

Rethinking Image Forgery Detection and Localization via Regression Perspective

Li Zhang^{ID}, Dong Li^{ID}, Yan Zhong^{ID}, Jiaying Zhu, Rujing Wang^{ID}, Xingyu Wu^{ID}, Xue Wang^{ID}, and Liu Liu^{ID}

Abstract—Image Forgery Detection and Localization is rapidly advancing in the field of computer vision. Most methods locate forged regions in the form of segmentation and subsequently perform detection, facing challenges such as false detections (i.e., FPs) and inaccurate boundaries. In this work, we suggest rethinking the Image Forgery Detection and Localization (IFDL) task from a regression perspective and propose the CatmullRom Splines-based Regression Network (CSR-Net) to address these issues. Specifically, we first design an adaptive CutmullRom splines fitting scheme to predict coarse forged regions. Subsequently, we develop a novel rescore mechanism that filters out samples with no response in both the classification and instance branches to reduce false positives. Besides, a learnable texture extraction module decouples horizontal and vertical forgery features, extracting more robust contour representations to further refine boundaries and suppress false detections. Compared to segmentation-based methods our method is simple but effective due to the unnecessity of post-processing. Extensive experiments conducted on several challenging benchmarks demonstrate that our method outperforms state-of-the-art methods qualitatively and quantitatively. Particularly, CSR-Net achieves optimal performance on three real-world datasets, indicating the applicability of our method to real scenarios such as social media and multi-tampered regions.

Received 10 September 2024; revised 8 November 2024 and 29 December 2024; accepted 27 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62302143, in part by Anhui Provincial Natural Science Foundation under Grant 2308085QF207, and in part by National Key Research and Development Program of China under Grant 2021YFD2000201. (Li Zhang and Dong Li contributed equally to this work.) (Corresponding authors: Xingyu Wu; Xue Wang; Liu Liu.)

Li Zhang is with the Hefei University of Technology, Hefei 230026, China, also with the University of Science and Technology of China, Hefei 230026, China, also with the Institute of Intelligent Machines, Chinese Academy Sciences, Hefei 230031, China, and also with the Hefei Institute of Physical Science, Chinese Academy Sciences, Hefei 230031, China (e-mail: zanly20@mail.ustc.edu.cn).

Dong Li and Jiaying Zhu are with the University of Science and Technology of China, Hefei 230026, China.

Yan Zhong is with the Peking University, Beijing 100000, China (e-mail: zhongyan@stu.pku.edu.cn).

Rujing Wang is with the University of Science and Technology of China, Hefei 230026, China, also with the Institute of Intelligent Machines, Chinese Academy Sciences, Hefei 230031, China, and also with the Hefei Institute of Physical Science, Chinese Academy Sciences, Hefei 230031, China (e-mail: rjwang@iim.ac.cn).

Xingyu Wu is with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University (PolyU), Hong Kong SAR, China (e-mail: xingy.wu@polyu.edu.hk).

Xue Wang is with the Institute of Intelligent Machines, Chinese Academy Sciences, Hefei 230031, China, and also with the Hefei Institute of Physical Science, Chinese Academy Sciences, Hefei 230031, China (e-mail: xwang@iim.ac.cn).

Liu Liu is with the Hefei University of Technology, Hefei 230026, China (e-mail: liuliu@hfut.edu.cn).

Recommended for acceptance by J. Wang.

Digital Object Identifier 10.1109/TETCI.2025.3543837

Index Terms—Image forgery, detection and localization, catmullrom splines, regression methods.

I. INTRODUCTION

THE rise of digital image editing tools has made image manipulation easy, enabling both creative and malicious uses. Previously, manipulating images required significant skill, typically involving copying or deleting objects. However, generative models now enable realistic, language-driven image edits that seamlessly integrate forgeries, affecting domains like news, forensics, and biometric recognition [1], [2]. This has led to a growing focus on Image Forgery Detection and Localization (IFDL), which aims to detect image authenticity and locate forged regions. Recent deep learning advancements have produced notable methods (Fig. 1), such as PSCC-Net [3] for multi-scale forgery representation, Objectformer for object-level consistency, and ERMPG [4] for decoupling forged and authentic features. Despite progress, challenges persist due to the complex attributes of forged regions and the sophisticated tools used by forgers to conceal tampered areas.

The first issue is **false positives (FPs)**, which occur when forgery localization results incorrectly identify authentic regions as tampered. This issue is akin to false alarms in image segmentation. Conventional segmentation methods often encounter such problems due to the binarization process, where inappropriate thresholds lead to the inclusion of unintended regions (Fig. 2). Previous approaches have generally prioritized detecting tampered regions while overlooking the false positive rate. However, misclassifying authentic regions as tampered can negatively impact the credibility of digital content, affecting news sources' profitability and limiting the progress of forgery detection techniques.

The second issue is **inaccurate boundaries**, as shown in Fig. 1(a). Traditional segmentation methods suffer from inconsistent mask predictions across decoder layers, leading to misaligned optimization objectives and weak feature coupling. Regression-based methods, when applied to IFDL tasks, also underperform due to their reliance on bounding boxes, which only localize target regions in a quadrilateral shape. This approach fails for forged regions with irregular boundaries, as seen in Fig. 1(b), where regression-based localization uses the minimum bounding quadrilateral as Ground Truth. Furthermore, the increasing complexity of manipulated images presents further challenges, as many methods struggle to model the boundaries of forged regions, often resulting in mixed predictions with other targets or incompatible backgrounds.

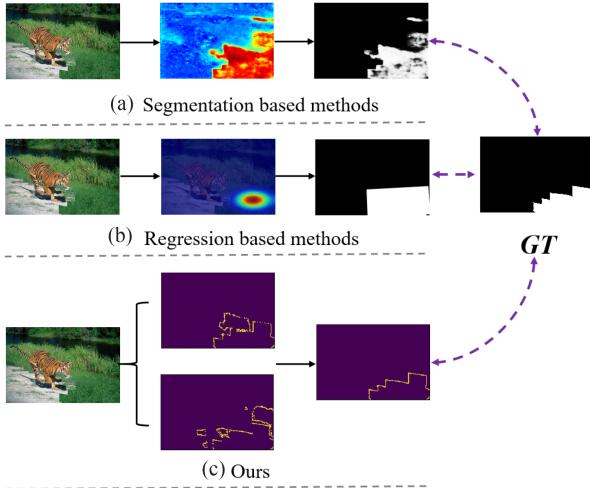


Fig. 1. The categorization of methods applied to IFDL. Please zoom in for better visualization.

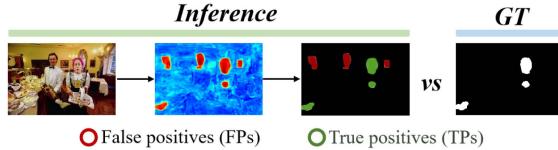


Fig. 2. The illustration of FPs in traditional segmentation-based methods.

The second issue is **inaccurate boundaries**, as shown in Fig. 1(a). Traditional segmentation methods suffer from inconsistent mask predictions across decoder layers, leading to misaligned optimization objectives and weak feature coupling. Regression-based methods, when applied to IFDL tasks, also underperform due to their reliance on bounding boxes, which only localize target regions in a quadrilateral shape. This approach fails for forged regions with irregular boundaries, as seen in Fig. 1(b), where regression-based localization uses the minimum bounding quadrilateral as Ground Truth. Furthermore, the increasing complexity of manipulated images presents further challenges, as many methods struggle to model the boundaries of forged regions, often resulting in mixed predictions with other targets or incompatible backgrounds.

In summary, our contributions can be summarized in three aspects:

- We tailor a **CatmullRom Splines-based Regression Network (CSR-Net)** to make the first attempt to introduce regression methods into the pixel-level task (referred to as IFDL in this paper).
- To suppress false positives and clarify region boundaries, we propose two complementary components: the Comprehensive Re-scoring Algorithm (CRA), which evaluates the confidence score of each region as a tampered area, and the Vertical Texture-interactive Perception (VTP), which refines the accuracy of region boundaries.
- Extensive experiments on several public datasets, including natural image and social media datasets, demonstrate

the superiority of our method over state-of-the-art approaches in IFDL. Notably, our method excels on the real-world IMD20 dataset, highlighting its effectiveness in handling forged images in open-world scenarios.

II. RELATED WORK

A. Classic Methods in IFDL

Image Forgery Detection and Localization (IFDL) aims to detect and pinpoint forged components in digital images. Traditional methods, such as color filter array analysis [5], photo-response noise [6], illumination analysis [7], and JPEG artifact detection [8], struggle with complex forgeries or seamless integrations. Recent approaches utilize local noise features, frequency domain analysis, and Camera Filter Array [9] analysis to identify manipulated image regions [10]. For example, PSCC-Net uses a two-path method for analysis, while Zhuo et al. introduced a self-adversarial strategy for improved localization.

However, these methods rely on segmentation and predefined hyperparameters for binarization, limiting further development.

B. Regression Based Methods

Regression-based methods are widely used in computer vision tasks like object detection and localization [11]. Algorithms such as Fast R-CNN [12] and Diffusion-Det [13] have evolved to tackle challenges like false positives in complex 3D scenes, often using detection frames to handle uneven segmentation [14]. When quadrilateral regions cannot be detected, regression problems are addressed using parametric curves via interpolation or approximation spline functions. For example, gesture recognition uses Bezier curves with constant memory, while B-splines aid in lane marking detection and 3D regression [15].

This paper introduces a customized Catmull-Rom detection method for IFDL, emphasizing the significance of optimal parameter settings. Our findings show that fine-tuning the tension factor (τ) enhances model fitting and improves Catmull-Rom spline characterization, highlighting the value of flexible parameter adjustment in real-world applications.

III. METHOD

A. CatmullRom Splines Detection

Segmentation-based methods [3] are widely used in Image Forgery Detection and Localization (IFDL), offering an intuitive approach by binarizing pixels to classify foreground (forged regions) and background. However, regression-based methods have not been fully explored in this area. The challenge lies in applying regression directly to irregular boundaries, where parameterized curves are required to describe polygonal regions. Yet, many existing methods involve complex parameters that hinder practicality. Recent advances in spline-based methods for tasks like autonomous driving lane detection, text detection, and fault detection demonstrate their efficacy. Among them, the Catmull-Rom spline is a classic interpolation function, well-suited for parameterizing tampered regions due to its fitting accuracy and low inference cost.

Catmull-Rom splines are fitted to irregular boundaries using cubic interpolation, with tangents calculated from adjacent points. Variants of Catmull-Rom splines can be adapted to any shape defined by control points. A key advantage over other spline functions is that cubic Catmull-Rom splines use only integer coefficients, reducing implementation cost, and leading to faster inference and lower computational cost (Flops). Mathematically, CatmullRom spline is defined as (1):

$$\mathbf{c}_i(t) = \sum_{j=0}^3 b_j(t) \mathbf{p}_{i+j}, \quad i = 0, 1, \dots, n-3. \quad (1)$$

where $0 \leq t \leq 1$, $\mathbf{p}_i (i = 0, 1, \dots, n-3; n \geq 3)$ are control points, $b_j(t)$ is the basis. For example, it can be expressed by (2) when the highest power of t in the function $b_j(t)$ is 3:

$$\begin{aligned} \mathbf{c}_i(t) &= \frac{1}{2} \cdot [1 \ t \ t^2 \ t^3] \cdot \\ &\left[\begin{array}{cccc} 0 & 2 & 0 & 0 \\ -\tau & 0 & \tau & 0 \\ 2\tau & \tau - 6 & -2(\tau - 3) & -\tau \\ -\tau & 4 - \tau & \tau - 4 & \tau \end{array} \right] \cdot \begin{bmatrix} \mathbf{p}_i \\ \mathbf{p}_{i+1} \\ \vdots \\ \mathbf{p}_{i+3} \end{bmatrix}. \end{aligned} \quad (2)$$

CatmullRom spline has a strong polygon fitting ability, thanks in large part to its adaptive tensile factor (τ). This is a key parameter, which is used to control the degree of curvature of the curve. More specifically, a higher value of the tension factor will cause the curve to bend more tightly between the control points, thus fitting closer to the given data points during the fitting process. Conversely, lower values of the tension factor will cause the curve to be smoother between the control points. In this paper, we carefully analyze the distribution of the degree of curvature of irregular regions in the curvature-oriented forged region dataset to determine the most suitable catmull curves for the IFDL task. Intuitively, the conventional CatmullRom spline (parameter $\tau=1$) is a poor fit for the IFDL task directly, so we seek to find the right balance between fitting accuracy and curve smoothness by adjusting τ . Ablation experiments (In Section IV-D) show that CatmullRom splines can be reliable for this task when τ is set to 16. It also allows the learned control points to be closer to the foreground (tampered) area (more analysis can be seen in ablation experiments in Section IV-D).

After careful selection of the τ (we make it to be 16), we can transform a polygonal region (in the dataset of the IFDL task, i.e. the region of the mask) into a region depicted by a CatmullRom spline interpolation curve expressed by 8 control points. Note that when the region to be processed is a parallelogram region with 4 points, two additional control points are introduced. Thus, we transform the task of localizing a polygonal region into a regression task of coordinate prediction for multiple control points.

To learn the coordinates of the control points sufficiently, we generate the CatmullRom splines ground truths described in the following paragraph and adopt a similar regression method as in [16] to regress the targets. For each region instance, we use:

$$\Delta_x = c_{ix} - x_{\min}, \Delta_y = c_{iy} - y_{\min}. \quad (3)$$

where x_{\min} and y_{\min} represent the minimum x and y values of the 4 vertexes, respectively. One of the key benefits of estimating the relative distance is its independence from the location of the CatmullRom spline control points relative to the image edges. Within the detection module, the task of learning the x and y offsets (denoted as Δ_x and Δ_y) is efficiently handled by a single convolutional layer equipped with 16 output channels. This approach enables the achievement of precise outcomes with minimal computational overhead.

1) *CatmullRom Ground Truth Generation*: As we know, most of the datasets used in IFDL consist of Mask or polygon-based images (more details can be referred to in the dataset instruction). To achieve CatmullRom splines-based regression location, we can simply apply the standard least square method, as shown in (4).

$$\begin{bmatrix} \mathbf{p}_{0,3}(t_0) & \cdots & \mathbf{p}_{3,3}(t_0) \\ \mathbf{p}_{0,3}(t_1) & \cdots & \mathbf{p}_{3,3}(t_1) \\ \vdots & \ddots & \vdots \\ \mathbf{p}_{0,3}(t_m) & \cdots & \mathbf{p}_{3,3}(t_m) \end{bmatrix} \begin{bmatrix} c_{x_0} & c_{y_0} \\ c_{x_1} & c_{y_1} \\ c_{x_2} & c_{y_2} \\ c_{x_3} & c_{y_3} \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{x_0} & \mathcal{P}_{y_0} \\ \mathcal{P}_{x_1} & \mathcal{P}_{y_1} \\ \vdots & \vdots \\ \mathcal{P}_{x_m} & \mathcal{P}_{y_m} \end{bmatrix} \quad (4)$$

where m represents the number of annotated points for a curved boundary, while t is calculated by using the ratio of the cumulative length to the perimeter of the polyline. $\mathbf{p}_{i,j}$ can be referred from (1), and we use \mathcal{P}_i represents the new coordinate points after the transformation. According to (1) and (4), we convert the original masked annotation to a parameterized CatmullRom spline. Illustration can be referred from Fig. 4.

Overall, given the annotated points $\{\mathbf{p}_i\}_{i=1}^n$ from the curved boundary where \mathbf{p}_i represents the i -th annotating point, our main goal is to obtain the optimal parameters for CatmullRom splines $c(t)$ in (1).

We give more specific processing algorithms here: in each tampered region, we first process the edges of the tampered region to obtain boundary points, and then further transform these boundary points into the corresponding control points via CatmullRom splines. We expect the model to regress these CatmullRom splines control points. The specific processing of the dataset steps can be concluded as follows:

- Selecting the boundary points of each tampered region;
- Getting the smallest outer rectangle of the region;
- Finding the boundary point with the shortest distance from the four vertices of the external rectangle, using the L_2 – norm as the distance metric;
- Performing a CatmullRom spline curve control point fitting solution for the boundary points corresponding to the long edges of the smallest outer rectangle, with the starting boundary point originating from the previous step;
- The CatmullRom control points of the dataset are obtained as Ground Truth.

In order to present the contribution of this section more clearly, we summarize the contribution of CatmullRom splines Detection as follows: 1) **Enhanced Shape Optimization**. We extend Catmull-Rom splines for boundary fitting by varying tensile factors, improving accuracy in complex IFDL tasks—a key innovation in our work. 2) **Application to Regression**. We develop

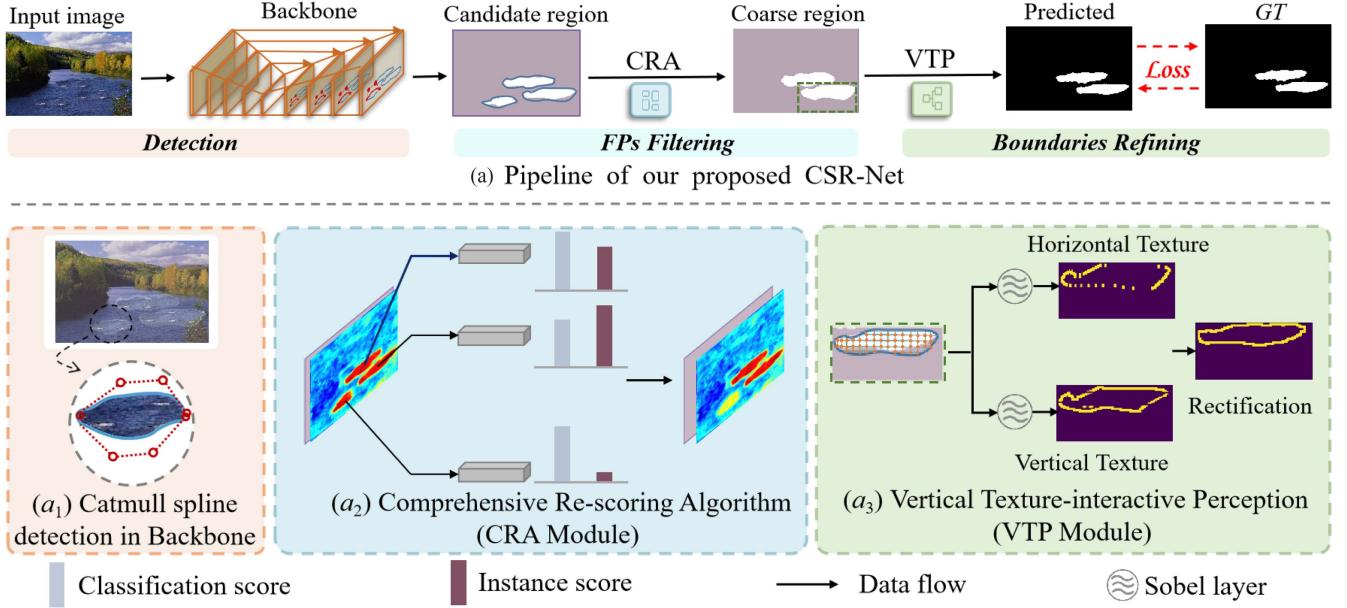


Fig. 3. Overall of our proposed CSR-Net. The top part is our pipeline.

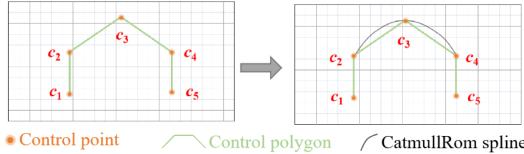


Fig. 4. An example of Cubic CatmullRom splines. Note that with only two end-points c_1 and c_5 the CatmullRom spline degenerates to a straight line.

a novel method using Catmull-Rom splines in regression tasks, predicting control points for precise boundary delineation and enhanced target localization. 3) **Polygon-to-Spline Conversion.** We present a technique for converting polygon-based boundaries to Catmull-Rom spline control points, crucial for accurate boundary representation in pixel-based IFDL tasks.

B. Comprehensive Re-Scoring Algorithm

First, let us briefly review the classic Mask R-CNN algorithm: During inference, the top-k (e.g., 1000) predicted bounding boxes after sorting (based on classification confidence) are subjected to standard NMS processing, and the top-M (e.g., 300) bounding boxes with the highest classification confidence are retained. These bounding boxes are provided to mask R-CNN as suggestions for generating predicted instance mappings. The core idea of the method is to consider the classification confidence of the resulting bounding box as a score, and then a pre-set threshold to filter out background boxes. However, even with progress, when a horizontal bounding box contains an instance of a clearly incompatible region, it is accompanied by a large amount of background information, and Mask R-CNN often filters out such low-scoring true positives, and, as a contrast, retains some confidence FPs. Based on the above observations, we perform a re-assignment of scores for each instance. In

concrete terms, the comprehensive score of region instance is composed of two parts: classification score (CLS) and instance score (INS). Mathematically, the comprehensive score for the i -th proposal, given the predicted n -class scores $\text{CLS} = \{s_{ij}^{\text{cls}} \mid j \in [0, \dots, n-1]\}$ and $\text{INS} = \{s_{ij}^{\text{ins}} \mid j \in [0, \dots, n-1]\}$ is computed via the customized softmax function (5).

$$s_{ij} = \frac{e^{s_{ij}^{\text{cls}} + s_{ij}^{\text{ins}}}}{\sum_{l=0}^{n-1} e^{s_{il}^{\text{cls}} + s_{il}^{\text{ins}}}}. \quad (5)$$

In IFDL, our target is to strictly distinguish between tampered (foreground) and authentic (background) regions in a suspicious image, which is obviously a pixel binary classification task (n is set to 2). In other words, the scores for the foreground class are our main concern. To be more specific, CLS is derived from a classification arm that parallels the architecture of Mask R-CNN. Meanwhile, INS represents the activation levels of the individual region instances within the holistic region segmentation map. For each instance, this score is mapped onto the corresponding tampered region segmentation map, denoted as $P_i = \{p_i^1, p_i^2, \dots, p_i^n\}$. The average of P_i over the area of the region instance can be expressed as:

$$s_{i1}^{\text{ins}} = \frac{\sum_j p_i^j}{N}. \quad (6)$$

where P_i refers to the collection of pixel values corresponding to the i -th region instance within the region segmentation map. The classification score is harmoniously combined with the instance score to yield a holistic score that serves to mitigate the false positive (FP) confidence in real-world applications. In practice, following a similar strategy to that used in Mask R-CNN, we retain regions where both the classification score and the instance score exceed a threshold of 0.5. The re-scoring algorithm ensures that only regions with sufficiently high confidence in

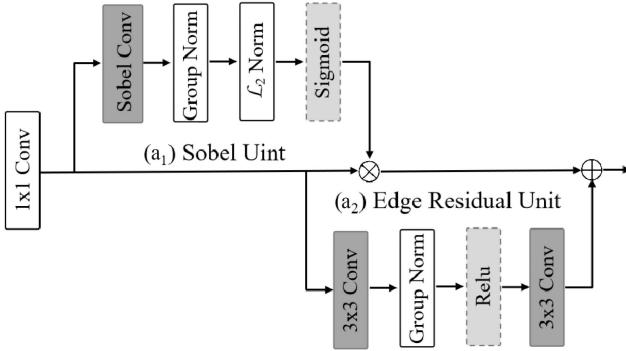


Fig. 5. Diagrams of Sobel layer, used in VTP for enhancing edge-related patterns and manipulation edge detection.

both classification and instance-level detection are considered as the forgery region. Overall, this approach is effective as false positives typically exhibit a less intense response compared to the regions represented on the segmentation map.

C. Vertical Texture-Interactive Perception

Traditional edge detection operators, such as Sobel, Roberts, and Prewitt, play a crucial role in extracting handcrafted features for natural image processing tasks. However, a notable limitation lies in their static nature, as they are unable to adapt dynamically to the specific requirements of a given task. Building upon the insights from [17], we integrate an adaptive edge detection operator into our model, termed the Sobel layer (refer to Fig. 5). To enhance the modeling of tampered area boundaries, we incorporate Vertical Texture-interactive Perception (VTP) into our network. In VTP, the tampered region is represented by a set of contour points. These points, characterized by robust texture features, enable precise localization of tampered regions with arbitrary shapes.

The VTP architecture encompasses two essential parallel branches. In the upper branch, a convolutional kernel of size $1 \times k$ traverses the feature maps, capturing local texture information along the horizontal axis. This approach emphasizes texture characteristics within a k -range region, demonstrating its efficacy in our preliminary experiments. Notably, this technique proves to be both straightforward and resource-efficient, contributing to its practical utility. Similarly, the lower branch adopts a comparable approach to model texture characteristics, but in the vertical direction. Here, a convolutional kernel with dimensions $k \times 1$ is employed, with the hyper-parameter k regulating the receptive field size for texture characteristics. In our practical experiments, we set $k = 3$ for optimal results. Subsequently, the normalization of heatmaps to the $[0, 1]$ range in both horizontal and vertical directions is achieved through the incorporation of two separate sigmoid layers. This process facilitates the detection of tampered regions along two orthogonal directions, leading to the representation of these regions using contour points in distinct heatmaps. Each heatmap selectively responds to texture characteristics specific to its designated direction.

To mitigate the impact of false positive predictions, an additional refinement is applied using the Point Re-scoring

Algorithm on the two heatmaps generated by VTP. Specifically, points within different heatmaps undergo Non-Maximum Suppression (NMS) to obtain a refined representation. To further suppress predictions characterized by strong unidirectional responses or weakly orthogonal responses, only points exhibiting distinct responses in both heatmaps are retained as candidates. Subsequently, the tampered region is accurately delineated by a polygon formed by these high-quality contour points.

D. Optimization

As described above, our network includes multi-task. Therefore, we calculate the loss function for the following components:

$$L = L_{rpn} + \lambda_1 \cdot L_{cls} + \lambda_2 \cdot L_{mask} + \lambda_3 \cdot L_{gts} + \lambda_4 \cdot L_{CR}. \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are set to be 0.3, 0.2, 0.2, 0.4, respectively in practical experiments. L_{rpn} , L_{cls} and L_{mask} are the standard loss derived from Mask R-CNN. The L_{gts} is used to optimize tampered region detection, defined as:

$$L_{gts} = \frac{1}{N} \sum_i -\log \left(\frac{e^{p_i}}{\sum_j e^{p_j}} \right). \quad (8)$$

The L_{gts} is Softmax loss, where p is the output prediction of the network.

The L_{CR} is used to optimize the fit of CatmullRom spline detection, defined as:

$$L_{CR} = L_{ctr} + L_{bias}. \quad (9)$$

The L_{ctr} and L_{bias} are all FCOS loss [18]. The former is used to optimize distance loss from the center of CatmullRom control points, while the offset distance of these control points from the center is constrained by the latter.

IV. EXPERIMENT

A. Experimental Setup

Pre-training Data: We create a sizable image tampering dataset and use it to pre-train our model. This dataset includes three categories: 1) splicing, 2) copy-move, and 3) removal. Our process on this dataset can be the following:

For splicing, we use the MS COCO [19] to generate spliced images. We adopt the same transformation as [20], including the scale, rotation, shift, and luminance changes. Since the spliced region is not always an object, we create random outlines using the curves and fill them to create splicing masks. For copy-move, the datasets from MS COCO and [21] are adopted. For removal, we adopt the SOTA inpainting method [22] to fill one annotated region that is randomly removed from each chosen MS COCO image. We randomly add Gaussian noise or apply the JPEG compression algorithm to the generated data to resemble the visual quality of images in realistic scenarios.

We compile a substantial image tampering dataset to pre-train our model, encompassing three distinct categories: 1) splicing, 2) copy-move, and 3) removal. Our dataset processing involves the following steps:

- For splicing, MS COCO [19] serves as the source for generating spliced images. Employing transformations akin to those outlined in [20], such as scale, rotation, shift, and luminance changes, we create spliced masks by generating random outlines, recognizing that the spliced region may not always represent a distinct object.
- Copy-move scenarios are addressed by utilizing datasets from both MS COCO and [21].
- Removal instances involve applying a state-of-the-art inpainting method [22] to fill in randomly removed annotated regions from selected MS COCO images.
- To simulate realistic visual scenarios, we introduce randomness to the generated data by adding Gaussian noise or applying the JPEG compression algorithm, aligning with the visual quality typical of real-world images.

Testing Datasets: Following [23], [24], we evaluate our model on CASIA [25], Columbia [26], NIST16 [27], COVER [28] and Openforensics [29]. Here is a more detailed introduction of this dataset.

CASIA is proposed to evaluate the authenticity and integrity of digital images. Fixed-size tampered images that make up the database v1.0, which are generated only by using crop-and-paste operation under Adobe Photoshop. While the database v2.0 is more comprehensive and challenging due to the post-processing applied in most tampered examples.

Columbia shares 180 tampered images manipulated only by splicing, which mainly focuses on uncompressed images.

NIST16 is a high-quality and challenging dataset. All three tampering techniques(copy-move, splicing, and removal) are included. This dataset focuses on evaluating the performance of algorithms for detecting and localizing image forgery in various image formats (PEG, TIFF, PNG, and BMP).

COVER was created in collaboration between Stanford University and Princeton University. In this dataset, There are 100 forged images which are all edited by copy-move.

OpenForensics is tampered by deep generative models (DGMs), OpenForensics contains 18895 tampered images of several facial images, generated via GAN and incorporating both genuine and tampered facial images in the latter category.

Besides, to further ensure the generalization of the method, a public Real-World IFL (RIFL) dataset [30] named MTRL (Multi-Tampered-Region Localization) is specifically introduced for training and testing. In **RIFL**, a mixture of multiple, carefully designed tampering patterns(including copy-move, splicing, and removal) appears in different locations on the same image. Meanwhile, some classical means of processing digital images for social media are imposed on this dataset, including Compression (various algorithms such as JPEG and PNG), Cropping(crop user-uploaded images into different sizes to fit different devices), Adjusting brightness and contrast(make images more vivid and bright), Removing metadata(remove metadata to protect user privacy and reduce file size). These modifications make RIFL a more challenging dataset. More details can be seen in Table I. Note that, we utilize RIFL22 [30] for training and RIFL21 [30] for testing.

Evaluation Metrics To quantify the localization performance, following previous works [24], we use pixel-level Area

TABLE I
THE DATASETS INVOLVED IN OUR EXPERIMENTS

Task	Type	Dataset	Total	C-M Sp Re
STRL	Training	CASIAv2 [25]	5063	3235 1828 0
	Testing	COVER [28]	100	100 0 0
	Testing	Columbia [26]	180	0 180 0
	Testing	NIST16 [27]	564	68 288 208
	Testing	CASIAv1 [25]	920	459 461 0
MTRL	Training	RIFL22 [30]	4000	-
	Testing	RIFL21 [30]	2005	-

C-M: copy-move; Sp: splicing; Re: removal. Note that every image in RIFL dataset contains multiple manipulation types.

Under Curve (AUC) and F1 score on manipulation masks. Since binary masks are required to compute F1 scores, we adopt the Equal Error Rate (EER) threshold to binarize them.

Implementation Details: Specifically, we set the batch size to 4 for each dataset, with a crop size of 512×512 . The model is optimized using the Stochastic Gradient Descent (SGD) optimizer, following a poly learning rate schedule with an initial learning rate of 0.007, momentum of 0.9, and a weight decay of 5e-4. Our model is trained end-to-end, without staged pre-training of individual components, for a maximum of 200 epochs. Additionally, the total loss is backpropagated as a whole. All hyperparameters remain fixed during evaluation.

B. Comparison With the SOTA Methods

Following classic methods [23], [24], our model is compared with other state-of-the-art tampering localization methods under two settings: 1) training on the synthetic dataset and evaluating the full test datasets, and 2) fine-tuning the pre-trained model on the training split of test datasets and evaluating on their test split. The pre-trained model will demonstrate each method's generalizability, and the fine-tuned model will demonstrate how well each method performs locally once the domain discrepancy has been significantly reduced.

Aligned with classical approaches [23], [24], our model undergoes comprehensive comparisons with leading tampering localization methods in two distinct scenarios: Initial evaluation involves training on a synthetic dataset and subsequently assessing performance on complete test datasets. Further investigation includes fine-tuning the pre-trained model on the training split of test datasets, followed by evaluation on their respective test splits. The pre-trained model serves to showcase each method's generalizability, while the fine-tuned model elucidates the local performance of each method after a substantial reduction of domain discrepancies.

Pre-trained Model: Table II shows the localization performance of pre-trained models for different SOTA methods on five datasets under pixel-level AUC. Our CSR-Net achieves the best localization performance on Coverage, CASIA, NIST16 and IMD20, ranking third on Columbia. Especially, It achieves 94.3% on the copy-move dataset (COVER), whose image forgery regions are indistinguishable from the background. This validates our model owns the superior ability to suppress the FPs and generates more accurate edges. Yet, we fail to achieve the

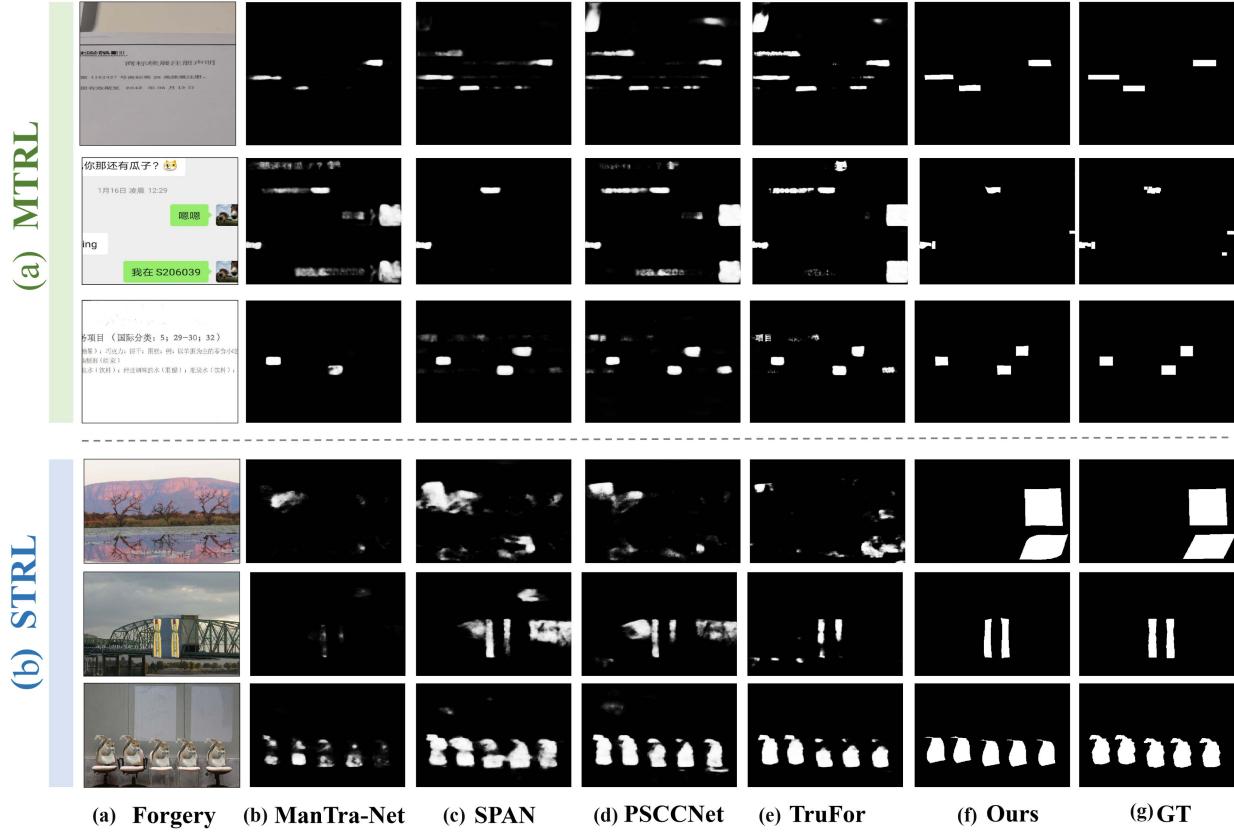


Fig. 6. Visualization of the predicted manipulation mask by different methods. From left to right, we show forged images, predictions of ManTra-Net, SPAN, PSCCNet, TruFor and GT masks.

TABLE II
COMPARISONS OF MANIPULATION LOCALIZATION AUC (%) SCORES OF DIFFERENT PRE-TRAINED MODELS

Method	Data	Columbia	Coverage	CASIA	NIST16	IMD20
SPAN [24]	96k	93.6	92.2	79.7	84.0	75.0
TruFor [31]	100k	97.7	85.4	83.3	83.9	81.8
PSCCNet [3]	100k	98.2	84.7	82.9	85.5	80.6
ObjectFormer [23]	62K	95.5	92.8	84.3	87.2	82.1
ManTraNet [32]	64K	82.4	81.9	81.7	79.5	74.8
NCL [33]	64K	95.1	83.5	80.1	80.2	76.6
PCL [34]	64K	94.5	82.7	78.9	79.6	74.8
MTCAM [35]	64K	94.5	91.2	83.3	83.5	78.6
GLSTR [36]	64K	93.6	91.3	84.5	83.7	79.8
Ours	60K	96.8	94.3	88.1	88.3	85.4

best performance on Columbia, with a gap of 1.4 % AUC score lower than that of PSCCNet. We conjecture that the explanation may be the distribution of their synthesized training data closely resembles that of the Columbia dataset. This is further supported by the results in Table III, which shows that CSR-Net performs better than PSCCNet in terms of both AUC and F1 scores. Furthermore, it is worth pointing out that we achieve decent results with less pre-training data.

Fine-tuned Model: The pre-trained model's network weights serve as the initial parameters for fine-tuning on the training splits of the Coverage, CASIA, and NIST16 datasets. Concretely, following the setting from [21], [37], [39], we fine-tune

TABLE III
COMPARISON OF MANIPULATION LOCALIZATION RESULTS USING FINE-TUNED MODELS

Methods	Coverage		CASIA		NIST16		Openforensics	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
J-LSTM [37]	61.4	-	-	-	76.4	-	63.2	59.1
H-LSTM [38]	71.2	-	-	-	79.4	-	66.5	63.1
TruFor [31]	77.0	68.0	85.3	56.3	85.3	79.5	80.1	67.2
SPAN [24]	93.7	55.8	83.8	38.2	96.1	58.2	79.3	63.4
PSCCNet [3]	94.1	72.3	87.5	55.4	99.1	74.2	91.8	73.2
ObjectFormer [23]	95.7	75.8	88.2	57.9	99.6	82.4	93.5	74.9
RGB-N [39]	81.7	43.7	79.5	40.8	93.7	72.2	85.3	66.8
NCL [33]	84.1	57.8	79.5	48.8	93.9	77.0	92.5	71.3
PCL [34]	83.5	56.9	79.5	46.9	93.6	75.8	91.3	70.2
MTCAM [35]	85.3	61.6	85.6	51.2	94.2	78.2	93.5	74.2
GLSTR [36]	84.9	63.5	84.9	50.6	94.6	80.2	93.2	73.2
Ours	97.9	78.0	90.4	58.5	99.7	83.5	94.2	75.3

our model on the training splits of four standard benchmarks after initial training on synthetic data. For NIST16 and Coverage, we use the exact splits provided by RGB-N [39]. For CASIA, we utilize CASIAv2 for training and CASIAv1 for testing, as specified in RGB-N [39]. The stopping criteria for fine-tuning are as follows: For NIST16 and Coverage, we use cross-validation instead of a fixed validation set due to the limited size of the training splits. For CASIA, we randomly sample 300 images from the training split to create a validation set. The model with the highest validation AUC score is selected for evaluation on the test set. It is important to note that the pre-trained feature

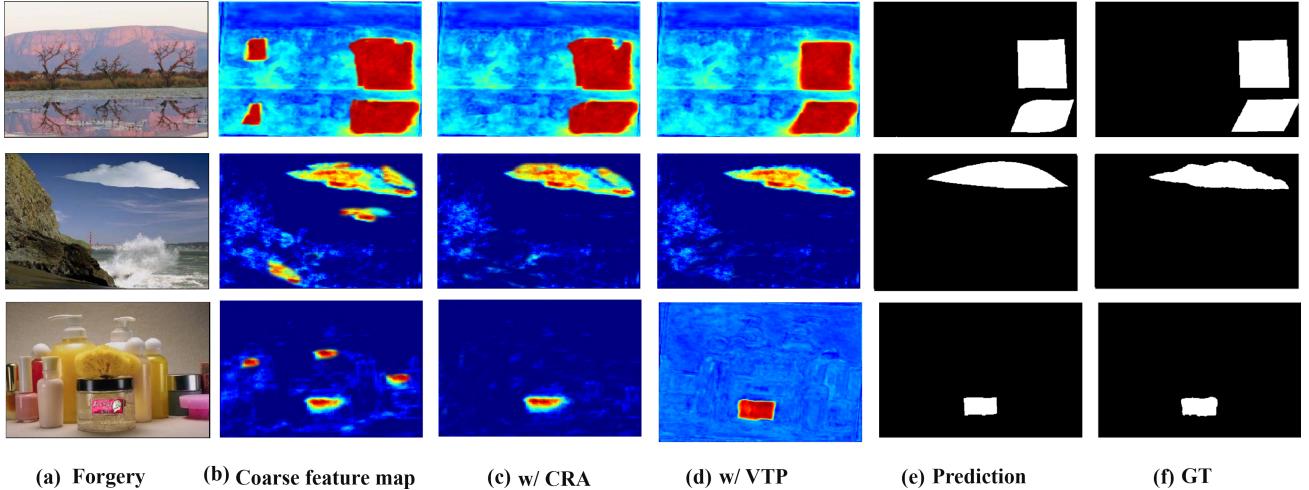


Fig. 7. Visualization of heatmap. From left to right, they are (a) forgery images (input); (b) the output feature maps of CatmullRom splines-based method; (c) the input feature maps of CRA; (d) the fusion feature maps of VTP; (e) the predicted masks; (f) the corresponding ground truth masks.

extractor performs optimally on images of the original size. The performance of various methods' fine-tuned models is assessed, as detailed in Table III. Notably, our model demonstrates substantial improvements in AUC and F1 metrics. It can be seen that our method achieved 94.2% AUC and 75.3% F1-score (i.e., 81.3% precision and 70.1% recall), which surpassed the baseline model J-LSTM with a larger advantage (31.0% precision, 16.2% recall). It is noted that J-LSTM achieved only 59.1% F1-score with 63.4% precision and 55.3% recall. This underscores the efficacy of the CRA module in effectively mitigating false positive instances and enhancing the precision of predicted region locations and boundaries by VTP.

Upon synthesizing the information presented in Tables II and III, our methodology convincingly demonstrates the effectiveness of incorporating regression techniques for pixel-level tasks, aligning with the expectations outlined in Section I.

C. Visualization Results

Qualitative Results: Fig. 6 presents predicted forgery masks from various methods. Due to the unavailability of ObjectFormer's source code [23], its predictions are not included. Our CSR-Net outperforms state-of-the-art methods, reducing false positives and more accurately delineating tampered regions. This improvement is attributed to the Comprehensive Re-scoring Algorithm (CRA), which identifies subtle differences between tampered and authentic areas, and Vertical Texture-Interactive Perception (VTP), which models texture boundaries to precisely depict target regions. Image forgeries commonly occur in political propaganda, evidence tampering, product promotion, internet fraud, special effects, and espionage, with social media being a major platform. Fig. 10 shows that our method maintains superior performance on a social media dataset.

Visualization of Feature Maps: In order to show our pipeline more clearly, visualization of feature maps is shown in Fig. 7. From left to right, they are (a) forgery images (input); (b) the output feature maps of the CatmullRom splines-based method;

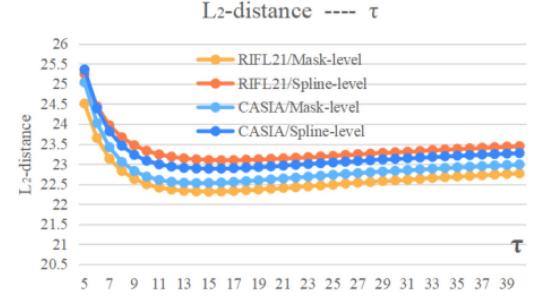


Fig. 8. L_2 distance with different value of τ . We show the results through the format of “dataset/label”, for example, RIFL21/Mask-level means the average distance of control points in RIFL21 from the Ground Truth with Mask-level.

(c) the input feature maps of CRA; (d) the fusion feature maps of VTP; (e) the predicted masks; (f) the corresponding ground truth masks.

D. Ablation Analysis

In this section, we conduct ablation experiments to demonstrate the effectiveness of our proposed method CSR-Net.

Impact on each module: Formally, we introduce Catmull-Rom Splines-based Regression (CSR) to enhance the precision in delineating tampered regions, surpassing the capabilities of traditional regression methods. The Comprehensive Re-scoring Algorithm (CRA) is designed to select regions with not only high classification scores but also superior instance scores. Concurrently, Vertical Texture-Interactive Perception (VTP) is employed to model horizontal and vertical texture features, refining the target region. To assess the efficacy of CSR, CRA, and VTP, each component is individually removed, and the resulting impact on forgery localization performance is evaluated on the CASIA and NIST16 datasets. Quantitative results are presented in Table IV. The baseline (I) represents the utilization of the traditional regression method [40]. Ablation experiments reveal a 1.9% decrease in F1 scores on CASIA and a 1.7% decrease on

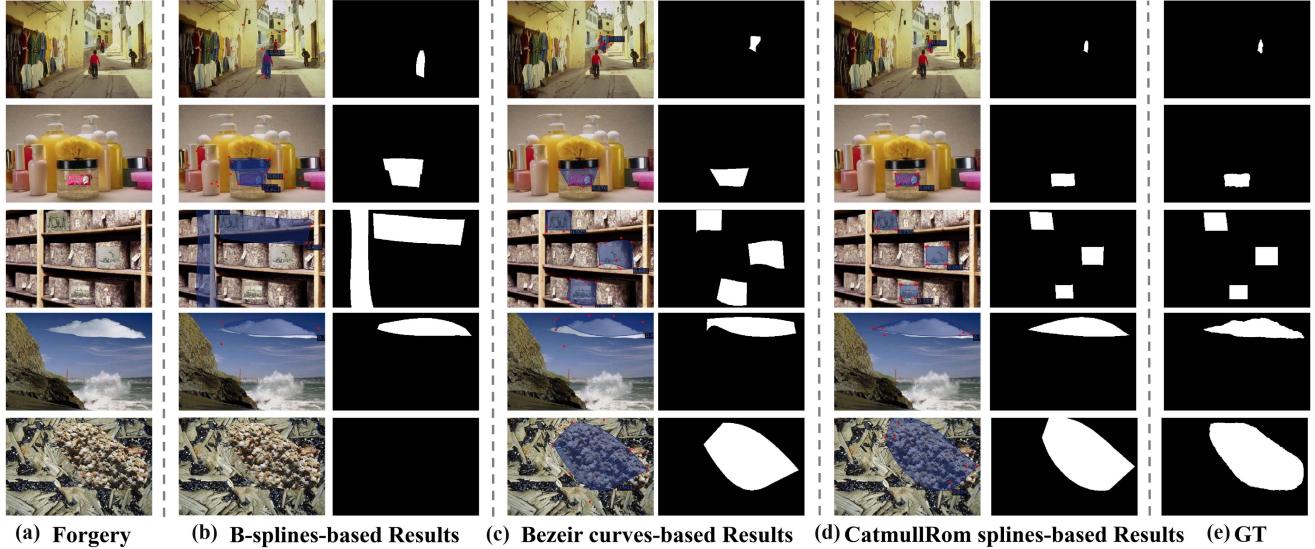


Fig. 9. Visualization of the results by different splines-based regression.

TABLE IV
ABLATION RESULTS ON CASIA AND NIST16 DATASET USING DIFFERENT VARIANTS OF CSR-NET

Index	Variants	CASIA		NIST16	
		AUC	F1	AUC	F1
I	Baseline	68.5	35.3	75.9	51.2
II	w/o CSR	78.1	47.6	86.1	62.1
III	w/o CRA	86.3	55.5	95.9	78.8
IV	w/o VTP	88.9	56.9	97.8	81.8
V	Ours	90.4	58.8	99.7	83.5

AUC and F1 scores (%) are reported.

NIST16 when VTP is omitted. In the absence of CRA, there is a more pronounced decrease in AUC scores, particularly observed in (IV). Notably, when CRA is excluded, a substantial performance degradation is evident in (II), with a 12.3% decrease in AUC and an 11.2% decrease in F1 on CASIA.

Value of τ : In Fig. 8, diverse values of the parameter τ in CatmullRom Ground Truth Generation are depicted to assess their impact on prediction outcomes across two distinct datasets: the natural image dataset (CASIA) and the social media dataset (RIFL21). Observably, as τ incrementally rises, the Euclidean distance between the fitted CatmullRom control points and the Ground Truth with mask-level in various datasets progressively diminishes, indicating an improved fit. However, once τ surpasses 16, there is an evident tendency for the Euclidean distance to increase, suggesting a diminishing fitting effect. Conclusively, selecting $\tau = 16$ emerges as an optimal choice for generating CatmullRom-based Ground Truth that aligns well with the data characteristics.

Different splines-based Regression: Interpolation functions like Catmull-Rom splines and Bezier curves are crucial in various applications. Catmull-Rom splines interpolate data points through nodes, while Bezier curves approximate them. In Image Forgery Detection and Localization (IFDL), datasets with varied shapes, such as those from natural images and social media,

TABLE V
ABLATION EXPERIMENTS

Backbone	CASIA		NISIT16	
	AUC	F1	AUC	F1
MobileNet	87.6	57.8	98.1	81.9
ShuffleNet	88.2	58.3	98.3	82.5
ResNet-50 (Ours)	90.4	58.6	99.7	83.6
Threshold	AUC	F1	AUC	F1
0.25	88.9	56.8	97.1	82.4
0.75	89.3	56.3	97.5	81.3
0.5 (Ours)	90.4	58.6	99.7	83.6

(I-II) quantitative results about different edge feature extraction methods. (III-V) different backbones

benefit from Catmull-Rom splines, which outperform Bezier curves in handling diverse curvatures (Fig. 11). We also incorporate B-splines into Image Forgery Localization (IFL), which are widely used in curve and surface modeling. Our method excels in locality and fitting by enabling precise local adjustments to curve shapes without altering the entire curve. Results are shown in Fig. 11.

Different Backbones: To evaluate the impact of different backbone networks on model performance, we conducted ablation experiments. Our state-of-the-art (SOTA) model utilizes ResNet-50 as the backbone network and was compared with models employing MobileNet and ShuffleNet. As shown in Table V (Top), ResNet-50 outperformed the other backbones, achieving an AUC of 90.4 and an F1 score of 58.6 on the CASIA dataset. In contrast, the baseline model using MobileNet showed a decrease of 2.8 and 0.8 percentage points in AUC and F1 score, respectively.

Different Thresholds: To evaluate the impact of different thresholds, we conducted ablation experiments. We can conclude that when the threshold is set to 0.5, we achieve optimal performance. Initially, our model employs a sigmoid activation function in its final layer, which yields outputs between 0 and 1. This feature is particularly beneficial for binary classification

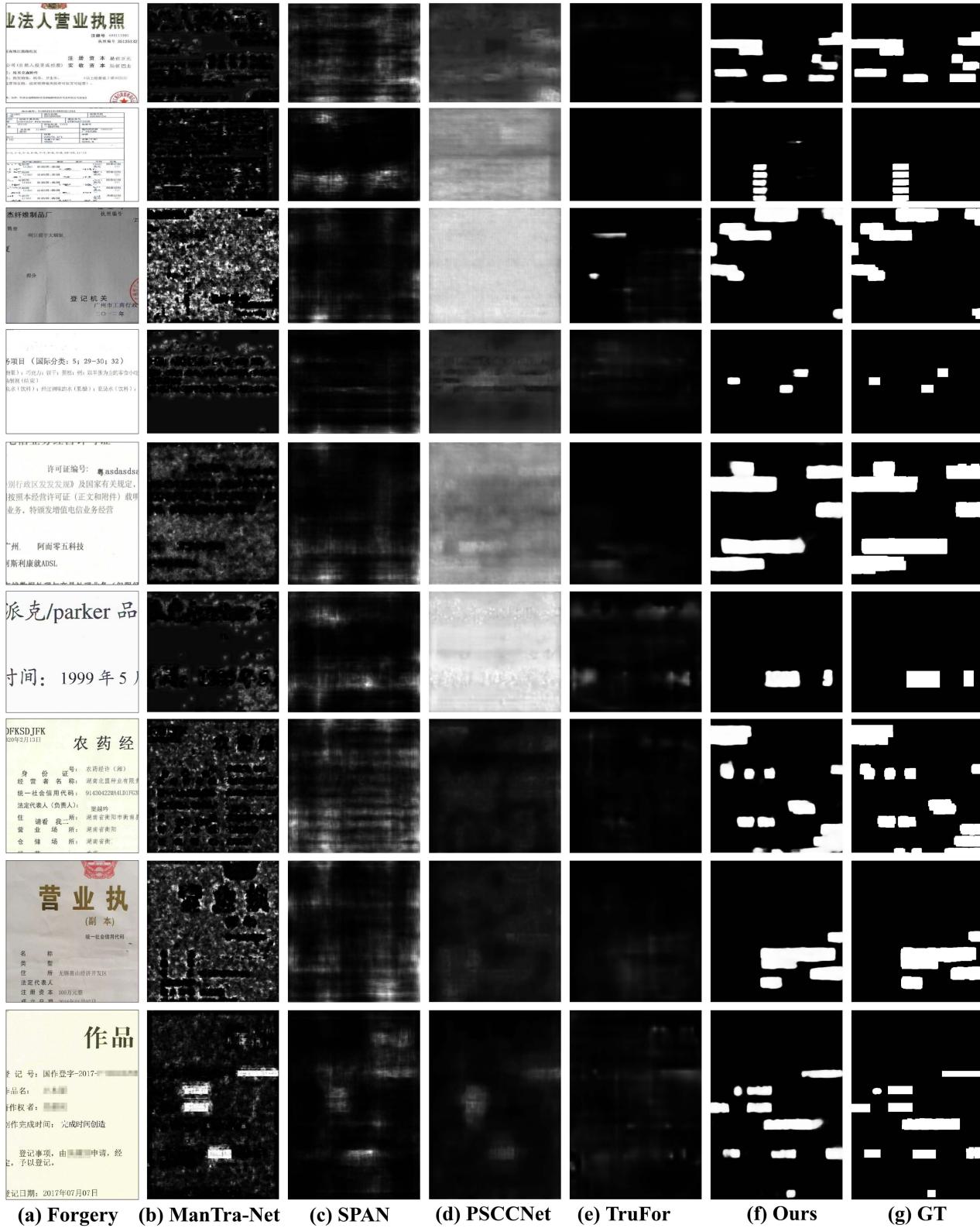


Fig. 10. Visualization of the predicted manipulation mask by different methods on Rifl21. From left to right, we show forged images, predictions of ManTra-Net, SPAN, PSCCNet, TruFor, Ours and GT masks. It is noted that our method maintains competitive performance, performs better against false positives and misses among the methods, and keeps almost the same pace as GT.

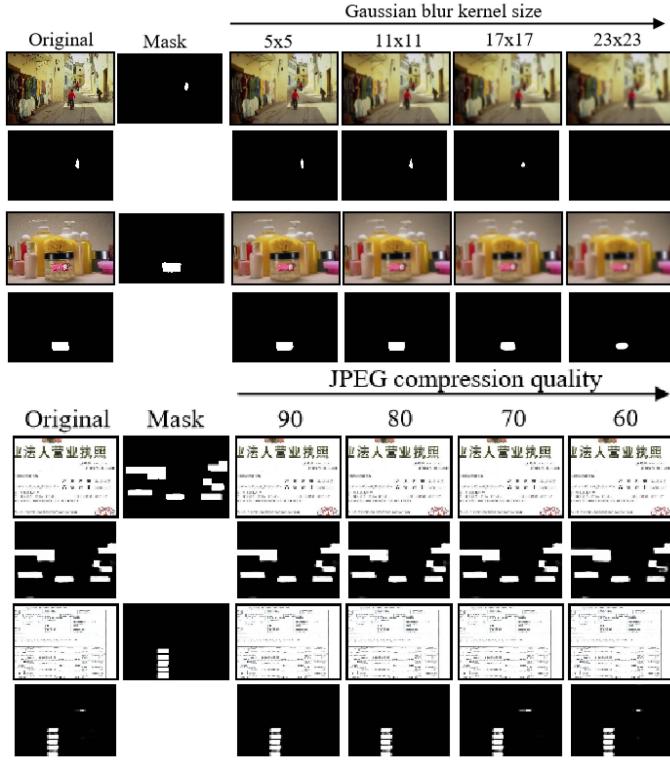


Fig. 11. Qualitative results against Gaussian blur and JPEG compression.

tasks, such as identifying manipulated images. The sigmoid function's crossing point at 0.5 offers a natural threshold for binary classification.

E. Robustness Evaluation

In this section, we apply different image distortion methods followed by SPAN [24] on raw images from NIST16. Several distortions types are included: 1) image scaling with different scales (*Resize*), 2) applying Gaussian blur with kernel size k (*GSBlur*), 3) Gaussian noise with a standard deviation and σ (*GSSNoise*), 4) performing JPEG compression with quality factor q (*JPEGComp*). Also, we offer a mixed version (*Mixed*) of the aforementioned distortions, where the resizing scale, kernel size k , standard deviation, and quality factor q are all randomly chosen from the intervals $[0.25, 0.78]$, $[3, 15]$, $[3, 15]$, and $[50, 100]$, respectively. The robustness study results are displayed in Table VI. Under all distortions, our CSR-Net is more robust than the SPAN and the HP-FCN. It is noted that when posting photographs to social media, resizing is frequently done. Our model performs significantly better on compressed images than other methods, which mainly attributes the improvement to VTP through fine-grained texture modeling.

Gaussian Blur: Going further, we apply Gaussian blur to each test image in CASIA following [24]. The qualitative results under different degradations are shown in Fig. 11 (Top), we set the Gaussian blur kernel size to be 5×5 , 11×11 , 17×17 , and 23×23 respectively. Note that 17×17 belongs to a higher degradation level in real social media, and our method can still maintain accurate localization performance.

TABLE VI
LOCALIZATION PERFORMANCE ON NIST16 DATASET UNDER VARIOUS DISTORTIONS

Distortion	SPAN	ObjectFormer	Ours
no distortion	83.95	87.18	89.32
Resize($0.78 \times$)	83.24	87.17	89.23 0.09 ↓
Resize($0.25 \times$)	80.32	86.33	87.74 1.58 ↓
Blur($k = 3$)	83.10	85.97	89.01 0.31 ↓
Blur($k = 15$)	79.15	80.26	86.88 2.44 ↓
Noise($\sigma = 3$)	75.17	79.58	88.10 1.22 ↓
Noise($\sigma = 15$)	67.28	78.15	83.38 5.94 ↓
Compress($q = 100$)	83.59	86.37	89.26 0.06 ↓
Compress($q = 50$)	80.68	86.24	88.63 0.69 ↓

AUC scores are reported (in %), (blur: gaussianblur, noise: gaussiannoise, compress: JPETGCompress.)

JPEG Compression: From Table VI, we can infer that our model has better robustness than the SOTA method ObjectFormer. However, these degradations still significantly affect the performance in some cases. We all know that there are various degradations in real social media, especially JPEG compression, which will bring greater challenges to the IFDL task. Qualitative results can be seen in Fig. 11 (down). It can be seen that our method has good resistance to JPEG compression. Note that JPEG compression is usually found in social media, we chose the RIFL dataset for evaluation.

V. CONCLUSION

This paper presents the Catmull-Rom Splines-based Regression Network (CSR-Net) for Image Forgery Detection and Localization (IFDL), which moves beyond traditional bounding box-based detection by integrating the Catmull-Rom fitting technique for contour modeling of target regions. This approach enhances the precision and efficiency of tampered region localization at the pixel level. To address false positives (FPs), we introduce the Comprehensive Re-scoring Algorithm (CRA), which filters tampered regions based on classification and instance scores. Additionally, the Vertical Texture-Interactive Perception (VTP) module is proposed for refined edge detection. CSR-Net demonstrates minimal false positives and precise localization, outperforming state-of-the-art methods across both natural image and social media datasets in extensive experiments.

REFERENCES

- [1] B. DeCann and K. Trapeznikov, “Comprehensive dataset of face manipulations for development and evaluation of forensic tools,” 2022, *arXiv:2208.11776*.
- [2] A. Nichol et al., “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” in *Proc. Int. Conf. Mach. Learn. PMLR*, 2022, pp. 16784–16804.
- [3] X. Liu, Y. Liu, J. Chen, and X. Liu, “PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7505–7517, Nov. 2022.
- [4] D. Li, J. Zhu, M. Wang, J. Liu, X. Fu, and Z.-J. Zha, “Edge-aware regional message passing controller for image forgery localization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8222–8232.
- [5] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, “Image forgery localization via fine-grained analysis of CFA artifacts,” *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.

- [6] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "A Bayesian-MRF approach for PRNU-based image forgery detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 4, pp. 554–567, Apr. 2014.
- [7] T. Carvalho, F. A. Faria, H. Pedrini, R. d. S. Torres, and A. Rocha, "Illuminant-based transformed spaces for image forensics," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 4, pp. 720–733, Apr. 2016.
- [8] C. Iakovovidou, M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Content-aware detection of jpeg grid inconsistencies for intuitive image forensics," *J. Vis. Commun. Image Representation*, vol. 54, pp. 155–170, 2018.
- [9] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. L. Dugelay, and P. Marc, "Defacto: Image and face manipulation dataset," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.
- [10] C. Yang, H. Li, F. Lin, B. Jiang, and H. Zhao, "Constrained R-CNN: A general image manipulation detection model," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6.
- [11] C. Nie, Z. Ju, Z. Sun, and H. Zhang, "3D object detection and tracking based on LiDAR-camera fusion and IMM-UKF algorithm towards highway driving," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 4, pp. 1242–1252, Aug. 2023.
- [12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [13] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19830–19843.
- [14] Y. You et al., "Canonical voting: Towards robust oriented bounding box detection in 3D scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1193–1202.
- [15] M. Pittner, A. Condurache, and J. Janai, "3D-SpLineNet: 3D traffic line detection using parametric spline representations," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 602–611.
- [16] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2020.
- [17] K. S. Holla and B. Lee, "Convolutional residual blocks with edge guidance for image denoising," in *Proc. IEEE 13th Int. Conf. Inf. Commun. Technol. Convergence*, 2022, pp. 645–647.
- [18] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
- [19] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Springer, Sep. 2014, pp. 740–755.
- [20] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14185–14193.
- [21] Y. Wu, W. Abd-Almageed, and P. Natarajan, "BusterNet: Detecting copy-move image forgery with source/target localization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 168–184.
- [22] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7760–7768.
- [23] J. Wang et al., "Objectformer for image manipulation detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2364–2373.
- [24] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 312–328.
- [25] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *Proc. 2013 IEEE China Summit Int. Conf. Signal Inf. Process.*, 2013, pp. 422–426.
- [26] J. Hsu and S. Chang, "Columbia uncompressed image splicing detection evaluation dataset," DVMM Lab, Columbia Univ. CalPhotos Digit Libr., 2006.
- [27] H. Guan et al., "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *Proc. 2019 IEEE Winter Appl. Comput. Vis. Workshops*, 2019, pp. 63–72.
- [28] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "Coverage—A novel database for copy-move forgery detection," in *Proc. 2016 IEEE Int. Conf. Image Process.*, 2016, pp. 161–165.
- [29] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10117–10127.
- [30] Alibaba, "Real-world image forgery localization dataset," 2021/2022. [Online]. Available: <https://tianchi.aliyun.com/competition/entrance/531945/introduction?spm=5176.12281957.0.0.1aa2448THhl4>
- [31] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 20606–20615.
- [32] Y. Wu, W. AbdAlmageed, and P. Natarajan, "Mantra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9543–9552.
- [33] J. Zhou, X. Ma, X. Du, A. Y. Alhammadi, and W. Feng, "Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 22346–22356.
- [34] Y. Zeng, B. Zhao, S. Qiu, T. Dai, and S.-T. Xia, "Toward effective image manipulation detection with proposal contrastive learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4703–4714, Sep. 2023.
- [35] D. Zhu et al., "MTCAM: A novel weakly-supervised audio-visual saliency prediction model with multi-modal transformer," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 2, pp. 1756–1771, Apr. 2024.
- [36] S. Ren, N. Zhao, Q. Wen, G. Han, and S. He, "Unifying global-local representations in salient object detection with transformers," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 4, pp. 2870–2879, Aug. 2024.
- [37] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4970–4979.
- [38] J. H. Bappy, C. Simons, L. Nataraj, B. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and encoder-decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, Jul. 2019.
- [39] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1053–1061.
- [40] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented repoints for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1829–1838.



Li Zhang is currently working toward the Ph.D. degree in computer science and application with the University of Science and Technology of China, Hefei, China. He has authored or coauthored several top-tier conference papers at ECCV, NeurIPS, and ICML. His research interests include image processing and image forgery localization.



Dong Li received the B.E. degree from the Hefei University of Technology, Hefei, China, in 2020. He is currently working toward the Ph.D. degree with the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA), University of Science and Technology of China, Hefei, China. His research interests focus on visual security, multimodal learning, and image enhancement.



Yan Zhong received the B.Sc. degree from the University of Northwestern Polytechnical University (NWPU), Xi'an, China, in 2019, and the M.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2022. He is currently working toward the Ph.D. degree with Peking University, Beijing, China. His research interests include multi-label feature selection, machine learning, and computer vision.



Jiaying Zhu received the B.Eng. degree in 2021 from the Sun Yat-sen University, Guangdong, China, where she is currently working toward the Ph.D. degree with the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA). Her research interests include image forgery localization and detection.



Xue Wang received the Ph.D. degree in biophysics from the University of Science and Technology of China, Hefei, China, in 2019. She is currently the Deputy Senior Researcher with Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Science. She has authored or coauthored more than 30 papers in this field, applied for 26 invention patents, and obtained 29 software copyrights. Her research interests include soil Big Data modeling, smart farm knowledge modeling, crop phenotype data modeling, and pest and disease data modeling.



Rujing Wang received the B.E. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 1987, and the M.S. degree in electronic engineering from the Dalian University of Technology, Dalian, China, in 1990, and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005. He is currently with the Institute of Intelligent Machinery, Chinese Academy of Sciences, as a Professor and Researcher. His research interests mainly include intelligent agriculture, agricultural Internet of Things, and agricultural knowledge engineering.



Liu Liu received the B.S. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2015, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2020. From 2020 to 2022, he was a Postdoctoral Researcher with Shanghai Jiao Tong University, Shanghai, China. He is currently a Associate Professor with the Hefei University of Technology, Hefei, China. His research interests include computer vision, deep learning, and embodied AI.



Kengyu Wu received the B.Sc degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2018, and the Ph.D. degree from the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China, in 2023. He is currently a Postdoctoral Fellow with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University (PolyU), Hong Kong SAR, China. His research interests include causality-based machine learning and automatic machine learning.