used in the development of the curated data for CompleteMe. Source: https://huanngzh.github.io/Parts2Whole/

The authors state:

> *'To meet our requirements, we [rebuilt] the training pairs by using occluded images with multiple reference images that capture various aspects of human appearance along with their short textual labels.*
>
> *'Each sample in our training data includes six appearance types: upper body clothes, lower body clothes, whole body clothes, hair or headwear, face, and shoes. For the masking strategy, we apply 50% random grid masking between 1 to 30 times, while for the other 50%, we use a human body shape mask to increase masking complexity.*
>
> *'After the construction pipeline, we obtained 40,000 image pairs for training.'*

Rival prior *non-reference* methods tested were Large occluded human image completion (LOHC) and the plug-and-play image inpainting model BrushNet; reference-based models tested were Paint-by-Example; AnyDoor; LeftRefill; and MimicBrush.

The authors began with a quantitative comparison on the previously-stated metrics:

Method	CLIP-I ↑	CLIP-T ↑	DINO ↑	DreamSim [6] ↓	LPIPS ↓	PSNR ↑	SSIM ↑
LOHC [41]	96.03	29.46	82.52	0.0732	0.0709	28.4884	0.9264
BrushNet [12]	95.90	**30.69**	95.08	0.0576	0.0600	28.5764	0.9224
Paint-by-Example [35]	95.04	29.79	94.98	0.0611	0.0601	28.6441	0.9222
AnyDoor [4]	89.65	28.14	88.80	0.1454	0.0812	28.1807	0.9089
LeftRefill [1]	96.33	29.74	95.12	0.0574	0.0598	**28.8657**	**0.9283**
MimicBrush [3]	96.98	29.48	94.37	0.0651	0.0694	28.3598	0.9174
CompleteMe (Ours)	**97.18**	29.83	**96.29**	**0.0419**	**0.0588**	28.7020	0.9239

Results for the initial quantitative comparison.

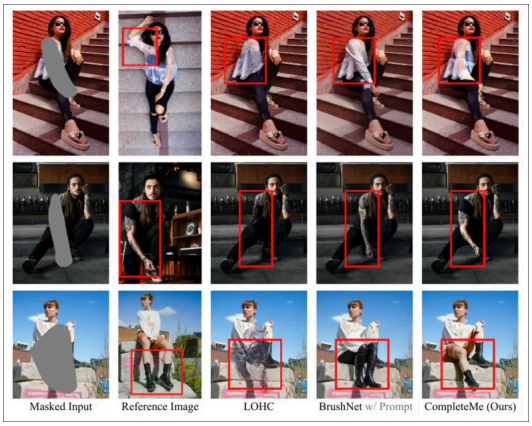
Regarding the quantitative evaluation, the authors note that CompleteMe achieves the highest scores on most perceptual metrics, including CLIP-I, DINO, DreamSim, and LPIPS, which are intended to capture semantic alignment and appearance fidelity between the output and the reference image.

However, the model does not outperform all baselines across the board. Notably, BrushNet scores highest on CLIP-T, LeftRefill leads in SSIM and PSNR, and MimicBrush slightly outperforms on CLIP-I.

While CompleteMe shows consistently strong results overall, the performance differences are modest in some cases, and certain metrics remain led by competing prior methods. Perhaps not unfairly, the authors frame these results as evidence of CompleteMe's balanced strength across both structural and perceptual dimensions.

Illustrations for the qualitative tests undertaken for the study are far too numerous to reproduce here, and we refer the reader not only to the source paper, but to the extensive supplementary PDF, which contains many additional qualitative examples.

We highlight the primary qualitative examples presented in the main paper, along with a selection of additional cases drawn from the supplementary image pool introduced earlier in this article:

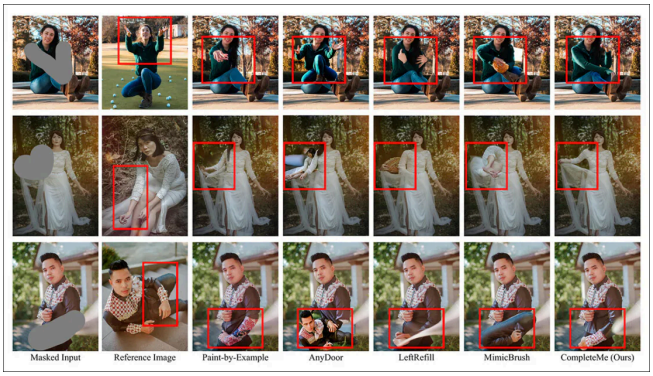


Initial qualitative results presented in the main paper. Please refer to the source paper for better resolution.

Of the qualitative results displayed above, the authors comment:

> *'Given masked inputs, these non-reference methods generate plausible content for the masked regions using image priors or text prompts.*
>
> *'However, as indicated in the Red box, they cannot reproduce specific details such as tattoos or unique clothing patterns, as they lack reference images to guide the reconstruction of identical information.'*

A second comparison, part of which is shown below, focuses on the four reference-based methods Paint-by-Example, AnyDoor, LeftRefill, and MimicBrush. Here only one reference image and a text prompt were provided.



Qualitative comparison with reference-based methods. CompleteMe produces more realistic completions and better preserves specific details from the reference image. The red boxes highlight areas of particular interest.

The authors state:

> *'Given a masked human image and a reference image, other methods can generate plausible content but often fail to preserve contextual information from the reference accurately.*
>
> *'In some cases, they generate irrelevant content or incorrectly map corresponding parts from the reference image. In contrast, CompleteMe effectively completes the masked region by accurately preserving identical information and correctly mapping corresponding parts of the human body from the reference image.'*

To assess how well the models align with human perception, the authors conducted a user study involving 15 annotators and 2,895 sample pairs. Each pair compared the output of CompleteMe against one of four reference-based baselines: Paint-by-Example, AnyDoor, LeftRefill, or MimicBrush.

Annotators evaluated each result based on the visual quality of the completed region and the extent to which it preserved identity features from the reference – and here, evaluating overall quality and identity, CompleteMe obtained a more definitive result:

Evaluation	Quality	Identity
Method	CompleteMe	
Paint-by-Example [35]	6.12%/**93.88%**	3.48%/**96.52%**
AnyDoor [4]	0.55%/**99.45%**	1.13%/**98.87%**
LeftRefill [1]	14.97%/**85.03%**	4.58%/**95.42%**
MimicBrush [3]	5.14%/**94.86%**	6.05%/**93.95%**

Results of the user study.

Conclusion

If anything, the qualitative results in this study are undermined by their sheer volume, since close examination indicates that the new system is a most effective entry in this relatively niche but hotly-pursued area of neural image editing.

However, it takes a little extra care and zooming-in on the original PDF to appreciate how well the system adapts the reference material to the occluded area in comparison (in nearly all cases) to prior methods.

We strongly recommend the reader to carefully examine the initially confusing, if not overwhelming avalanche of results presented in the supplementary material.

** It is interesting to note how the now severely-outmoded V1.5 release remains a researchers' favorite – partly due to legacy like-on-like testing, but also because it is the least censored and possibly most easily trainable of all the Stable Diffusion iterations, and does not share the censorious hobbling of the FOSS Flux releases.*

† *VRAM spec not given – it would be either 40GB or 80GB per card.*

First published Tuesday, April 29, 2025

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