Image Blending Algorithm with Automatic Mask Generation

Haochen Xue[1], Mingyu Jin[2], Chong Zhang[1], Yuxuan Huang[1], Qian Weng[1], and Xiaobo Jin[1](✉)

[1] Department of Intelligent Science, School of Advanced Technology,
Xi'an Jiaotong-Liverpool University, Suzhou, China
{Haochen.Xue20,Chong.zhang19,Yuxuan.Huang2002,
Qian.Weng22}@student.xjtlu.edu.cn, Xiaobo.Jin@xjtlu.edu.cn

[2] Electrical and Computer Engineering Northwestern University, Evanston, IL, USA
u9o2n2@u.northwestern.edu

Abstract. In recent years, image blending has gained popularity for its ability to create visually stunning content. However, the current image blending algorithms mainly have the following problems: manually creating image blending masks requires a lot of manpower and material resources; image blending algorithms cannot effectively solve the problems of brightness distortion and low resolution. To this end, we propose a new image blending method with automatic mask generation: it combines semantic object detection and segmentation with mask generation to achieve deep blended images, while based on our proposed new saturation loss and two-stage iteration of the PAN algorithm to fix brightness distortion and low-resolution issues. Results on publicly available datasets show that our method outperforms other classical image blending algorithms on various performance metrics including PSNR and SSIM.

Keywords: Image Blending · Mask Generation · Image Segmentation · Object Detection

1 Introduction

Image blending is a versatile technique used in various applications [19,20] where different images must be combined to create a unified and visually appealing final image. It involves taking a selected part of an image (usually an object) and seamlessly integrating it into another image at a specified location. The ultimate goal of image fusion is to obtain a uniform and natural composite image. This task poses two significant challenges: relatively low localization accuracy in

H. Xue and M. Jin—Equal contribution.

B. Luo et al. (Eds.): ICONIP 2023, CCIS 1962, pp. 234–248, 2024.
https://doi.org/10.1007/978-981-99-8132-8_18

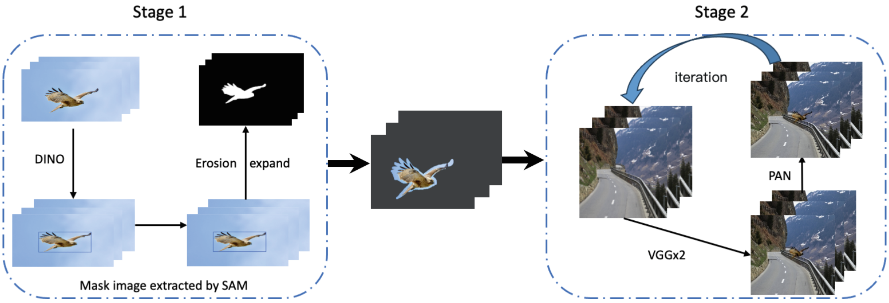


Fig. 1. Image blending process based on automatic mask generation: the first stage generates basic blending results through SAM; the second stage generates more detailed blending results after fusing PAN

cropped regions of objects and consistency issues between cropped objects and their surroundings.

GP-GAN and Poisson image editing is currently popular image blending methods [16]. In this method, the user selects an object in the source image with an associated mask to generate a high-resolution image and uses GP-GAN or Poisson to generate high-quality versions of the source and target images. However, the images generated by GP-GAN and Poisson image editing are not too realistic, where the composite image suffers from brightness distortion and often exhibits excessive brightness in small pixel clusters, thus compromising the overall realism of the image.

All image fusion algorithms require a mask as input to cut out the object to be fused from the source image, but the mask images of the previous algorithms are all handmade. These handcrafted mask images are insufficient to accurately represent the location of the foreground, which may lead to poor image fusion. Traditional segmentation methods for automatically generating masks mainly include RCNN [5], which was subsequently replaced by more powerful methods, such as the Segment Anything Model (SAM) method proposed by Meta [8]. However, SAM has its limitations in image blending, as it tends to capture all objects [17] in a particular picture, whereas image blending requires a mask for one specific object in an image.

In our work, we reconstruct the deep image blending algorithm [20] using Pixel Aggregation Network (PAN) and a new loss function, which iteratively improves the image blending process [11,15]. To address the limitations of artificial clipping masks, we apply DINO and use target text to distinguish our desired objects, resulting in better image blending. However, there remains a potential problem here that other researchers may not have noticed, namely that precise segmentation of objects may not always yield the best results. The blended image may lose important details if the mask image does not contain relevant information about the original image. We apply a classic erosion-dilation step to address this challenge, which helps preserve important details in the original image for better-blending results. Evaluation metrics including PSNR, SSIM, and MSE on multiple image datasets show that our hybrid image can outperform previous models GP-GAN, Poisson Image, etc. Our experiments show that combining DINO (DETR with improved denoising anchor boxes) [18] and SAM can generate more accurate masks than RCNN. The images generated by the hybrid algorithm have consistent brightness, higher resolution rate and smoother gradients. The whole process can be simple as shown in Fig. 1.

Our work has mainly contributed to the following three aspects:

- We propose an automatic mask generation method based on object detection and SAM segmentation, where erosion and dilation operations are used to manipulate the resulting mask for better image blending.
- We propose a new loss function, called saturation loss, for deep image blending algorithms to account for sudden changes in contrast at the seams of blended images.
- We use PAN to process blended images, solving the problem of low image resolution and distortion of individual pixel grey values of blended images.

The remainder of this paper is organized as follows: in Sect. 2, we introduce previous related work on image segmentation and detection and image blending; Sect. 3 gives a detailed introduction to our method; in Sect. 4, our algorithm will be compared with other algorithms. Subsequently, we summarize our algorithm and possible future research directions.

2 Related Work

2.1 Image Blending

The simplest approach to image blending (Copy-and-paste) is to directly copy pixels from the source image and paste them onto the destination image. Still, this technique can lead to noticeable artefacts due to sudden intensity changes at the composition's boundaries. Therefore, researchers have proposed advanced methods that use complex mathematical models to integrate source and destination images and improve the overall aesthetics of the blended image.

A traditional approach to image blending is Poisson image editing, first cited by Perez et al. [13], which exploits the concept of solving Poisson's equation to blend images seamlessly and naturally. This method transforms the source and target images into the gradient domain, thus obtaining the gradient of the target image. Another image blending technique is the gradient domain blending algorithm, proposed by Perez et al. [10]. The basic idea is to decompose the source and target images into gradient and Laplacian domains and combine them using weighted averaging. Deep Image Blending [20] refers to the gradient in Poisson image editing, and turns the gradient into a loss function, plus the loss of texture and content fidelity, resulting in higher image quality than Position image editing. Our work further optimizes Deep Image Blending to generate more realistic images.

Image inpainting is a technique [9] that uses learned semantic information and real image statistics to fill in missing pixels in an image, done by deep learning models trained on large datasets. However, this technique does not perform well in large-scale image mixing. Besides image blending, several other popular image editing tools include image denoising, image super-resolution, image inpainting, image harmonization, style transfer, etc. With the rise of generative adversarial networks (GANs), these editing tasks have become increasingly important in improving the quality of generated results, such as GP-GAN [1,2,7]. Image super-resolution involves using deep learning models to learn image texture patterns and upsampling low-resolution images to high-resolution images. PAN is often used to continuously refine and enhance the details in the image through a series of iterations. This process involves leveraging the power of neural networks to make predictions and complement the image's resolution. This trained model then predicts high-resolution details missing from the low-resolution input. We use PAN to increase the image's resolution and make the blending image more realistic.

2.2 Image Segmentation and Detection

In the past, Regions with CNN Features (RCNN) [5] was the best region-based method for semantic segmentation based on object detection. RCNN can be used on top of various CNN structures and shows significant improvements over traditional CNN structures. Fast R-CNN is an improved version of RCNN that improves speed and accuracy by sharing the feature extraction process. The YOLO [14] algorithm treats the target detection task as a regression problem and realizes real-time target detection by dividing the image into grids and predicting the bounding box and category in each grid cell. Still, the accuracy of its detection target localization is relatively low.

DINO is an advanced object detection and segmentation framework used by our pipeline to identify the most important objects from segmented images via SAM [4]. DINO introduces improved anchor box and mask prediction branches to implement a unified framework to support all image segmentation tasks, including instance, bloom and semantic segmentation. Mask DINO is an extension of DINO that leverages this architecture to support more general image segmentation tasks. The model is trained end-to-end on a large-scale dataset and can accurately detect and segment objects in complex scenes. Mask DINO extends DINO's architecture and training process to support image segmentation tasks, making it an effective tool for segmentation applications.

Essentially, an ideal image segmentation algorithm should be able to recognize unknown or new objects by segmenting them from the rest of the image. SegNet [3] achieves pixel-level image segmentation through an encoder-decoder structure, which is difficult to handle small objects and requires a lot of computation. The core idea of CRFasRNN [12] is to combine conditional random fields (CRF) and recurrent neural networks (RNN). Compared with SegNet, CRFasRNN has less calculation and finer segmentation results. Facebook Meta AI has developed a new advanced artificial intelligence model called Segmented Anything Model (SAM) [8] that can extract any object in an image with a single click. SAM leverages cutting-edge deep learning techniques and computer vision algorithms to accurately segment any object in an image. SAM can efficiently cut out objects from any type of image, making the segmentation process faster and more precise. This new technology is a breakthrough in computer vision and image processing because it can save a lot of time and effort when editing images and videos.

3 Proposed Approach

In this section, we first describe how to automatically generate a synthetic mask; then how to blend the source and target images to produce the initial result; finally, we propose a new saturation loss to refine the blended result image. The overall framework is shown in Fig. 2.

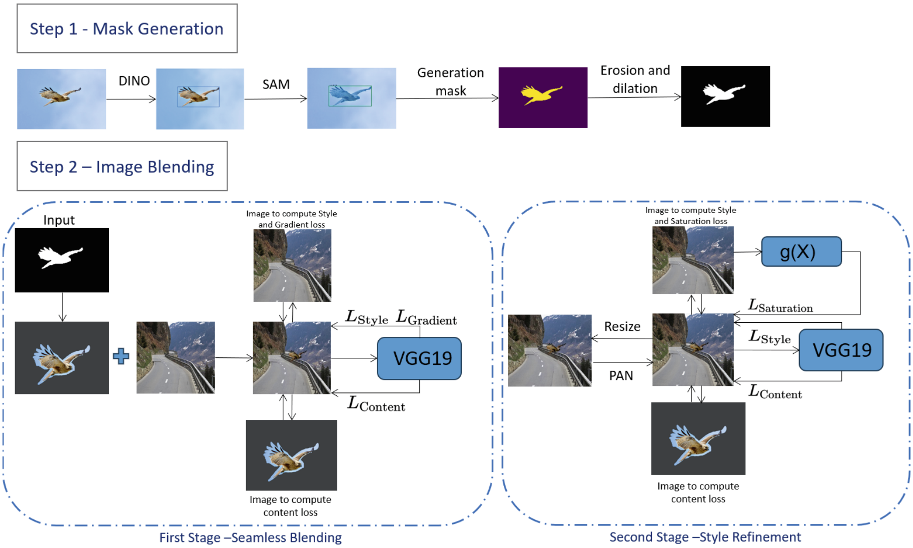

Step 1 - Mask Generation
DINO
SAM
Generation mask
Erosion and dilation
Step 2 – Image Blending
Input
Image to compute Style and Gradient loss
L_{Style} L_{Gradient}
VGG19
L_{Content}
Image to compute content loss
First Stage –Seamless Blending
Image to compute Style and Saturation loss
g(X)
$L_{\text{Saturation}}$
Resize
PAN
L_{Style}
VGG19
L_{Content}
Image to compute content loss
Second Stage –Style Refinement

Fig. 2. The overall framework of our method: 1) generate the mask of the object through DINO and SAM; 2) use VGG19 to optimize the content loss, gradient loss and style loss to obtain an initial blending image; 3) integrate the PAN algorithm, and replace the gradient loss at the same time into saturation loss for more realistic blending image

3.1 Mask Generation

Below we describe in detail how to automatically generate high-quality masks. First, we use DINO to detect a specific region in an image based on a textual description and generate a box around that object, as shown in Fig. 4 (the input word is "bird"). We then feed the frame into the SAM and extract the mask of the region. Combining DINO and SAM, we solve the problem that SAM can only segment objects but cannot select specific objects. In Fig. 5, we can observe that our algorithm can precisely identify the desired object in the image and generate an accurate mask. This method saves time and effort compared to traditional manual editing methods. It is worth noting that after obtaining the yellow-purple Mask, the Mask image needs to be converted into a black-and-white image.

In object detection, DINO has better performance than RCNN due to its self-supervised learning method and the advantages of the Transformer network, which enables it to better capture global features. For semantic segmentation, SAM can better capture key information in images and achieve more accurate pixel-level object segmentation through its multi-scale attention mechanism and attention to spatial features. Therefore, the mask generation achieved by the combination of these two methods outperforms traditional convolutional neural networks, as shown in Fig. 3. We use IOU to measure the quality of the mask. It can be seen from the figure that the combination of DINO and SAM not only has a better segmentation effect visually, but also outperforms RCNN in terms of IOU.

Finally, an erosion and dilation operation [6] is performed on the mask to better refine the mask. The overall process of the mask operation is as follows: First, an etch operation is applied to shrink the sharp and edge areas of the mask. Second, a dilation operation is performed on the eroded mask to expand its edges to ensure that the mask completely covers the target object and maintains a smooth boundary. Through the above operations on the mask, we can improve the coverage of the mask and make the mask input of the image blending algorithm more accurate. In the image fusion stage, the processed mask can also carry part of the source image information, making the final fused image more natural.

3.2 Seamless Blending

Seamless blending is the first stage of deep image mixing [20], the style loss L_{style} is used to transfer the style information of the background image to the resulting image, it can calculate the texture difference between the generated image and the background, making the generated image style unified, more harmonious and authentic. The content loss L_{cont} measures the pixel difference between the object in the fusion image and the object in the source image, which is used to ensure the fidelity of the content in the image, and avoid the content smearing caused by the style transfer and cause the loss of details. Gradient loss L_{grad} is used to compute the pixel-wise difference between the source image and the target image on the edge for smooth blending of edges. At this stage, through continuous iteration, the fusion edge will gradually become smoother, and the texture of the fusion object will gradually resemble the background image without losing any details. Specially, we will optimize the following loss function in the first stage

$$\mathcal{L}_1 = \lambda_{grad}L_{grad} + \lambda_{cont}L_{cont} + \lambda_{style}L_{style}, \tag{1}$$

where λ_{grad}, λ_{cont} and λ_{style} respectively represent the weight of each loss.

After the first stage of processing, the blended edge of the object in the blending image is very smooth, but there are still significant differences between the blended object and the background in terms of similarity and illuminance, which may affect the quality and realism of the image. To solve this problem, we need to continue to optimize the image in terms of style refinement.

3.3 Style Refinement

In the second stage, although we use the same network architecture, it achieves a different task: to achieve the consistency of the background and source image in the blended image and to ensure that the generated image is more realistic, so we propose the new Saturation loss to replace the previous gradient loss, which computes the difference in saturation gradient between the background image and the blended image to account for blended images with unrealistic lighting and large contrast differences in the medium. The texture of the object in the final generated result image is consistent with the source image.

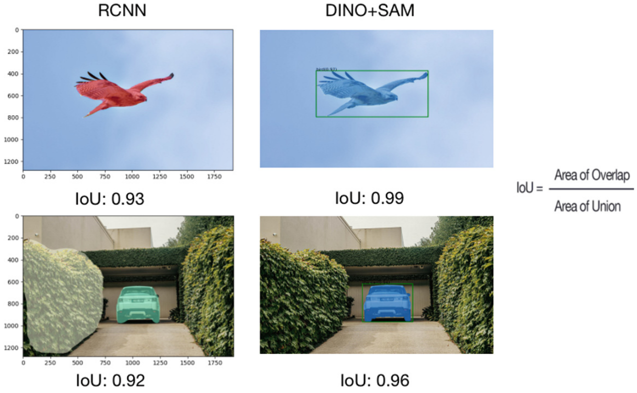
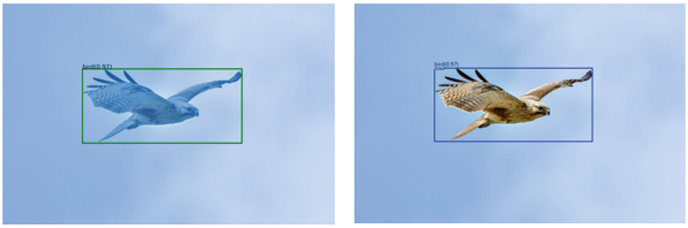


Fig. 3. Comparison of traditional RCNN algorithm and DINO+SAM algorithm in segmentation tasks

Since the basic brightness of the mixed background image is different from that of the target image, there will be a certain contrast difference after mixing, so that the naked eye can perceive the existence of the mixing operation. At different colour coordinates, each layer of the fused image behaves differently. We observed obvious differences at the fusion seams of R, G, and B channels under RGB colour coordinates. However, after converting the fused image to the HSV colour model, the Saturation channel of the blended image will have a sudden change in the saturation value at the edge where the source image and the target image are blended, as shown in Fig. 6.

To solve the above-mentioned sudden saturation problem of the blending boundary, we propose a new saturation loss to make the saturation change of



DINO's Frame

Mask image extracted by SAM

Fig. 4. Object detection by DINO and segmentation by SAM

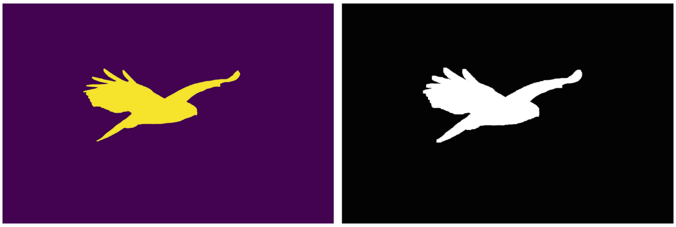
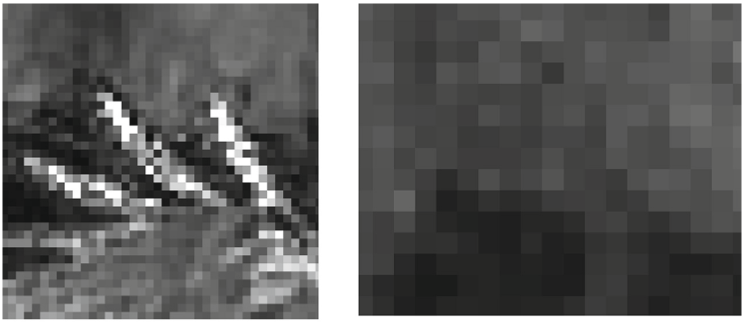


Fig. 5. Mask generation by SAM: 1) left: before dilation-corrosion operation; 2) right: after dilation-corrosion operation



Blended Image

Original Image

Fig. 6. Pixel comparison of the S channel of both the fusion image and the original image on the blending boundary

the blended image more realistic and natural. The detailed process is shown in the Fig. 7: First, we convert the RGB colour coordinates of the background image and the blended image to HSV colour coordinates, and extract their saturation channels; then calculate the saturation gradient difference of the mixed image and the background image (H and W are height and width respectively)

$$\mathcal{L}_{sat} = \frac{g(M) - g(B)}{HW}, \tag{2}$$

where the function $g(X)$ represents the sum of the gradient values along the row and column directions on the saturation channel X_s of the image X

$$g(X) = \sum_{i=1}^{H}\sum_{j=1}^{W} |X_s(i+1,j) - X_s(i,j)| + |X_s(i,j+1) - X_s(i,j)|. \tag{3}$$

. Finally, we optimize the following loss under the original framework

$$\mathcal{L}_2 = \lambda_{sat}L_{sat} + \lambda_{cont}L_{cont} + \lambda_{style}L_{style}, \tag{4}$$

where λ_{sat}, λ_{cont}, λ_{style} are the weight coefficients of each item.

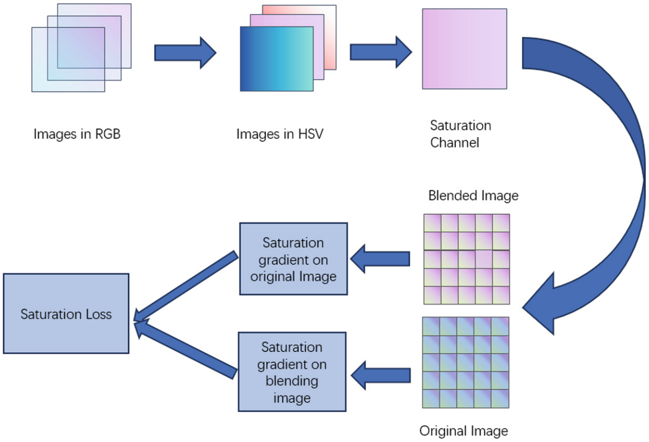


Fig. 7. Calculation steps of saturation loss: 1) convert the image from RGB space to HSV space and extract the S channel; 2) calculate the saturation gradient on the original image and the blending image respectively; 3) solve the average difference of the two saturation gradients value

4 Experiments

In this section, we describe the experimental setup, compare our method with other classical methods, and conduct ablation experiments on our method. The SSIM and PSNR scores of each image are tested multiple times to ensure data stability.

4.1 Experiment Setup

The experimental settings are shown in Table 1 below. In the loss function, the weight of the gradient loss in the first stage is set higher, because the main goal of the first stage is to solve the gradient problem and make the blending edge smoother. In the second stage, in order to solve the lighting and texture problems of the fused image, the weight of the style loss and saturation loss functions is set to 10^5.

Table 1. Experimental hyperparameter settings

Parameter	λ_{grad}	λ_{style}	λ_{cont}	λ_{sat}
Stage 1	10^4	10^3	1	0
Stage 2	0	10^5	1	10^5

4.2 Result Comparison

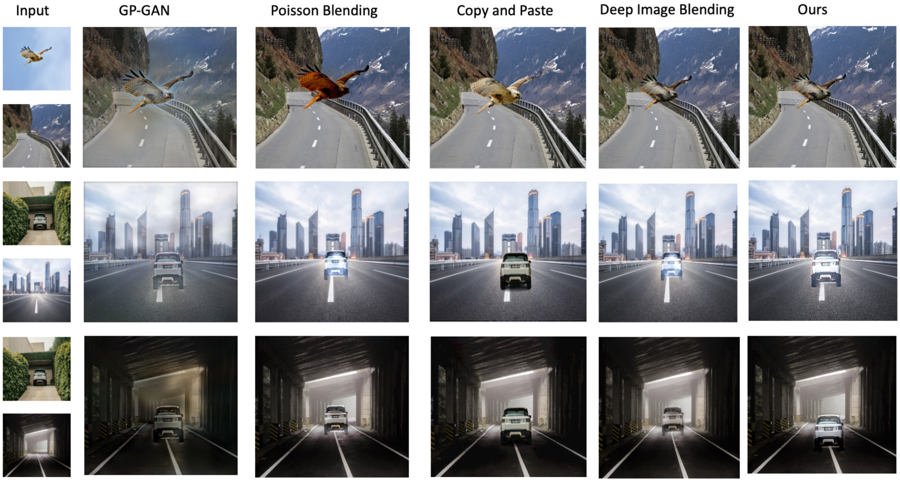


Fig. 8. Comparison of the effect of image blending between our method and other methods on the same input

As shown in Fig. 8, the size of all the images is 512 × 512, the copy-paste method copies the source image to the corresponding location on the destination. Deep reconciliation requires training a neural network to learn visual patterns from a set of images and use them to create realistic compositions or remove unwanted elements. Poisson blending computes the gradients of two images and minimizes their difference to seamlessly blend two images. The technique preserves the overall structure of an image while describing the flow of colours from one image to another. Finally, GP-GAN is a generative adversarial network (GAN) that uses a pre-trained generator network to generate high-quality images similar to the training data. Generator networks are pretrained on large image datasets and then fine-tuned on smaller datasets to generate higher-quality images. Unfortunately, these methods produce results with unrealistic borders and lighting. As can be seen from the qualitative results in Fig. 8, our algorithm produces the most visually appealing results for mixing borders, textures, and colour lighting.

Copy-paste methods for image fusion require precise control over the alignment of the target image to the background image, otherwise, it will lead to obvious incongruity and artefacts, while the inconsistency of image style will make the result have obvious artificial boundaries. GP-GAN produces worse visual results under mixed boundary and coloured lighting, where overall colours are darker and edges are not handled well. While it brings rich colour to the raw edges of an image, it introduces inconsistencies in style and texture. Compared with these algorithms, our method erodes and dilates edges, which prevents jagged edges or visible artefacts, and the two-stage algorithm adds more style and texture to the blended image.

Furthermore, we refer to the experimental protocol of the deep hybrid image algorithm and conduct comparative experiments on 20 sets of data. We compared the results of these methods on indicators such as PSNR (peak signal-to-noise ratio), SSIM (structural similarity) and MSE (mean square error), as shown in Table 2. It can be seen that the average performance of our method achieves the best results on PNSR and SSIM, while MSE is slightly worse than Poisson Blending. This is mainly because our method does not simply migrate the source image to the target image, but further refines the blended result of style and saturation consistency to make the generated picture more realistic, resulting in a slightly larger fitting error MSE. Compared with images from Deep Image Blending, images generated by our optimized model perform better in terms of PSNR, SSIM, and MSE, which also illustrates the superiority of our model.

Table 2. Quantitative comparison of average results between our method and other methods on PSNR, SSIM and MSE metrics

Method	PNSR	SSIM	MSE
GP-GAN	19.94	0.73	833.81
Poisson Blending	21.77	0.71	**472.46**
Deep Image Blending	22.11	0.72	712.64
Copy and Paste	17.98	0.57	866.41
Ours	**23.59**	**0.78**	617.51

4.3 Ablation Study

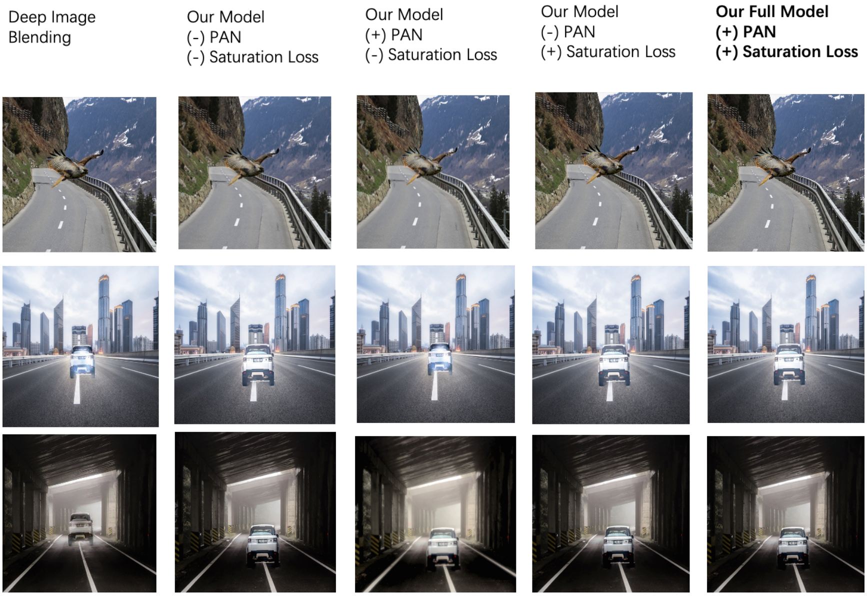

Deep Image Blending
Our Model (-) PAN (-) Saturation Loss
Our Model (+) PAN (-) Saturation Loss
Our Model (-) PAN (+) Saturation Loss
Our Full Model (+) PAN (+) Saturation Loss

Fig. 9. The results (512 × 512) of the ablation experiment: (+) and (-) respectively indicate that a certain part of the algorithm participates or does not participate

We take three images as an example to conduct ablation experiments to analyze the role of PAN composition and saturation loss in our method. From Fig. 9, we can observe:

PAN in blending refinement We will keep only the saturation loss and remove the PAN component from the model. Experimental results show that the image resolution will be obviously reduced, resulting in a loss of clarity.

Saturation loss in the second stage If you remove the saturation loss and keep the PAN module, you will find that the original image still maintains the original saturation in the resulting image, which is not consistent with the background.

Both PAN and saturation loss When using both components at the same time, you will find that the generated results are more realistic, and the style of the source and target images is more consistent, especially in the blending edge area.

Table 3. Quantitative results of our algorithm's ablation experiments on 3 sets of images, where the three values in each cell represent the results on different images

Metrics	PSNR	SSIM	MSE
Baseline	17.85/18.32/17.99	0.57/0.61/0.67	718.20/661.94/594.32
+PAN	20.77/20.16/22.09	**0.74**/0.69/0.79	**543.98**/721.69/**401.49**
+Saturation Loss	19.79/19.94/22.29	0.6/0.62/0.81	681.91/658.20/383.38
+PAN+Saturation Loss	**29.57/24.58/23.64**	0.72/**0.83/0.79**	612.95/**568.47**/402.93

Table 3 further gives the quantitative results of the ablation experiments. The three numbers in the grid represent the experimental results for 3 images. As far as PNSR is concerned, the combination of PAN+Saturation performed best. Regarding SSIM, just using PAN results may be better. Finally, MSE using both PAN and Saturation does not always perform best, since our method does not simply fit a merge of source and target images.

5 Conclusion and Future Work

In our work, we address the low accuracy and low efficiency of manually cutting masks by generating masks through object detection and segmentation algorithms. Specifically, we combine DINO and SAM algorithms to generate masks. Compared with the traditional RCNN algorithm, the mask of this algorithm can cover objects better and has a stronger generalization ability. We perform erosion and dilation operations on the mask to avoid sharp protrusions in the mask. However, there may be a limitation with this part. If one object A overlaps with another object B, performing image blending after corroding and expanding the mask of A may introduce B's information into the blending process, which may lead to unreal results. Finally, we also propose a new loss, called saturation loss, to address brightness distortion in generated images. Results on multiple image datasets show that our method can outperform previous image fusion methods GP-GAN, Poisson Image, etc. Future work includes proposing new evaluation criteria to better reflect human perception and aesthetics to improve the objectivity and accuracy of the model. Another potential research direction is how to deal with object occlusion in image fusion.

6 Authorship Statement

Haochen Xue proposed the idea of automatic mask generation and constructed the overall framework of the image fusion project. Mingyu Jin uses PAN to process images, resulting in higher-quality images. Chong Zhang proposed a new loss and used erosive dilation to optimize the mask. Yuxuan Huang completed the comparison experiment and ablation experiment. Qian Weng participated in the revision of the article and undertook the polishing of the article. Xiaobo Jin supervised this work and made a comprehensive revision and reconstruction.

Acknowledgments. This work was partially supported by the "Qing Lan Project" in Jiangsu universities, National Natural Science Foundation of China under No. U1804159 and Research Development Fund with No. RDF-22-01-020.

References

1. Alsaiari, A., Rustagi, R., Thomas, M.M., Forbes, A.G., et al.: Image denoising using a generative adversarial network. In: 2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT), pp. 126–132. IEEE (2019)
2. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. arXiv:1701.04862 (2017)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
4. Chen, J., Yang, Z., Zhang, L.: Semantic segment anything (2023). www.github.com/fudan-zvg/Semantic-Segment-Anything
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
6. Gonzalez, R.C.: Digital image processing. Pearson Education, India (2009)
7. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
8. Kirillov, A., et al.: Segment anything. arXiv:2304.02643 (2023)
9. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
10. Leventhal, D., Gordon, B., Sibley, P.G.: Poisson image editing extended. In: Finnegan, J.W., McGrath, M. (eds.) International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2006, Boston, Massachusetts, USA, July 30 - August 3, 2006, Research Posters, p. 78. ACM (2006)
11. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
12. Monteiro, M., Figueiredo, M.A.T., Oliveira, A.L.: Conditional random fields as recurrent neural networks for 3d medical imaging segmentation. arXiv:1807.07464 (2018)
13. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Trans. Graph. **22**(3), 313–318 (2003)
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
15. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: few-shot image semantic segmentation with prototype alignment. In proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)
16. Wu, H., Zheng, S., Zhang, J., Huang, K.: GP-GAN: towards realistic high-resolution image blending. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2487–2495 (2019)
17. Xie, G., et al.: SAM: self-attention based deep learning method for online traffic classification. In: Proceedings of the Workshop on Network Meets AI & ML, pp. 14–20 (2020)

18. Zhang, H., et al.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv:2203.03605 (2022)
19. Zhang, H., Han, X., Tian, X., Jiang, J., Ma, J.: Image fusion meets deep learning: a survey and perspective. Inf. Fusion **76**, 323–336 (2021)
20. Zhang, L., Wen, T., Shi, J.: Deep image blending. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 231–240 (2020)