# HCF-Net: Hybrid Coarse-to-Fine Network for Forgery Reconstruction

Long Zhuo[1], Shunquan Tan*[1,2]

[1]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Shenzhen Institute of Artificial Intelligence and Robotics for Society, China

[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

## Abstract

Due to the ubiquity of photo editing software, it is convenient and prevalent to create fake images which may cause terrible misunderstandings. To address this issue, we introduce forgery reconstruction, a novel image reconstruction that automatically reconstructs a forged image into its original appearance. To achieve this, we present a new two-phase system, including a forgery localization phase and a reconstruction phase, where we first localize forged regions and then inpaint the natural texture. The proposed approach, called HCF-Net, is a hybrid two coarse-to-fine networks. Unlike the existing solutions of image reconstruction that relies on the prior knowledge of the areas to be recovered, our proposed system is free from the mask information. To guide the reconstruction, we propose a new strategy for extracting noise features in the forgery localization phase, which leverages the relationships among the image channels and localizes the tampered regions more precisely. For better performance, we equip a dual attention module in the reconstruction phase. Our automatic system is applicable to localize the tampered areas and reconstruct the missing part in a single pass. The extensive experiments demonstrate that our system achieves state-of-the-art performance in forgery localization and generates higher-quality and more flexible results than traditional inpainting methods.

## Keywords

HCF-Net, forgery reconstruction, coarse-to-fine network

## 1. Introduction

Low-cost tampering generation processes have negatively affected many aspects of real life, e.g., Internet rumors, fake news, and even academic publications [1]. Due to the widespread image forgery, one may be curious about the truth, which makes image restoration become a promising application. However, because of the mature and low-cost editing software, it is a tough task for human beings to discriminate the forged regions, not to mention reconstructing the original appearance automatically. In this paper, we introduce a new image reconstruction application called forgery reconstruction. Given a forged image, forgery reconstruction reconstructs the tampered areas into the natural regions (see Figure 1) by predicting its original appearance automatically. It detects the forged areas and then replace them with the natural contextual information synthesized by neural networks that learn from the visual semantics. Although forgery reconstruction is similar to image inpainting, there is no solution for localizing and reconstructing the forged regions of the image automatically in a single pass yet.

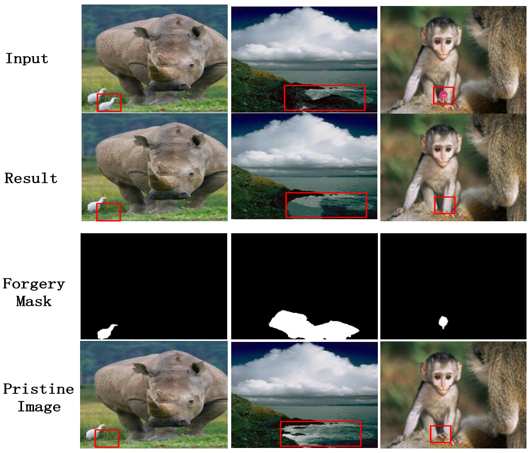

**Figure 1:** Forgery Reconstruction automatically reconstructs the tampered regions into original appearance. Input images are forgery images. The forgery masks contain the location information of tampered regions and the pristine images show the original appearance of the forgery images.

Image inpainting synthesizes semantically plausible contents to fill in the missing regions. Plenty of solutions for image inpainting have been proposed [2, 3, 4]. The typical image inpainting requires a manual binary mask to indicate the regions to be reconstructed, which means that the inpainting algorithms need high-level semantics. A joint system [5] is proposed to alleviate this issue. However, it only considers the removal of text with a

*IJCAI'21: the 30th International Joint Conference on Artificial Intelligence, 19th -26th August, 2021, Montreal-themed Virtual Reality*

zhuolong@email.szu.edu.cn (L. Zhuo); tansq@szu.edu.cn (Corresponding Author) (S. Tan*)

CEUR Workshop Proceedings (CEUR-WS.org)

small patch in the image. Text in the image is also a type of high-level semantics and one could potentially utilize human intuition to reconstruct the image manually. Small patch and finite application limit the existing joint system.

Real world is full of image forgeries that human intuition cannot resist due to tricksy manipulations. We believe that a more applicable strategy is to simultaneously detect the forged regions and synthesize the natural areas corresponding to the original images. For this purpose, our system makes good use of end-to-end strategy without any hand-made tricks, such as masking the tampered areas in the image.

There are two main challenges that we need to tackle to enable forgery reconstruction: (1) The reconstruction algorithms lack prior knowledge of location information that indicates the regions to be reconstructed. That's because the forged regions could be tricksy and difficult to be revealed, as shown in Figure 1. Therefore we need to design a promising forgery localization model to guide the reconstruction. (2) Removing the forged areas creates holes in the background, so we need to reconstruct the holes while maintaining a smooth transition between the tampered regions and the background.

Due to these reasons, we propose a hybrid and end-to-end system to address these challenges. We call it hybrid coarse-to-fine network (HCF-Net) that contains two phases, i.e., forgery localization phase and reconstruction phase.

The forgery localization phase provides the locations of manipulated areas to the reconstruction phase. Inaccurate localization could mislead the reconstruction stage, so it is necessary to develop a reliable model for this phase. Steganalysis rich models filter layer (SRM) [6, 7] can effectively reveal the noise inconsistency between pristine and tampered regions. Motivated by [8] and [7], we propose a new strategy that extracts noise features distributed in each channel to leverage the relationship among these channels. In particular, we split the image channels into red, green, and blue channels (RGB), and apply an SRM filter to extract the noise features respectively. This Split-SRM strategy enhances the inter-channel relationships of noise features. The coarse-to-fine manner is adopted to enhance the detection on these features. This addresses Challenge (1) by jointly introducing the outstanding image forgery localization model into the system.

The reconstruction phase is to synthesize the natural contents in the forged regions, which is similar to image inpainting task that reconstructs the hollowed images. To address Challenge (2), we adopt a dual attentive module [9] based coarse-to-fine network to predict the original appearance of the tampered regions and maintain smooth transitions between the reconstructed regions and background.

Overall, to address the forgery reconstruction task, we propose a two-phase pipeline shown in Figure 2. It first localizes the tampered regions, then reconstructs and refines the original appearance smoothly.

To the best of our knowledge, this work is the first to target the problem of reconstructing forgery images, which could be a practical application in real life. The main contributions of this paper are summarized as follows:

- We propose the first, to the best of our knowledge, forgery reconstruction system in an end-to-end manner to reconstruct the contextual information of the tampered regions in images. The hybrid coarse-to-fine network, HCF-Net, is proposed to integrate two different phases of image forgery reconstruction, i.e., forgery localization phase and reconstruction phase.
- We introduce a novel coarse-to-fine network with Split-SRM strategy to localize the forged regions with high precision and effectively guide the reconstruction. And our reconstruction phase is also based on the coarse-to-fine fashion.
- Extensive experiments demonstrate that HCF-Net has the advantages of high-precision forgery reconstruction, and outperforms the advanced inpainting methods in forgery reconstruction.

## 2. Related Work

### 2.1. Image Forgery Localization

There are three popular image forgery techniques: splicing [1], copy-move [10, 11, 12, 13, 14] and removal [15]. Many methods, including low pass filter [16, 17], long short-term memory (LSTM) architecture [18] and steganalysis rich models filter (SRM) layer [7, 6], are derived from these traces. SRM is a powerful filter originating from steganalysis, which is the most challenging image forensics algorithm to detect slight distributions of images. SRM, proposed by [19], extracts local noise features from adjacent pixels, capturing the inconsistency between tampered and pristine areas. However, the existing methods using SRM kernel do not consider the connection among the channels of the image, which extracts few features from the tampered areas. Although these methods could be used in the detection stage of image forgery reconstruction, unavoidable inaccuracies are prone to significantly decrease the quality of the reconstructed natural regions. Different from the previous work, our proposed Split-SRM filter layer leverages the channel-wise noise features.

WISERNet [8] applies SRM filters layer to extract the noise features in each channel and it succeeds in color im-

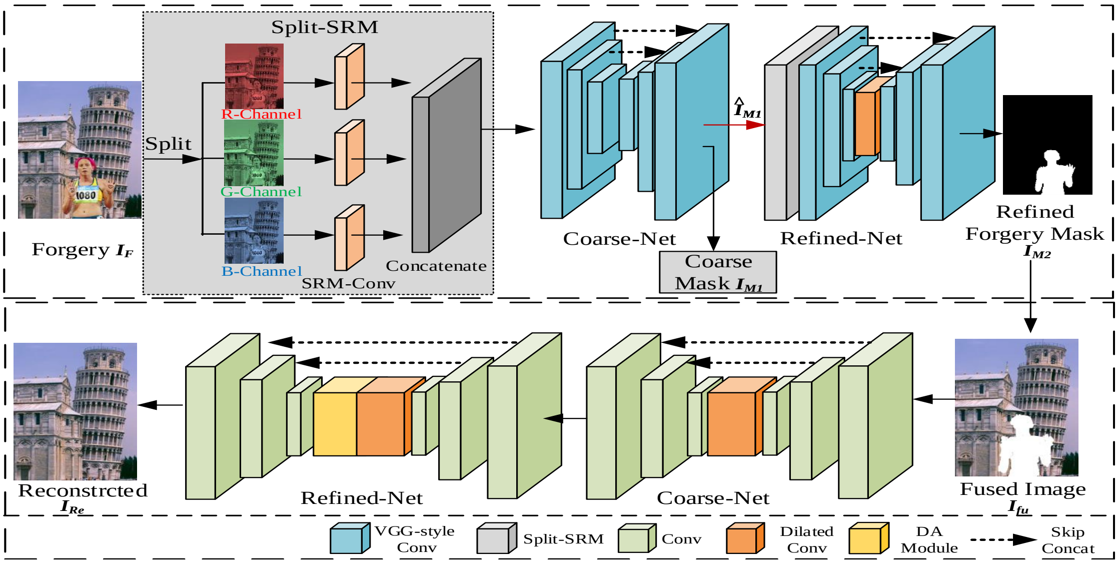

**Figure 2:** An overview of our HCF-Net architecture. The top one is the forgery localization phase and the bottom one is the reconstruction phase. The red arrow( →) denotes feeding the features instead of the prediction.

age forensics. Different from that, we adopt a simplified SRM filter [6] to perform Split-SRM filter layer.

## 2.2. Image Inpainting

Various methods have been introduced for image inpainting. Early methods use correlation of image statistics [20, 21] while recent methods are based on CNNs [22, 23] or GANs [24, 2, 25]. Latterly, the advanced method include Gated Conv [3], PEN-Net [4] and RFR-Net [26]. Gated Conv is a novel convolutional mechanism to filter useless information. PEN-Net applies pyramid structures to gain more contextual information. RFR-Net utilizes the constrains of the hole center. Our system will compare with these three state-of-the-art networks. The coarse-to-fine structures have been widely used in the above mentioned methods, as well as in our algorithm. Most recent methods [2, 3, 4, 26] employ the attention mechanism. Different from their attention modules, we use a dual attention module in our refinement network because of its great success in segmentation [9].

# 3. Methodology

## 3.1. Overall Framework

Our goal is to convert a forgery image $I_F$ into its original appearance $I_{Re}$, as shown in 2. Our two-phase system, HCF-Net, is based on two coarse-to-fine networks. The coarse-to-fine network, whose effectiveness is verified in image inpainting [3, 2], consists of two subnetworks: a coarse network and a refinement network. Firstly, it generates an initial coarse prediction with the coarse network and refines the coarse results by extracting the related features with the refinement network. The proposed two-phase HCF-Net is an end-to-end manner without preprocessing or postprocessing.

The first phase is the forgery localization phase that predicts the a coarse forgery mask $I_{M1}$ and a refined mask $I_{M2}$ of a forgery image $I_F$. $I_{M1}$ and $I_{M2}$ contain the location information of forgery regions. We then fuse the pixels of $I_{M2}$ into $I_F$ to create a fused image $I_{fu}$. This design make the remaining problem easier, since most of the pixels can be directly borrowed from the input forgery image and thus fewer pixels need to be reconstructed by the the following phase. The reconstruction phase inputs $I_{fu}$ and synthesizes the original appearance, and refines the details to generate the reconstructed image $I_{Re}$.

## 3.2. Forgery Localization Phase

The forgery localization phase firstly separates the R, G, and B channels of the image. Each channel's low-level noise features are extracted by a convolution layer initialized by SRM filter of $5 \times 5 \times 3$ kernel size and then concatenated. This operation, called Split-SRM, is a non-trainable channel-wise convolution layer with the same initiation as [6]. After Split-SRM, the features are fed into a regular convolutional module in VGG-style. VGG-style module is a convolution block composed of a series of convolution layers of kernel size $3 \times 3$. Specifically, each module contains three convolution layers of kernel size $3 \times 3$ followed by ReLU. We concatenate the encoding features with decoding features in the coarse stage. The coarse results contain two branches. One branch predicts the coarse mask $I_{M1}$, and the other one produces the coarse features $\hat{I}_{M1}$ for the refining stage. The reason why we do not use the coarse result as the input of refined net is that $\hat{I}_{M1}$ remains more trainable features. This design delivers complete feature information for the following subnetwork. The refinement network also starts with the Split-SRM scheme, basically the same as the coarse net. The difference is that a series of dilated convolution layers with different dilation rates connect the encoder and decoder to extract features with large receptive fields. In particular, the dilated convolution layers are of $3 \times 3$ kernel sizes and the dilation rates of them are 2,4,8, and 16.

The intuition behind Split-SRM is that when an object is removed from the source image and pasted into another, the noise features in each channel between the source and target images are unlikely to match. WISERNet [8] proposes the split-channel strategy for color image forensics and proves to be effective. The coarse-to-fine framework enhances the representation of the network. The VGG-style structure outperforms other block styles in the experiments of [7]. To utilize these features, we transform each channel into the noise domain using an SRM kernel followed by VGG-style structures in the coarse network with skip connections to provide more relevant information in the network [27].

## 3.3. Reconstruction Phase

This phase is based on the coarse-to-fine network as well, where the coarse stage produces an initial coarse prediction, and the refine stage takes the coarse prediction as inputs and predicts refined results $I_{Re}$. In terms of the layer implementations, we use the convolution layer of $3 \times 3$ kernel size followed by the ELUs activation function. A series of dilated convolution layers connect the encoder and decoder. The structures of dilated convolution layers are same as those in forgery localization phase. Skip concatenation links the encoding and decoding layers. The refinement network exploits the dual attention module, proposed by [9], to adaptively integrate local features with their global dependencies. The architecture of a dual attention (DA) module is shown in Figure 3. Specifically, DA module has two branches, i.e., position attention branch (PAB) and channel attention branch (CAB). In the PAB, given the input feature $F \in \mathbb{R}^{H \times W \times C}$, we firstly perform spatial matrix operation (SMO), and obtain a position attention matrix $PAM$. SMO uses convolution layers to generate three feature maps $F_1, F_2$, and $F_3$, respectively, where $F_1, F_2, F_3 \in \mathbb{R}^{H \times W \times C}$. All of them are reshaped to $\mathbb{R}^{L \times C}$ where $L = H \times W$. Then the product of ${F_1}^T$ and $F_2$ passes through a softmax layer and the result of softmax is then multiplied by $F_3$. We then reshape the product result and generate $PAM \in \mathbb{R}^{H \times W \times C}$. On the other hand, CAB firstly performs channel matrix operation (CMO) to generate channel attention matrix $CAM$. Like PMO, CMO reshapes $F$ to $F' \in \mathbb{R}^{L \times C}$ and the product of $F'$ and $F'^T$ goes through softmax to generate a softmax feature map which is in $\mathbb{R}^{C \times C}$. We then multiply the softmax feature map and $F'$, and reshape the result to generate $CAM \in \mathbb{R}^{H \times W \times C}$. We set the trainable parameters to $PAM$ and $CAM$ and add them with $F$, respectively. Finally, we sum them with weights and produce dual attention feature $DAF$ as,

$$DA(F) = (W_0 \times PAM + F) + (W_1 \times CAM + F), \quad (1)$$

where $W_0$ and $W_1$ are trainable parameters.

## 3.4. Loss

We train our system in an end-to-end manner. We adopt two different losses during training. In the forgery localization (FL) phase, we use binary cross-entropy loss (BCE):

$$Loss_{FL} = BCE(y_{gt}, y_C) + BCE(y_{gt}, y_F) \quad (2)$$

where $y_{gt}$ denotes ground-truth masks, $y_C$ denotes coarse masks, and $y_F$ denotes refine masks.

In the reconstruction (Re) phase, for better qualitative results, we define a mix-loss similar to [28] as the reconstruction loss:

$$\begin{aligned} Loss_{Re} = &\sum_{i=1}^{n} \alpha[(1 - ssim(y'_{C_i}, y_f))/2 \\ &+ (1 - ssim(y'_{F_i}, y_f))/2] \\ &+ (1 - \alpha)[h(y'_{C_i}, y_f) + h(y'_{F_i}, y_f)], \end{aligned} \quad (3)$$

$$h(y, y') = \begin{cases} \frac{1}{2}(y - y')^2, & |y - y'| \leq \delta, \\ \delta \cdot (|y - y'| - \frac{1}{2}\delta), & otherwise. \end{cases} \quad (4)$$

where $ssim$ denotes structural similarity, $h$ denotes the Huber-loss function, $y_t$ denotes the targeted images, $y'_C$ denotes coarse predictions, and $y'_F$ denotes refined results. We set $\alpha$ = 0.86 and $\delta$ = 1.0 to enhance the impact on ssim.

The total training loss is a weighted sum of the individual losses presented above:

$$Loss_{total} = \lambda_R Loss_{Re} + \lambda_F Loss_{FL}, \quad (5)$$

where we set the weights $\lambda_R = 100$ for the reconstruction loss and $\lambda_F = 1$ for the localization loss to focus more on the reconstruction.

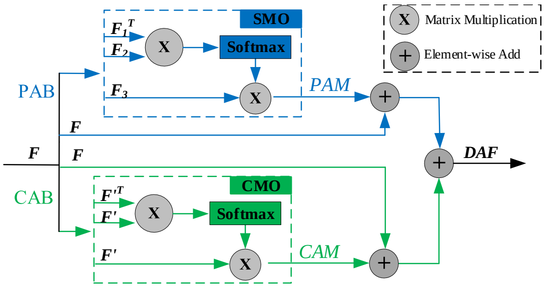

**Figure 3:** Illustration of Dual attention (DA) module.

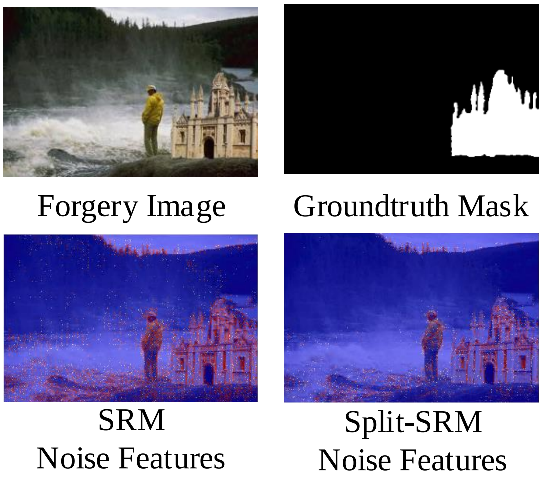

**Figure 4:** Visualization of the noise features extracted by different SRM strategies. The filter is expected to respond to the white region while ignoring the black region in the mask. The red regions have high response of noise while blue indicates low ones, and the heatmaps are super-imposed on the forgery image. We can clearly see that the SRM filter extracts false positive noise features in many areas while our proposed Split-SRM filter works more precisely.

## 4. Understanding our Model

We investigate the benefit of Split-SRM scheme in handling the inter-channel relationships of noise features and explore the advantages of the DA module in filling the hole smoothly. We visualize the internal feature representations to gain more insight.

### 4.1. Noise Feature Visualization for Split-SRM

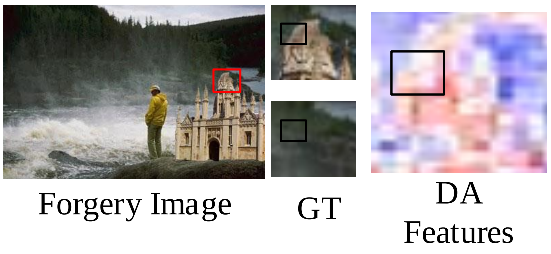

**Figure 5:** Visualization of dual attention (DA) features. We enlarge the region in the red rectangle in the forgery image and show it in the top center subfigure. Compared with the groundtruth at the bottom center, the color and texture are very different in the contours. The right subfigure shows the heatmap of DA features, which indicates that DA module focuses on the contours.

Figure 4 visualizes the focus of noise features extracted by the SRM filter and our Split-SRM filter. Obviously, Split-SRM focuses more on the regions of noise inconsistency instead of global noise features. This shows that our Split-SRM is capable of capturing the unnatural noise features more accurately at a low false positive rate.

[8] reveals that the relationships among channels could be leveraged to enhance the slight disturbance. In a forgery image, the tampered regions can be seen as the disturbance of the pristine image, causing the noise inconsistency. Split-SRM seeks the relationship among channels, which strengthens this noise inconsistency and makes it easier for capturing these noise.

### 4.2. Feature Visualization for the Dual Attention Module

Reconstructing the tampered regions needs to consider about the edges between the regions and background. Figure 5 visualizes the features generated by the DA module. As we can see, the DA module pays attention to the edges between the hole and the background and results in smooth transitions. Seen as Figure 5, there is a significant gap of texture and colors between holes and background. DA module contains two effective attention branches, i.e., PAB and CAB. The possible reason why DA module focus on the transitions is that PAB provides the positions of salient texture while CAB focuses on the channels of colors. Capturing rich contextual dependencies, the DA module guides the network to pay attention to the edges. Thus the network can learn from these information and generate natural transitions.

| Method | DEFACTO |
|---|---|
| Baseline$_{\mathrm{FL}}$ | 78.3 |
| SplitSRM | 85.5 |
| **SplitSRM + Refine** | **90.5** |

**Table 1**
Ablation study of our forgery localization phase using the Split-SRM strategy and coarse-to-fine manner on the DEAFCTO dataset. Results are reported in AUC(%). Both Split-SRM and Refinement network boost the performance.

| Method | NIST | Columbia | COVER | CASIA |
|---|---|---|---|---|
| ELA | 42.9 | 58.1 | 58.3 | 61.3 |
| EOI1 | 48.7 | 54.6 | 58.7 | 61.2 |
| CFA1 | 50.1 | 72.0 | 48.5 | 52.2 |
| J-LSTM⋆ | 76.4 | N/a | 61.4 | N/a |
| RGB-N⋆ | 93.7 | 85.8 | 81.7 | 79.5 |
| ManTra-Net ⋆ | 79.5 | 82.4 | 81.9 | 81.7 |
| **Ours-FL⋆** | **95.2** | **89.1** | **83.6** | **82.7** |

**Table 2**
Comparison of our approach with SOTA methods on several datasets in forgery localization. ⋆ indicates that the method uses a large-scale dataset and N/a means that the result is not reported. Results are reported in AUC(%). Our model shows comparable ability when trained on the small dataset and achieves the state-of-the-art performance when trained on the large dataset.

# 5. Experiments

## 5.1. Setup

We conduct extensive experiments to demonstrate our approach's superiority on standard datasets such as NIST 2016 [29], CASIA [30], COVER [31], and Columbia [32]. The four datasets contain 564, 6044, 100, and 180 samples, respectively. We select 42000 image pairs from the synthetic dataset DEFACTO [33] for pre-training as well. The number of pretrained dataset is the same as [6]. NIST and DEFACTO contain all three tampering technique. CASIA provides spliced and copy-move images. COVER is a relatively small dataset focusing on copy-move. Columbia focuses on splicing. NIST, CASIA, COVER, and Columbia consist of a relatively small amount of data, so we category them as small datasets. We do not operate any augmentations on the dataset to prove the effectiveness of our proposed networks. We train our model with input images resized to $256 \times 256$ on a single Tesla P100 GPU. The output size of reconstructed images and masks are $256 \times 256$. An ADAM solver with a learning rate of 0.0002 is used. We adopt the well-known peak signal to noise ratio (PSNR) and structural similarity (SSIM) as the metrics to evaluate our model.

## 5.2. Ablation Study

### 5.2.1. Forgery Localization Phase

The methods are compared in terms of the widely used area-under-the-curve (AUC) measure. We assign a confidence score to every pixel for pixel-level AUC evaluation. Meanwhile, we compare our proposed approach's performance against the results reported in [7], in which the score of ELA [34], NOI1 [35], CFA1 [36], MFCN [37], J-LSTM [38], RGB-N [6], and ManTra [7]. The recent state-of-the-art (SOTA) methods use the synthesized large-scale datasets to ensure the performance, while our method is trained with two strategies. One is to train on the small dataset without the extra data for validating our approach, and the other is to train on the DEAFCTO and finetune on the small dataset to achieve the most excellent performance. Like [6], the small datasets are divided into 75% finetuning and 25% testing. When we train our model on the large dataset, the Columbia dataset is only used for testing.

**Baseline$_{\mathrm{FL}}$**: As a baseline, we train the proposed architecture without the Split-SRM scheme or refinement network, but with an SRM filter layer before the coarse network.

**Split-SRM**: The coarse network with Split-SRM.

**Split-SRM + Refine**: Split-SRM scheme and coarse-to-fine network, our full model for the forgery localization phase.

We adopt DEAFCTO [33] for ablation study of Split-SRM, and the networks use 70% for training and 30% for testing. As one can see in Table 1, our proposed components improve over the baseline model. The possible reason is that the Split-SRM strategy provides more richer noise features and the relationships among channels, and the refinement network polishes the coarse results by visual correlation.

We compare our model with SOTA methods, as shown in Table 2. It is safe to conclude that our model, trained on the small dataset, is comparable to those SOTA methods. When our model is trained by large datasets, it outperforms the existing methods and achieves the state-of-the-art performance.

### 5.2.2. Reconstruction Phase

In this phase, we conduct the ablative study on the standard dataset and show the results in the final Table 3. Note that our model is trained together with forgery localization phase, and produces a forgery mask and a reconstructed image in a single forward pass. The reconstruction phase is fed with fused images of the forgery images and the refined forgery masks from the forgery localization phase.

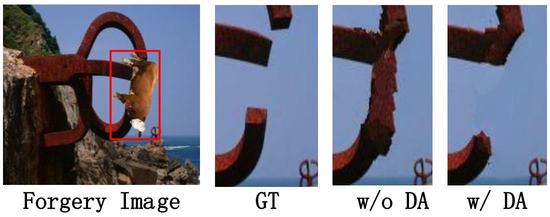

**Figure 6:** Visual comparison for our model with and without the dual attention module. Our model with DA smooths the contours of the reconstruction and shows image more realistically.

| Method | NIST | Columbia | COVER | CASIA |
|---|---|---|---|---|
| PEN-Net | 24.5/0.84 | 19.1/0.82 | 21.2/0.81 | 18.9/0.60 |
| GatedConv | 24.9/0.90 | 18.7/0.83 | 22.6/0.88 | 19.2/0.63 |
| RFR-Net | 28.6/0.92 | 20.1/0.86 | 24.3/0.91 | **22.3/0.72** |
| Baseline$_{\text{Re}}$ | 25.1/0.91 | 18.6/0.82 | 21.6/0.89 | 19.6/0.69 |
| **HCF-Net** | **31.1/0.94** | **20.3/0.89** | **24.7/0.91** | 20.6/0.71 |

**Table 3**
Comparison of our approach with SOTA methods on several datasets in forgery reconstruction. Results are reported in (PSNR(dB)/SSIM(%)). Other traditional inpainting methods are fed with the segmentation masks generated by our forgery localization phase. Our proposed HCF-Net is comparable to the SOTA level.

| | NIST |
|---|---|
| HCF-Net>PEN-Net | 96% |
| HCF-Net>GatedConv | 90.6% |
| HCF-Net>RFR-Net | 81.8% |

**Table 4**
Results of user studies. Each entry is the percentage of cases where the participants choose the images by our results as the more natural one than another method.

**Baseline$_{\mathbf{Re}}$**: The coarse-to-fine network without the DA module.

**HCF-Net**: The coarse-to-fine model with the DA module.

We define **Baseline$_{\mathbf{Re}}$** by simply removing the DA module of our network. As shown in Table 3, the results show that the DA module improves PSNR and SSIM significantly on each dataset. Figure 6 presents the qualitative comparison of the model without the DA module (w/o DA). The proposed model with DA generates visually smoother transition with less artifacts. It demonstrates that the DA module alleviates the gap in transitions and helps generate natural appearance, which benefits for generating high-quality images.

## 5.3. Comparison with the State-of-the-art

Note that our model is trained together with forgery localization phase, and produces a forgery mask and a reconstructed image in a single forward pass. The reconstruction phase is fed with fused images of the forgery images and the refined forgery masks from the forgery localization phase. We define **Baseline$_{\mathbf{Re}}$** by simply removing the DA module of our network. We define **Baseline$_{\mathbf{Re}}$** by simply removing the DA module of our network.

For fair comparison, we only consider deep-learning based approaches whose masks are provided by the refined results generated by the forgery localization phase. The SOTA methods include PEN-Net [4], GatedConv [3] and RFR-Net [26]. We train our model on four small datasets with 70% training and 30% testing.

For quantitative evaluation, Table 3 summarizes the overall performance. We use the official pretrained models provided by the authors. Note that, although our network is only trained on small datasets, it can still generalize well and show outstanding performance. We observe significant gains in terms of both PSNR and SSIM on all of the standard evaluation benchmarks, which indicates that our joint system can reconstruct the forged regions with high quality by learning the cascade features. We have to emphasize that CASIA is a very complex dataset with various scenes. RFR-Net is better than HCF-Net in CASIA dataset because it benefits from the complicated network structure while our network is concise. Furthermore, other methods need to be fed with the masks from our forgery localization phase while our approach does not receive any external hand-made mask during reconstruction.

Figure 7 shows qualitative results for comparison, where we compare with three recent methods. We show the examples of reconstructing three main types of manipulations, including splicing, copy-move, and removal. The results of our method are visually more natural.

We also construct user studies on the NIST dataset, as shown in Table 4. Deployed on Google Forms platform, our protocol depends on large batches of blind randomized A/B tests. Every survey contains a batch of 20 pairwise comparisons. Every pair involves two images generated by two different methods. We invite 25 participants, and the participants are supposed to select the more natural one in the pair. The pairs and orders are randomized. We limit ten seconds in one pair selection. The results show that the choices of participants border on random guesses. The results demonstrate that our system reconstructs forgery images more realistically.

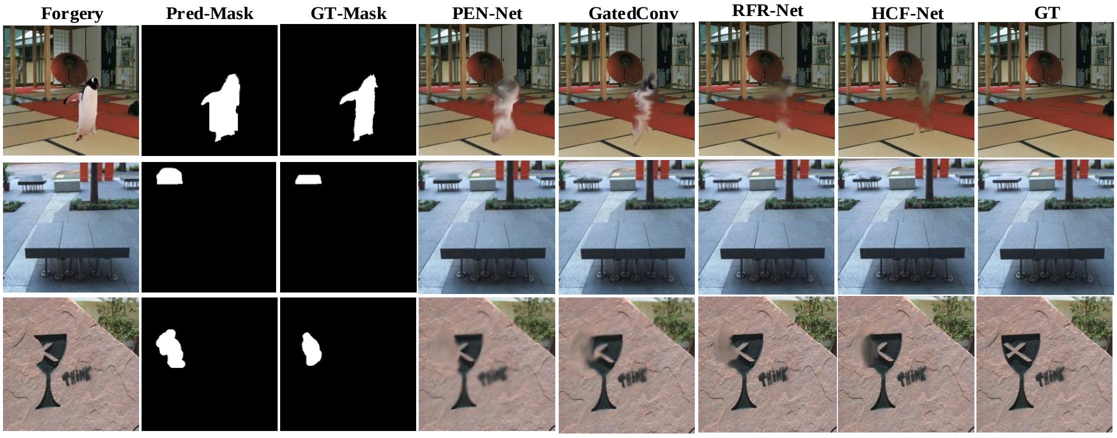

**Figure 7:** Visualization of samples of the results. We show three kinds of forgery reconstruction, i.e. splicing, copy-move and removal. Note that the second row consists of the masks predicted by our forgery localization phase and from the fourth to seventh row show the reconstruction result by different methods. The first row is splicing, the second row is copy-move and the third row is removal.

## 6. Conclusion

In this paper, we propose the first forgery reconstruction system, called HCF-Net, that automatically localizes and reconstructs the forged regions of complex manipulations in images. HCF-Net operates a hybrid coarse-to-fine network in an end-to-end manner that integrates a forgery localization network and a reconstruction network. We also introduce a novel coarse-to-fine network using the Split-SRM strategy for image forgery localization and makes precise restoration in the reconstruction network. Extensive experiments show that our proposed HCF-Net for image forgery localization achieves state-of-the-art performance and our joint system provides better results compared with traditional inpainting algorithms.

## Acknowledgments

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province (2019B010139003), NSFC (61772349, U19B2022, 61872244), Guangdong Basic and Applied Basic Research Foundation (2019B151502001), and Shenzhen R&D Program (GJHZ20180928155814437, JCYJ20180305124325555). This work was also supported by Alibaba Group through Alibaba Innovative Research (AIR) Program.

## References

[1] M. Zampoglou, S. Papadopoulos, Y. Kompatsiaris, Large-scale evaluation of splicing localization algorithms for web images, Multimedia Tools and Applications 76 (2017) 4801–4834.

[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, 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.

[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. S. Huang, Free-form image inpainting with gated convolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4471–4480.

[4] Y. Zeng, J. Fu, H. Chao, B. Guo, Learning pyramid-context encoder network for high-quality image inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 1486–1494.

[5] S. Qin, J. Wei, R. Manduchi, Automatic semantic content removal by learning to neglect, arXiv preprint arXiv:1807.07696 (2018).

[6] P. Zhou, X. Han, V. I. Morariu, L. S. Davis, Learning rich features for image manipulation detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1053–1061.

[7] Y. Wu, W. AbdAlmageed, P. Natarajan, Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features, in: Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition, 2019, pp. 9543–9552.

[8] J. Zeng, S. Tan, G. Liu, B. Li, J. Huang, Wisernet: Wider separate-then-reunion network for steganalysis of color images, IEEE Transactions on Information Forensics and Security 14 (2019) 2735–2748.

[9] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[10] D. Cozzolino, G. Poggi, L. Verdoliva, Efficient dense-field copy–move forgery detection, IEEE Transactions on Information Forensics and Security 10 (2015) 2284–2297.

[11] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, S. Winkler, Coverage—a novel database for copy-move forgery detection, in: 2016 IEEE International Conference on Image Processing, IEEE, 2016, pp. 161–165.

[12] Y. Rao, J. Ni, A deep learning approach to detection of splicing and copy-move forgeries in images, in: 2016 IEEE International Workshop on Information Forensics and Security, IEEE, 2016, pp. 1–6.

[13] Y. Wu, W. Abd-Almageed, P. Natarajan, Busternet: Detecting copy-move image forgery with source/target localization, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 168–184.

[14] Y. Wu, W. Abd-Almageed, P. Natarajan, Image copy-move forgery detection via an end-to-end deep neural network, in: 2018 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2018, pp. 1907–1915.

[15] X. Zhu, Y. Qian, X. Zhao, B. Sun, Y. Sun, A deep learning approach to patch-based image inpainting forensics, Signal Processing: Image Communication 67 (2018) 90–99.

[16] J. Chen, X. Kang, Y. Liu, Z. J. Wang, Median filtering forensics based on convolutional neural networks, IEEE Signal Processing Letters 22 (2015) 1849–1853.

[17] B. Bayar, M. C. Stamm, A deep learning approach to universal image manipulation detection using a new convolutional layer, in: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, 2016, pp. 5–10.

[18] J. H. Bappy, C. Simons, L. Nataraj, B. Manjunath, A. K. Roy-Chowdhury, Hybrid lstm and encoder–decoder architecture for detection of image forgeries, IEEE Transactions on Image Processing 28 (2019) 3286–3300.

[19] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Transactions on Information Forensics and Security 7 (2012) 868–882.

[20] S. Roth, M. J. Black, Fields of experts: A framework for learning image priors, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2, IEEE, 2005, pp. 860–867.

[21] D. Zoran, Y. Weiss, From learning models of natural image patches to whole image restoration, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 479–486.

[22] J. S. Ren, L. Xu, Q. Yan, W. Sun, Shepard convolutional neural networks, in: Advances in Neural Information Processing Systems, 2015, pp. 901–909.

[23] K. Zhang, W. Zuo, S. Gu, L. Zhang, Learning deep cnn denoiser prior for image restoration, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3929–3938.

[24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[25] Y. Wang, X. Tao, X. Qi, X. Shen, J. Jia, Image inpainting via generative multi-column convolutional neural networks, in: Advances in neural information processing systems, 2018, pp. 331–340.

[26] J. Li, N. Wang, L. Zhang, B. Du, D. Tao, Recurrent feature reasoning for image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7760–7768.

[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[28] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for neural networks for image processing, arXiv preprint arXiv:1511.08861 (2015).

[29] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, J. Fiscus, Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation, in: IEEE Winter Applications of Computer Vision Workshops, IEEE, 2019, pp. 63–72.

[30] J. Dong, W. Wang, T. Tan, Casia image tampering detection evaluation database, in: IEEE China Summit and International Conference on Signal and Information Processing, IEEE, 2013, pp. 422–426.

[31] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, S. Winkler, Coverage—a novel database for copy-move forgery detection, in: IEEE International Conference on Image Processing, IEEE, 2016, pp. 161–165.

[32] T.-T. Ng, J. Hsu, S.-F. Chang, Columbia image splicing detection evaluation dataset, DVMM lab. Columbia Univ CalPhotos Digit Libr (2009).

[33] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. L. Dugelay, P. Marc, Defacto: Image

and face manipulation dataset, in: European Signal Processing Conference, IEEE, 2019, pp. 1–5.

[34] N. Krawetz, H. F. Solutions, A picture's worth, Hacker Factor Solutions 6 (2007) 2.

[35] B. Mahdian, S. Saic, Using noise inconsistencies for blind image forensics, Image and Vision Computing 27 (2009) 1497–1503.

[36] P. Ferrara, T. Bianchi, A. De Rosa, A. Piva, Image forgery localization via fine-grained analysis of cfa artifacts, IEEE Transactions on Information Forensics and Security 7 (2012) 1566–1577.

[37] R. Salloum, Y. Ren, C.-C. J. Kuo, Image splicing localization using a multi-task fully convolutional network, Journal of Visual Communication and Image Representation 51 (2018) 201–209.

[38] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, B. Manjunath, Exploiting spatial structure for localizing manipulated image regions, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4970–4979.