A Multi-Domain Camera Model Identification Feature Restoration Network to Counter AI Compression Attacks

Jinkai Zhang†, Zijuan Han‡, Yunxia Liu‡ and Yang Yang†*
† School of Information Science and Engineering, Shandong University, Qingdao, China
‡ Center for Optics Research and Engineering (CORE), Shandong University, Qingdao, China
* Corresponding Author
Email:202232729@mail.sdu.edu.cn

***Abstract*—Camera model identification(CMI) has been a well-established passive forensic technique in recent decades. Previous research primarily focused on unaltered images, while real-world scenarios often involve malicious attacks. With the rapid advancement of AI compression technology, it is becoming increasingly popular and prevalent in various applications. However, our experimental study demonstrates that the camera model identification drops dramatically when applied to AI-compressed images without any processing. In order to enhance the robustness of the camera model identification algorithm against AI compression attacks, we propose a multi-domain camera model identification feature restoration Network. Firstly, a spatial and frequency domain dual branch architecture is proposed to restore forensic features from AI compressed image patches. The frequency feature restoration module employs an unbalanced restoration strategy in amplitude and phase for efficient network training and improved identification robustness. Afterwards, a feature fusion and selection module are utilized to fuse and filter restored features, where double pooling cross fusion strategy is adopted in channel attention mechanism for making full use of channel statics. Experimental results show that the proposed method can efficiently restore camera model related forensic features and significantly improve the performance of the CMI model on AI compressed images.**

I. Introduction

The widespread of imaging devices and advancements in image generation algorithms have significantly enhanced the accessibility and quality of images. However, this presents substantial challenges in image authenticity verification [1]. Being a passive forensic technique, Camera model identification(CMI) aims to determine the origin of images by analyzing features such as photo response non-uniformity, sensor noise patterns [2] and so on. It has found wide applications in image forensics, copyright protection, and privacy protection [3].

In recent years, deep learning-based image compression techniques have made significant progress. The Jpeg Association is working on the JPEG AI standard [4]. AI compression is considered as a strong candidate for the next generation of image compression techniques, and is highly likely to receive widespread deployment in industrial applications in the future [5], [6]. However, malicious use of AI compression can seriously disrupt forensic and investigative efforts[4]. To the best of our knowledge, how AI compression affects the accuracy and effectiveness of CMI has not been reported in literatures.

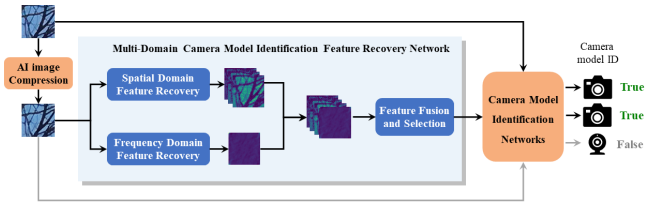


Fig. 1. AI compression significantly decreases the accuracy of camera model identification(CMI). The proposed MDFRNet aims to restore CMI related forensic features from multiple domains to adapt AI compressed patches to prevalent CMI networks for correct camera model prediction.

In this study, we carry out experimental evaluation of the camera model identification performance taken AI compressed image patches as network input to current prevalent CMI networks. Results reveal that most CMI networks fail without any processing. To address this issue, a Multi-Domain camera model identification Feature Restoration Network(MDFRNet) is proposed. As depicted in Fig.1, a dual branch architecture is adopted to restore complementary features from the frequency and spatial domains and successfully improves the accuracy of the CMI network on AI compressed images. The primary contributions of this study are summarized as follows:

- This study pioneers the study of AI compression attack in camera model identification, highlighting that while AI compression achieves high-quality image compression, it often alters critical source camera features, making it susceptible to malicious attacks.
- A multi-domain camera model identification feature restoration network(MDFRNet) is proposed to restore forensic features from both spatial and frequency domains, and successfully improve CMI accuracy on AI-compressed images.
- Experimental validation across multiple CMI methods demonstrate the method's effectiveness in restoring identification features distorted by AI compression. The results affirm the robustness and versatility of proposed network.

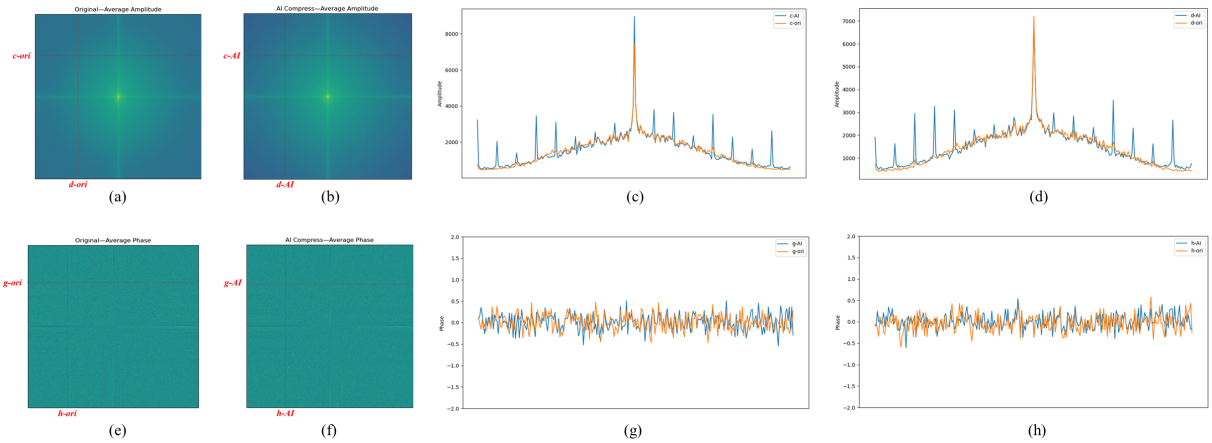


Fig. 2. Frequency domain comparison of images with and without HIFIC compression. (a) Average amplitude of 100 images. (b) Average amplitude of HIFIC compressed images. (c) A typical visualization of lateral average amplitude. (d) A typical visualization of longitudinal average amplitude. (e) Average phase of original images. (f) Average phase of HIFIC compressed images. (g) A typical visualization of lateral average phase. (h) A typical visualization of longitudinal average phase.

II. Related Works

How to extract camera model related features from noise residual to suppress the influence of image content plays an important role in camera model identification. To address this issue, Davide et al. [7] developed a noise residual extraction method using frequency domain constraints. Rafi et al. [8] proposed the Remnet to enhance intrinsic camera fingerprint by suppressing image content, and then combined a dynamic CNN-based preprocessing block with a shallow classifier and verified the applicability of residual blocks for CMI tasks. Bennabhaktula et al. [9] applied MobileNet with pre-trained weights from ImageNet for transfer learning, and study the relationship between deep learning network complexity and recognition accuracy. Our previous work [10] improved VGG with a fine-grained multi-scale residual prediction module to boost recognition accuracy and reduce scene content influence. Huan et al. [11] proposed ConvNeXt with a dual-path attention mechanism and classical residual extraction to capture crucial model features for camera provenance. Although literatures have reported satisfying CMI results, for instance, the state-of-the-art Remnet [8] has reached 94.6% accuracy on Dresden dataset, and DPEC ConvNeXt [11] has achieved 83.1% on VISION [12] dataset. Images inevitably undergo various post-processing during transmission in practice, which will greatly affect its application in real scenarios.

In recent years, with the continuous development of AI compression technology, deep learning-based image compression techniques have made significant progress. Some recent studies [5], [13] have greatly outperformed traditional JPEG in terms of PSNR and SSIM metrics. Toderici et al.[14] introduced a full-resolution image compression method using RNNs and neural networks for entropy coding. Balle et al.[15] proposed an end-to-end variational autoencoder model, incorporating hyperprior and spatial autoregressive models to optimize latent representation encoding. GAN-based compression algorithms now achieve high-quality images at very low bit rates. Mentzer et al. [13] developed HiFiC, a GAN-based technique that achieves perceptual fidelity close to the original image quality at half the original bit rate. AI compression is undergoing rapid development and is expected to be key technology of the next generation of image compression standards in the near future. Consequently, study of potential risks it may bring to related applications is necessary.

Although visually consistent with the original image, AI compression brings certain artifacts as a lossy compression method. Bergmann et al.[6] were the first to study the compression artifacts of HiFiC in the frequency domain, and the detectability of these artifacts was evaluated. Berthet et al. [4] have realized that AI compression is a novel unintended attack method, and used HiFiC to attack the advanced tamper detection network CatNet[16] to verify the aggressiveness of AI compression in the field of tamper detection. To the best of our knowledge, there is no in-depth study on the impact of AI compression in the field of CMI at present. However, existing research on source camera identification mainly focuses on the detection of raw images, ignoring the threats and challenges that AI compression may bring to the source camera identification task.

III. The Proposed Method

A. Frequency domain analysis of AI compressed images

Frequency domain analysis is very important to study the underlying characteristics of artifacts introduced by AI compression. Referring to the work of Bergmannet [6], we experimentally compare the amplitude and phase difference between original and AI compressed images which forms the basis of our later proposed network and loss function design.

We randomly selected 100 images from the VISION dataset [12], which is a widely acknowledged dataset in camera model identification field. Central cropping is applied to both original and HIFIC compressed sets to obtain series of

256×256 patches. Fast Fourier Transform with N = 256 is applied and the average amplitude is calculated for enhanced feature robustness. Average amplitude of the original and AI compressed images are shown in Fig.2(a) and (b), respectively. To further compare their differences intuitively, we depict typical lateral values in Fig.2(c) and longitudinal values in Fig.2(d). Obviously, we can observe that the spectrogram of the AI compressed images regularly shows peaks at 16 bisection positions [6]. Furthermore, the average amplitude maintains stable with respect to image contents when number of images is greater than 50 in our experiment. Consequently, it can serve as a template to regularize frequency domain features in network training.

The average phase of the original and AI compressed images are shown in Fig.2(e) and (f), respectively. Similarly, to compare the difference introduced by HIFIC compression, we also depict a typical lateral and longitudinal slice in Fig.2(g) and (h). As compared with amplitude, phase variations are more disordered and complex. Due its underlying modulus 2π characteristic of phase information, there is no regularity observed in either average phase nor phase of any specific image in our experiments. In contrast to the amplitude which is intuitive and easy to understand, the complex of phase has been reported in literatures [17], [18], and has long been the reason that limits its application in feature representation. However, phase contains rich information of nearby pixels and is of vital importance to restore camera model related subtle features. We will report our effort in exploiting phase information in CMI feature restoration in subsequent sections.

B. Overview of proposed network

Given the fact that camrea model identificaiton (CMI) resutls is significantly affected by artifacts introduced by AI compression, our motivation is to design a CMI feature restoration network to restore camera model identification features. The overall architecture of the proposed multi domain feature restoration network (MDFRNet) is illustrated in Fig.3.

Since the state of the art CMI networks have been able report satisfying performance on patches of size 64×64, we also follow this setting. For the given input patch of 64×64, the spatial domain feature restoration (SFR) module and frequency domain feature restoration (FFR) module are applied to restores camera model related forensic features from the detection features. Then, concatenated features are fused and enhanced by the feature fusion and selection(FFS) module with specifically designed channel attention mechanism. Finally, the fused features are fed into the CMI network to obtain final camera identification results.

C. Multi-domain feature restoration module

To make better use of advantages of both spatial domain and frequency domain methods, a dual branch multi-domain feature restoration module is proposed. As depicted in Fig.3, taking RGB patches as input, the spatial feature restoration(SFR) module aims to restore spatial domain camera model related forensic features with a UNet architecture. SFR outputs are more intuitive as it directly operates on pixel values and is able to provide better local relationship characterization. While the frequency feature restoration (FFR) module has more advantages in dealing with global features and periodic components. As restored features of SFR and FFR modules are finally concatenated for further fusion and selection, special attention is paid to network design so that they have same dimensions of feature maps.

In the encoding stage of the spatial feature restoration (SFR) module, the 3-channel RGB patch is firstly passed through a 3×3 convolutional layer to adjust the number of channels to 64. Each of the next four convolutional layers utilizes a 3×3 convolution kernel to double the number of feature channels to 128 , whose feature map sizes are then halved by Max pooling. In the decoding stage, the feature map is passed through a convolution layer with a kernel size of 3×3 to halve the number of channels. At the same time, the feature map of the encoding layer is spliced with the feature map of the decoding layer, and the number of channels of the feature map is gradually halved. After four convolution and pooling operations, we adjust the number of feature channels to 3 through an output convolution layer with a kernel size of 1×1.

In the frequency feature restoration(FFR) module, different restoration strategies are adopted for amplitude and phase. As for amplitude restoration, we intend to remove the effect of AI compression by subtracting the average amplitude increment (shown as Amplitude Mask in Fig.3). The average amplitude increment is a residual obtained by subtracting the average amplitude of the original patches from those of the AI compressed patches. As for phase restoration, phase of AI compressed patches are fed into the same Unet structure for feature restore, where the original phase is utilized in a supervised manner to guide the network learning of complex phase features. Finally, Inverse Fast Fourier Transform is used to obtain FFR restoration features. Considering the frequency domain stability, FFR also contributes to network training difficulty reduction.

D. Feature fusion and selection module

The dual branch FFR and SFR modules offers complementary benefits. However, it can't be ignored that there is certain redundancy in these two features. A feature fusion and selection module(FFS) is proposed to fully exploit their complementarity and minimize redundancy. As shown in Fig 3, based on spatial correspondence and complementarity between spatial and frequency domain restoration features, the upper branch fuses these features using two 3×3 convolution layers. For the lower branch, we employ an improved channel statistical feature calculation method to reduce feature redundancy. Specifically, we employ two pooling modalities commonly used in channel attention mechanisms: Max pooling and average pooling. Pooling values are fed into two convolution layers with kernel size of 1×1, to obtain the inner features that will be added to their own pooling values and the outer features that will be added to other pooling values. The inner

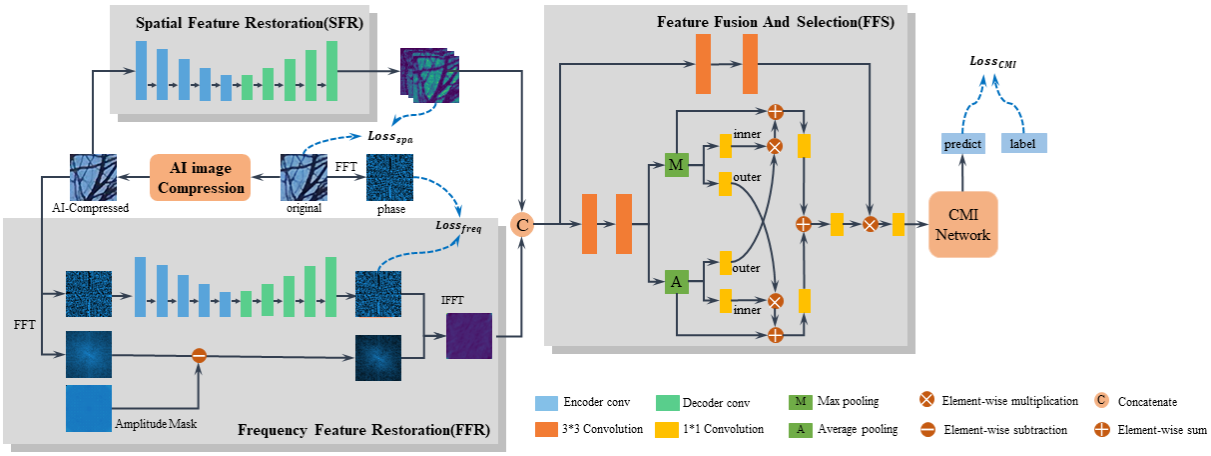


Fig. 3. Overall network structure of the proposed multi domain feature restoration network(MDFRNet).

features are multiplied with the outer features of the other pooling, and then added to the pooling value to output the cross values of each pooling. The maximum cross values and the average cross values are added and convolved to obtain the channel pooling information. The channel pooling information and the features output by the feature restoration network are multiplied channel by channel to obtain the fused and selected restored features.

After a 1×1 convolutional layer to adjust the number of channels, the restored features are sent to the CMI network for detection. This dual-pooling cross-fusion strategy maximizes the utilization of channel statistical information, significantly enhancing feature representation.

E. Loss Function

To guide efficient learning of CMI related features, we utilize three kinds of loss functions in the proposed network, which are the spatial feature restoration loss, the frequency feature restoration loss, and the camera model identification loss.

(1)Spatial feature restoration loss: $Loss_{spa}$ evaluates the spatial feature restoration capability of SFR module by comparing restored spatial features with the original patch:

$$Loss_{spa} = \frac{1}{M}\sum_{j=1}^{M}(y_j - p_j)^2 \tag{1}$$

where y and p denotes the ground truth pixel value in the original patch and restored output of the SFR module, M represents the number of pixels within a patch.

(2)Frequency feature restoration loss: $Loss_{freq}$ evaluates the frequency feature restoration capability by comparing the consistency of restored phase features with phase features of the original image. Inspired by Noiseprint [10], we utilize the geometric mean to arithmetic mean ratio of image phases to constrained training of the FFR. For the original image $I_{ori}(x, y)$, the frequency domain representations $F_{ori}(x, y)$ is obtained by Fourier transform:

$$I(x, y) \overset{FFT}{\rightarrow} F(u, \nu) \tag{2}$$

The phase $P_{ori}(u, \nu)$ is calculated by the phase formula as follows:

$$P(u, v) = \arctan\left(\frac{Im\left(F\left(u, \nu\right)\right)}{Re\left(F\left(u, \nu\right)\right)}\right) \tag{3}$$

where $Im\left(F\left(u, \nu\right)\right)$ denotes the imaginary part of $F\left(u, \nu\right)$ and $Re\left(F\left(u, \nu\right)\right)$ denotes the real part of $F\left(u, \nu\right)$.

Calculate the residual between the $P_{ori}\left(u, \nu\right)$ and phase restoration features $P_{AI}\left(u, \nu\right)$ output to get $R\left(u, \nu\right)$:

$$R\left(u, \nu\right) = P_{ori}(u, v) - P_{AI}(u, v) \tag{4}$$

Assuming that the input batchsize is N, the average phase of each batch can be expressed as $S\left(u, \nu\right)$:

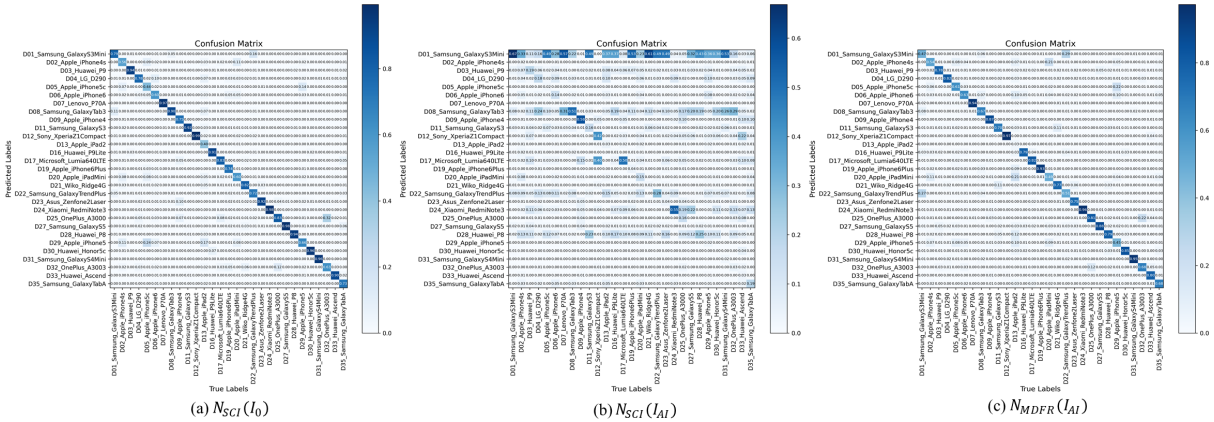
$$S\left(u, \nu\right) = \frac{1}{N}\sum_{k=1}^{N}\left|R_k(u, \nu)\right|^2 \tag{5}$$

The average loss per batch can be obtained by calculating the ratio of the geometric mean to the arithmetic mean of $S\left(u, \nu\right)$ and then taking the logarithm:

$$\begin{aligned} Loss_{freq} &= \log\left[\frac{S_{GM}}{S_{AM}}\right] \\ &= \left[\frac{1}{K^2}\sum_{u,\nu}\log\left(S\left(u, \nu\right)\right)\right] - \log\left[\frac{1}{K^2}\sum_{u,\nu}S\left(u, \nu\right)\right] \end{aligned} \tag{6}$$

Where K represents the input feature size. The ratio of the geometric mean to the arithmetic mean reflects the characteristics of the data distribution: The closer the ratio is to 1, the better the phase feature is restored, and the closer the ratio is to 0, the worse the restored phase feature is.

Source camera identification loss $Loss_{CMI}$ is utilized to provide supervised information by labeled source camera ground truth. It is calculated based on the difference between



(a) $N_{SCI}(I_0)$ (b) $N_{SCI}(I_{AI})$ (c) $N_{MDFR}(I_{AI})$

Fig. 4. Confusion matrix comparison of AI compressed images with/without feature restoration. (a) CMI of the original images (b) CMI of compressed images without feature restoration. (c) CMI of compressed images with feature restored by the proposed MDFRNet.

the camera identification predicted by the network and the actual label of the original patch as:

$$Loss_{CMI} = -\sum_{i=1}^{K} q\left(x_i\right) \log p\left(x_i\right) \tag{7}$$

where q denotes the target distribution and p denotes the predicted matching distribution, and K is total number of camera models involved.

The total training loss L can be expressed as follows:

$$Loss = \lambda_1 Loss_{spa} - \lambda_2 Loss_{freq} + \lambda_3 Loss_{CMI} \tag{8}$$

where λ_1, λ_2 , and λ_3 represent hyper-parameter controlling contributions of different losses that can be empirically determined.

IV. EXPERIMENTAL RESULT

To evaluate the effectiveness of the proposed method, we conduct series of experiments based on the Vision dataset, which is the most well acknowledged dataset in camera model identification field. We followed the experimental settings of our previous work[11] that we selected 29 camera models from them, and each camera model selected 100 images to segment the patch, and the shooting scene was not restricted. Our experiments are based on the deep learning framework Pytorch 1.13.0 and Python 3.8.18. We used the RMSprop optimizer with an initial learning rate of 0.0002 and weight decay of 0.00001. The λ_1 is set to 0.6, λ_2 to 0.5, and λ_3 to 0.8. We performed all the experiments reported in the paper using a Nvidia A4000 graphics card.

In this paper, we verify the effectiveness of our proposed MDFRNet on five CMI models. The results are shown in TABLE 1. We use I_0 to represent the original image, N_{CMI} represents the CMI network, I_{AI} represents the image compressed by AI algorithm, N_{MDFR} means patches are restored by MDFRNet before being sent to CMI network. The right three columns represent the results for different cases. The results show that after AI compression, the accuracy of all CMI models decreases significantly, indicating that AI compression is an effective attack method for the CMI task. After feature restoration via the proposed MDFRNetwork, the detection accuracy of all models is significantly improved. This shows that the proposed MDFRNetwork can effectively resist the AI compression attack for CMI.

TABLE I
IDENTIFICATION ACCURACY COMPARISON OF STATE-OF-THE-ART CAMERA MODEL IDENTIFICATION NETWORKS WITH/WITHOUT FEATURE RESTORATION.

Methods	$\mathbf{N_{CMI}(I_0)}$	$\mathbf{N_{CMI}(I_{AI})}$	$\mathbf{N_{MDFR}(I_{AI})}$
RemNet, Rafi	79.9	10.1	69.7
Res2Net, Liu	70.7	6.8	65.1
ResNet50, Bennabhaktula	78.2	10.1	66.3
MobileNet, Bennabhaktula	76.4	16.3	68.0
DPEC ConvNext, Huan	83.1	16.2	79.0

In Fig. 4, we show the confusion matrices corresponding to $N_{CMI}(I_0)$,$N_{CMI}(I_{AI})$ and $N_{MDFR}(I_{AI})$ using Huan's model. By comparing Fig.4(a) with Fig.4(b), it can be seen the identification accuracy for most camera models drops below 50%. It proves that AI compression causes damage to the camera fingerprint. However, as shown in Fig.4(c), when the identification features are restored by MDFRNet, the identification accuracy of each camera model is significantly improved.

We also compare with the fine-tuning, and do ablation experiments on the network proposed in this paper. The results are shown in Table 2. The fine-tuning uses weight trained using original images, and the corresponding AI compressed patches are used for training.

Setup#2 denotes that the average amplitude increment is excluded, with channel attention relying solely on max pooling. Setup#3 similarly indicates that the average amplitude increment is not utilized. Setup#4 specifies that only the max-pooling strategy is employed. Setup#5 signifies that the feature restoration network is not applied for phase reconstruction. Setup#6 represents our proposed method. The results

TABLE II
HUAN'S MODEL IS USED AS THE CMI NETWORK. EVEN WITH THE REDUCTION OF SOME PROCESSING STEPS, OUR PROPOSED METHOD CAN STILL OUTPERFORM THE COMMON FINE-TUNING(SETUP#1).

Setup	Accuracy
#1 Fine-tune	72.5
#2 backbone	72.9
#3 wo/res	76.8
#4 wo/at	74.3
#5 wo/phase_restore	76.2
#6 proposed	79.4

of these Settings provide further strong evidence to support the effectiveness of our proposed method in restoring the identification features.

V. CONCLUSIONS

In this paper, we experimentally evaluated current camera model identification networks on AI compressed images for the first time and found that AI compression is a powerful attack against the CMI task. To counter this new attack, a multi-domain camera model identification feature restoration network is proposed. Comprehensive experiments are conducted to evaluate the effectiveness of the proposed method. Significant identification performance improvement is observed on AI compressed images. At present, our work is only based on the AI compression algorithm. In the future, we will further investigate the impact of new attacks such as using large models to generate images on the camera model recognition task.

REFERENCES

[1] X. Lin, J. H. Li, S. L. Wang, A. W. C. Liew, F. Cheng, and X. S. Huang, *Recent advances in passive digital image security forensics: A brief review. engineering 4 (1): 29–39*, 2018.

[2] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 259–263, 2016.

[3] V. U. Sameer and R. Naskar, "Deep siamese network for limited labels classification in source camera identification," *Multimedia Tools and Applications*, vol. 79, no. 37, pp. 28 079–28 104, 2020.

[4] A. Berthet and J.-L. Dugelay, "Ai-based compression: A new unintended counter attack on jpeg-related image forensic detectors?" In *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 3426–3430.

[5] H. E. Egilmez, A. K. Singh, M. Coban, *et al.*, "Transform network architectures for deep learning based end-to-end image/video coding in subsampled color spaces," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 441–452, 2021.

[6] S. Bergmann, D. Moussa, F. Brand, A. Kaup, and C. Riess, "Frequency-domain analysis of traces for the detection of ai-based compression," in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, 2023, pp. 1–6.

[7] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2019.

[8] A. M. Rafi, T. I. Tonmoy, U. Kamal, Q. J. Wu, and M. K. Hasan, "Remnet: Remnant convolutional neural network for camera model identification," *Neural Computing and Applications*, vol. 33, pp. 3655–3670, 2021.

[9] G. S. Bennabhaktula, D. Timmerman, E. Alegre, and G. Azzopardi, "Source camera device identification from videos," *SN Computer Science*, vol. 3, no. 4, p. 316, 2022.

[10] Y. Liu, Z. Zou, Y. Yang, N.-F. B. Law, and A. A. Bharath, "Efficient source camera identification with diversity-enhanced patch selection and deep residual prediction," *Sensors*, vol. 21, no. 14, p. 4701, 2021.

[11] S. Huan, Y. Liu, Y. Yang, and N.-F. B. Law, "Camera model identification based on dual-path enhanced convnext network and patches selected by uniform local binary pattern," *Expert Systems with Applications*, vol. 241, p. 122 501, 2024.

[12] D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, "Vision: A video and image dataset for source identification," *EURASIP Journal on Information Security*, vol. 2017, pp. 1–16, 2017.

[13] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 913–11 924, 2020.

[14] G. Toderici, D. Vincent, N. Johnston, *et al.*, "Full resolution image compression with recurrent neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5306–5314.

[15] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[16] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Learning jpeg compression artifacts for image manipulation detection and localization," *International Journal of Computer Vision*, vol. 130, no. 8, pp. 1875–1895, 2022.

[17] A. Leclaire and L. Moisan, "No-reference image quality assessment and blind deblurring with sharpness metrics exploiting fourier phase information," *Journal of Mathematical Imaging and Vision*, vol. 52, pp. 145–172, 2015.

[18] A.-C. Li, S. Vyas, Y.-H. Lin, Y.-Y. Huang, H.-M. Huang, and Y. Luo, "Patch-based u-net model for isotropic quantitative differential phase contrast imaging," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 3229–3237, 2021.