remote sensing

MDPI

Article

Residual Attention Mechanism for Remote Sensing Target Hiding

Hao Yuan [1], Yongjian Shen [2], Ning Lv [3], Yuheng Li [1], Chen Chen [4,5,6,*] and Zhouzhou Zhang [1]

[1] Beijing Research Institute of Telemetry, Beijing 100076, China; haoyuan_1@stu.xidian.edu.cn (H.Y.); liyuheng@mail.nwpu.edu.cn (Y.L.); 18901257065zhang-zz14@tsinghua.org.cn (Z.Z.)
[2] Electronic Information Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China; shenyongshen@buaa.edu.cn
[3] School of Electronic Engineering, Xidian University, Xi'an 710071, China; nlv@mail.xidian.edu.cn
[4] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
[5] Xidian Guangzhou Institute of Technology, Guangzhou 510555, China
[6] Science and Technology on Communication Networks Laboratory, Shijiazhuang 050000, China
* Correspondence: cc2000@mail.xidian.edu.cn

Abstract: In this paper, we investigate deep-learning-based image inpainting techniques for emergency remote sensing mapping. Image inpainting can generate fabricated targets to conceal real-world private structures and ensure informational privacy. However, casual inpainting outputs may seem incongruous within original contexts. In addition, the residuals of original targets may persist in the hiding results. A Residual Attention Target-Hiding (RATH) model has been proposed to address these limitations for remote sensing target hiding. The RATH model introduces the residual attention mechanism to replace gated convolutions, thereby reducing parameters, mitigating gradient issues, and learning the distribution of targets present in the original images. Furthermore, this paper modifies the fusion module in the contextual attention layer to enlarge the fusion patch size. We extend the edge-guided function to preserve the original target information and confound viewers. Ablation studies on an open dataset proved the efficiency of RATH for image inpainting and target hiding. RATH had the highest similarity, with a 90.44% structural similarity index metric (SSIM), for edge-guided target hiding. The training parameters had 1*M* fewer values than gated convolution (Gated Conv). Finally, we present two automated target-hiding techniques that integrate semantic segmentation with direct target hiding or edge-guided synthesis for remote sensing mapping applications.

Keywords: emergency remote sensing mapping; image inpainting; residual attention mechanism; target hiding

Citation: Yuan, H.; Shen, Y.; Lv, N.; Li, Y.; Chen, C.; Zhang, Z. Residual Attention Mechanism for Remote Sensing Target Hiding. *Remote Sens.* **2023**, *15*, 4731. https://doi.org/10.3390/rs15194731

Academic Editors: Wei Chen, Paraskevas Tsangaratos, Ioanna Ilia and Haoyuan Hong

Received: 22 July 2023
Revised: 18 September 2023
Accepted: 20 September 2023
Published: 27 September 2023

1. Introduction

Remote sensing images contain abundant surface features that can support governments and rescue agencies in emergency decision making [1], disaster assessment, and rescue deployment [2]. There are many research projects and application directions of remote sensing mapping for urban development planning and emergency disaster response. To protect privacy, especially that of personal buildings, sensitive targets must be processed before public release and use. Current methods mainly rely on the manual or semiautomatic labeling of sensitive targets and image editing tools to cover and fill target areas. These methods cannot meet the timeliness requirements of mapping tasks. Moreover, the results depend on the operators' skills and thus lack control. Therefore, an automatic method for sensitive target hiding is needed.

Artificial intelligence approaches have become indispensable for remote sensing data analysis, thus catalyzing breakthroughs across the field [3–5]. A new multisensor dataset of very-high-resolution satellite imagery from diverse landscapes was recently introduced

for the super-resolution restoration of high-resolution (HR) remote sensing images from low-resolution (LR) inputs [6]. In addition, to automatically capture the positions and contours of sensitive targets, semantic segmentation was introduced [7]. Its process is the same as that of remote sensing interpretation tasks [8]. After detecting sensitive targets, Qiu et al. proposed an image inpainting model to remove and fill target areas [7]. This method combines object detection with image inpainting to achieve the fast, automatic hiding of sensitive targets. Image inpainting is closely related to hiding, but the latter aims to completely hide targets.

The image inpainting model known as contextual attention (Cont Atten) [9], which was used in the combined method, aims to hide sensitive targets. Cont Atten is a coarse-to-refinement image inpainting network. Its efficiency is limited by using regular masks, which are exactly the missing areas of the images, in the training step. For this reason, a new model named Gated Convolution (Gated Conv) [10] that trains on free-form masks with better inpainting performance was proposed. Yu et al. used masks combining irregular and regular masks [10] to reduce the computation of hard-gating masks proposed by Partial Conv [11]. Gated convolution is another innovation of their research. It provides a learnable, dynamic feature selection mechanism at both the channel and spatial location levels. Their work provided significant inspiration for our research. However, the extensive use of elementwise products can cause unstable gradients. The method of generating two branches with one convolution operation also has the problem of having too many parameters. Furthermore, as the coarse-to-refinement network permits only two chances of original image input, the inpainted objects can exhibit incongruence regarding the surrounding contextual semantics. The techniques that simply replace original targets with fabricated ones can still retain residual traces of the original information.

While residual connections can propagate original information more deeply, capturing object distributions is critical when synthesizing fake targets for believable target hiding. Thus, we introduce a Residual Attention Target-Hiding (RATH) model incorporating residual attention. Our key contributions are the following:

- We proposed a residual attention module that bifurcates the gated convolution into two branches utilizing concatenated convolutions. This extracts the features representing object distributions while enabling adjustable kernel sizes within the gated convolutions, thereby conferring greater flexibility. Additionally, the residual attention mechanism ameliorates gradient vanishing and explosion issues.
- To expand the fusion patch size, we substituted the complex operation with two 3×3 convolutional layers utilizing an all-in-one kernel. This can elevate low similarities based on neighboring element values, thus providing more global context.
- We extended the edge-guided approach [12] to synthesize fabricated targets with higher realism, thereby utilizing edges derived from semantic segmentation. This technique is better suited for hiding targets when provided with highly confounding artificial edges that match the target spatial distribution.
- Finally, we performed ablation experiments on benchmark datasets to validate the proposed RATH model, thus achieving a state-of-the-art structural similarity index metric (SSIM) of 90.44% for edge-guided [13] target hiding using fewer parameters than Gated Conv. Additionally, this paper presents two automated frameworks integrating semantic segmentation with direct or edge-guided target hiding for remote sensing mapping applications.

The remainder of this paper is organized as follows. Section 1 introduces the development of target hiding in emergency remote sensing mapping and the contributions of our research. Section 2 reviews the advancements in image inpainting using deep learning, which the target hiding is based on. Section 3 elaborates on the principal framework and methodology of the proposed approach. Section 4 presents extensive experiments on diverse datasets to evaluate image inpainting, target hiding, and edge-guided inpainting capabilities. Section 5 provides a discussion and introduces our automated application for

target hiding utilizing the proposed techniques. Finally, Section 6 concludes the paper and discusses directions for future work.

2. Relate Work

2.1. Target Hiding Based on Image Inpainting

To protect the privacy and the inviolability of personal buildings, information containing these structures must be anonymized during remote sensing mapping. Traditional approaches rely on manual identification and image editing, which prove to be inefficient. Recently, deep learning has assumed an increasingly prominent role in remote sensing for disaster management and mitigation. A new Conv-Trans Dual Network (CTDNet) based on Swin-Unet was proposed for landslide detection, which was motivated by the powerful global modeling capability of the Swin Transformer [14]. Additionally, a building target detection model integrating convolutional block attention modules (CBAM) into YOLO V5 [15] improved classification accuracy and detection speed. This CBAM-enhanced framework enhances performance for time-critical building detection tasks. Ref. [16] introduced a specially weighted crossentropy contour loss to constrain residual attention U-Nets for optical remote sensing interpretation. A deeply supervised generative adversarial network (D-sGAN) [17] was proposed to enable the semantic interpretation of remote sensing data. Such deep learning approaches help overcome limitations such as low timeliness and the inconsistent performance of conventional remote sensing image processing techniques.

The image inpainting network Cont Atten [9] calculates the contribution of external features to each location in the missing region, and it was applied to the inpainting areas after removing airplanes [7]. This pioneered an automated approach for target hiding. Despite poor performance in hiding irregularly shaped targets such as airplanes when trained on regular masks, their work provided key inspiration that image inpainting networks could be adapted for target hiding applications.

2.2. Image Inpainting

Image inpainting involves reconstructing lost or corrupted regions in images and videos, whereas target hiding entails removing and inpainting designated objects in images. Early image inpainting methods are relatively simple. Nitzberg et al. proposed an algorithm using image segmentation to remove the objects in front of the foreground [18]. Using the combined frequency with location information, Hirani and Totsuka selected a similar texture to fill the target areas [19]. This simple technology produced incredibly good results at that time. However, this technology is only responsible for analyzing image texture. Whether the texture is used or not depends on the users. The target area needs to be segmented by users, which is complex and time-consuming. In 1998, an algorithm based on Nitzberg's was proposed by Masnou and Morel [20]. The main idea was to perform the inpainting by connecting the points of equal rays (lines with equal gray values) that reached the boundary of the area to be inpainted, while the area needed to have a simple topology. Then, a new static image restoration algorithm was introduced by Colomba Ballister and Marcelo Bertalmio [21]. After the user selected the areas to be restored, the algorithm would automatically fill these areas with the information around them, which achieved considerable success.

In 2016, the first image inpainting model known as Context Encoder (CE) [22] and based on generative adversarial networks (GANs) [23] was proposed. The core idea is the channelwise fully connected layer, which is similar to the standard fully connected layer, but each channel handles its characteristics separately. Next, multiscale neural patch synthesis (MSNPS) [24] was regarded as the enhanced CE. The researchers introduced local texture loss to ensure that the fine details of the missing area were similar to other parts. Then, another classical model of image inpainting was Globally and Locally Consistent Image Completion (GLCIC) [25]. It used dilated convolution instead of the fully connected layer for a larger receptive field, and then the global discriminator and local discriminator were introduced in the training process. PGGAN [26] embedded residual mechanisms and

PatchGAN [27] in the GLCIC to enhance the performance. In contrast to the discriminator of GAN, the output of the PatchGAN discriminator was the matrix of the prediction labels. Shift-Net [28] introduced the shift–connection layer to U-Net for filling in missing regions of any shape with sharp structures and fine-detailed textures. In the same year, Partial Convolutions (Partial Conv) [11] used partial convolution to predict the area, and the prediction did not depend on the initial value of the hole. It was the first model to train in irregular masks. The results proved the effectiveness of the irregular mask training strategy.

Recently, an image inpainting model with adversarial edge learning known as EdgeConnect [29] was proposed. It divided the image inpainting task into two steps: edge prediction and image inpainting based on the edge. The second step is similar to the CGAN. Then, Gated Conv [10] followed the idea of EdgeConnect. The authors of Gated Conv extended the image inpainting to user-guided inpainting. By providing a sketch, the model would generate an image that had the same edge as the sketch. They also proposed a gated convolution to replace the convolution operation to learn the effectiveness of each feature in each location. The primary distribution of the recent methods is summarized in Table 1.

Table 1. Comparison of different approaches, including Cont Atten [9], Partial Conv [11], Gated Conv [10], and our approach.

Methods	Cont Atten	Partial Conv	Gated Conv	RATH (Ours)
Nonlocal	✓		✓	✓
Free-Form		✓	✓	✓
Edge-Guided			✓	✓
Residual Attention				✓

3. Method and Materials

3.1. Coarse-to-Refinement Network

While target hiding and image inpainting are similar, they produce distinct outputs. Image inpainting imposes no constraints on the modality of the generated image except for textural consistency. In contrast, target hiding treats the object as the foreground and the rest as the background. Its objective is to synthesize new images containing only the background. Therefore, training a target-hiding model primarily involves learning the textural characteristics of the background. In practice, however, the core capability of target hiding is filling missing regions with surrounding pixels. As such, when the missing area in the image inpainting corresponds to the target region in the target hiding, the two methods share the same goal.

The architecture of the proposed Residual Attention Target-Hiding (RATH) model is summarized in Figure 1. To obtain optimal hiding results, this paper removed targets before inputting images into the network such that the excised regions were excluded from midnetwork processing. This compelled the network to infer the missing sections based on the remaining pixels. This paper implemented a coarse-to-refinement network architecture akin to Gated Conv [10], which proved that the encoder–decoder architecture is better suited for image inpainting compared to the U-Net [30] employed in Partial Conv [11], especially when masks are centrally located. We replaced the gated convolution with residual attention modules, which are depicted in the red rectangle in the dashed box in Figure 1 to reduce calculations and parameters. The engineered module can also furnish original features postconvolution. Consequently, the inpainting outcomes more closely resemble the original images. In the bottom module of the coarse encoder–decoder, this paper opted for dilated networks to extract deeper features to inpaint the missing segments.

After the coarse network, a skip connection layer is implemented to furnish the input mask and the incomplete original images. The mask region in the coarse network output is retained, while the remaining region is replaced with the original image. Subsequently, the refinement network persists in honing the inpainted pixels, which contain two branches: one is the same as the coarse network, and the other uses contextual attention layers to fuse both the inpainting and original features. Furthermore, this paper expanded the

fusion patch to provide a more global context. The outcomes from the two branches undergo channelwise concatenation before connection with the identical decoder present in the coarse network. The discriminator employed in the RATH model adopts a patch-based architecture. The outputs are prediction matrices, where each element indicates the probability of the patch being "real" or "fake".

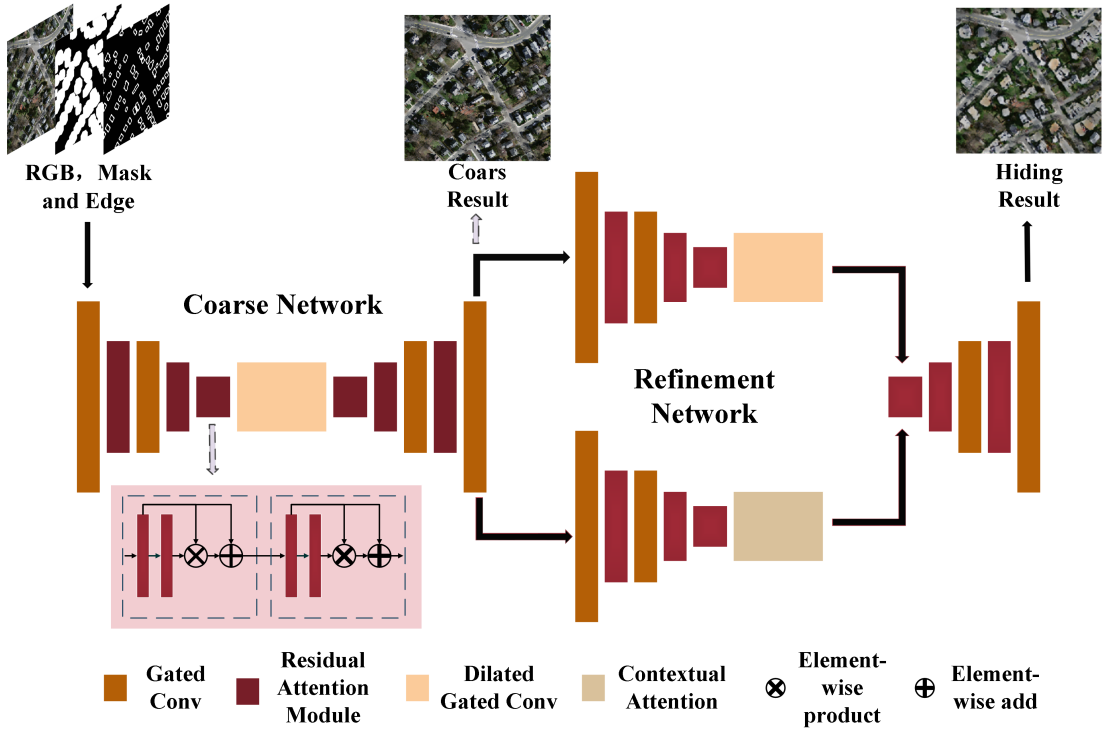

Figure 1. The two-stage generator. Refinement Network refines the preliminary inpainting results of the Coarse Network.

3.2. Methodology

3.2.1. The Proposed Residual Attention Module

Convolutional modules are critical to the overall performance of the model. Compared to Partial Conv [11] and Gated Conv [10], gated convolution offers the following advantages: (1) The parameters in gated convolution are meaningful and trainable. Partial Conv heuristically classifies all spatial locations as either valid or invalid. This means that all spatial locations within the mask region are assigned weights according to the partial mechanism, and these weights cannot be trained. (2) Gated convolution is more flexible, since the effectiveness of the pixels can be learned during training without being limited to the weight set, which is specifically designed for image restoration. However, we observed a substantially prolonged training time and significant instability in the loss function during our experiments when applying Gated Conv for image inpainting. Gated convolution achieves two branches with the same channel number through a single convolution operation. One branch acts as a control gate for the other branch. The parameters in both the "gate" and those it controls are fully trainable, since they are updated during training. Let I denote the input to the module. This operation can be formulated as

$$\begin{aligned} H_G &= \sigma(Gating(I)) \odot \phi(Feature(I)) \\ &= \sigma(\sum\sum W_g \cdot I) \odot \phi(\sum\sum W_f \cdot I)' \end{aligned} \tag{1}$$

where H_G represents the output of the gated convolution, $Gating(I)$ represents the "gate", and $Feature(I)$ represents the features waiting to be selected by $Gating(I)$. The parameters W_g and W_f have the same value and denote the kernels in the gated convolution. The function ϕ can be any activation function, while σ is confined to the sigmoid function

to limit the output within $(0, 1)$. The core of gated convolution lies in the elementwise product operation, which consumes substantial computational resources. At the same time, the complexity of the model increases with its depth, thereby leading to overfitting when trained for an extended period. Therefore, we proposed a residual attention module for the image inpainting network. An illustration of our module and other alternatives is provided in Figure 2.

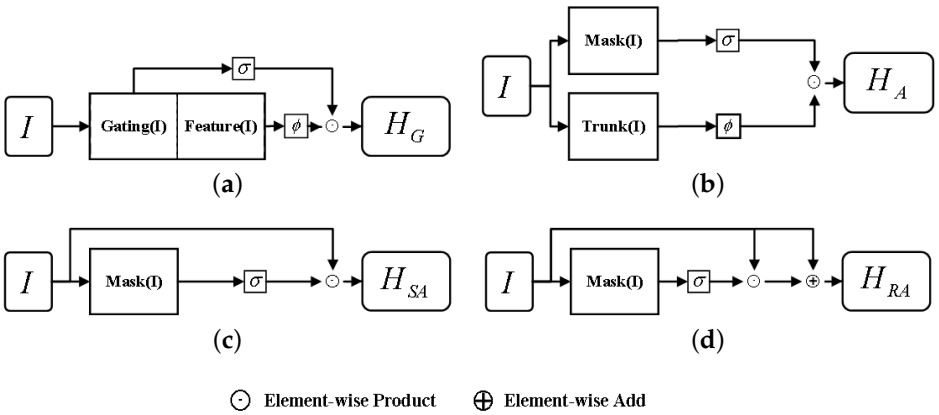

Figure 2. Illustration of four modules. The gated convolution could be regarded as a special attention mechanism. (**a**) Gated convolution. (**b**) Attention. (**c**) Self-Attention. (**d**) Residual attention.

The "gate" mechanism is proposed early in the attention mechanism [31]. The output of the attention module H_A can be formulated as

$$H_A = \sigma(Mask(I)) \odot \phi(Trunk(I)), \tag{2}$$

where $Mask(I)$ represents the gate controlling the output of $Trunk(I)$. In the application of attention modules, the $Mask(I)$ and $Trunk(I)$ have various structures. Although gated convolution and attention modules have the same expression, their implementation methods are different. The $Mask(I)$ and $Trunk(I)$ could be obtained with the same convolutional kernels as in gated convolution. Inspired by this process, this paper replaced the $Trunk(I)$ in the attention mechanism, or $Feature(I)$ in gated convolution, with the input, which is called the self-attention mechanism. It was formulated as follows:

$$H_{SA} = \sigma(Mask(I)) \odot I \tag{3}$$

As shown in Figure 2, the inputs are connected directly to the elementwise product. In this paper, one convolutional layer is inserted before the attention modules to extract input features during the forward pass of the training. The kernel size of each of the self-attention modules is 1 to obtain channelwise features. This significantly decreases the overall number of model parameters and accelerates training.

However, the gradient transfer problem still exists. Let θ denote the parameters of the mask branch and ϕ denote the parameters used in the previous layer to extract information I. The gradient of the self-attention module can be computed as follows:

$$\frac{\partial H_{SA}(I, \theta, \phi)}{\partial \phi} = \sigma(Mask(I, \theta)) \frac{\partial I(\phi)}{\partial \phi} \tag{4}$$

Due to the sigmoid function, the values in $\sigma(Mask(I, \theta))$ always stay within $(0, 1)$. Then, the gradient value has the risk of developing toward near zero, which is also called the gradient disappearance. Therefore, this paper introduced a residual mechanism to mitigate the problem of gradient disappearance. The output of the residual attention mechanism H_{RA} can be calculated by Formula (5). In addition, the calculation of the gradient of H_{RA} follows the Formula (6).

$$H_{RA} = \sigma(Mask(I)) \odot I + I = (1 + \sigma(Mask(I))) \odot I \tag{5}$$

$$\frac{\partial H_{RA}(I, \theta, \phi)}{\partial \phi} = (1 + \sigma(Mask(I, \theta))) \frac{\partial I(\phi)}{\partial \phi} \tag{6}$$

Given that 1 exists in the gradient, the weights of ∂I remain confined between $(1, 2)$. To mitigate the risk of gradient explosion, rectified linear unit (ReLU) activation functions are applied after each convolution operation. Additionally, we insert one convolutional layer before the residual attention module to preclude the retention of the original veridical information in the inpainting outcomes. The kernel size of the residual attention module is set to 1 to extract channelwise features. As depicted in Figure 2, if n denotes the number of channels of $Trunk(I)$, then $Gating(I)$, $Feature(I)$, and $Mask(I)$ share the same value. However, by factorizing the convolution operation in the gated convolution into two successive steps, the residual attention module affords greater flexibility. The total number of parameters can be modulated contingent upon available computational resources. The elementwise product enables the learning of a dynamic feature selection mechanism for each channel and spatial location, thereby conferring the same benefits as gated convolution. For a substantial quantity of convolution kernels, the self-attention mechanism and gated convolution yield nearly equivalent outcomes.

In this work, we substituted the intermediate layers with residual attention modules. Due to having more parameters, gated convolution exhibits superior performance in extracting global and local information. The gated mechanism can support a soft mask training strategy in the initial stage of the network. Consequently, we retained gated convolution in the upsampling layer, downsampling layer, and the initial and final convolution layers of the coarse and refinement networks. This strategy aims to balance the computational cost across layers, thereby reducing the number of weight parameters and simplifying the model. Relative to Gated Conv, our model contains 1 MB fewer parameters and achieved improved performance.

3.2.2. The Larger Fusion Patch Size of the Contextual Attention Layer

The primary innovation of Cont Atten [9] resides in the contextual attention layer. To synthesize more photorealistic images, contextual attention extracts image patches from the foreground and background to serve as convolution kernels. This process is illustrated in Figure 3.

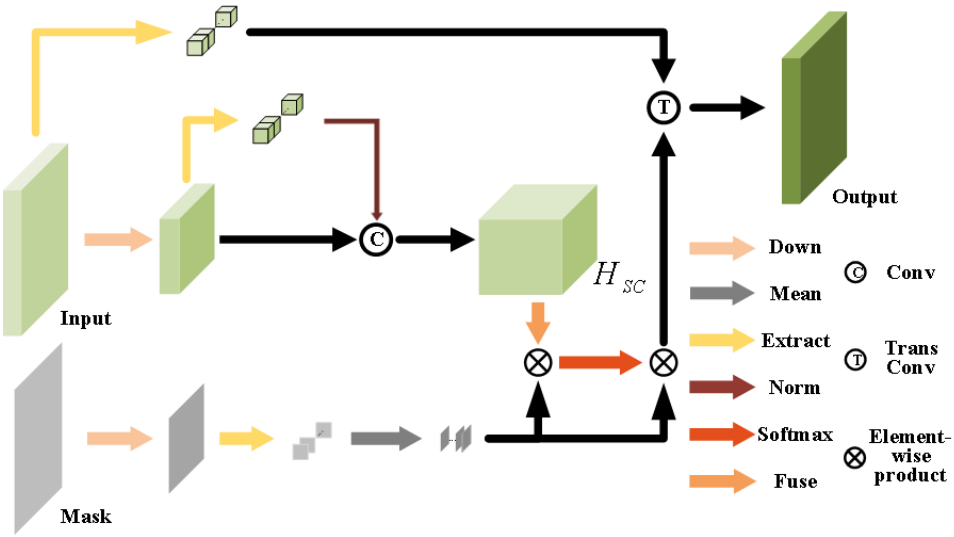

Figure 3. The structure of the contextual attention layer. To mitigate the substantial computational burden, this layer resizes the input dimensionality by half while maintaining the original input shape for the output.

The contextual attention layer departs from the conventional approach of employing direct convolution and elementwise multiplication in the attention module. To mitigate the substantial computational burden, this work resizes the input dimensionality by half while maintaining the original input shape for the output. Two image extraction operations are performed on the input: the first is to obtain the original information and the second is to acquire information after downsampling. The extraction mechanism can preserve the native input values. With these primordial input values serving as kernels, the convolution operation can be regarded as a self-attention mechanism. The model furnishes similarity scores between each patch. Subsequently, convolution operations integrate these scores within the fusion module. The mask branch is intended to suppress probabilities outside the same mask area. Upon completion of the result, the extracted features are utilized as transformation convolution kernels to realign the input matrix to its original shape. The resulting output is then merged with the other branch in a channelwise fashion.

The prior fusion module conducts two transform operations and two convolutions with a unit matrix of size 3. By letting M_{eye} represent the identity matrix, the fusion module can be formulated as follows:

$$Fuse(H_{SC}) = (Conv((Conv(H_{SC}, M_{eye}))^T, M_{eye}))^T \tag{7}$$

This paper replaced the formula with the following:

$$Fuse(H_{SC}) = Conv(Conv(H_{SC}, M_{eye}), M_{eye}) \tag{8}$$

As shown in Figure 4, to facilely contrast the convolution transformations, the original image undergoes division by 8; absent this preprocessing, the ensuing pixel values would exceed 255. "Conv with Trans" represents the outcomes obtained via Equation (7), whereas "Only Conv" denotes the results yielded by Equation (8). Our experiment validated the equivalency of both formulations under simplified conditions. In addition, the identity matrix only fuses the two neighborhoods. Therefore, for simplifying the calculation and enlarging the patch size, this paper replaces the fusion module with two convolutional operations that employ unit matrices consisting entirely of 1 s. Then, the fusion module can cover eight neighborhoods, whose formula is as follows:

$$Fuse(H_{SC}) = Conv(Conv(H_{SC}, M_{one}), M_{one}) \tag{9}$$

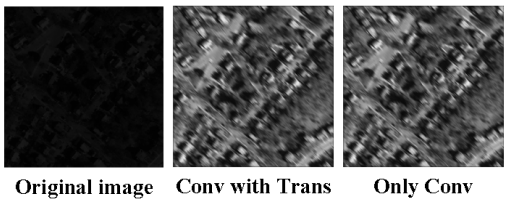

Figure 4. Comparison of the two fusion modules.

3.2.3. Free-Form Mask

Before Partial Conv [11], the conventional method for image inpainting involved the use of a rectangular mask at the center. However, this approach lacked flexibility and controllability, as the masks were not considered a pivotal component of the overall model. Despite attempts at introducing random rotations, dilations, and cropping, the resulting irregular masks ended up being mere transformations of the original mask, with limited effectiveness in image inpainting. Furthermore, when applied to target hiding, the unpredictability and uncontrollable nature of the target shapes posed a significant challenge. Therefore, it became necessary to devise a more sophisticated algorithm that could effectively address these limitations.

To address the limitations of the previous approaches, this paper proposes a novel algorithm that generates randomized masks with irregular shapes during training. The

shape of the mask can cover various forms such as lines, circles, and rectangles, and, by limiting the available range, the algorithm randomly applies these graphics onto a zero board of the same size as the original images. The shape of the resulting mask is illustrated in Figure 5. The input comprised images with masked regions. Figure 6 delineates the distinctions between regular and irregular masks. By introducing unique masks for each training instance, overfitting is avoided, and the random irregular masks significantly improve the model's ability to handle target shapes with nonconventional geometries. Experimental results demonstrated the efficacy of the proposed free-form mask training strategy for image inpainting.

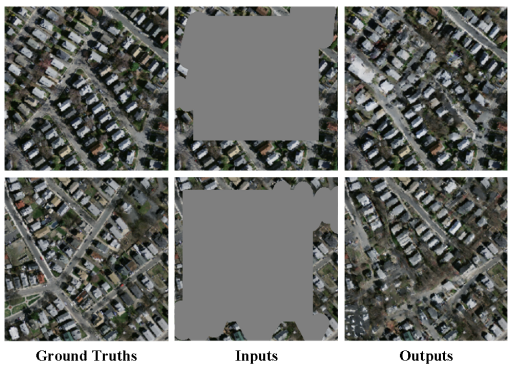

Figure 5. The examples in the training process. The inputs are images with masks covered, and the outputs are results inpainted by models.

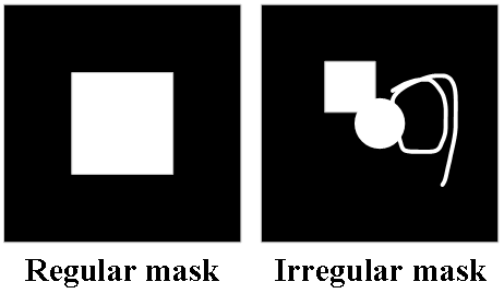

Figure 6. Comparison of the two mask generation strategies.

3.2.4. Edge Extracted by Semantic Segmentation

The proposed model also includes edge-guided target hiding as one of its functions. In previous research, Yu et al. [10] introduced sketches or edges into the image inpainting model using gated convolutions, which guide the model in inpainting the image based on the edges. In this paper, the sketch mainly represents the edge of the target, thus serving as the conditional label in the CGAN. However, traditional edge extraction algorithms often perform poorly on remote sensing images. These images contain more refined details compared to ordinary images [32], thereby making it difficult for traditional methods that rely on contour continuity and gradient changes. Targets within remote sensing images often have various colors with large gradients in both value and channel differences, thus causing edge extraction algorithms based on these factors to lose their effectiveness. To address this issue, Figure 7 demonstrates a comparison between the edges extracted from the Canny operator and the proposed method.

Therefore, this paper leverages image dilation and erosion techniques to accurately extract building edges from the results of semantic segmentation. This methodology is notably more straightforward and precise compared to using either the holistically nested edge detection(HED) [33] or the Canny operator. Despite the exemplary edge extraction performance of HED, antecedent training on pertinent imagery remains imperative before utilization. Moreover, for edge-guided [34] target hiding, the network needs only acquire

representations of the target boundaries rather than comprehensive edges. As illustrated in Figure 7, the edges produced by the Canny operator suffer from many noise points that cannot be utilized for training an image inpainting network. Moreover, by contouring these edges with a semantic segmentation network, we enrich the adaptability of our image inpainting approach towards automatic target hiding. Thus, this paper replaced the HED with semantic segmentation in Gated Conv [10].

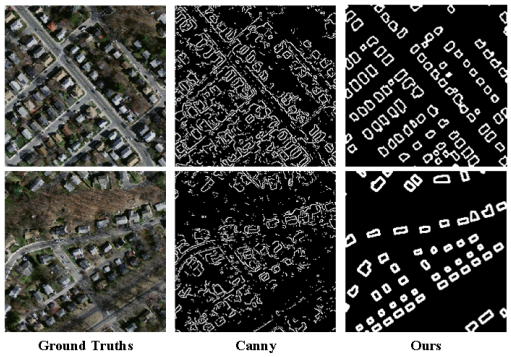

Figure 7. Comparison of edge extraction methods. The Canny operator results exhibit many noise points, while ours are significantly clearer and more accurate.

For training the edge-guided target-hiding function, the edge channel is incorporated into the discriminator's loss calculation. Ultimately, the edge-guided image inpainting model requires three inputs of equal shape (image, mask, and edge channel).

3.3. Materials

All ablation experiments were performed on a 64-bit Linux system that was equipped with a single NVIDIA GeForce RTX 2080 Ti graphics card and 12 GB of memory. All models were trained using Tensorflow v1.14. This paper utilized the Adam optimizer in both the generator and discriminator, with a learning rate of 1×10^{-4}. The batch size for all experiments was set to 16, and the iterations were limited to 10^8. The experiments saved the model at every 2000 iterations. The loss function used for our experiments remained unchanged and was the same as for Gated Conv.

The Mnih Massachusetts Building dataset [35],consisting of 151 aerial images with a size of 1500×1500, was chosen for evaluation in this paper. The main foreground in the dataset is buildings, and the images were cut into 256×256 for analysis purposes. Two different input types were utilized for the image inpainting and target-hiding experiments in this paper. In the case of image inpainting, the focus was on inpainting random missing regions across the input image. Alternatively, for target hiding, inpainting was made specifically to regions in the input images where the targets were located. Additionally, the paper sought to extend its work to include edge-guided target-hiding tasks, which required a different training process than that used for image inpainting and traditional target-hiding tasks. The datasets used in the following experiments contain 4464 pictures, with 3960 pictures used for the training set, 144 pictures used for the validation set, and 360 pictures used for the test set.

4. Experiment and Result

4.1. Experimental Comparison for the Image Inpainting Task

The main objective of this experiment was to evaluate the performance of various models in inpainting images. The focus of image inpainting is on repairing or reconstructing the overall appearance of the image, thus ensuring that it appears visually consistent. Thus, the masks used in this experiment were irregular and randomly generated. This section compares the computational cost of the models, along with their respective results, evaluation indices, and loss curves. Specifically, the free-form mask strategy was only applied to the Gated Conv with an attention mechanism and a residual attention mechanism,

while the other models employed their respective training strategies. In both the training and testing phases, each sample consisted of a single image and an irregular mask. During testing, researchers drew random masks to assess the proposed strategy's ability to inpaint irregular regions.

4.1.1. Computational Cost

The training speed and weight parameters for each model are presented in Table 2. Notably, the self-attention (Self-Atten) network and residual attention network had reduced parameters by 1 MB and exhibited faster training speeds compared to other models. The Self-Atten network represents our model with gated convolutions substituted by self-attention modules, though it achieved inferior performance compared to RATH and suffered from unstable training dynamics. Consequently, the experiments focused our analyses on the proposed RATH architecture.

Table 2. Comparison of three image inpainting methods for calculation cost. The minimal values within each indicator are denoted by boldface font.

Methods	Gated Conv	Self-Atten	Res Atten (Ours)
Parameters	9M548K958B	**8M400K414B**	**8M400K414B**
Training Speed (sec/batch)	0.705	**0.66**	**0.66**

4.1.2. Image Inpainting Results

The foremost objective comprises demonstrating the efficacy of the proposed methodology for image inpainting. The results of the various inpainting models are presented in Figure 8. All models demonstrated the ability to inpaint missing image regions, but their performance outcomes varied. Notably, the distribution of pixels around the boundary of the missing region remained inconsistent, particularly in the case of Partial Conv. Additionally, while the objects in the generated images appeared valid, their contours were imperfect. Small objects such as houses were adequately reconstructed by all models, but larger objects suffered from varying levels of distortion. In particular, the buildings inpainted by Partial Convo and Cont Atten exhibited suboptimal performance. Another significant difference among the various models was in their ability to learn object relationships, most notably the correlation between buildings and roads. Typically, there exists a well-established relationship between these two objects, whereby buildings are situated along the edges of roads on either the left or right side. Our proposed models exhibited consistent and remarkable performance in generating accurately positioned roads flanked by buildings that conformed to this known relationship. In terms of the overall performance, our models demonstrated superiority compared to the other approaches evaluated in this study.

This paper also assessed the performance of image inpainting using similarity indicators, which are presented in Table 3. Ablation studies were conducted with the novel fusion module (termed new fusion) and the residual attention module (Res Atten). Empirical evaluation revealed that the former yielded only marginal performance improvements, whereas inclusion of the latter engendered substantial gains. Our proposed method exhibited the highest values regarding the peak signal-to-noise ratio (PSNR), ℓ_2 Sim indices, thereby indicating that images generated by the residual attention mechanism had the highest similarity in pixel distribution with the original images. Furthermore, our method achieved higher values on both a ℓ_1 similarity (Sim) and universal quality image index (UQI), thereby verifying its efficacy for image restoration tasks.

Table 3. Evaluation index of models on image inpainting. The maximal values within each indicator are denoted by boldface font.

Methods	Cont Atten	Partial Conv	Gated Conv	Gated Conv (New Fusion)	Gated Conv (Res Atten)	RATH (Ours)
ℓ_1 Sim (%)	98.47	98.54	98.59	98.56	**98.67**	98.61
ℓ_2 Sim (%)	87.98	88.30	88.50	88.54	88.59	**88.62**
PSNR	18.81	19.10	19.29	19.36	19.42	**19.43**
SSIM (%)	91.86	**92.04**	81.49	90.21	91.94	91.72
UQI (%)	90.98	91.38	91.62	91.45	91.69	**91.70**

Figure 8. Example cases of qualitative comparison on image inpainting. The images from left to right represent the inputs, original images, and the results of Cont Atten, Partial Conv, Gated Conv, and ours. The orange boxes illustrate the variances in inpainting outcomes among the respective models.

4.1.3. Loss Curves

In Figures 9 and 10, we compare the loss curves of the generator and discriminator with those of the Gated Conv. The x axis represents the number of epochs, measured in units of 10^4, and the y axis corresponds to the loss value. A smoothing technique was applied to the loss curve to present a more consistent representation of the training trend. The initial loss curve exhibited significant divergence, which may have obscured the underlying pattern. By employing this approach, we aimed to enhance the clarity of the displayed training trend. The G_{loss} values of both models showed a decrease over time. In contrast, our D_{loss} curve demonstrated an initial increase followed by a subsequent decline.

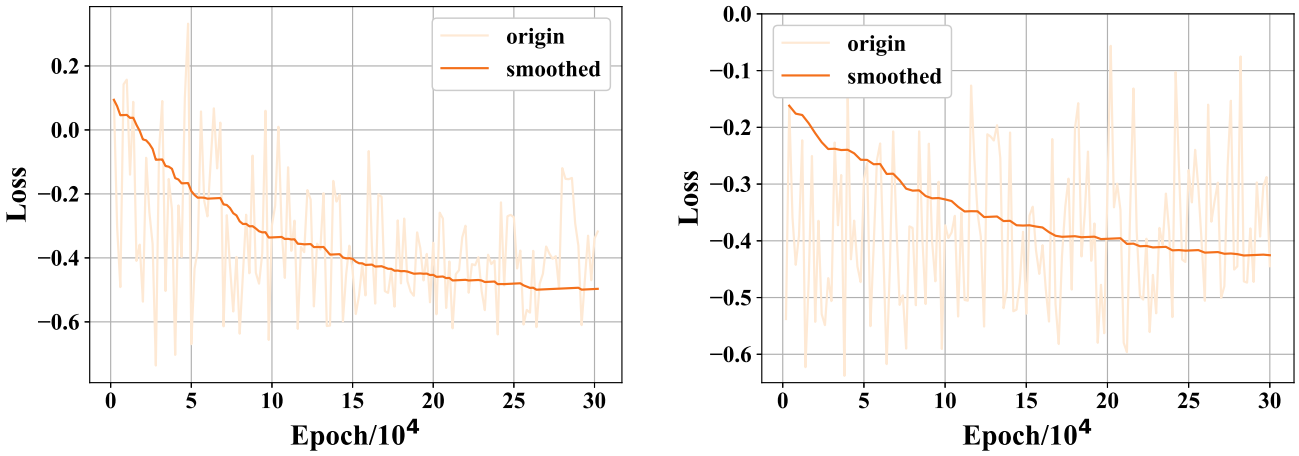

Figure 9. The loss curves of generators. Our method demonstrated narrower ranges of change on G_{loss}. (**a**) Gated Conv. (**b**) Ours.

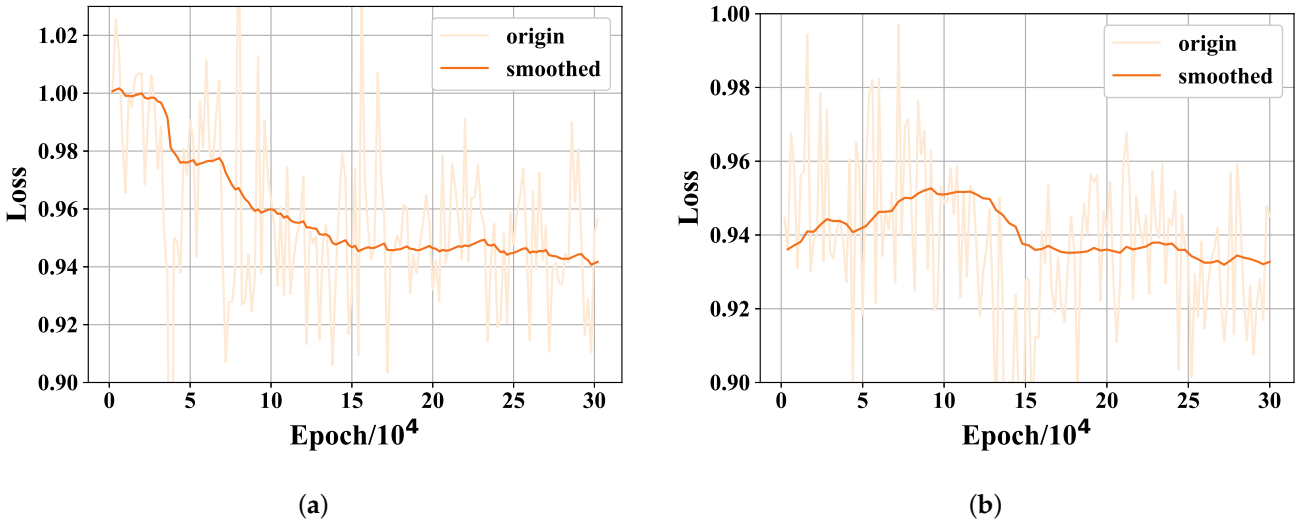

Figure 10. The loss curves of discriminators. Our D_{loss} curve exhibited a trend of initial growth followed by decline and demonstrated narrower ranges of change in the D_{loss}. (**a**) Gated Conv. (**b**) Ours.

To further demonstrate the convergence properties of our model, Figure 11 depicts the discriminator loss curve over an extended training period. As can be observed, the loss consistently decreased and eventually plateaued, thereby indicating that the model reached a stable equilibrium. Our proposed methodology exhibited smaller generator and discriminator loss fluctuations compared to the Gated Conv, as is evidenced by the loss curves. This demonstrates a more stable and effective training process. The superior convergence properties of our framework, with narrower loss ranges, indicate that it is better optimized and outperforms Gated Conv for image inpainting. Our loss trajectory analysis provides quantitative verification that the training stability afforded by our approach translates to improved model performance.

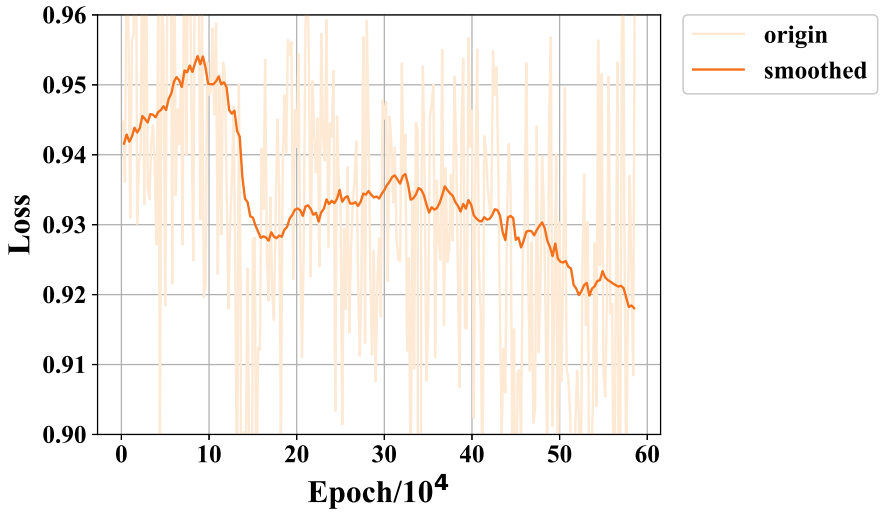

Figure 11. The training loss curve of our discriminator over an extended training period. It exhibited a trend of initial growth followed by a decline.

4.2. Experimental Comparison for the Target-Hiding Task

The emphasis in target hiding is on selectively suppressing or obscuring specific content within the image while maintaining overall image quality and coherence. Therefore, the masks in this experiment covered the targets completely. The results and evaluation indexes are displayed in this section.

In comparison with the method presented in [7] for the autodetection and hiding of targets, we further enhanced the effectiveness of the results obtained through semantic segmentation. Predictions made by a semantic segmentation network yield more accurate contours, but they may not cover objects entirely, thereby leading to exposed regions that reveal information about hidden objects. In light of this drawback, we dilated the labels of our datasets, thereby utilizing the resulting masks as missing portions. The input of the network consisted of images containing four channels (R, G, B, and mask) with hiding applied to the outputs. Figure 12 displays the results of our hiding models.

In this task, we considered several crucial aspects that influenced the effectiveness of our approach. The first point was similar to image inpainting, where the focus was on maintaining continuity between the inpainted and original regions. Inconsistency in pixel distribution can cause discontinuities, thereby making it difficult to distinguish differences in pixel values visually, especially for high-density target areas. However, evaluating the mean image pixel values can help address these challenges. The second aspect concerned whether objects were entirely hidden, and although Gated Conv and our model demonstrated good hiding performance overall, large-object hiding remained an issue. Lastly, attention must be paid to the characteristics of the generated objects, such as inpainting a region covering part of a road, which must be restored while preserving continuity with the original road. As depicted in Figure 12, the original characteristics of the roads remained intact when there were only a few missing parts. Our model successfully maintained the width and high-continuity of the original roads, thus highlighting its effectiveness in this regard.

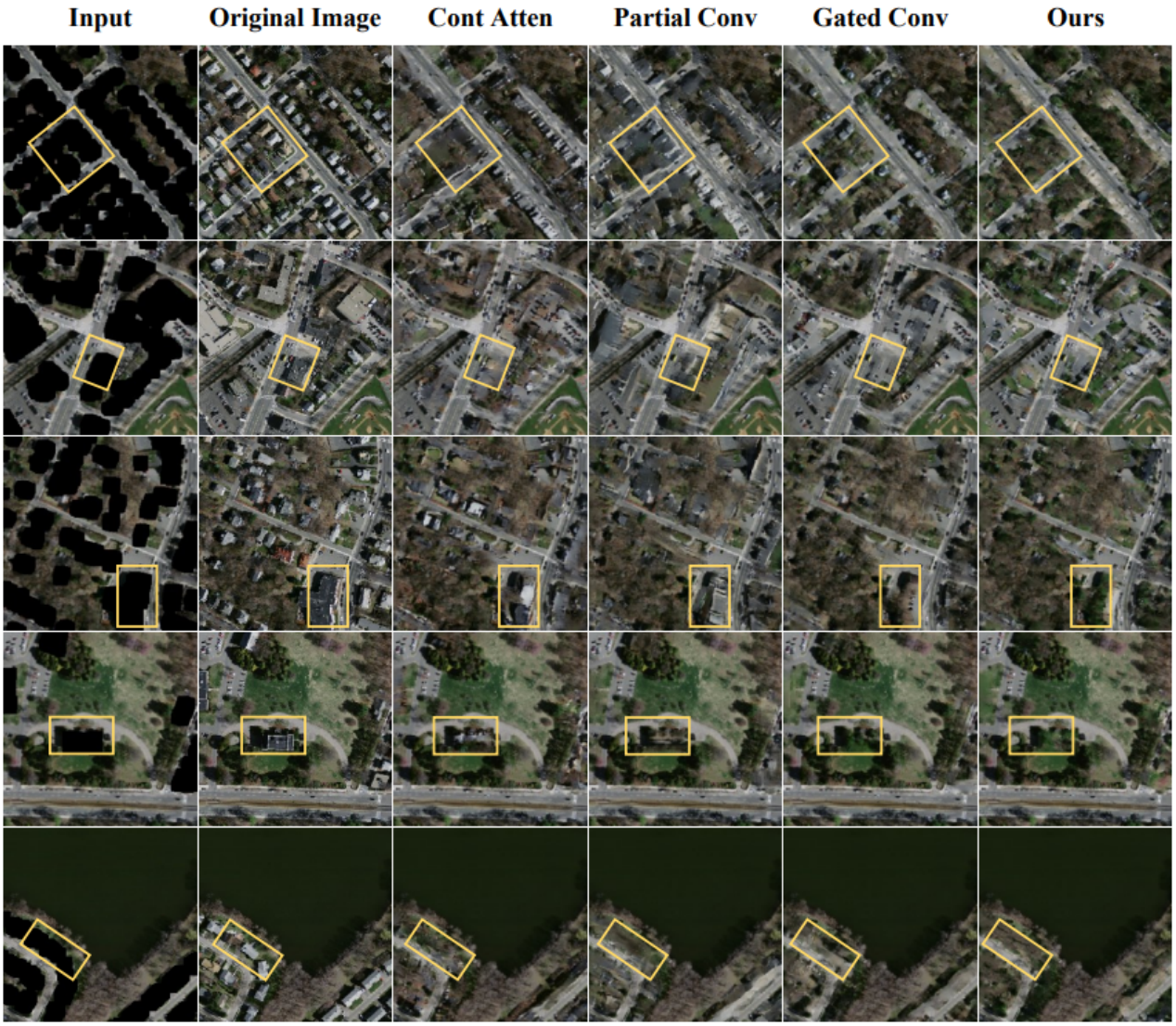

Figure 12. Target-hiding case study with comparison. The orange boxes illustrate the variances in hiding outcomes among the respective models.

Table 4 shows the evaluation indexes of these models. Due to the lack of accepted evaluation methods for target hiding, similarities were selected as evaluation indices to ascertain the extent to which the generated results adhered to the underlying features of the original images, including the ℓ_1 Sim, ℓ_2 Sim, PSNR, the structural similarity index (SSIM), and the UQI. The ℓ_1 Sim and ℓ_2 Sim reflect the overall pixel distribution difference between the original and inpainted images. The similarity values for these methods were relatively close, thereby indicating that they could remove targets with a similar pixel distribution as the original images. However, since buildings have different colors than the background, the similarity between the results and the original will generally be lower when removing targets. The target-hiding task aims to comprehensively remove designated objects. Lower values for the evaluation metrics indicate fewer residual traces of the targets. The empirical results demonstrated that our proposed approach achieved maximal suppression of the target buildings compared with existing methods. Thus, our model had the minimum value in both the SSIM and UQI indicators compared with the other models, thus proving its efficacy in target hiding. To further substantiate the superiority of our method, additional experimental validations were conducted as follows.

Table 4. Evaluation indexes of models on target hiding. Each indicator's maximal and minimal values are denoted by boldface and red font, respectively.

Methods	Cont Atten	Partial Conv	Gated Conv	RATH (Ours)
ℓ_1 **Sim (%)**	97.45	97.51	**97.68**	97.52
ℓ_2 **Sim (%)**	85.32	85.26	**86.02**	85.54
PSNR	18.19	18.32	**18.60**	18.33
SSIM (%)	**88.92**	88.71	88.71	88.18
UQI (%)	86.41	86.74	**87.31**	86.40

4.3. Experimental Comparison for the Edge-Guided Target-Hiding Task

Although the above method could, to some extent, effectively conceal targets, the location and contour information of the inpainted images were still discernible. To address this issue, we proposed an extension to our model by incorporating edge guidance for image generation. Specifically, given that our primary objective is to hide objects, this paper trained the models using only object edges. By changing the original object edge, the proposed model effectively obscured an object's location information to mislead the viewer. In addition, the target-hiding model's robust inpainting ability enabled the preservation of valid detail characteristics of the targets, thereby concealing their location information. There were two edge-generating methods for the edge-guided target hiding: one was the automatic method by semantic segmentation, and the other was drawn by hand.

4.3.1. Edge Generated by Semantic Segmentation

This section compares the proposed method with the other models, and the results are presented in Figure 13. The edges used in this experiment were generated by semantic segmentation. The Partial Conv and Cont Atten methods did not incorporate the edge-guided function. Comparative evaluations were thus restricted to gated convolution and the proposed approach. Our model yielded results that were more similar to the original images, as are demonstrated by the results in the yellow region of the figure. In particular, the fake buildings generated by the two methods varied in color, whereas our model reproduced the same color as the original buildings. However, we observed that the existing methods struggled to handle large objects, with particularly poor results for large buildings. Additionally, they were unable to deal with the connection part between the foreground and background, such as the red region.

The comparison between our proposed method and the gated convolution model on the edge-guided target-hiding task is presented in Table 5. To demonstrate the efficacy of our method, this paper computed the similarity between the results and the original images. Specifically, this paper replaced the true targets with fake targets at their respective locations. A higher similarity evaluation index indicates a greater ability to perform this task. Our method achieved higher values in all five indexes, which demonstrates its superior suitability for the edge-guided target-hiding task.

Table 5. Evaluation indexes of models regarding edge-guided target hiding. The maximal values within each indicator are denoted by boldface font.

Methods	Gated Conv	RATH (Ours)
ℓ_1 **Sim (%)**	97.45	**97.84**
ℓ_2 **Sim (%)**	85.31	**86.44**
PSNR	18.19	**18.80**
SSIM (%)	89.50	**90.44**
UQI (%)	88.21	**89.01**

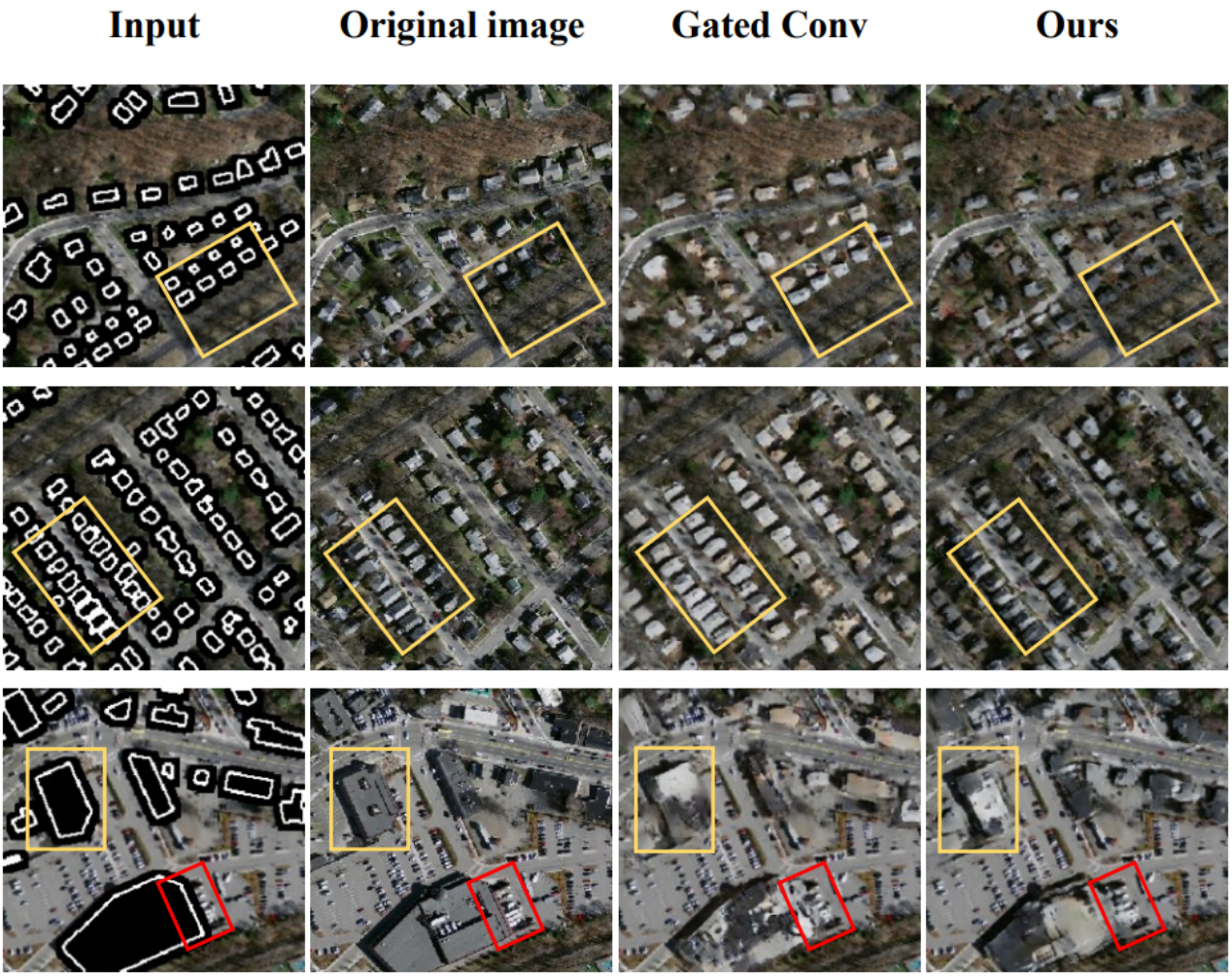

Figure 13. Edge-guided target-hiding case study with comparison. The images from left to right represent the inputs, original images, the results of Gated Conv, and the results of our method. The orange boxes illustrate the variations in edge-guided hiding outcomes, while the red boxes illustrate the deficiencies in different models.

4.3.2. Edge Generated by Hand Drawing

In this section, we drew some edges that differed from the original building. By using masks to cover the buildings that were intended to be concealed, our models were able to generate new objects on the incomplete images. These hidden results have been illustrated in Figure 14. The ability to generate fake objects in both models is undeniable; however, there are still differences regarding the relationship between the foreground and background. We made the building have a regular arrangement and painted edges and masks, thereby giving the impression that there were roads in the images. Our model was able to recognize the regular pattern and generate the expected roads between the fake buildings, as is shown in the second row. Although the new roads in the first row were similar to the original image, the crossover point of the two roads was not executed well. The two images did not connect as 'T' roads as we had expected. Our model was highly effective in dealing with the connection points between the generated roads and the original roads, as is demonstrated in the third row. As Figure 14 indicates, the finely designed mask and edge images were able to guide the models in generating new and significantly different complete images.

Overall, our proposed method produced significant improvements over the existing approaches, thus effectively concealing the target location and contour information while maintaining the visual fidelity of the original images.

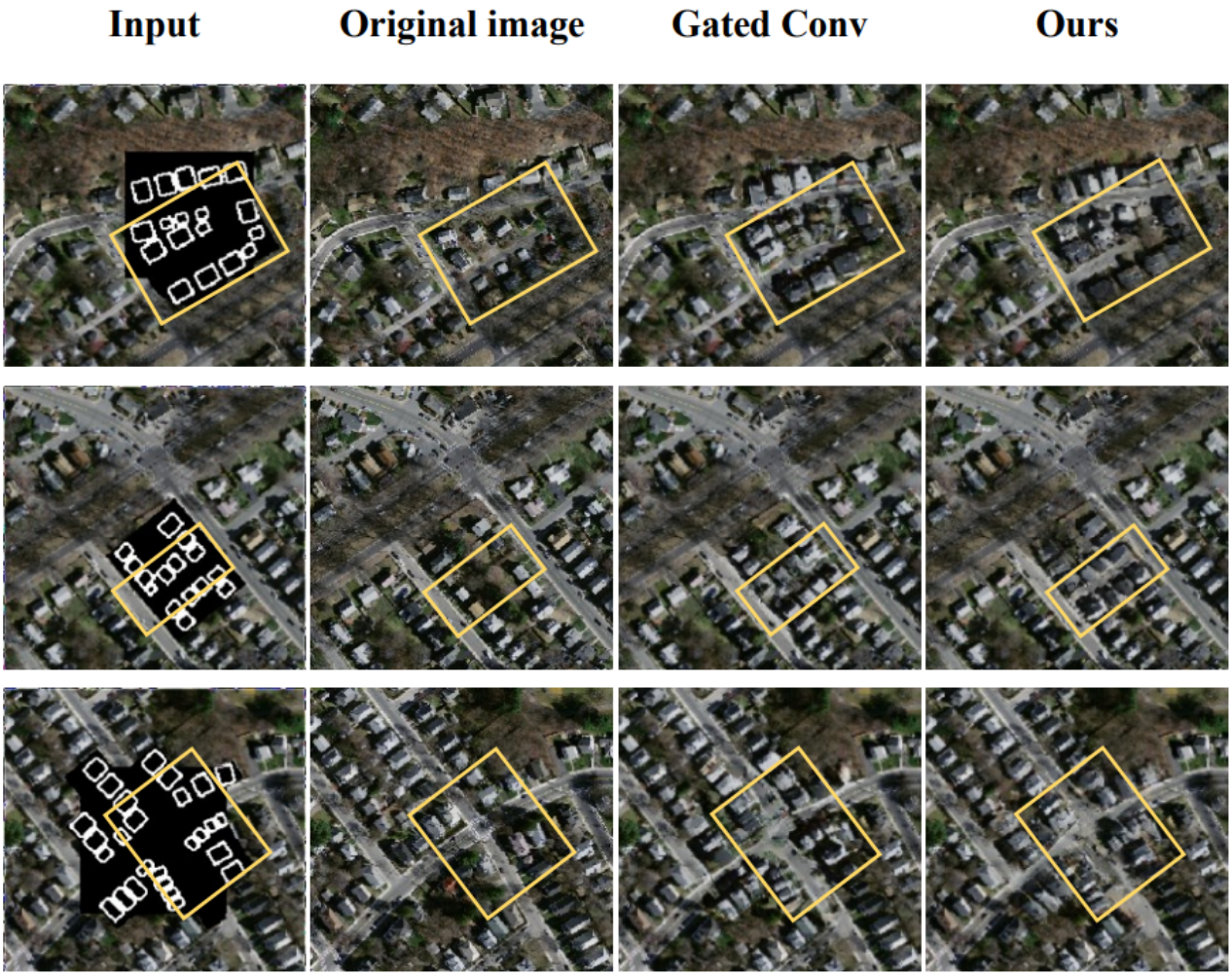

Figure 14. Hand-drawn edge-guided target-hiding case study with comparison. The buildings were regularly arranged in edges and masks. The orange boxes illustrate the variances in hiding outcomes among the respective models.

5. Discussion

The proposed method demonstrates a particular aptitude for the target obfuscation task, especially in confounding viewers through hand-drawn edge manipulation. This was achieved through several key contributions. Firstly, the gated convolution was replaced with a residual attention module during the core stages spanning downsampling to upsampling. While the residual attention module shares a similar underlying mechanism with gated convolution, the latter possesses a greater number of parameters, thereby conferring superior representational capacity for feature extraction. Moreover, this module transmitted the original features between successive gated layers, thereby constraining the inpainting to better resemble the underlying image distribution. Consequently, our inpainting results exhibited greater verisimilitude with the original images, as was evidenced by the optimal quantitative similarity metrics.

Secondly, the fusion patch size was enlarged by substituting the identity matrix with an all-ones matrix. In contrast to the identity matrix, which aggregates over two local neighborhoods, the all-ones matrix can fuse scores across a broader eight-neighborhood region. This module can thus be viewed as implementing a blurring operation. Consequently,

the inpainting and original features are blended more seamlessly through this smoothing process, thereby improving the qualitative integrity of the inpainted outputs.

Thirdly, the edge extraction operation was optimized through the integration of semantic segmentation, which also conferred advantages for automated target obfuscation. By training exclusively on target boundaries, the proposed method exhibited a particular aptitude for hiding targets that were delineated by hand-drawn contours. This allowed realistic synthetic camouflage to be generated that accorded with the original data distribution, thereby effectively confounding viewers.

In the end, this paper integrated semantic segmentation and target hiding into an automated framework to handle large-batch data. This section describes the workflow of our proposed application for automated target hiding by leveraging deep learning approaches. As illustrated in Figure 15, the framework first performs semantic segmentation on the input image to generate semantic maps. Based on the maps, target regions can be identified and concealed through two alternative techniques—direct replacement or edge-guided synthesis—depending on the desired hiding effects. This automated pipeline enables efficient batch processing for target hiding in large datasets.

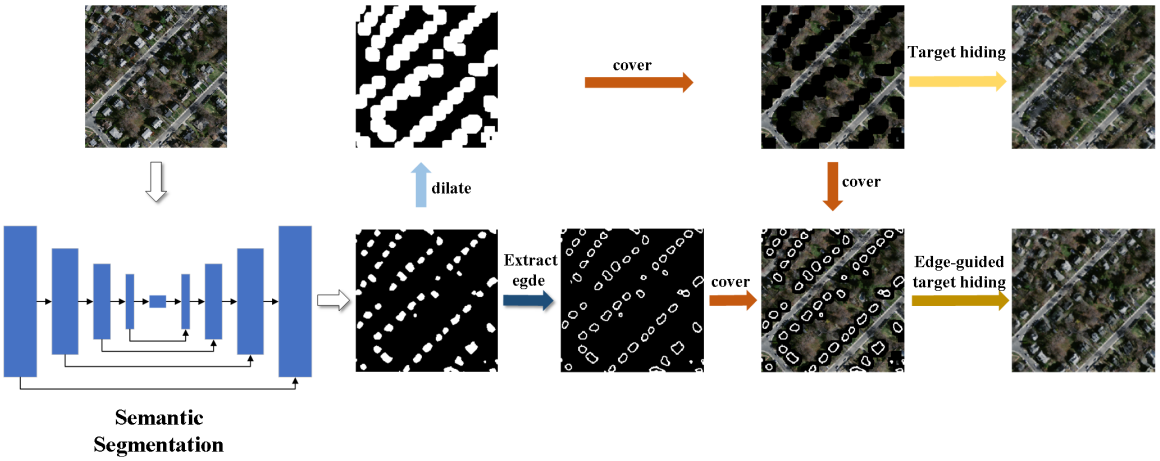

Figure 15. Schematic diagram of the automated target-hiding application workflow.

We utilized Inception-v3 U-Net, which substituted the decoder with Inception-v3 as the semantic segmentation network. This approach achieves a relatively balanced trade-off between training efficiency and segmentation performance. The dilate operation expands the prediction area, thereby reducing the impact of semantic segmentation errors on target-hiding outcomes. Our proposed framework focused on achieving effective target hiding. This paper presents two techniques to process original images containing targets, depending on the desired hiding effects.

The first employs a direct target-hiding method, which takes an image and mask as inputs, where the mask delineates the coverage area and is derived from dilating the segmentation outputs. This method aims to completely remove targets from the original image. The second leverages an edge-guided approach with the image, mask, and edge as inputs. The edge also originates from segmentation but preserves more spatial details. This edge-guided technique generates simulated targets for more natural integration rather than removal. As discussed in Sections 4.2 and 4.3, the two methods achieve very different hiding effects—one eliminates real targets, while the other synthesizes fake targets. Our framework provides the flexibility to produce varied results based on the desired concealment goals.

Figure 16 compares the outputs of our direct target-hiding and edge-guided target-hiding techniques. From left to right are the following: original images, semantic segmentation results, target-hiding inputs, edge-guided inputs, direct hiding outputs, and edge-guided outputs. The target-hiding inputs provide only location information, thus resulting in random synthesized objects such as trees, grass, buildings, or parking lots in

the filled regions. Thus, direct hiding is better suited for completely removing any targets from the images.

In contrast, the edge-guided inputs contain contour information delineating foreground objects. Thus, the edge-guided model training is more targeted and learns to generate replacements that are consistent with the input contours. When the input edges match the typical foreground patterns seen during training, such as buildings, the model synthesizes building-like structures in the missing areas. Furthermore, the edge-guided approach captures relationships between foreground and background elements such as buildings and roads. This domain-specific knowledge enables more semantically coherent and natural scene completion compared to direct target hiding. The edge guidance provides critical spatial cues to generate plausible foreground objects that blend with the original backgrounds.

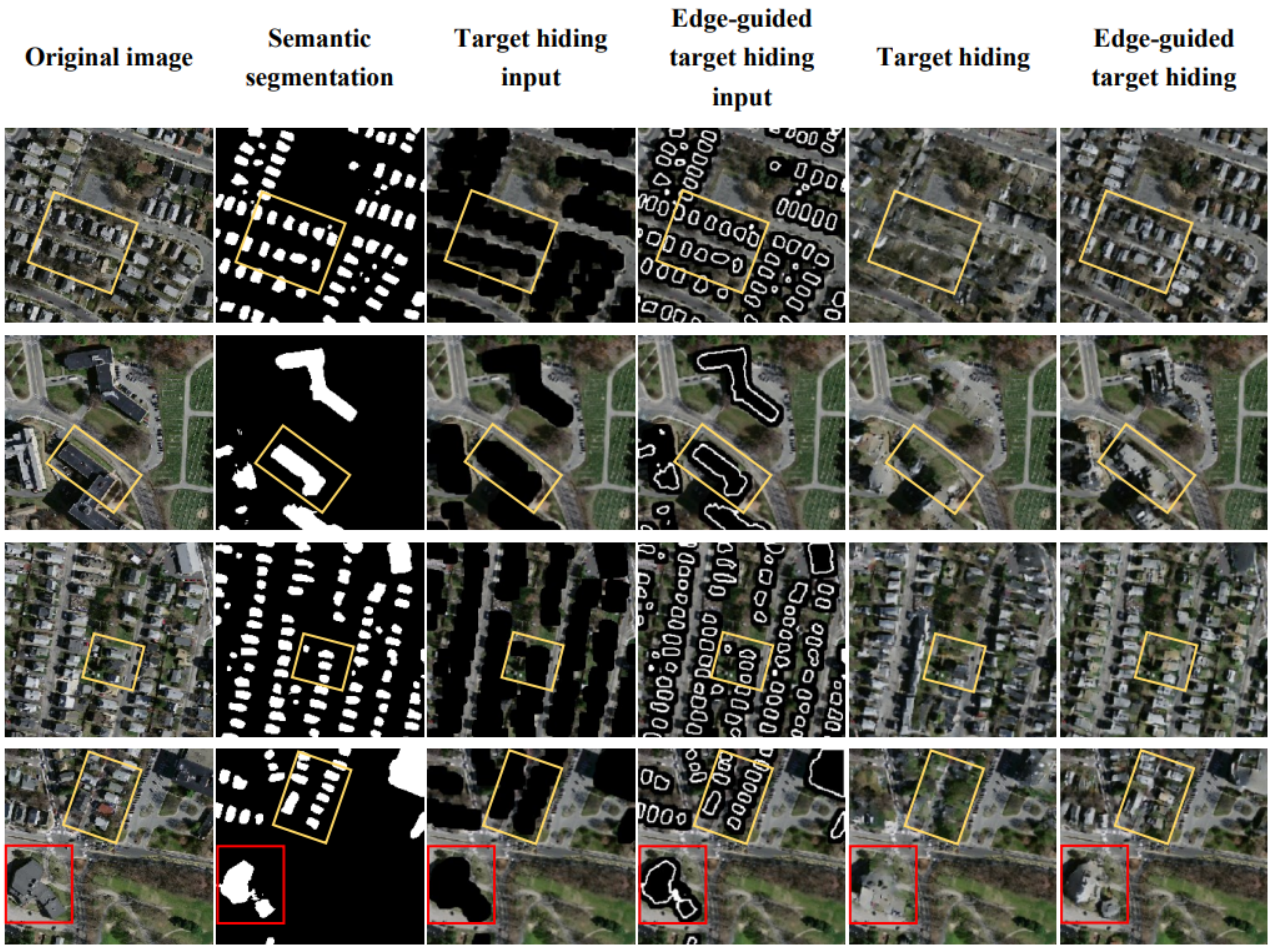

Figure 16. Comparison of target hiding and edge-guided target hiding. From left to right: original images, semantic segmentation results, target-hiding inputs, edge-guided inputs, direct hiding outputs, and edge-guided outputs. The orange boxes illustrate the variations in hiding outcomes, while the red boxes illustrate the deficiencies in different methods.

Furthermore, the edge utilized is a simple binary image of white and black pixels that can be easily generated. As shown in Figure 14, a random mask and hand-drawn outline could also be supplied to our framework. However, the hand-drawn outline should reflect similar spatial distributions as the targets, such as buildings in a regular grid layout. When provided with a reasonable edge, our proposed edge-guided target-hiding technique can effectively conceal the desired target regions. The ability to use crude inputs such as hand-drawn outlines highlights the robustness and flexibility of our edge-guided approach for target hiding under varied conditions.

However, for both inpainting and hiding outcomes, the boundary between the inpainted and original pixels exhibited discontinuities in the final rendered results. While not always conspicuous to casual human observation, these artifacts such as ghosting nevertheless persist, even though they may not be readily apparent upon cursory visual inspection. The underlying cause stems from reliance on the ℓ_1 or ℓ_2 loss functions during optimization. Despite minimizing aggregate pixelwise deviation, such losses fail to explicitly encode spatial smoothness constraints between real and inpainting areas. An important direction for future work involves developing edge-aware losses that are attuned to both target boundaries and missing region contours.

6. Conclusions

In this work, we proposed the RATH model for hiding targets in emergency remote sensing mapping. The distinctive feature of our approach is the substitution of gated convolutions with a residual attention mechanism, thereby enabling the propagation of original features between the downsampling and upsampling stages. This allows synthetic targets to be generated that conform to the true data distribution. Our model demonstrated a high aptitude for target-hiding tasks while maintaining computational efficiency. The residual attention mechanism also resolved issues of gradient instability without compromising the hiding effect. Furthermore, this paper replaced the kernels used for fusing contextual attention layers with full one matrices to enlarge the patch size. In addition, this paper extended the edge-guided function to preserve target contours and positions, thus misleading viewers with fabricated targets. The edges came from the semantic segmentation results. Compared with other methods, our method had $1M$ fewer training parameters than Gated Conv, as well as the highest similarity, at 90.44% SSIM, for edge-guided target hiding. Experiments proved that our model is well-suited for target-hiding tasks. Finally, by integrating semantic segmentation, our framework can efficiently process large batches of remote sensing data in an automated manner.

Although our proposed model demonstrated effectiveness for both image inpainting and target-hiding tasks, artifacts such as ghosting and discontinuities remained evident along the boundaries between the inpainted and original regions. This stems from the current approaches failing to integrate loss functions that explicitly promote the edge congruity between synthesized and authentic areas. This could result in artifacts or discontinuities along the boundaries. To address these limitations, weighting the ℓ_1 or ℓ_2 loss function, which is analogous to the contour loss approach proposed in [16], will be explored in future work to enhance boundary smoothness and enable seamless transitions between synthesized and authentic areas.

Author Contributions: H.Y.: Conceptualization, Methodology, Software, Writing—Original Draft Preparation, and Visualization. Y.S.: Conceptualization, Resources, Writing—Original Draft, Supervision, Project Administration, and Methodology. N.L.: Data curation, and Writing—review. Y.L.: Validation and Resources. C.C.: Editing and Formal Analysis. Z.Z.: Investigation and Data Curation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (2020YFB1807500), the National Natural Science Foundation of China (62072360, 62001357, 62172438, 61901367), the Key Research and Development Plan of Shaanxi Province (2021ZDLGY02-09, 2023-GHZD-44, 2023-ZDLGY-54), the Natural Science Foundation of Guangdong Province of China (2022A1515010988), the Key Project on Artificial Intelligence of Xi'an Science and Technology Plan (23ZDCYJSGG0021-2022, 23ZDCYYYCJ0008, 23ZDCYJSGG0002-2023), the Xi'an Science and Technology Plan (20RGZN0005), and the Proof-Of-Concept Fund from the Hangzhou Research Institute of Xidian University (GNYZ2023QC0201).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, X.; Li, Z.; Fu, X.; Yin, Z.; Liu, M.; Yin, L.; Zheng, W. Monitoring House Vacancy Dynamics in The Pearl River Delta Region: A Method Based on NPP-VIIRS Night-Time Light Remote Sensing Images. *Land* **2023**, *12*, 831. [CrossRef]
2. Zhu, Q.; Cao, Z.; Lin, H.; Xie, W.; Ding, Y. Key technologies of emergency surveying and mapping service system. *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomat. Inf. Sci. Wuhan Univ.* **2014**, *39*, 551–555. [CrossRef]
3. Zhang, L.; Zhang, L. Artificial Intelligence for Remote Sensing Data Analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [CrossRef]
4. Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 104110. [CrossRef]
5. Zhang, J.; Liu, Y.; Wang, B.; Chen, C. A Hierarchical Fusion SAR Image Change-Detection Method Based on HF-CRF Model. *Remote Sens.* **2023**, *15*, 2741. [CrossRef]
6. Dang, W.; Xiang, L.; Liu, S.; Yang, B.; Liu, M.; Yin, Z.; Yin, L.; Zheng, W. A Feature Matching Method based on the Convolutional Neural Network. *J. Imaging Sci. Technol.* **2023**, *67*, 1–11. [CrossRef]
7. Qiu, T.; Liang, X.; Du, Q.; Ren, F.; Lu, P.; Wu, C. Techniques for the Automatic Detection and Hiding of Sensitive Targets in Emergency Mapping Based on Remote Sensing Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 68. [CrossRef]
8. Lv, N.; Zhang, Z.; Li, C.; Deng, J.; Su, T.; Chen, C.; Zhou, Y. A hybrid-attention semantic segmentation network for remote sensing interpretation in land-use surveillance. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 395–406. [CrossRef]
9. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting with Contextual Attention. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514. [CrossRef]
10. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T. Free-Form Image Inpainting with Gated Convolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4470–4479. [CrossRef]
11. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image Inpainting for Irregular Holes Using Partial Convolutions. *arXiv* **2018**, arXiv:1804.07723.
12. Chen, C.; Yao, G.; Liu, L.; Pei, Q.; Song, H.; Dustdar, S. A cooperative vehicle-infrastructure system for road hazards detection with edge intelligence. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 5186–5198. [CrossRef]
13. Chen, C.; Yao, G.; Wang, C.; Goudos, S.; Wan, S. Enhancing the robustness of object detection via 6G vehicular edge computing. *Digit. Commun. Netw.* **2022**, *8*, 923–931. [CrossRef]
14. Chen, X.; Liu, M.; Li, D.; Jia, J.; Yang, A.; Zheng, W.; Yin, L. Conv-trans dual network for landslide detection of multi-channel optical remote sensing images. *Front. Earth Sci.* **2023**, *11*, 1182145. [CrossRef]
15. Ding, W.; Zhang, L. Building Detection in Remote Sensing Image Based on Improved YOLOV5. In Proceedings of the 2021 17th International Conference on Computational Intelligence and Security (CIS), Chengdu, China, 19–22 November 2021; pp. 133–136. [CrossRef]
16. Yang, P.; Wang, M.; Yuan, H.; He, C.; Cong, L. Using contour loss constraining residual attention U-net on optical remote sensing interpretation. *Vis. Comput.* **2023**, *39*, 4279–4291. [CrossRef]
17. Lv, N.; Ma, H.; Chen, C.; Pei, Q.; Zhou, Y.; Xiao, F.; Li, J. Remote Sensing Data Augmentation Through Adversarial Training. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9318–9333. [CrossRef]
18. Nitzberg, M.; Mumford, D.; Shiota, T. *Filtering, Segmentation and Depth*; Springer: Berlin/Heidelberg, Germany, 1993; Volume 662. [CrossRef]
19. Hirani, A.N.; Totsuka, T. Combining frequency and spatial domain information for fast interactive image noise removal. In Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 4–9 August 1996.
20. Masnou, S.; Morel, J.M. Level lines based disocclusion. In Proceedings of the 1998 International Conference on Image Processing, ICIP98 (Cat. No.98CB36269), Chicago, IL, USA, 7 October 1998; Volume 3, pp. 259–263. [CrossRef]
21. Bertalmío, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000.
22. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Los Alamitos, CA, USA, 2016; pp. 2536–2544. [CrossRef]
23. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. In Proceedings of the 28th Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014. [CrossRef]
24. Yang, C.; Lu, X.; Lin, Z.; Shechtman, E.; Wang, O.; Li, H. High-Resolution Image Inpainting Using Multi-scale Neural Patch Synthesis. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4076–4084. [CrossRef]
25. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and Locally Consistent Image Completion. *ACM Trans. Graph.* **2017**, *36*, 107. [CrossRef]

26. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* **2018**, arXiv:1710.10196.
27. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976. [CrossRef]
28. Yan, Z.; Li, X.; Li, M.; Zuo, W.; Shan, S. Shift-Net: Image Inpainting via Deep Feature Rearrangement. In Proceedings of the European Conference on Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19.
29. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. *arXiv* **2019**, arXiv:1901.00212.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
31. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Training Very Deep Networks. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Nice, France, 2015; Volume 28, pp. 2377–2385.
32. Liao, Z.; Chen, C.; Ju, Y.; He, C.; Jiang, J.; Pei, Q. Multi-controller deployment in SDN-enabled 6G space–air–ground integrated network. *Remote Sens.* **2022**, *14*, 1076. [CrossRef]
33. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1395–1403. [CrossRef]
34. Chen, C.; Wang, C.; Liu, B.; He, C.; Cong, L.; Wan, S. Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* 2023, *early access*. [CrossRef]
35. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.