# ArtTraceNet: Manipulated Region Localization in Artwork Images

Tran Hoang Ha Nguyen
International University
VNU-HCM
Ho Chi Minh, Viet Nam
Email: ITITIU21127@student.hcmiu.edu.vn

Van Sinh Nguyen
International University
VNU-HCM
Ho Chi Minh, Viet Nam
Email: nvsinh@hcmiu.edu.vn

Thi-Ngoc-Hanh Le*
International University
VNU-HCM
Ho Chi Minh, Viet Nam
Email: ltnhanh@hcmiu.edu.vn

***Abstract*—Local manipulation in artwork images is difficult to detect because edited regions may closely follow surrounding brush strokes, colors, and stylized textures. Existing manipulation-localization methods are mainly developed for natural photographs, while artwork-specific datasets with pixel-level annotations remain limited. This paper presents ArtTraceNet, a segmentation framework designed to localize subtle digital manipulation traces in artwork images. To enable supervised training, we construct a synthetic dataset containing clean images, diffusion-inpainted samples, shape-overlay edits, and aligned binary masks. ArtTraceNet combines normalized RGB channels with high-frequency residual channels to expose local inconsistencies that may be weak in appearance space alone. The resulting six-channel representation is progressively encoded, refined, and decoded to preserve both contextual information and spatial manipulation cues, while an attention module emphasizes informative channel and spatial responses. An exponential moving average model and validation-based threshold selection are further used to stabilize checkpoint selection and mask generation. Experimental results and evaluations demonstrate that combining controlled synthetic supervision with high-frequency trace learning is effective for manipulated-region localization in complex artwork images.**

***Index Terms*—arttracnet, artwork manipulation localization, image forensics, semantic segmentation, high-frequency residual, attention mechanism**

## I. Introduction

Digitized artworks have become an important source of visual data for cultural access, research, preservation, and creative reuse [1], [2]. At the same time, modern image-editing and generative tools make localized modification increasingly accessible. In particular, diffusion-based inpainting can replace selected regions while maintaining surrounding colors, structures, and textures [3]. Consequently, manipulated regions may remain visually plausible and difficult to recognize by manual inspection.

Existing image manipulation localization methods are mainly developed for natural photographs and commonly exploit camera traces, compression inconsistencies, abnormal boundaries, or learned forensic features [4]–[7]. Artwork images present a substantially different setting. Brush strokes naturally contain strong high-frequency variations, stylized contours may resemble editing boundaries, and irregular color transitions can obscure locally generated content. Therefore, manipulation cues that are useful for photographs may become weak or ambiguous in artwork images. A localization model must not only recognize suspicious regions but also distinguish genuine manipulation traces from legitimate artistic structures.

This observation exposes two technical limitations of conventional RGB-based segmentation. First, subtle editing traces can be overwhelmed by dominant appearance information. When a manipulated region follows the surrounding painting style, its RGB values may remain highly consistent with neighboring pixels, although weak local residual differences still exist. Second, deep feature extraction alone may not sufficiently preserve these weak cues during reconstruction. Strong semantic or visually salient structures can dominate the representation, causing the model to activate on legitimate artwork content or incompletely recover the actual manipulated region.

To address these issues, we propose **ArtTraceNet**, a pixel-level manipulation localization framework designed around two complementary mechanisms. The first is *high-frequency residual (HFR) enhanced input construction*. Instead of relying solely on RGB appearance, ArtTraceNet explicitly extracts local high-frequency information and concatenates it with the RGB channels. The resulting representation preserves normal visual content while exposing subtle texture, edge, and color inconsistencies that may indicate local editing. This provides the network with manipulation-sensitive evidence from the beginning of feature extraction rather than expecting such traces to emerge implicitly from RGB features.

The second mechanism is an *Attentive Feature Bridge*. Features extracted from the input are refined through sequential channel and spatial attention before being passed to the decoding stage. Channel attention emphasizes manipulation-sensitive feature responses, while spatial attention identifies locations that deserve stronger consideration. This intermediate refinement is particularly important for artwork images, where visually dominant objects and complex brush patterns may otherwise distract the localization process. The decoder then reconstructs these selectively refined features into a pixel-wise manipulation probability map. A validation-selected threshold converts the probability map into the final binary localization mask.

A further challenge is the limited availability of artwork-specific manipulation datasets with pixel-level annotations. We

therefore construct a controlled synthetic dataset containing clean images and localized manipulations generated using diffusion-based inpainting and shape-overlay operations. Each edited image is paired with its corresponding binary mask, providing direct supervision for manipulation localization without manual pixel-level annotation. The data generation stage is used only for training preparation, while ArtTraceNet itself operates directly on a single artwork image at inference time.

The main contributions of this work are summarized as follows:

- We formulate digital artwork manipulation analysis as a pixel-level localization problem, targeting subtle edited regions embedded in complex artistic textures.
- We introduce an HFR-enhanced input representation that complements RGB appearance with explicit high-frequency evidence to expose weak manipulation traces.
- We propose an attention-guided feature connection that sequentially refines channel and spatial responses between feature extraction and decoding, suppressing irrelevant artistic patterns while preserving manipulation-sensitive information.
- We construct a synthetic artwork manipulation dataset with aligned pixel-level masks and demonstrate that ArtTraceNet provides improved localization over representative segmentation baselines.

## II. Related Work

Image segmentation provides the foundation for pixel-level manipulation localization. U-Net introduced encoder-decoder segmentation with skip connections [8], while ResNet enabled deeper feature extraction through residual learning [9]. UNet++ refined skip pathways for better multi-scale fusion [10], and DeepLabV3+ combined encoder-decoder refinement with atrous context modeling [11]. A residual-attention UNet++ variant further integrated nested decoding, residual features, and attention mechanisms [12], [13]. Dice-based losses have also been developed to address class imbalance in segmentation [14], while Mean Teacher improves training stability through weight-averaged model parameters [15].

Image manipulation localization has been widely studied in digital forensics. Noiseprint suppresses scene content to emphasize camera-related forensic traces [4]. MVSS-Net combines multiple forensic views with multi-scale supervision [5], whereas PSCC-Net progressively models spatial and channel correlations [6]. ObjectFormer combines RGB and high-frequency representations to capture subtle manipulation traces [16], while TruFor fuses learned forensic traces, RGB evidence, and confidence estimation [7]. TrainFors further provides a large training benchmark for manipulation detection and localization [17]. For inpainting localization, HP-FCN shows that high-pass residuals can reveal weak manipulation artifacts [18], while enriched attentive patterns improve detection of near-original inpainted regions [19]. These findings motivate the high-frequency residual input adopted in our model.

Artwork-related anomaly localization remains less explored. Reconstruction-based methods have been applied to X-ray images of paintings [20], while the Inpainting Transformer detects anomalies from reconstruction discrepancies [21]. MVTec AD has also provided an important benchmark for unsupervised anomaly detection and pixel-level localization [22]. However, these methods target conservation imagery or unsupervised anomaly settings rather than supervised manipulation localization in ordinary artwork images.

Deep image inpainting is closely related because modern methods can reconstruct masked regions with realistic structures and textures. Contextual attention improves long-range information use during reconstruction [23], while latent diffusion further enables high-quality conditional inpainting [3]. Such realism complicates manipulation localization in artworks, where brush strokes and stylized boundaries may resemble editing traces. We therefore use diffusion-based inpainting to generate supervised image-mask pairs for training ArtTraceNet.

## III. Methodology

### A. System Overview

Given an artwork image, our goal is to localize regions that have been digitally modified rather than only determining whether the entire image is manipulated. Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input artwork image and $M \in \{0,1\}^{H\times W}$ its manipulation mask, where manipulated pixels are assigned 1. ArtTraceNet learns a mapping:

$$P = f_\theta(I), \qquad P \in [0,1]^{H\times W}, \tag{1}$$

where $P$ is the pixel-wise manipulation probability map. The final localization result is obtained by:

$$\widehat{M}_{i,j} = \begin{cases} 1, & P_{i,j} \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\tau$ is selected on the validation set. Since the presence of manipulation can be inferred directly from $\widehat{M}$, no separate image-level classification branch is required.

As illustrated in Fig. 1, ArtTraceNet is designed around two complementary mechanisms for preserving weak manipulation evidence. First, the normalized RGB image is augmented with a high-frequency residual to form a six-channel *HFR-enhanced input*. This representation retains the original visual appearance while explicitly exposing local texture, edge, and color inconsistencies that may be difficult to observe in RGB space alone. Second, the extracted features are passed through an *Attentive Feature Bridge*, where channel and spatial responses are selectively refined before decoding. This helps suppress dominant but irrelevant artistic patterns while preserving manipulation-sensitive cues. The refined representation is then reconstructed into a pixel-wise probability map and thresholded to produce the final manipulation mask.

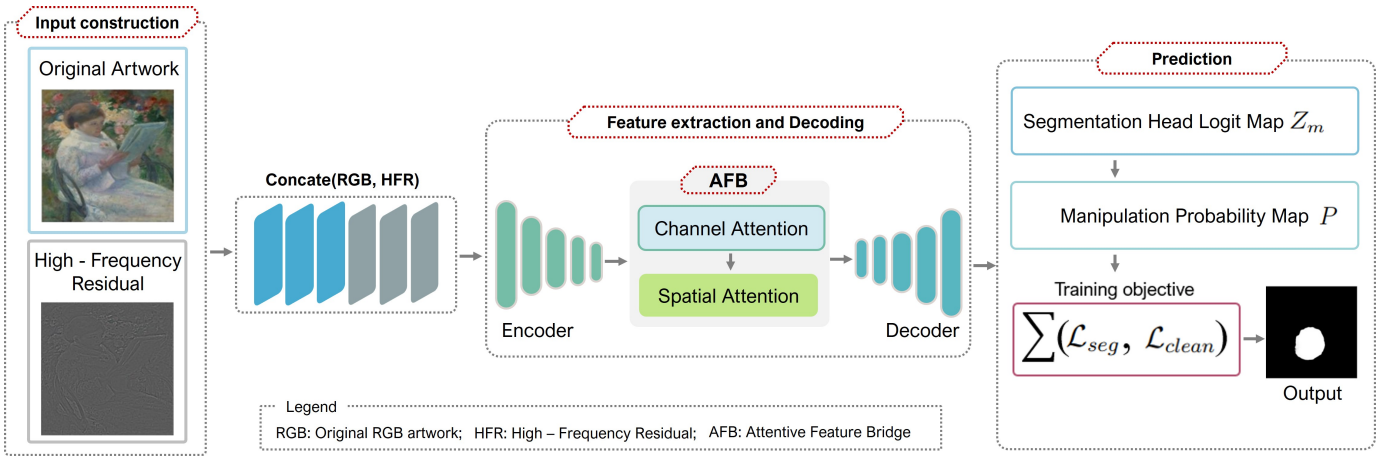



Fig. 1. Overview of the proposed ArtTraceNet framework

### B. Synthetic Dataset Generation

Pixel-annotated datasets for artwork manipulation localization are difficult to obtain because public artwork collections primarily contain original images, while manually creating plausible manipulations and their pixel-level annotations is expensive. We therefore construct a synthetic dataset in which each manipulated artwork is paired with its original image and a spatially aligned binary mask. This strategy is visualized in Fig. 2.

Given an original artwork image $A^o$ and a randomly generated binary mask $M$, we use a pretrained Stable Diffusion Inpainting model [24] to synthesize an edited image:

$$\tilde{A} = \mathcal{G}(A^o, M, c), \tag{3}$$

where $\mathcal{G}$ denotes the inpainting model and $c$ is the text condition. To diversify the manipulated regions, circular, elliptical, and rounded-rectangular masks are generated with randomized locations and areas. Accepted masks occupy approximately $4\%$–$20\%$ of the image, producing localized rather than global modifications.

Since diffusion-based inpainting may also modify pixels outside the supplied mask, directly using $\tilde{A}$ could introduce inconsistencies between the manipulated image and its annotation. We therefore restore all pixels outside $M$ from the original artwork:

$$A^m = M \odot \tilde{A} + (1 - M) \odot A^o, \tag{4}$$

where $\odot$ denotes element-wise multiplication. Consequently, $A^m$ and $A^o$ differ only inside $M$, ensuring pixel-level alignment between the manipulated region and its ground-truth mask.

Some inpainting results may introduce negligible changes or excessively visible artifacts. We therefore measure the mean absolute difference within the manipulated region:

$$D = \frac{\sum_{i,j} M_{i,j} \left\| A^m_{i,j} - A^o_{i,j} \right\|_1}{3 \sum_{i,j} M_{i,j}}, \tag{5}$$

and retain only samples whose differences fall within a predefined range. Additional percentile and changed-pixel criteria are used to reject samples whose modifications are concentrated in only a few pixels. Small brightness, contrast, or color perturbations may also be introduced within $M$ to generate measurable but visually subtle local inconsistencies; pixels outside the mask are restored again afterward.

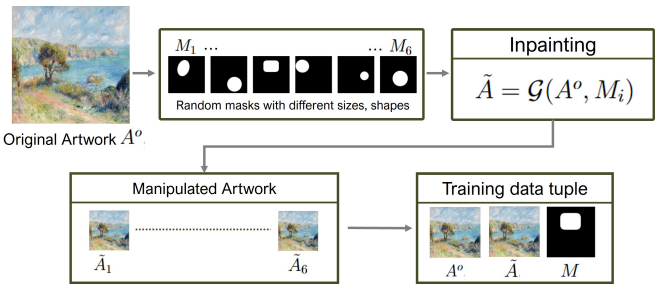



Fig. 2. Illustration of our synthetic data generation strategy

Clean artwork images with all-zero masks are additionally included as negative samples to discourage the model from interpreting complex artistic textures as manipulation. The resulting dataset is represented as

$$\mathcal{D} = \{(A^m_n, A^o_n, M_n)\}_{n=1}^{N}, \tag{6}$$

where $A^m_n = A^o_n$ and $M_n = \mathbf{0}$ for clean samples. This controlled generation process provides pixel-aligned supervision without requiring manual manipulation annotation.

### C. ArtTraceNet Architecture

ArtTraceNet is designed to address two difficulties in artwork manipulation localization: weak editing traces can be obscured by dominant RGB appearance, and these cues may be further diluted during feature extraction and reconstruction. We therefore introduce two complementary components: an *HFR-enhanced input* that explicitly exposes local residual evidence, and an *Attentive Feature Bridge (AFB)* that selectively refines manipulation-sensitive features before decoding.

**HFR-enhanced Input.** Artwork images naturally contain strong brush strokes, irregular contours, and non-uniform color transitions. When an edited region follows the surrounding artistic style, its RGB appearance may remain visually consistent with neighboring pixels. To expose weak local inconsistencies, ArtTraceNet augments the normalized RGB image $I_n$ with a high-frequency residual (HFR). A locally smoothed image is first estimated by average pooling:

$$R = (I_n - \mathrm{AvgPool}_k(I_n))\, s, \tag{7}$$

where $k$ denotes the pooling kernel size and $s$ controls the residual magnitude. We use $k = 5$ and $s = 1$. The HFR is then concatenated with the RGB image:

$$X = \mathrm{Concat}(I_n, R), \qquad X \in \mathbb{R}^{H\times W\times 6}. \tag{8}$$

The resulting six-channel representation preserves the original appearance while explicitly introducing texture, edge, and local color deviations that can serve as manipulation evidence.

Given $X$, a hierarchical encoder extracts multi-level features

$$\{E_1, E_2, E_3, E_4, E_5\} = \mathcal{E}(X), \tag{9}$$

where shallow features retain spatial details and deeper features encode stronger contextual information. In our implementation, ArtTraceNet employs a ResNet34 encoder within a U-Net++-based segmentation backbone; these components provide the underlying multi-scale feature extraction and reconstruction.

**Attentive Feature Bridge.** Although HFR exposes local editing traces at the input, such weak responses may still be suppressed by semantically dominant artwork structures during encoding. We therefore introduce AFB between the deepest encoder representation and the decoder. AFB sequentially applies channel and spatial attention:

$$F_c = \mathcal{A}_c(E_5) \odot E_5, \text{ and } E_5' = \mathcal{A}_s(F_c) \odot F_c, \tag{10}$$

where $\mathcal{A}_c$ and $\mathcal{A}_s$ denote channel and spatial attention, respectively. Channel attention identifies feature responses that are more informative for manipulation analysis, while spatial attention emphasizes suspicious locations within these responses. This refinement suppresses dominant but irrelevant artistic patterns and preserves manipulation-sensitive information before spatial reconstruction.

The refined feature $E_5'$, together with multi-level encoder features, is decoded to recover the manipulation region:

$$Z = \mathcal{S}\left(\mathcal{D}(E_1, E_2, E_3, E_4, E_5')\right), \tag{11}$$

where $\mathcal{D}$ denotes the decoder and $\mathcal{S}$ is the segmentation head. The final manipulation probability map is:

$$P = \sigma(Z), \tag{12}$$

where each $P_{i,j}$ represents the probability that pixel $(i, j)$ belongs to a manipulated region. Thus, HFR provides explicit local evidence at the input, while AFB preserves and selectively strengthens this evidence before decoding.

**Training Objective.** Manipulated regions generally occupy only a small portion of an artwork image, resulting in substantial foreground-background imbalance. For manipulated samples, we combine pixel-wise supervision with region-overlap objectives:

$$\mathcal{L}_{\mathrm{seg}} = \lambda_b \mathcal{L}_{\mathrm{BCE}} + \lambda_d \mathcal{L}_{\mathrm{Dice}} + \lambda_t \mathcal{L}_{\mathrm{FT}}, \tag{13}$$

in which we empirically set $\lambda_b = 0.30$, $\lambda_d = 0.40$, and $\lambda_t = 0.30$. Binary cross-entropy supervises individual pixels, Dice loss promotes region overlap, and Focal Tversky loss emphasizes difficult foreground predictions. They are defined as:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[m_i \log p_i + (1 - m_i)\log(1 - p_i)\right], \tag{14}$$

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2\sum_i p_i m_i + \epsilon}{\sum_i p_i + \sum_i m_i + \epsilon}, \tag{15}$$

and

$$\mathcal{L}_{\mathrm{FT}} = \left(1 - \frac{TP + \epsilon}{TP + \alpha FP + \beta FN + \epsilon}\right)^{\gamma}, \tag{16}$$

with $\alpha = \beta = 0.5$ and $\gamma = 0.75$.

Clean images are used as negative samples to discourage false manipulation responses:

$$\mathcal{L}_{\mathrm{clean}} = \lambda_c\, \mathrm{BCE}(Z, \mathbf{0}) + \lambda_m \frac{1}{N}\sum_{i=1}^{N} p_i, \tag{17}$$

where $\lambda_c = 0.10$ and $\lambda_m = 0.05$. The final objective is:

$$\mathcal{L}_{\mathrm{total}} = 1_{|M|>0}\mathcal{L}_{\mathrm{seg}} + 1_{|M|=0}\mathcal{L}_{\mathrm{clean}}. \tag{18}$$

This objective encourages ArtTraceNet to recover subtle manipulated regions while suppressing false responses on unmodified artwork content.

## IV. Experimental Results

### A. Implementation Details

Our framework was implemented in PyTorch and evaluated on an NVIDIA A100 GPU. Our model ArtTraceNet is trained with AdamW using a learning rate and weight decay of $10^{-4}$, batch size 4, and at most 40 epochs. The training time takes approximately 19 hours. Once trained, our model spends about 0.067 second to estimate and output the result.

The training dataset contains 1102 artwork images, 879 for training and 223 for validation. The training split includes 746 manipulated and 133 clean images, while the validation split contains 185 manipulated and 38 clean images. RGB inputs are normalized with a mean and standard deviation of 0.5, while masks use nearest-neighbor interpolation. Augmentation includes image flipping and mild brightness and contrast changes; stronger geometric transforms and noise are avoided to preserve forensic cues.

### B. Results and Evaluation

**Evaluation Metrics.** Since the primary objective is to localize manipulated regions, evaluation emphasizes the overlap between the predicted mask $\widehat{M}$ and the ground-truth mask $M$. We report the Dice coefficient and Intersection over Union (IoU) on manipulated samples, while clean images are assessed separately using the clean false-positive rate. Mean prediction probability is recorded as a diagnostic measure, and a weighted validation score is used for threshold and checkpoint selection. More specifically, they are formulated as:

$$\text{Dice} = \frac{2|\widehat{M} \cap M| + \epsilon}{|\widehat{M}| + |M| + \epsilon}, \qquad \text{IoU} = \frac{|\widehat{M} \cap M| + \epsilon}{|\widehat{M} \cup M| + \epsilon}. \tag{19}$$

Dice measures region overlap, whereas IoU provides a stricter evaluation by penalizing both missed and falsely detected pixels.

**Quantitative and qualitative comparisons.** To evaluate performance of our ArtTraceNet, we compare it with three typical baseline model: UNet++ [10], FPN [25], and DeepLabV3 [11]. Table I reports the quantitative comparison with representative segmentation baselines. Among the competitors, ArtTraceNet achieves the best performance, reaching a Dice score of $0.8853$ and an IoU of $0.8289$. Compared with the RGB-only U-Net++ baseline, this corresponds to improvements of $0.1745$ in Dice and $0.2152$ in IoU. These gains indicate that explicitly introducing high-frequency residual cues and refining features through the AFB provide complementary evidence for subtle manipulation localization beyond RGB-based segmentation alone.

TABLE I
COMPARISON WITH SEGMENTATION BASELINES

| Model | Input | HFR | AFB | Dice | IoU |
|---|---|---|---|---|---|
| U-Net++ | RGB | No | No | 0.6875 | 0.5823 |
| FPN | RGB | No | No | 0.7108 | 0.6137 |
| DeepLabV3+ | RGB | No | No | 0.8455 | 0.7833 |
| **ArtTraceNet** | RGB+HFR | Yes | Yes | **0.8853** | **0.8289** |

Fig. 3 further visualizes these quantitative differences. We can see that the compared methods show unstable localization behavior. UNet++ frequently shifts its response toward visually salient but unmodified objects, or detects only a small portion of the edited region. FPN is more conservative and often produces weak or missing responses, particularly when the manipulation is subtle or embedded in highly textured areas. DeepLabV3 generally identifies the correct neighborhood more reliably, but its predictions are often incomplete, spatially shifted, or accompanied by false-positive regions.

In contrast, ArtTraceNet produces masks that are more consistently aligned with the ground-truth regions across portraits, landscapes, still-life scenes, and highly textured paintings. The improvement is especially visible in cases where the edited region overlaps strong brush strokes or complex color transitions, where the proposed model better preserves both the location and approximate extent of the manipulation. This

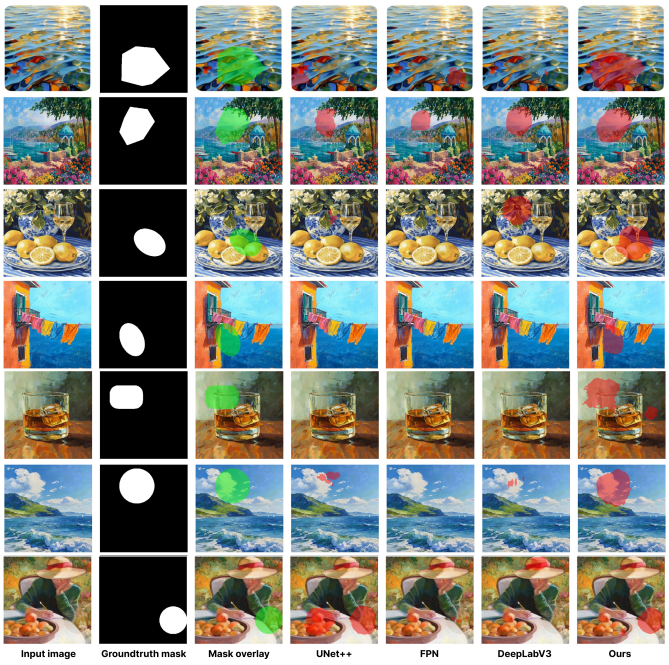



Fig. 3. Visual comparisons with prior works

behavior supports the role of the HFR-enhanced input in exposing weak local inconsistencies and the AFB in suppressing irrelevant artistic responses before decoding. Some boundary overestimation remains in several examples, but the predicted regions are generally more complete and spatially consistent than those of the compared baselines.

### C. Ablation Study

Fig. 4 and Table II evaluates the contribution of each component in our proposed framework. Obviously, removing both HFR and AFB leads to a strongly misplaced prediction, indicating that RGB appearance alone is insufficient to identify the manipulated region. Introducing HFR substantially shifts the response toward the correct area, although the mask remains fragmented and incomplete. Without AFB, localization becomes more stable but still contains scattered false responses. The full configuration produces the most spatially coherent mask and best matches the ground-truth region, demonstrating that HFR and AFB play complementary roles in exposing and refining subtle manipulation evidence.

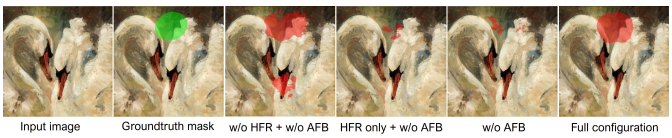



Fig. 4. Ablated results

**Limitation.** Fig. 5 shows a typical failure case in our experiments. In some cases, the predicted mask may cover only part of the edited region. In other cases, the prediction may extend beyond the ground-truth mask or produce a slightly different

TABLE II
ANALYSIS ON ABLATED RESULTS

| Variant | HFR | AFB | Dice | IoU |
|---|---|---|---|---|
| w/o HFR + w/o AFB | No | No | 0.7752 | 0.7128 |
| HFR only | Yes | No | 0.8653 | 0.8235 |
| w/o AFB | Yes | No | 0.8771 | 0.8245 |
| **Full ArtTraceNet** | Yes | Yes | **0.8851** | **0.8334** |

shape. These errors usually occur when the manipulated region has similar color, texture, or brush stroke patterns to the surrounding artwork. As a result, the model may recognize the general manipulated area but still have difficulty producing an exact boundary.




Fig. 5. A sample fail case in our prediction

## V. CONCLUSION

This paper presented ArtTraceNet for pixel-level localization of digitally manipulated regions in artwork images. By combining controlled synthetic supervision with high-frequency residual cues and attentive multi-scale feature learning, the proposed method effectively distinguishes subtle edits from complex artistic textures. Experimental results and evaluations show that our ArtTraceNet outperformed representative segmentation baselines on manipulated samples. In the near future, we plan to extend the framework to broader manipulation types, more diverse artwork domains, and real-world edits with weaker or unknown forensic traces.

## REFERENCES


[1] T.-N.-H. Le, K. G. Pham, and S. Van Nguyen, "Material-aware style transfer for lacquer paintings," in *2025 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2025, pp. 102–107.

[2] T.-N.-H. Le, Y.-H. Chen, and T.-Y. Lee, "Structure-aware video style transfer with map art," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3s, pp. 1–25, 2023.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[4] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2020.

[5] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 185–14 193.

[6] X. Liu, Y. Liu, J. Chen, and X. Liu, "Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7505–7517, 2022.

[7] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 606–20 615.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, vol. 9351. Springer, 2015, pp. 234–241.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2018, pp. 801–818.

[12] Z. Li, H. Zhang, Z. Li, and Z. Ren, "Residual-attention UNet++: A nested residual-attention U-Net for medical image segmentation," *Applied Sciences*, vol. 12, no. 14, p. 7149, 2022.

[13] S. Van Nguyen, V. X. Nguyen, and T.-N.-H. Le, "An improved method for enhancing images quality based on convolution efficient transformer," *SN Computer Science*, vol. 7, no. 4, p. 347, 2026.

[14] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, vol. 10553. Springer, 2017, pp. 240–248.

[15] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 1195–1204.

[16] J. Wang, Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim, and Y.-G. Jiang, "ObjectFormer for image manipulation detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2364–2373.

[17] S. Nandi, P. Natarajan, and W. Abd-Almageed, "TrainFors: A large benchmark training dataset for image manipulation detection and localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023, pp. 403–414.

[18] H. Li and J. Huang, "Localization of deep inpainting using high-pass fully convolutional network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8301–8310.

[19] W. Yang, R. Cai, and A. C. Kot, "Image inpainting detection via enriched attentive pattern with near original image augmentation," in *Proceedings of the 30th ACM International Conference on Multimedia*. Association for Computing Machinery, 2022, pp. 2816–2824.

[20] A. Mezina, R. Burget, and M. Kotrly, "A deep learning approach for anomaly detection in x-ray images of paintings," *npj Heritage Science*, vol. 13, p. 127, 2025.

[21] J. Pirnay and K. Chai, "Inpainting transformer for anomaly detection," in *Image Analysis and Processing – ICIAP 2022*, ser. Lecture Notes in Computer Science, vol. 13232. Springer, 2022, pp. 394–406.

[22] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9592–9600.

[23] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[24] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2022, pp. 11 451–11 461.

[25] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 6392–6401.