Boosting Geometric Invariants for Discriminative Forensics of Large-Scale Generated Visual Content

Shuren Qi, Chao Wang, Zhiqiu Huang, Yushu Zhang, Xiangyu Chen, Yi Zhang, Tieyong Zeng, and Fenglei Fan*

***Abstract*—Generative artificial intelligence has shown great success in visual content synthesis such that humans struggle to distinguish between real and synthesized images. Forensic research seeks to reveal artifacts in such generated images, ensuring information security or improving generation capability. In this regard, the robustness and interpretability are important for the trustworthy purpose of forensic tasks. However, typical forensic models and their underlying data representations rely on empirical learning algorithms, which cannot effectively handle the high robustness and interpretability requirements beyond experience. As an effective solution, we extend the classical geometric invariants to the forensic research of large-scale generated images. Invariants are handcrafted representations with robust and interpretable geometric principles. However, their discriminability is far from the large scale of today's forensic tasks. We boost the discriminability by extending the classical invariants to the hierarchical architecture of convolutional neural networks. The resulting overcompleteness allows for an automatic selection of task-discriminative features, while retaining the previous advantages of robustness and interpretability.**

From generative adversarial networks to diffusion models, the forensic with our boosted invariants demonstrates state-of-the-art discriminability against large-scale content diversity. It also exhibits high efficiency on training examples, intrinsic invariance to geometric variations, and better interpretability of the forensic process.

***Index Terms*—Artificial intelligence generated content, forensics, representations, invariants, geometric deep learning.**

I. Introduction

Generative artificial intelligence (AI) has made great progress in text, image, and video synthesis, as seen in the commercial ChatGPT, DALL·E, and Sora [1]. In particular, AI-generated images present a high degree of realism versus natural images captured by cameras – humans struggle to distinguish between the real and fake [2]. Thus, forensic is important in terms of ensuring information security and improving generation capability. On the one hand, the forensic results help the public better fight off the misleading information generated by malicious actors [3]. On the other hand, the forensic results help generative AI developers better understand the weakness of their models in simulating natural images [4]. We highlight that robustness and interpretability [5] play a particularly important role in achieving the above practical goals, compared to typical computer vision scenarios.

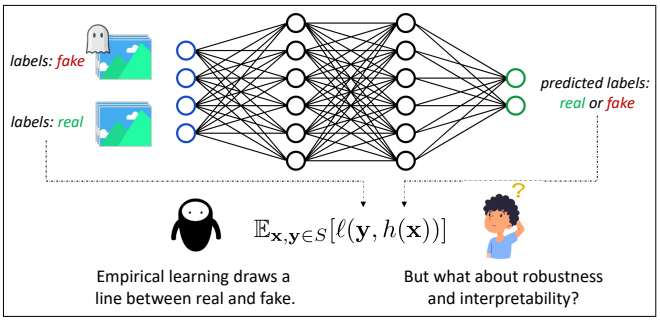


Fig. 1. Illustration for the motivation of this paper. We highlight the needs for robustness and interpretability in forensic tasks, rethink the flaws of empirical learning in this regard, and propose a solution from geometric invariance theory.

- *Robustness.* For information security, robustness means the ability to defend against potential attacks by malicious actors. For generation quality, developers are more interested in the robust pattern of differences between real and fake images, rather than individual cases.
- *Interpretability.* For information security, forensic needs to meet human interpretability if results are considered as publicly credible evidence. For generation quality, interpretability determines the potential for developers to understand and overcome such differences.

Robust and interpretable forensics rely on robust and interpretable representations [6]. Technically, the current forensics typically employ the *representation learning* in computer vision – convolutional neural networks (CNN) with empirical risk minimization training on large-scale real/fake image datasets [7]. The community has revealed that forensic systems with such representations can draw the line between real and fake [8]. However, such empirical learning neither naturally satisfies the robustness and interpretability requirements beyond experience nor is approximating the two with the empirical risk minimization efficient and protected.

One natural question is how to endow the forensic representations with principles of robustness and interpretability? For this problem, let us recall an early wisdom in feature

Shuren Qi and Tieyong Zeng are with the Department of Mathematics, The Chinese University of Hong Kong, Hong Kong, China (e-mail: shurenqi@cuhk.edu.hk, zeng@math.cuhk.edu.hk).

Fenglei Fan is with the Department of Data Science, City University of Hong Kong, Hong Kong, China (e-mail: fenglfan@cityu.edu.hk).

Chao Wang and Zhiqiu Huang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China (e-mail: c.wang@nuaa.edu.cn, zqhuang@nuaa.edu.cn).

Yushu Zhang is with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China, and also with the Jiangxi Provincial Key Laboratory of Multimedia Intelligent Processing, Nanchang, China (e-mail: zhangyushu@jxufe.edu.cn).

Xiangyu Chen is with the Institute of Artificial Intelligence (TeleAI), China Telecom, Shanghai, China (e-mail: chxy95@gmail.com).

Yi Zhang is with the School of Cyber Science and Engineering, Sichuan University, Chengdu, China (e-mail: yzhang@scu.edu.cn).

* *Corresponding author: Fenglei Fan*

engineering: *geometric invariants* [9], [10]. With the *symmetry* idea from mathematics and physics [11], invariants are designed with the algebraic structure to be unchanged w.r.t. image coordinate transformations like translation, rotation, and scaling. Reprensentations derived from geometric invariants are intrinsically robust and interpretable. So far, they have been well explored in low-level vision problems with geometric variations such as image registration [12]. However, extending robust and interpretable invariants to today's large-scale tasks, e.g., forensics of AI-generated images, is challenging. Mathematically, they are typically (under-)complete, leading to very limited discriminability compared to modern learning networks [13].

In this work, we identify the potential of geometrical invariants in the forensic field. Our motivation is to leverage the concept of geometrical invariants to design a general strategy for achieving robustness, interpretability, and discriminability in today's forensic tasks of AI-generated images. Our contributions are twofold:

- *Representation.* We mathematically redesign typical CNN modules from the perspective of geometric invariance theory rather than empirical learning. Here, the symmetry principle for translation, rotation, flipping, and scaling holds across all layers, serving as a basis for robustness and interpretability. Without the need for training, the resultant invariants can exhibit a similar discriminative power to learned CNNs, due to the high degree of overcompleteness in deep cascading representations.
- *Forensic.* We achieve state-of-the-art forensic scores from generative adversarial networks to diffusion models, even with large-scale real/fake content diversity like ImageNet. Here, a naive classifier equipped with our invariants shows direct gains over typical forensics of empirical learning – high efficiency on training examples, intrinsic invariance to geometric variations, and better interpretability of the forensic process.

II. Related Works

We briefly review the topics of geometric invariant representations and AI-generated image forensics that are closely related to our work.

A. Geometric Invariant Representations

The quest for geometric invariance is as old as the field of computer vision itself.

In the era of feature engineering, geometric invariant designs are ubiquitous in a variety of image representations, detectors, and descriptors [14], with milestones ranging from global moment invariants [10], to sparse SIFT [15], and to dense DAISY [16]. Here, moment invariants are a very systematic theory, covering almost all other invariant practices. However, this theory is more suitable for single-layer global representations, and extending its successes to SIFT-like localities and CNN-like hierarchies remains an open question.

In the era of deep learning, geometric invariant designs are relatively less common, due to the new philosophy of being data-driven [17]. Nevertheless, the geometric invariance still serves as a cornerstone in the successful deep representation, and is also known as a regularization against overfitting [18]. Here, the translation invariance of the convolution operator in CNNs is the key to achieving a better representation of images than typical neural networks [19]. However, for other geometric transformations, it is very difficult to consider their invariance in deep cascading CNNs. Currently, the main solution in this regard is data augmentation, i.e., covering geometric variants in the training. As can be expected, such empirical invariance, driven only by data and accuracy, is inefficient and unprotected [20]. More recently, the design of more elegant deep learning invariants has become a topic called geometric deep learning [11], to which our work belongs.

An over-simplified but popular practice in geometric deep learning is group equivariant convolution [21]–[26]. The input feature map is convolved with symmetry versions of a learned filter to obtain multi-channel features, where the distortion on the input (e.g., rotation) corresponds to the cyclic shift between channels, and hence the invariance is achieved by pooling across channels. We argue that this practice is computationally inefficient, where the invariance is limited by the sampling rate on symmetry [27]. This is especially true when seeking the joint invariance of multiple geometric transformations. On the contrary, in this paper we try to give a more efficient practice towards forensic tasks. Here, we achieve *one-shot* invariance for translation, rotation, flipping, and scaling in a learning-free manner, which is not available in the current community of geometric deep learning.

B. AI Generated Image Forensics

Generative AI is a hot research field with numerous commercial applications whose abuse sparks new challenges in forensics due to the high fidelity and intractability of AI-generated content.

In generating visual content, generative adversarial network (GAN) [28], variational autoencoder (VAE) [29], and diffusion model (DM) [30] are the three of the most successful techniques, playing fundamental roles in commercial applications such as Deepfake[1], DALL·E[2], and Sora[3]. Here, GAN and DM are superior in fidelity [31] of generated images, whereas DM further outperforms GAN in diversity [31] of generated content. Therefore, though we underscore that our method is general-purpose, various commercial variants of GAN and DM are reasonably considered as our main targets for forensics.

As a safeguard against the abuse of generative AI, forensic research is carried out in active and passive manners [8], depending on whether the action is taken before or after data distribution. Active forensics are typically performed by embedding robust patterns in the image, i.e., digital watermarking [32], or extracting image fingerprints as registration, i.e., hashing and blockchain [33]. Clearly, the embedded or extracted information serves as strong priors for high-confidence forensics. However, their deployment needs additional action before each image distribution, which is difficult to acquire.

[1]https://deepfakesweb.com/
[2]https://openai.com/index/dall-e-3/
[3]https://openai.com/index/sora/

Passive forensics relies exclusively on the given image itself. They all attempt to discover artifacts at digital [34], physical [35], and semantic [36] levels, which are inevitably introduced in certain manipulations by models. Clearly, the representation of such artifacts is a central concern of passive forensics, with discriminability for large-scale real/fake images, robustness to natural/artificial perturbations, and interpretability during the process, to which our work belongs.

III. Methodology

Our methodology consists of the following four propositions:

- *Feature extraction.* The hierarchical invariants we extract form a huge space of image features with a high degree of overcompleteness.
- *Feature selection.* Such overcompleteness allows further feature selection for data adaptivity and task discriminability, while compressing the computational overhead.
- *Feature classification.* With the selected features, a naive machine learning model can efficiently divide the feature space to classify real and fake images.
- *Feature interpretation.* For the sake of human understanding and further improving, the crucial differences between real and fake images can be highlighted by feature interpretation.

We would first like to highlight the simplicity and generality of the above pipeline compared to end-to-end learning. Although the four propositions appear to be more complex than end-to-end learning, they are easy to implement because there is no tricky optimisation for a large number of parameters; moreover, the four propositions are abstract enough to cover a very wide range of downstream tasks beyond forensics.

A. Feature Extraction

In this part, we discuss how to extract the geometric invariant features with hierarchical architectures. Before formally giving definitions, we highlight some high-level intuitions for the sake of understanding.

- Our representation can be regarded as a non-learning CNN with predefined modules: *convolutional*, *activation*, *pooling*, and *invariant* layers. Here, "non-learning" means all these modules are directly obtained from geometric invariants without the need of training.
- The convolutional layer captures *local* features of the image, while the invariant layer forms *global* features from such local features as the final output, both have robust structures to translation, rotation, flipping, and scaling.
- Here, inspired by the idea of deep learning, the combination of the convolutional, activation, and pooling layers can be cascaded multiple times to extract features in a *hierarchical* manner, which can greatly boost the discriminative ability of geometric representation.
- Note that we will define the convolutional (local) and invariant (global) layers using only *harmonic* functions, which are well understood in harmonic analysis and thus serve as a good basis for interpretability from the perspective that its mathematical property is crystal clear.

Let us define convolutional, activation, and pooling layers.

Definition 1 (Convolutional Layer)**.** *For the input feature map $M(i,j;k)$ with the pixel position $(i,j) \in \Omega = \{1,2,...,N_i\} \times \{1,2,...,N_j\}$, the channel $k \in \{1,2,...,K\}$, and the pixel value $M \in \mathrm{H} = \mathbb{C}^K$, the convolutional layer $\mathbf{C}$ is defined channel-wise as a local transformation [6]:*

$$\mathbf{C}M \triangleq M(i,j;k) \otimes (H^w_{nm}(i,j))^T, \tag{1}$$

where $\otimes$ is the convolution over the Ω, $(\cdot)^T$ denotes the matrix transpose, and H^w_{nm} is a convolution kernel defined as

$$H^w_{nm}(i,j) = \{h^{uvw}_{nm}(i,j) : u,v = w, (i,j) \text{ s.t. } D_{ij} \cap D \neq \emptyset\}, \tag{2}$$

where h^{uvw}_{nm} is the integral value of the basis function over a valid pixel region:

$$h^{uvw}_{nm}(i,j) = \iint_{D_{ij}\cap D} (V^{uvw}_{nm}(x,y))^* dxdy, \tag{3}$$

with the (i,j)-centered pixel region $D_{ij} = \{(x,y) \in [i - \frac{\Delta i}{2}, i + \frac{\Delta i}{2}] \times [j - \frac{\Delta j}{2}, j + \frac{\Delta j}{2}]\}$. The basis function V^{uvw}_{nm} is defined with the order parameters $(n,m) \in \mathbb{Z}^2$, the position parameters $(u,v) \in \mathbb{R}^2$, and the scale parameter $w \in \mathbb{R}^+$ [6]:

$$V^{uvw}_{nm}(x,y) \triangleq R_n\left(\frac{\sqrt{(x-u)^2 + (y-v)^2}}{w}\right) A_m\left(\arctan\left(\frac{y-v}{x-u}\right)\right), \tag{4}$$

on the domain $D = \{(x,y) : (x-u)^2 + (y-v)^2 \leq w^2\}$, where a family of complex exponential functions serves as the angular basis $A_m(\theta) = \exp(\boldsymbol{j}m\theta)$ ($\boldsymbol{j} = \sqrt{-1}$), and the radial basis $R_n(r)$ is defined by families of cosine functions [37]:

$$R^{(\mathrm{C})}_n(r;\alpha) = \begin{cases} \sqrt{\frac{\alpha r^{\alpha-2}}{2\pi}} & n = 0 \\ \sqrt{\frac{\alpha r^{\alpha-2}}{\pi}} \cos(\pi n r^\alpha) & n > 0 \end{cases}, \tag{5}$$

sine functions [37]:

$$R^{(\mathrm{S})}_n(r;\alpha) = \sqrt{\frac{\alpha r^{\alpha-2}}{\pi}} \sin(n\pi r^\alpha), n > 0, \tag{6}$$

and complex exponential functions [37]:

$$R^{(\mathrm{CE})}_n(r;\alpha) = \sqrt{\frac{\alpha r^{\alpha-2}}{2\pi}} \exp(\boldsymbol{j}2n\pi r^\alpha), \tag{7}$$

respectively, with a hyperparameter $\alpha \in \mathbb{R}^+$.

We would like to highlight the motivations for defining the discrete convolution kernel (2) and its continuous basis functions (4) as the above structure: 1) computational simplicity of $\sin$ or $\cos$, 2) orthogonality of $A_m(\theta)$ and $R_n(r)$, 3) interpretable analytic meanings of (n,m), 4) time-frequency diversity of α, and 5) symmetry properties for translation, rotation, flipping, and scaling of (u,v,w) or θ.

For the symmetry properties, we have $< f(x+\Delta x, y+\Delta y), V_{nm}^{uvw} >=< f(x,y), V_{nm}^{(u+\Delta x)(v+\Delta y)w} >$ as equivariance of translation; $< f(r,\theta+\phi), V_{nm}^{uvw} >=< f(r,\theta), V_{nm}^{uvw} > A_m^*(-\phi)$ with $(u,v)=(0,0)$ as covariance of center-aligned rotation; $< f(r,-\theta), V_{nm}^{uvw} >= (< f(r,\theta), V_{nm}^{uvw} >)^*$ with $(u,v)=(0,0)$ as covariance of center-aligned flipping, and it is straightforward that the joint invariance of center-aligned rotation and flipping holds when taking the magnitude; $< f(sx,sy), V_{nm}^{uvw} >=< f(x,y), V_{nm}^{uv(ws)} >$ with $(u,v)=(0,0)$ as covariance of center-aligned scaling.

For the representation properties when $(u,v)\neq(0,0)$, they can be derived from the composite of translation with center-aligned rotation, flipping, and scaling, respectively. Hence, the magnitude of the representation has *joint equivariance for any translation, rotation, and flipping* on (u,v) domain, as well as *covariance for any scaling* on w domain.

Definition 2 (Activation Layer)**.** *For the input feature map $M(i,j;k)$ with the pixel position $(i,j)\in\Omega=\{1,2,...,N_i\}\times\{1,2,...,N_j\}$, the channel $k\in\{1,2,...,K\}$, and the pixel value $M\in\mathrm{H}=\mathbb{C}^K$, the activation layer $\mathbf{S}$ is defined channel-wise as a magnitude operation:*

$$\mathbf{S}M=\sigma(M(i,j))\triangleq|M(i,j;k)|, \tag{8}$$

where $M(i,j;k)$ is complex-valued, and (8) can be written explicitly as $\sqrt{(\mathrm{Re}M(i,j;k))^2+(\mathrm{Im}M(i,j;k))^2}$.

Definition 3 (Pooling Layer)**.** *For the input feature map $M(i,j;k)$ with the pixel position $(i,j)\in\Omega=\{1,2,...,N_i\}\times\{1,2,...,N_j\}$, the channel $k\in\{1,2,...,K\}$, and the pixel value $M\in\mathrm{H}=\mathbb{C}^K$, the pooling layer $\mathbf{P}$ is defined channel-wise as a downsampling operation:*

$$\mathbf{P}M=M', \tag{9}$$

with $M':\Omega'\to\mathrm{H}$, where the local pooling layer $\mathbf{P}$ downsamples the plane dimensions of feature maps to reduce computational complexity, such that $\Omega'\subseteq\Omega$.

With the above definitions, the convolutional, activation, and pooling layers can be cascaded multiple times, all satisfying the following symmetry properties w.r.t. the group of geometric transformations (see Appendix for proofs).

Property 1 (Equivariance for Translation, Rotation, and Flipping)**.** *For a representation unit $\mathbf{U}\triangleq\mathbf{P}\circ\mathbf{S}\circ\mathbf{C}$ with an arbitrary parameter $\lambda=(n,m,w)$ (in the convolutional layer), any composition of $\mathbf{U}$ satisfies the joint equivariance for translation, rotation, and flipping (ignoring edge effects and sampling errors):*

$$\mathbf{U}_{[L]}\circ\cdots\circ\mathbf{U}_{[2]}\circ\mathbf{U}_{[1]}(\mathfrak{g}_1M)\equiv\mathfrak{g}_1\mathbf{U}_{[L]}\circ\cdots\circ\mathbf{U}_{[2]}\circ\mathbf{U}_{[1]}(M), \tag{10}$$

for any composition length $L\geq1$, any $\mathfrak{g}_1\in\mathfrak{G}_1$, and any $M\in\{\Omega\to\mathrm{H}\}$, where $\mathfrak{G}_1$ is the translation/rotation/flipping symmetry group.

Property 2 (Covariance for Scaling)**.** *For a representation unit $\mathbf{U}$, where the scale parameter of its convolutional layer is specified as w with a notation $\mathbf{U}^w\triangleq\mathbf{P}\circ\mathbf{S}\circ\mathbf{C}^w$, any composition of $\mathbf{U}^w$ satisfies the covariance for scaling (ignoring edge effects and sampling errors):*

$$\begin{aligned}&\mathbf{U}_{[L]}^w\circ\cdots\circ\mathbf{U}_{[2]}^w\circ\mathbf{U}_{[1]}^w(\mathfrak{g}_2M)\\&\equiv\mathfrak{g}_2'\mathbf{U}_{[L]}^w\circ\cdots\circ\mathbf{U}_{[2]}^w\circ\mathbf{U}_{[1]}^w(M)\\&=\mathfrak{g}_2\mathbf{U}_{[L]}^{ws}\circ\cdots\circ\mathbf{U}_{[2]}^{ws}\circ\mathbf{U}_{[1]}^{ws}(M),\end{aligned} \tag{11}$$

for any composition length $L\geq1$, any $\mathfrak{g}_2\in\mathfrak{G}_2$, and any $M\in\{\Omega\to\mathrm{H}\}$, where $\mathfrak{g}_2'$ is a predictable operation corresponding to $\mathfrak{g}_2$ with explicit form $\mathfrak{g}_2'\mathbf{U}^w\triangleq\mathfrak{g}_2\mathbf{U}^{ws}$, s is the scaling factor w.r.t. $\mathfrak{g}_2$, and $\mathfrak{G}_2$ is the scaling symmetry group.

With the above beneficial properties in line with the symmetry principle, we can design the last invariant layer on the *deep feature map*, exactly as the early wisdom of moment invariants directly on the *input image*.

Definition 4 (Invariant Layer)**.** *For the input feature map $M(i,j;k)$ with the pixel position $(i,j)\in\Omega=\{1,2,...,N_i\}\times\{1,2,...,N_j\}$, the channel $k\in\{1,2,...,K\}$, and the pixel value $M\in\mathrm{H}=\mathbb{C}^K$, the invariant layer $\mathbf{I}$ is defined channel-wise as a global transformation [13]:*

$$\mathbf{I}M=\mathcal{I}(\{\langle M(i,j;k),V_{nm}^{001}(x_i,y_j)\rangle\}), \tag{12}$$

with $\mathcal{I}$ maps image moments [13] to global invariants w.r.t. the symmetry group of interest $\mathfrak{G}_0\subseteq\mathfrak{G}_1\times\mathfrak{G}_2$, where $\mathfrak{G}_1$ is the translation/rotation/flipping symmetry group and $\mathfrak{G}_2$ is the scaling symmetry group.

Note that we have not restricted $\mathcal{I}$ to a specific mathematical operation in the above definition, allowing the generality of our core theoretical result (to be seen later). In the monograph [13], a number of strategies for directly constructing global invariants in image domains have been presented. They can be naturally used to define $\mathcal{I}$, with the symmetry properties of deep feature maps.

Definition 5 (Path)**.** *Based on Definitions 1 $\sim$ 4, we define a path as $p=(\lambda_{[1]},\lambda_{[2]},\cdots,\lambda_{[L]})$, where $\lambda_{[z]}=(n,m,w)_{[z]}$ is a triplet specifying the parameters of the convolutional layer in the representation unit sorted by z. The representation $\mathcal{R}_p$ is defined as the following ordered cascading with corresponding parameters $p=(\lambda_{[1]},\lambda_{[2]},...,\lambda_{[L]})$:*

$$\mathcal{R}_p\triangleq\mathbf{I}\circ\mathbf{U}_{[L]}\circ\mathbf{U}_{[L-1]}\circ\cdots\circ\mathbf{U}_{[1]}. \tag{13}$$

With the above layer definitions and their properties, we draw the following core theoretical result: our representation satisfies *hierarchical invariance* along any *path* (see Appendix for proofs).

Property 3 (Hierarchical Invariance)**.** *The representation $\mathcal{R}_p$ satisfies the invariance w.r.t. the symmetry group of interest $\mathfrak{G}_0\subseteq\mathfrak{G}_1\times\mathfrak{G}_2$ (ignoring edge effects and sampling errors):*

$$\mathcal{R}_p(\mathfrak{g}_0M)=\mathcal{R}_p(M), \tag{14}$$

for any path p (i.e., any layer parameter λ and any compo-

sition length $L \geq 1$), any $\mathfrak{g}_0 \in \mathfrak{G}_0$, and any $M \in \{\Omega \to \mathrm{H}\}$.

B. Feature Selection

In this part, we discuss how to select the discriminative features for the forensic task from the set of overcomplete invariants. Note that our representation entails a very wide range of features, with a scale proportional to the number of all possible paths, i.e., the number of *combinations* of all possible values taken about the layer parameter λ and the composition length L. This highly overcomplete nature distinguishes our representation from classical geometric invariants in the discriminative potential. Next, a new challenge is how to select task-discriminative features from such a huge space.

We will respond to this challenge through the following three-step method:

- **Super network**. Building a super network that covers preferred and sufficiently diverse parameters under a given composition length.
- **Correlation analysis**. Computing the features on the images of the training set by this super network and filtering the most relevant features by the correlation analysis of features and labels for a given forensic task.
- **Concise network**. With the parameters behind these most relevant features, resampling the super network yields a concise network.

Regarding the super network, we adopt a simple yet effective strategy that enjoys a very diverse and complementary feature space.

- Specifically, for the parameters $(n, m)_{[z]}$ of the representation unit $\mathbf{U}_{[z]}$ sorted by z, we set $\{(n, m)_{[z]} : n + m = z, (n, m) \in \mathbb{N}^2\}$, i.e., they are equal under the ℓ_1 norm. This setting allows each level of the network to capture *complementary frequency* information. Moreover, low frequencies preferentially appear in the shallow layers, enabling the key information to be transmitted to deeper layers. To further enhance the *information transmissibility* for deeper layers, we also introduce an identity function for each layer with $z > 1$, similar to the skip connection trick in ResNet [38].
- For the parameter $w_{[z]}$ of the representation unit $\mathbf{U}_{[z]}$ sorted by z, we set $w_{[z]} = w_0$, i.e. a fixed scale w_0 for the entire network, as a *single-scale* representation; we can also set $\{w_{[z]} : w_{[z]} = 2^t, t \in \mathbb{Z}\}$ for each of the *multi-scale* networks, where the scaling covariance (w.r.t. w) is transformed into a linear translation pattern (w.r.t. t) between multi-scale networks.
- For the parameters (n, m) of the invariant layer $\mathbf{I}$, we set $\{(n, m) : n, m \leq T, (n, m) \in \mathbb{N}^2\}$ to capture *multi-frequency global features* of the final layer, instead of the often-used global average/maximum pooling. In all experiments of this paper, a class of global invariants is designed by frequency band pooling: $\mathcal{I}(\cdot) \triangleq \{I_i = \sum_{(n,m)\in\mathcal{B}_i} |\cdot| : i = 1, 2, \cdots, \#_B\}$, where $\mathcal{B}_i = \{(n, m) : \sqrt{2}T(i-1)/\#_B| \leq ||(n, m)||_2 \leq \sqrt{2}Ti/\#_B\}$ is the i-th frequency band, with the number of bands $\#_B$. Note that I_i offers global rotation and flipping invariance.
- As for the hyperparameter α of the radial basis $R_n(r; \alpha)$, it can control the distribution of zeros inside the convolutional kernel. Here, two values of α are important: $\alpha = 1$ exhibits a uniform zero distribution and the strongest representation capability, but may have computational stability issues due to singularities; $\alpha = 2$ avoids singularities, but with limited representation capability. Therefore, a *fixed* value of α can be chosen from the interval $[1, 2]$ to balance representation capability and computational stability. The *non-fixed* values of α, such as random sampling in the interval $[1, 2]$, can further provide over-complete and space-frequency features on such small scales, which may be useful for the boosting-like machine learning ideas.

Regarding the correlation analysis, a large number of techniques in the field of feature selection are optional, including *filter*, *wrapper*, and *embedded* types of feature selection [39]. In our implementation, the simplicity and efficiency of filter type are mainly considered, and features are ranked by the *chi-square test*.

Regarding the concise network, we identify fewer paths of the super network so that they largely cover the above discriminative features. Such paths along with their parameters will be recorded as the concise network. Thus, we can extract features efficiently for the deployment phase.

C. Feature Classification

In this part, we discuss how to learn the division of the feature space to classify real and fake images. There exist two important classes of models, i.e. support vector machine (SVM) and neural network (NN), realizing the nonlinear division of the feature space. Such classification models can be trained on features from the training image set, and the learned model parameters are recorded for deployment.

We would like to highlight the difference between the applicability of SVM and NN in forensic scenarios.

- **SVM** exhibits higher training stability and testing generalization/robustness in the case with fewer examples. However, its training complexity is high for large-scale datasets, and it also only weakly supports the multi-class classification.
- **NN** is conveniently accelerated by GPU parallelism. However, it requires sufficiently diverse examples and performs inconsistently in training stability and testing generalization/robustness.

Given the relatively complementary nature of the SVM and NN in scenarios, both models are considered in our implementation.

D. Feature Interpretation

In this part, we discuss how to interpret the features that play a decisive role in distinguishing real and fake images. Because our model is totally mathematically transparent, we can analyze the mode of feature extraction in our model straightforwardly. In this vein, we can provide the following two levels of interpretation w.r.t. the terms of harmonic analysis and frequency.

TABLE I
LABORATORY FORENSIC SCORES (F1, %) OF REPRESENTATION BACKBONES AND FORENSIC SOLUTIONS AGAINST COMPREHENSIVE GENERATORS.

	ADM	BGAN	GLIDE	Midjourney	SD 1.4	SD 1.5	VQDM	Wukong	*Avg.*	*Min.*
Handcrafted										
DCT NN	0	0	20.49	62.38	0	33.63	0	0	14.56	0
DCT SVM	99.52	99.92	99.61	89.65	98.66	98.39	99.42	98.73	97.99	89.65
DWT NN	65.84	66.63	66.7	0	0	66.90	0	13.33	34.93	0
DWT SVM	99.97	99.91	99.96	84.98	99.17	98.86	99.97	99.13	97.74	84.98
ScatterNet NN	99.22	99.53	1.03	80.86	0	0	96.77	89.77	58.40	0
ScatterNet SVM	96.09	99.78	97.02	88.96	96.46	97.04	98.71	95.96	96.25	88.96
Learning										
SimpleNet	98.25	97.87	92.98	68.00	73.52	74.37	74.88	76.32	82.02	68.00
ResNet	98.78	99.14	97.78	87.41	89.88	90.85	88.53	88.80	92.65	87.41
DenseNet	99.63	99.60	98.57	93.08	93.79	94.50	95.01	92.55	95.84	92.55
InceptionNet	97.69	99.41	98.32	90.07	89.40	92.55	92.72	88.35	93.56	88.35
ViT	99.43	99.22	98.24	35.38	76.47	82.09	98.03	88.02	84.61	35.38
Forensic										
CNN Spot	76.44	99.3	99.83	75.63	58.26	52.55	98.48	66.03	78.32	52.55
F3Net	98.02	99.79	98.93	89.63	87.41	86.54	93.94	87.8	92.76	86.54
Ours										
BGI NN	99.9	99.96	99.91	93.28	99.4	99.14	99.83	99.18	98.83	93.28
BGI SVM	99.88	99.95	99.93	93.58	99.55	99.36	99.85	99.31	98.93	93.58

- **Local feature interpretation.** We can identify the most important *convolution kernels* capturing *key local patterns*, with harmonic analysis meanings specified by the order parameters (n, m) in the *representation unit* $\mathbf{U}$. For human intuition, such important kernels are visualized layer by layer; the deep feature maps passing through these filters (i.e., the important path) can also be visualized.
- **Global feature interpretation.** We can identify the most important *frequency bands* capturing *key global patterns*, with harmonic analysis meanings specified by the order parameters (n, m) in the *invariant layer* $\mathbf{I}$. For human intuition, such important bands are visualized by histograms; the corresponding frequency components of the deep feature maps can be visualized by reconstruction.

The above feature interpretations promise interesting applications that are difficult to achieve with the end-to-end learning. Here, we list two applications, and more are yet to be discovered.

- **Generative model attribution.** Going beyond real/fake discrimination, we can achieve a more *fine-grained attribution analysis of fakes*, even revealing information about what generative models are used to create these fake images. Here, we subdivide the already selected features (see also Section III.B), training another multi-class classifier for attribution based on the corresponding generative model labels. Note that all models are represented by such a fixed set of local/global features (rather than end-to-end learning), so attribution results of different models are *comparable* for human intuition.
- **Visual artifact polishing.** With the above interpretations, we can obtain *explicit features that describe visual artifacts* in the real/fake discrimination and model attribution. Such knowledge can be directly used to generate more realistic visual media by polishing these explicit features during the training process. In addition, such knowledge is also instrumental to some subproblems of generative models, such as ablation analysis with feature comparison, which can further improve the human understanding of the model and boosting the generation quality.

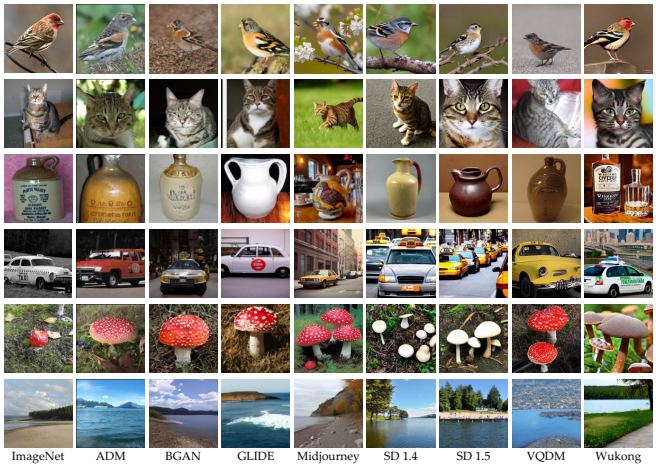


Fig. 2. Illustration for the experimental image dataset synthesized through a variety of generative methods ranging from generative adversarial networks to diffusion models.

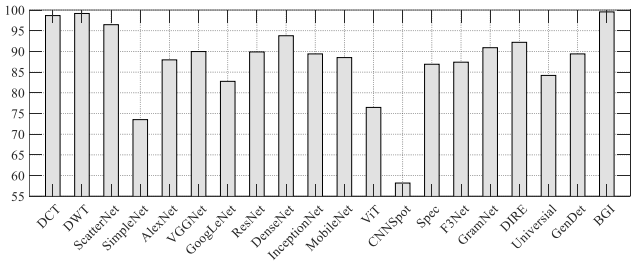


Fig. 3. Laboratory forensic scores (F1, %) of comprehensive representation backbones and forensic solutions against SD 1.4 generator.

IV. NUMERICAL AND VISUALIZATION EXPERIMENTS

In this section, the proposed forensic is evaluated on large-scale generated images, with experiments completely covering discriminability, efficiency, robustness, and interpretability.

A. Experimental Organization and Setup Details

The organization of the series of experiments and their setup details are sorted out here.

- **Organization.** We conduct *benchmark experiments* in ideal and realistic scenarios, i.e., with only laboratory discriminability factors and with also real-world efficiency and robustness factors, respectively. We also conduct *specific experiments* on the equivariant networks (theoretical competitors) and the updated generators (stronger benchmarks); however, due to their computational complexity, these comparisons are in-depth rather than large-scale. Going inside our forensic process, especially its representation, we also provide interpretability insights with *visualization experiments*.
- **Comparison.** We consider related works at both levels of *representation* and *forensic*. Regarding the competitors at representation level, we cover *1)* a classical frequency representation – DCT, *2)* a simple CNN trained from scratch – SimpleNet, *3)* a series of milestones in deep CNN with transfer training from ImageNet pre-trained

weights – AlexNet [40], VGGNet [41], GoogLeNet [42], ResNet [38], DenseNet [43], InceptionNet [44], MobileNet [45], and ViT [46], *4)* non-learning and learning deep CNNs with invariant design, as our direct competitors in theory – ScatterNet [47] and EquiNet [21]. Due to the high complexity of EquiNet, we conducted only specific comparisons. Regarding the competitors at forensic level, we cover a series of major developments in forensics for large-scale generated images – CNNSpot [48], Spec [49], F3Net [50], GramNet [51], DIRE [52], Universal [53], and GenDet [54]. Note that most of such forensic competitors use the above listed representations as their backbone, with also author's tricks on prior knowledge for improving the domain accuracy. Therefore, throughout the experiment section, we will mainly consider representation-level competitors for a truly fair comparison.

- **Dataset.** As shown in Fig. 6, we synthesize fake images through a variety of generative methods ranging from *generative adversarial networks* to *diffusion models*, with similar visual effect to the real images of *ImageNet* dataset. Here, generative methods include ADM [55], BGAN [56], GLIDE [57], Midjourney[4], SD 1.4 [58], SD 1.5 [58], VQDM [59], and Wukong[5], resulting in 8 benchmarks, with 100,000 pairs of fake and real images. Note that such benchmarks exhibit discriminative challenges, due to the very rich content diversity for both real and fake images. Considering the progress of diffusion models, we also included updated generators: SD 3.5[6] and FLUX[7]. Due to the high complexity of the two, we conducted only specific comparisons with 2000 pairs of fake and real images.
- **Implementation.** We implemente a super network with single-scale setting $w = 6$ and maximum depth $L = 7$, where the invariant layer is specialized by $T = (N_i + N_j)/2$ (see also Section III-B). The hyperparameter α for each convolution kernel is set by random sampling in the interval $[1, 2]$ for boosting over-complete and space-frequency features. The correlation analysis is performed by chi-aquare tests, retaining only the top-ranked 500-dimensional features along with their concise network (see also Section III-B). All such features are fed into both NN and SVM classifiers (see also Section III-C). Unless otherwise stated, the training and testing sets are formed without any crossover by random sampling at 50% and 50% ratios on the original dataset, respectively.

B. Benchmark Experiments with Laboratory Protocols

In this part, we conduct benchmark experiments with following protocols: the *discriminability* is measured by the 8 benchmarks of generated images, respectively. Note that such protocols are laboratory style due to the high degree of alignment between training and testing, where the forensic is trained and tested only on sufficient data by one generator at one time.

[4]https://www.midjourney.com/home
[5]https://xihe.mindspore.cn/modelzoo/wukong
[6]https://stability.ai/news/introducing-stable-diffusion-3-5
[7]https://github.com/black-forest-labs/flux

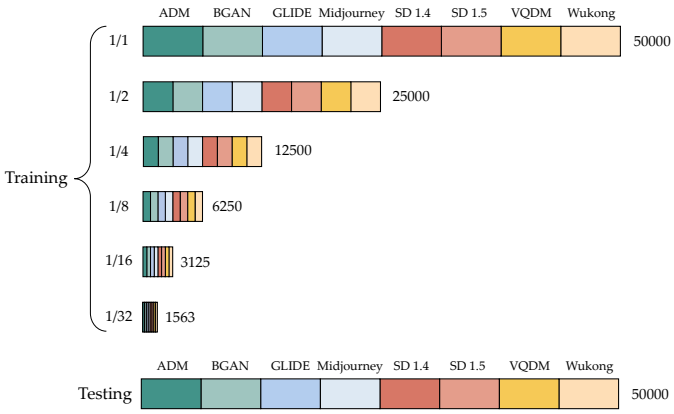


Fig. 4. Illustration for different ratios regarding number of examples for training v.s. testing in realistic experimental protocols.

TABLE II
REALISTIC FORENSIC SCORES (F1, %) OF OUR REPRESENTATION WITH COMPREHENSIVE RATIOS REGARDING NUMBER OF EXAMPLES FOR TRAINING V.S. TESTING.

	1/1	1/2	1/4	1/8	1/16	1/32
BGI NN	95.30	96.01	95.14	95.39	95.73	94.23
BGI SVM	96.23	96.06	96.49	96.09	95.71	95.08

As shown in Fig. 3, we first evaluate the discriminability of all comparison methods on a typical SD 1.4 benchmark with F1 scores. It can be seen that the majority of comparison methods achieve scores $> 80\%$; handcrafted representations, learning representations, and AIGC forensic solutions all have methods that achieve scores $> 90\%$. Their scores may continue to increase as the number of training rounds increases. In conclusion, such a laboratory scenario with a high degree of alignment between training and testing does not pose a significant challenge to most existing methods, based on the data adaptivity of the representations or classifiers. In this scenario, our boosted geometric invariants (BGI) achieve the highest forensic scores, initially confirming their effectiveness in the AIGC forensic task.

As shown in Table I, we then evaluate the discriminability of the focused comparison methods on 8 AIGC benchmarks respectively, with F1 scores and their average and minimum statistics.

- For the handcrafted representations, they scored well in most benchmarks, but the performance degrades significantly on a few benchmarks, e.g., Midjourney. In particular, they are almost all sensitive to the choice of classifiers, revealing the limitations on adaptability for AIGC forensics.
- For the learning representations, the direct-trained SimpleNet performs poorly on both average and minimum statistics; the transfer-trained ResNet, DenseNet, and InceptionNet perform significantly better, but the transfer-trained ViT also performs poorly. This phenomenon suggests that learning representations are sensitive to the choice of training strategies, parameters, and data. In

TABLE III
REALISTIC FORENSIC SCORES (%) OF REPRESENTATION BACKBONES AND FORENSIC SOLUTIONS WITH SMALLER RATIOS REGARDING NUMBER OF EXAMPLES FOR TRAINING V.S. TESTING.

	1/1			1/4			1/8		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Handcrafted									
DCT NN	0	0	0	50.00	99.97	66.66	35.82	40.41	37.98
DCT SVM	92.27	94.83	93.53	93.65	94.94	94.29	95.00	94.98	94.99
DWT NN	47.04	22.61	30.54	70.67	28.64	40.76	49.98	99.72	66.59
DWT SVM	97.21	95.04	96.13	96.77	94.51	95.62	96.75	93.61	95.15
ScatterNet NN	70.66	92.27	80.03	86.17	82.88	84.49	86.12	77.46	81.56
ScatterNet SVM	88.03	89.80	88.91	88.88	91.21	90.52	86.75	91.67	89.14
Learning									
SimpleNet	72.85	94.06	82.11	68.13	71.00	69.53	75.76	36.49	49.25
ResNet	95.65	95.31	95.48	88.96	87.84	88.40	82.11	86.66	84.32
DenseNet	99.68	99.62	99.64	92.82	94.35	93.58	89.67	90.34	90.01
InceptionNet	97.94	97.06	97.50	88.34	88.30	88.32	83.39	79.15	81.22
ViT	99.47	52.98	69.13	98.32	29.18	45.01	96.55	55.65	70.61
Forensic									
CNN Spot	61.63	87.04	72.10	66.44	84.75	74.41	66.57	82.93	73.74
F3Net	80.79	78.49	79.35	84.55	81.68	82.76	80.43	79.54	79.67
Ours									
BGI NN	94.17	96.65	95.30	95.87	94.98	95.14	94.67	96.40	95.39
BGI SVM	96.50	96.30	96.23	96.30	96.89	96.49	94.59	97.73	96.09

particular, the largest ViT exhibits significant instability in training. The DenseNet shows the best scores, while the training time is also significantly higher, second only to ViT. In general, learning representations perform well with high alignment between training and testing, due to their data adaptivity and empirical risk minimization principle.

- For the forensic solutions, their performance is very unstable on different benchmarks. Here, F3Net is better, but it does not reach the general level of learning representations. This phenomenon indicates the limitations on discriminability and generalizability for AIGC forensics.
- In this scenario, our BGI consistently achieves good forensic scores for different classifiers and AIGC benchmarks, on average and minimum statistics. This phenomenon further confirms its discriminability and generalizability for AIGC forensics. Note that BGI achieves a discriminability level of learned representations but in a nonlearned manner. It is a strong evidence for the advantages of embedding geometric invariant structures in forensic representations. In subsequent experiments, we will further demonstrate its benefits over handcrafted representations, learning representations, and forensic solutions.

C. Benchmark Experiments with Realistic Protocols

In this part, we conduct benchmark experiments with the following protocols: the *efficiency* and *robustness* are measured by a **hybrid testing** of the 8 benchmarks, may also with random **degradations** and **smaller ratios** regarding number of examples for training v.s. testing. Note that such protocols are real-world style, as they reflect the challenging factors of real forensic scenarios: 1) facing heterogeneous data sources, 2) being affected by natural or artificial degradations, and 3) lacking representative training examples for certain patterns.

As a visualization of the realistic protocols, we show the hybrid testing of the 8 benchmarks with different ratios regarding number of examples for training v.s. testing in Fig. 4. Note that the fake label covers images generated by different AIGC methods – the intra-class variations increase

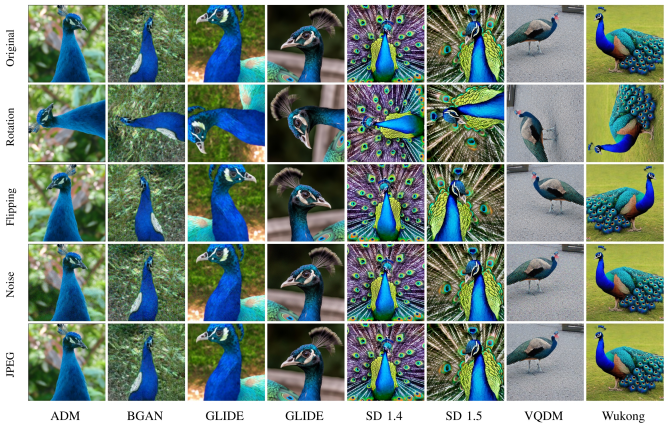


Fig. 5. Illustration for different geometric and signal degradations, i.e., random orientation, flipping, noise, and compression, in realistic experimental protocols.

TABLE IV
REALISTIC FORENSIC SCORES (%) OF REPRESENTATION BACKBONES AND FORENSIC SOLUTIONS WITH RANDOM GEOMETRIC AND SIGNAL DEGRADATIONS

	Clean			Geometric Degradation			Signal Degradation		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Handcrafted									
DCT NN	0	0	0	0	0	0	0	0	0
DCT SVM	92.27	94.83	93.53	80.86	95.03	87.38	78.50	91.35	84.44
DWT NN	47.04	22.61	30.54	50.62	25.59	33.99	52.40	26.11	34.86
DWT SVM	97.21	95.04	96.13	79.70	94.75	86.58	81.47	96.65	88.41
ScatterNet NN	70.66	92.27	80.03	69.34	88.58	77.79	67.97	95.97	79.58
ScatterNet SVM	88.03	89.80	88.91	90.67	80.23	85.13	92.37	90.38	91.36
Learning									
SimpleNet	72.85	94.06	82.11	65.03	85.90	74.02	66.13	92.61	77.16
ResNet	95.65	95.31	95.48	91.70	83.85	87.60	94.54	89.56	91.98
DenseNet	99.68	99.62	99.64	96.02	89.92	92.87	98.78	90.01	94.19
InceptionNet	97.94	97.06	97.50	92.00	92.24	92.12	96.77	84.06	89.97
ViT	99.47	52.98	69.13	61.53	99.17	75.94	64.96	97.60	78.00
Forensic									
CNN Spot	61.63	87.04	72.10	83.12	80.64	81.51	68.35	59.32	63.14
F3Net	80.79	78.49	79.35	79.83	77.57	77.96	80.96	74.83	77.13
Ours									
BGI NN	94.17	96.65	95.30	96.84	92.01	94.36	90.03	95.10	92.50
BGI SVM	96.50	96.30	96.23	96.45	93.40	94.90	92.52	95.84	94.15

significantly compared to the laboratory protocols. Also, when the ratio decreases, the number of training examples for each AIGC method decreases accordingly – the sample efficiency becomes important compared to the laboratory protocols.

As shown in Table II, we evaluate the efficiency and discriminability of our BGI on the hybrid benchmark with a variety of training/testing ratios. Here, our scores are quite stable for both ratio decreases and classifier choices. This phenomenon illustrates that BGI exhibits sufficient discriminability for complex and heterogeneous data. Note that at a ratio of 1/32, the total number of examples is only 1563, and the number for each AIGC method is < 200 – our BGI still maintains $\sim 95\%$ scores under such harsh conditions. It is an evidence of the excellent efficiency, implying a better adaptability to under-sampled AIGC methods practically.

As shown in Table III, we then evaluate the efficiency and discriminability of the focused comparison methods on the hybrid benchmark, with Precision, Recall, and F1 scores. Let us analyze the performance of each approach.

- For the handcrafted representations, the sensitivity to classifier choices remains, further confirming the limitations on data adaptability. Meanwhile, their scores are

very stable when the ratio decreases – non-data-driven designs exhibit good efficiency.

- For the learning representations other than ViT, they performed well when the ratio was 1/1, but when the ratio dropped to 1/4 and 1/8, their scores dropped significantly. The uniqueness of ViT arises from the instability of the optimization process for such large models. Here, direct training performed worse than transfer training in both average scores and score decreases. This phenomenon suggests that learning representations are highly dependent on the choice of training strategies, parameters, and data, exhibiting a poor example efficiency. Note that when the ratio continues to decrease to 1/32, learning representations fail completely, where our BGI can still maintain high scores (see also Table II).
- For the forensic solutions, when transferred from laboratory to real-world scenarios, their scores all decreased significantly, indicating insufficient discriminability for complex and heterogeneous data. On the other hand, they performed well in efficiency, with stable scores as the ratio decreased. This phenomenon illustrates the dual impact of modifying the representations with domain priors.
- In this scenario, our BGI consistently achieves good forensic scores for different classifiers and smaller ratios. In particular, when the ratio is 1/8, the efficiency advantage over learning representations is significant, which is crucial for real-world forensic problems with limited examples and resources.

As a visualization of the realistic protocols, we show the geometric and signal degradations, i.e., random orientation, flipping, noise, and compression in Fig. 5.

As shown in Table IV, we evaluate the robustness of the focused comparison methods on the hybrid benchmark, with Precision, Recall, and F1 scores. Let us analyze the performance of each approach.

- For the handcrafted representations, they exhibit varying score decreases after such post-processing, especially for geometric cases, revealing the limitations on robustness for AIGC forensics.
- For the learning representations, a similar degree of score decreases occurred, especially on Recall scores, meaning an increased probability of failing to identify AIGC images.
- For the forensic solutions, CNN Spot is less robust for post-processing at the signal level – its learning representations with training strategies fail to fully consider robust requirements. F3Net is robust for both signal and geometric post-processing, but with a low level of scores. Note that the overall performance of the two is still worse than the learning representations.
- In this scenario, our BGI consistently achieves good forensic scores for different geometric and signal degradations. The proposed invariant representation exhibits intrinsic robustness, which is crucial for real-world forensic problems with data variability.

Fig. 6. Illustration for the specific experimental images synthesized through updated generators SD 3.5 and FLUX.

TABLE V
FORENSIC SCORES (F1, %) COMPARISON ON EQUIVARIANT NETWORKS AND UPDATED GENERATORS WITH COMPREHENSIVE PROTOCOLS.

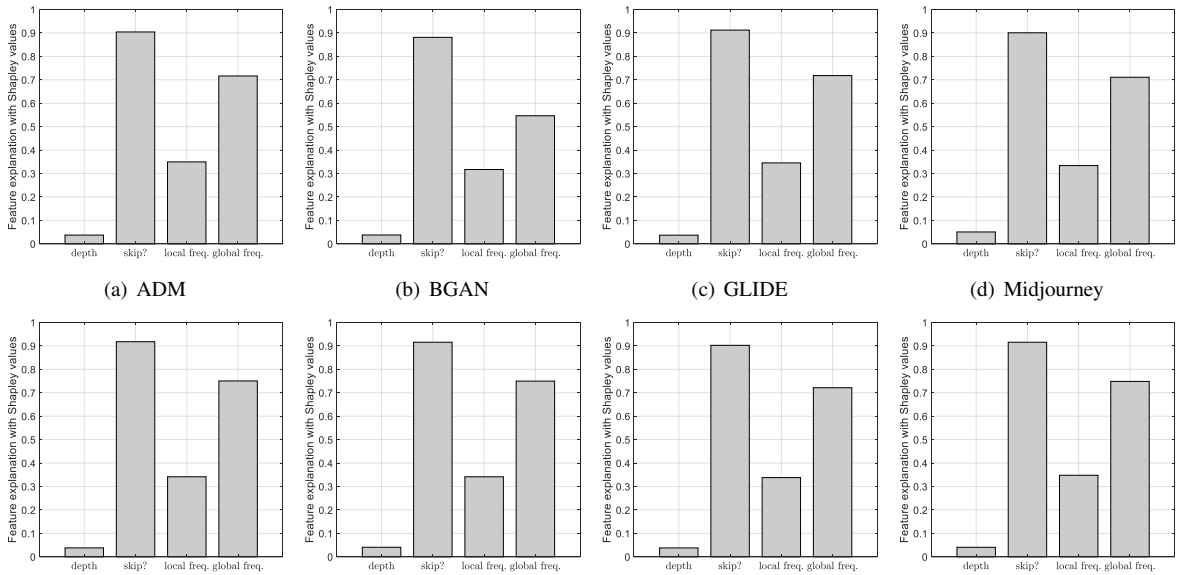
	Laboratory SD3.5	Laboratory FLUX	Realistic 1/1	Realistic 1/8	Realistic Geometric	Realistic Signal
ResNet	96.20	95.93	95.48	84.32	87.60	91.98
ViT	86.33	85.17	69.13	70.61	75.94	78.00
EquiNet	82.45	88.35	76.82	60.70	61.11	68.58
BGI	96.48	98.01	96.23	96.09	94.90	94.15

D. Specific Experiments on Equivariant Networks and Updated Generators

In this part, we conduct specific experiments on **equivariant networks**, also with updated generators **SD 3.5** and **FLUX**. Our focus is on revealing the flaws of equivariant networks by specific discriminability, efficiency, and invariance experiments, as a *motivation justification* for our representation. We also evaluate the effectiveness in the context of significant advances in generators, as a *stronger benchmarking* for our representation. Note that due to high computational and memory complexity, we cannot cover all of the benchmark experiments in Sections IV-B and IV-C.

As shown in Table V, we evaluate the discriminability of our BGI w.r.t. equivariant networks and updated generators, with also the two representations and four protocols in Sections IV-B and IV-C.

- For the updated generators, the listed representations basically maintain the performance described above. This phenomenon means that updated generators have not posed a serious challenge to the discriminative power of existing representations. The generated images still contain artificial patterns that are stable for machine learning but not obvious to human vision. On the two protocols, our BGI achieves the best scores, indicating a discriminability advantage even in the significant progress of diffusion models.
- For the equivariant networks, the discriminability is not good enough with relatively poor scores under all protocols. In fact, the EquiNet involved in this experiment is an equivariant version of the ResNet. This phenomenon means that the introduction of equivariant structures with symmetry group sampling results in a discriminability decrease. In addition, one can note that the discrete sampling does not achieve ideal efficiency and robustness in realistic protocols. Specifically, this design leads to increased memory consumption and optimization difficulty. Such problems will be further analyzed in the following experiments.



(a) ADM (b) BGAN (c) GLIDE (d) Midjourney

(e) SD 1.4 (f) SD 1.5 (g) VQDM (h) Wukong

Fig. 7. Visualization for parameters (normalized) of discriminative features regarding real/fake discrimination, by weighted average with Shapley values on each generator respectively.

TABLE VI
EXECUTION TIME (SECONDS, OOM = OUT OF MEMORY) COMPARISON ON EQUIVARIANT NETWORKS WITH DIFFERENT NUMBERS OF INPUT IMAGE PIXELS AND NUMBERS OF ROTATION GROUP SAMPLES.

Train.+Test. Time /1K+1K Images	Baseline 100^2, 4	# Pix. 200^2	# Pix. 300^2	# Pix. 400^2	# Rot. 8	# Rot. 32	# Rot. 64	# Rot. ∞
EquiNet (GPU)	182	1859	33235	OOM	682	4109	OOM	-
BGI (CPU)	64	305	1023	2027	-	-	-	64

TABLE VII
MNIST CLASSIFICATION SCORES (F1, %) ON EQUIVARIANT NETWORKS WITH ROTATIONS AS AN EMPIRICAL VERIFICATION OF INVARIANCE.

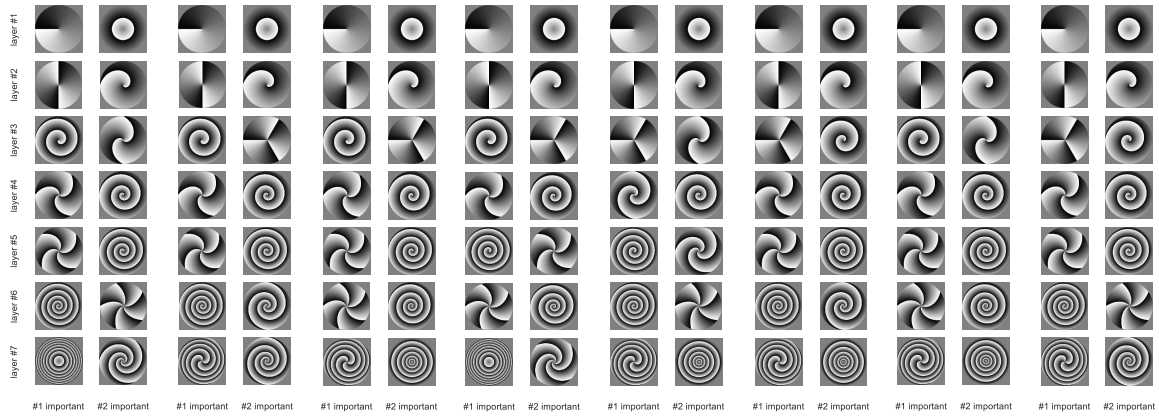
	# Rot. 4	# Rot. 8	# Rot. 16	# Rot. ∞
EquiNet (4)	85.20	87.12	83.77	85.42
EquiNet (8)	90.25	90.95	90.24	90.50
EquiNet (16)	89.99	91.41	89.38	90.17
BGI (∞)	97.34	97.30	97.45	96.70

As shown in Table VI, we directly evaluated the efficiency of our BGI w.r.t. equivariant networks, with different numbers of input image pixels and numbers of rotation group samples. Here, the execution time on 1000 training images and 1000 testing images is recorded as an efficiency metric. At the hardware level, the equivariant network runs on a single GPU: 3584 cores, 1.32 GHz, and 12 GB video RAM; our BGI runs on a single CPU: 16 cores, 3.40 GHz, and 32 GB RAM.

- The input size (number of pixels) is a major factor in efficiency. We first try to expand the input size and observe the corresponding efficiency behavior. The execution time of EquiNet expands rapidly (perhaps exponentially) when the number of pixels increases compared to the baseline. When the number of pixels reaches 400^2, EquiNet suffers from memory overflow. Note that this size is insufficient to support discriminability in many tasks. In contrast, our BGI maintains low execution time and low memory usage, and this advantage becomes more apparent as the number of pixels increases.
- For equivariant networks, the number of discrete samples of the symmetry group is another major factor affecting efficiency. As for our BGI, it has continuous invariance and therefore does not involve such efficiency factor. When the number of rotation group samples increases, the execution time of EquiNet also expands rapidly. When the numbers of rotation group samples reaches 64, EquiNet suffers from memory overflow. Clearly, EquiNet must face a trade-off between efficiency and invariance, and in theory, continuous invariance (i.e., infinite sampling) remains unachievable. In contrast, our BGI exhibits very low complexity while achieving the continuous invariance.

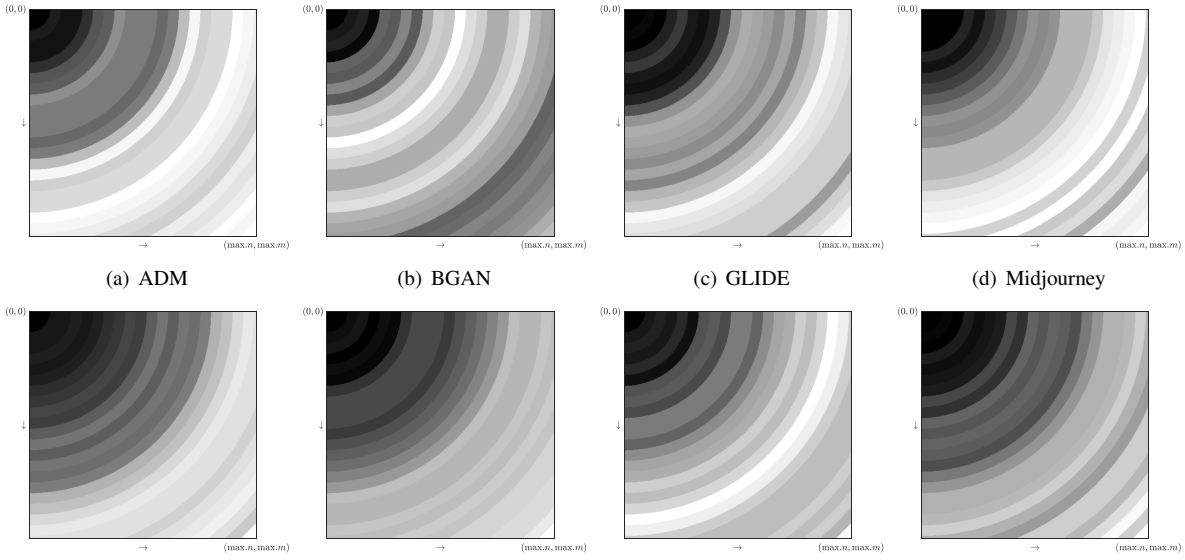
As shown in Table VII, we directly evaluate the invariance of our BGI w.r.t. equivariant networks by simulating rotations on a simple dataset. Here, the experiment is conducted on the MNIST benchmark, where images consist of clean digits and blank backgrounds – easy to simulate rotation and reflect the invariance itself (rather than being constrained by discriminability).

As already mentioned, the implementation of EquiNet is based on discrete sampling of the symmetry group, which is not involved in our BGI. Therefore, we considered different number of samples for both the implementation of EquiNet and the simulation of rotations in the testing images. During training, the original MNIST was used with relatively fixed orientations, and EquiNet easily achieved a training accuracy $\sim 100\%$. During testing, EquiNet exhibits a notable decline in scores for rotated versions, especially when the number of discrete samples of the symmetry group is small. Even with increased samples at a cost of complexity, the scores remain unsatisfactory, and this is true when training and



(a) ADM (b) BGAN (c) GLIDE (d) Midjourney (e) SD 1.4 (f) SD 1.5 (g) VQDM (h) Wukong

Fig. 8. Visualization for discriminative local features (convolution kernels) regarding real/fake discrimination, by statistic analysis on each generator respectively.



(a) ADM (b) BGAN (c) GLIDE (d) Midjourney

(e) SD 1.4 (f) SD 1.5 (g) VQDM (h) Wukong

Fig. 9. Visualization for discriminative global features (frequency bands) regarding real/fake discrimination, by statistic analysis on each generator respectively.

testing are perfectly aligned. As expected, our BGI achieves consistently good scores across various protocols, due to its continuous invariance. This phenomenon highlights the limitations of achieving invariance through symmetry group sampling: limited invariance and discriminability, with rapid increases in complexity.

E. Visualization Experiments for Interpretability

In this part, we conduct visualization experiments with following protocols: the *interpretability* is shown by **feature statistic analysis** at the local and global levels regarding real/fake discrimination, as well as more fine-grained **model attribution analysis**.

As shown in Fig. 7, we first analyze the parameter properties of discriminative features, including network topology and feature frequency properties. Here, we quantify the contribution of a BGI feature to the forensic decision-making process via the Shapley values [60]. We use Shapley values to perform a weighted average of the parameters, and then normalize them according to their value ranges, quantifying the values that tend to be used for discriminative features.

In addition, we visualize the local features (convolution kernels) and global features (frequency bands) in Fig. 8 and Fig. 9, respectively. Here, the features are sorted by Shapley values. Then, the top 2 features are selected on each composition length $l = 1, ..., L$ to show thier convolutional kernels as a visualization of local features; the top 500 features are selected on all invariant layers to calculate a frequency histogram as a visualization of global features.

One can note that the above visualizations provide some interesting insights into feature interpretation.

- In general, the statistical analysis of features on different AIGC benchmarks shows some common patterns, while also differing in detail.
- Regarding network topology, the most discriminative features typically come from the shallow layers, with a clear tendency to the skip connection. This interpretation is quite instructive for designing forensic representations

TABLE VIII
MODEL ATTRIBUTION SCORES (F1, %) OF COMPREHENSIVE REPRESENTATION BACKBONES BY TRAINING ANOTHER MULTI-CLASS CLASSIFIER OVER THE FAKE CATEGORY.

	ADM	BGAN	Midjourney	VQDM	GLIDE	SD	*Avg.*	*Min.*
Handcrafted								
DCT NN	21.94	0	0	0	0.92	0	3.81	0
DCT SVM	94.49	99.93	92.04	94.31	98.38	98.65	96.30	92.04
DWT NN	19.43	21.00	0	0.03	1.23	1.43	7.19	0
DWT SVM	98.55	100	92.07	97.22	97.99	98.76	97.43	92.07
ScatterNet NN	37.41	87.37	32.40	39.35	75.70	88.01	60.04	32.4
ScatterNet SVM	96.61	91.62	76.06	87.29	87.74	90.30	88.27	76.06
Learning								
SimpleNet	99.70	94.61	90.10	88.08	90.73	97.94	93.53	88.08
ResNet	99.58	97.76	89.79	93.33	96.70	97.74	95.82	89.79
DenseNet	99.92	99.85	99.10	99.78	99.67	99.73	99.68	99.1
InceptionNet	99.13	98.55	88.33	94.53	96.54	97.23	95.72	88.33
ViT	94.80	97.39	78.85	84.23	89.52	89.95	89.12	78.85
Ours								
BGI NN	91.33	98.39	56.17	90.85	93.35	96.34	87.74	56.17
BGI SVM	99.21	99.77	96.62	98.86	99.16	98.84	98.74	96.62

of AIGC images – we should not increase the depth of the network as in the case of image classification tasks (e.g., 201-layer DenseNet), where the main difference between AIGC images and natural images is reflected in the appearance details.

- Regarding feature frequency, the most discriminative local and global features typically come from the low-to-medium frequency bands and the medium-to-high frequency bands, respectively. Previous forensics usually focused on global high-frequency information – our interpretation is consistent with this. Also, the newly discovered tendency of local frequencies is instructive for designing forensic representations of AIGC images – we should not use high-frequency filters as in the case of steganalysis tasks (e.g., Spatial Rich Model).
- In addition to the above common patterns, the feature analysis between different AIGC benchmarks also shows instructive differences.
- The discriminative features of BGAN tend to appear in lower frequency bands that are more perceptible to humans, meaning a significant weakness in generative quality than other AIGC methods; it is true when observing such images.
- The discriminant features of SD 1.4, SD 1.5, and Wukong are similar in network topology and feature frequency. This is intuitive, as they are all based on the same SD model – SD 1.5 is the result of increasing training rounds from SD 1.4; Wukong is a version of the Chinese prompts. We will analyze this phenomenon further in subsequent experiments.
- The discriminative features are more different at the global level than at the local level. This explains why AIGC forensics based on classical frequency transforms (e.g., F3Net) have poor generalization — the intra-class variation of such global features is greater for heterogeneous data scenarios, and the forensic algorithm has difficulty adapting to this variation.

As shown in Table VIII, we perform a model attribution analysis of the focused comparison methods on the hybrid benchmark, with the multi-classification F1 scores.

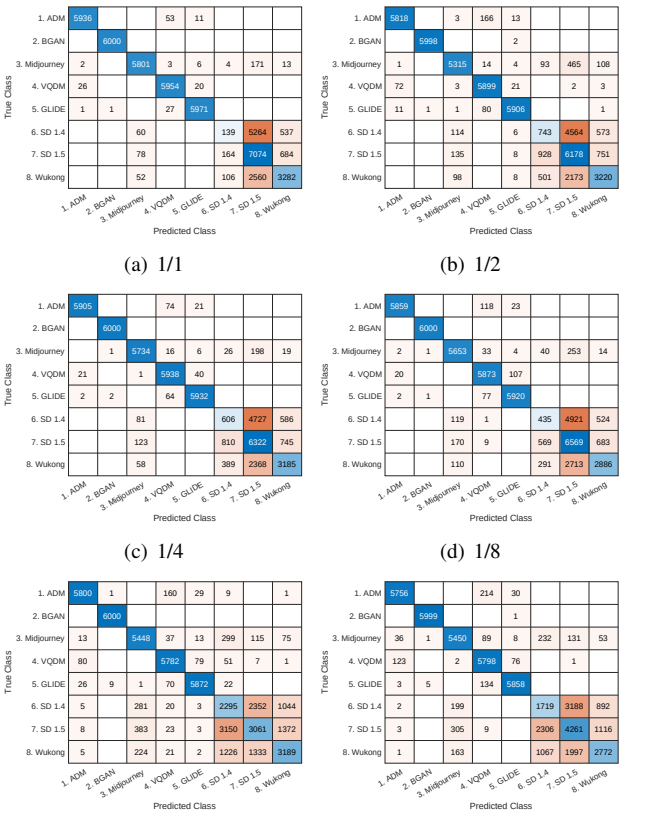
Note that the model attribution analysis aims to further reveal whether the learned features reflect the artifact patterns of interest in forensic problems. In other words, we need to ensure that the forensic captures the inherent features of the generative model, rather than simply memorizing the semantic distribution in the training set; even a score close to 100% could be misleading. Here, the model attribution analysis is performed on the hybrid benchmark of 8 AIGC methods, excluding real images. Note that SD 1.4, SD 1.5, and Wukong are in fact all based on the exact same diffusion model, differing only in the training rounds or prompt languages. Therefore, images generated by SD 1.4, SD 1.5, and Wukong should be classified as the same SD model from the perspective of model attribution analysis. Let us analyze the performance of each approach.

- For the handcrafted representations, they still exhibit the sensitivity to classifier choices. In particular, ScatterNet has a significant decrease in both average and minimum scores compared to the laboratory scenario. This phenomenon indicates the limitation of separability between classes in their feature spaces.
- For the learning representations, their overall scores are very high – SimpleNet's score has increased significantly compared to the laboratory scenario, and DenseNet's minimum score is also as high as 99%. Does such scores mean that learning representations have been perfected in model attribution analysis? The answer is no. In fact, SimpleNet and DenseNet are highly overfit. Overfitting is likely to occur when the model capacity is significantly larger than the task or when the training data is insufficiently diverse. We will confirm this proposition in the subsequent visualization experiments.
- In this scenario, our BGI achieves high scores (average and minimum) when using the SVM classifier, indicating a good discriminability in multi-class attribution problems. When using the NN classifier, the performance decreases but is still better than the most relevant ScatterNet.

As shown in Fig. 10, we directly visualize the confusion matrices of our BGI for model attribution analysis with a variety of training/testing ratios.

Note that BGI exhibits reasonable attribution behavior at the 1/1 ratio. For ADM, BGAN, GLIDE, Midjourney, and VQDM, which are based on different model structures, our BGI show good discriminability with few misclassified cases. As for SD 1.4, SD 1.5, and Wukong, which are based on the same model structure, our BGI shows the expected confusion. Here, the images from SD 1.4 were largely attributed to SD 1.5 – SD 1.5 is a variation of SD 1.4 with more training rounds, may covering the generation patterns of SD 1.4; many images from Wukong were attributed to SD 1.5, with only difference on types of language for prompts.

As the ratio decreases, the performance of BGI remains stable, inheriting the reasonable attribution behavior at the 1/1 ratio. It is true even in the extreme case of 1/32 ratio, where the number of examples for each AIGC method is < 200. This phenomenon demonstrates that the symmetry



(a) 1/1 (b) 1/2

(c) 1/4 (d) 1/8

(e) 1/16 (f) 1/32

Fig. 10. Visualization for confusion matrix (generator labels) regarding fine-grained model attribution of our representation, with comprehensive ratios regarding number of examples for training v.s. testing.

principle is very valuable in AIGC forensics – our BGI with geometric invariance avoids overfitting at the 1/1 ratio and also avoids underfitting at the 1/32 ratio, exhibiting superiority in its structure.

As shown in Fig. 11, we then directly visualize the confusion matrices of learning representations for model attribution analysis with a variety of training epochs and ratios. Here, the main purpose is to further analyze the nature of learning representations in model attribution analysis (Table VIII), as a comparison baseline for our work (Fig. 10). The protocol consisted of 20 epochs with 1/1 ratio to visualize potential overfitting behavior and 10 epochs with 1/1, 1/4, and 1/8 ratios to visualize potential underfitting behavior. Let us analyze the performance of each approach.

- The direct-learned SimpleNet exhibits strong overfitting behavior at 20 epochs, even distinguishing between SD 1.4, SD 1.5, and Wukong, which have the same structure. An important reason for this phenomenon is that the training examples are sufficient but lack intra-class diversity, while the representations are not designed with regularization of task priors. In fact, such overfitting phenomenon is very common in AIGC forensics. Whereas, when at 10 epochs with 1/1, 1/4, and 1/8 ratios, SimpleNet exhibits increasingly underfitting behavior – it is unable to distinguish well between ADM, BGAN, GLIDE, Midjourney, and VQDM, which are based on different structures. This phenomenon proves its poor efficiency.
- The transfer-learned ResNet and InceptionNet are slightly better than SimpleNet in overfitting and underfitting. An important reason for this phenomenon is that the pre-training on ImageNet with sufficient diversity reduces the risk of overfitting, while the larger capacity of the model reduces the risk of underfitting.
- The transfer-learned DenseNet is slightly better than ResNet and InceptionNet in underfitting, but is more severe in overfitting, exhibiting complete overfitting behavior at 20 epochs.
- As can be seen, the attribution performance of the above learned representations is significantly weaker than that of our work – it has trouble achieving a reasonable balance between overfitting and underfitting, relying on complex and empirical training rules. This dilemma further illustrates the necessity of introducing priors and invariants, where empirical learning cannot naturally fit realistic AIGC forensic problems.

V. Conclusion

Forensic representations still lack a general strategy for achieving robustness, interpretability, and discriminability. They rely on 1) representation learning with empirical risk minimization, leading to inherent weaknesses in the robustness and interpretability requirements beyond experience; or 2) handcrafted representations of robust and interpretable designs, but also along with very limited discriminability.

In this paper, we attempt to define a general strategy of image representations, satisfying robustness, interpretability, and discriminability principles in today's forensic tasks of AI-generated images.

The *representation* ingredients of our work are as follows.

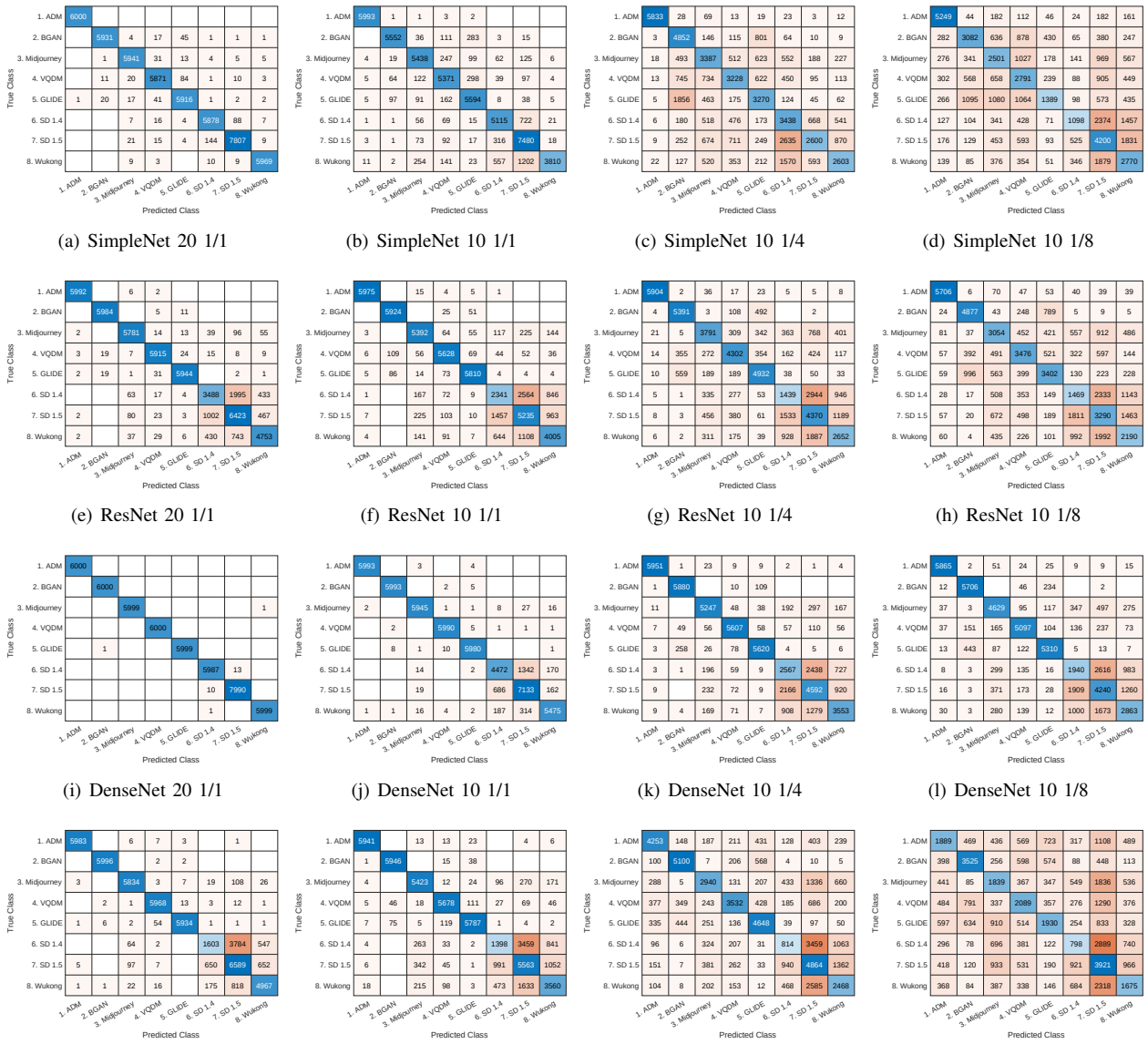
- We have redesigned typical CNN modules from the perspective of geometric invariance theory rather than empirical learning. Here, the symmetry principle for translation, rotation, flipping, and scaling holds across all layers (Section III-A).
- We have boosted the geometric invariance theory to a similar discriminative power of learned CNNs. Here, a high degree of overcompleteness is achieved by the deep cascading representations (Section III-B).

The *forensic* ingredients of our work are as follows.

- We have achieved state-of-the-art forensic scores from generative adversarial networks to diffusion models, even with large-scale real/fake content diversity like ImageNet (Section IV-B).
- We have explored some better insights over typical forensics of empirical learning, such as high efficiency (Sections IV-C and IV-D), intrinsic invariance (Sections IV-C and IV-D), and better interpretability (Sections III-D and Section IV-E).

VI. Acknowledgment

This work was supported in part by the Young Talent Support Project (Hong Kong and Macau) of Guangzhou As-



(a) SimpleNet 20 1/1 (b) SimpleNet 10 1/1 (c) SimpleNet 10 1/4 (d) SimpleNet 10 1/8

(e) ResNet 20 1/1 (f) ResNet 10 1/1 (g) ResNet 10 1/4 (h) ResNet 10 1/8

(i) DenseNet 20 1/1 (j) DenseNet 10 1/1 (k) DenseNet 10 1/4 (l) DenseNet 10 1/8

(m) InceptionNet 20 1/1 (n) InceptionNet 10 1/1 (o) InceptionNet 10 1/4 (p) InceptionNet 10 1/8

Fig. 11. Visualization for confusion matrix (generator labels) regarding fine-grained model attribution of comprehensive representation backbones, with different training epochs and ratios.

sociation for Science and Technology under Grant QT-2025-047, the Lagrange Mathematics and Computing Research Center of Huawei Technologies France, the Startup Fund of the City University of Hong Kong, the Ganpo Talent Program of Jiangxi Province under Grant gpyc20240012, the Outstanding Youth Fund Program of Jiangxi Province under Grant 20252BAC220008, and the National Natural Science Foundation of China under Grant 62522112.

REFERENCES

[1] Z. Epstein, A. Hertzmann, I. of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach *et al.*, "Art and the science of generative ai," *Science*, vol. 380, no. 6650, pp. 1110–1111, 2023.

[2] S. J. Nightingale and H. Farid, "Ai-synthesized faces are indistinguishable from real faces and more trustworthy," *Proc. Natl. Acad. Sci.*, vol. 119, no. 8, p. e2120481119, 2022.

[3] M. Boháček and H. Farid, "Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms," *Proc. Natl. Acad. Sci.*, vol. 119, no. 48, p. e2216035119, 2022.

[4] K. Schwarz, Y. Liao, and A. Geiger, "On the frequency bias of generative models," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 18 126–18 136, 2021.

[5] H. Liu, M. Chaudhary, and H. Wang, "Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives," *arXiv preprint arXiv:2307.16851*, 2023.

[6] S. Qi, Y. Zhang, C. Wang, J. Zhou, and X. Cao, "A principled design of image representation: Towards forensic tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5337–5354, 2022.

[7] D. Tariang, R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Synthetic image verification in the era of generative artificial intelligence: What works and what isn't there yet," *IEEE Secur. Priv.*, vol. 22, no. 03, pp. 37–49, may 2024.

[8] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020.

[9] D. Mumford, J. Fogarty, and F. Kirwan, *Geometric invariant theory*. Springer Science & Business Media, 1994, vol. 34.

[10] J. Flusser, B. Zitova, and T. Suk, *Moments and moment invariants in pattern recognition*. John Wiley & Sons, 2009.

[11] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[12] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[13] S. Qi, Y. Zhang, C. Wang, J. Zhou, and X. Cao, "A survey of orthogonal moments for image representation: Theory, implementation, and evaluation," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–35, 2021.

[14] V. Balntas, K. Lenc, A. Vedaldi, T. Tuytelaars, J. Matas, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2825–2841, 2020.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[16] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, 2009.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[19] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern Recognit.*, vol. 15, no. 6, pp. 455–469, 1982.

[20] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of image degradation and degradation removal to cnn-based image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1239–1253, 2019.

[21] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2990–2999.

[22] T. S. Cohen and M. Welling, "Steerable CNNs," in *Proc. Int. Conf. Learn. Representations*, 2016.

[23] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5028–5037.

[24] M. Weiler, F. A. Hamprecht, and M. Storath, "Learning steerable filters for rotation equivariant CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 849–858.

[25] I. Sosnovik, M. Szmaja, and A. Smeulders, "Scale-equivariant steerable networks," in *Proc. Int. Conf. Learn. Representations*, 2019.

[26] D. Worrall and M. Welling, "Deep scale-spaces: Equivariance over scale," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[27] Z. Sun and T. Blu, "Empowering networks with scale and rotation equivariance using a similarity convolution," in *Proc. Int. Conf. Learn. Representations*, 2022.

[28] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018.

[29] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[30] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[31] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, "Multimodal image synthesis and editing: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[32] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*. Morgan kaufmann, 2007.

[33] L. Du, A. T. Ho, and R. Cong, "Perceptual hashing for image authentication: A survey," *Signal Process.-Image Commun.*, vol. 81, p. 115713, 2020.

[34] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," *IEEE Trans. Inf. Forensic Secur.*, vol. 15, pp. 144–159, 2019.

[35] F. Matern, C. Riess, and M. Stamminger, "Gradient-based illumination description for image forgery detection," *IEEE Trans. Inf. Forensic Secur.*, vol. 15, pp. 1303–1317, 2019.

[36] Y. Li, Y. He, C. Chen, L. Dong, B. Li, J. Zhou, and X. Li, "Image copy-move forgery detection via deep patchmatch and pairwise ranking learning," *IEEE Trans. Image Process.*, vol. 34, pp. 425–440, 2025.

[37] T. V. Hoang and S. Tabbone, "Fast generic polar harmonic transforms," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2961–2971, 2014.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[39] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2017.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[45] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenet-v2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. Int. Conf. Learn. Representations*, 2021.

[47] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, 2013.

[48] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8695–8704.

[49] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *Proc. IEEE Int. Workshop Inf. Forensic Secur.* IEEE, 2019, pp. 1–6.

[50] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 86–103.

[51] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8060–8069.

[52] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proc. IEEE Int. Conf. Comput. Vis*, 2023, pp. 22 445–22 455.

[53] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 24 480–24 489.

[54] M. Zhu, H. Chen, M. Huang, W. Li, H. Hu, J. Hu, and Y. Wang, "Gendet: Towards good generalizations for ai-generated image detection," *arXiv preprint arXiv:2312.08880*, 2023.

[55] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 8780–8794, 2021.

[56] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[57] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[58] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 684–10 695.

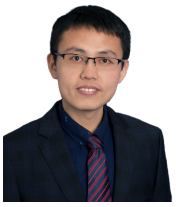
[59] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 696–10 706.

[60] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, "Algorithms to estimate shapley value feature attributions," *Nature Mach. Intell.*, vol. 5, no. 6, pp. 590–601, 2023.

Shuren Qi is currently a Postdoctoral Fellow with Department of Mathematics, The Chinese University of Hong Kong (CUHK). His research interests cover computer vision, deep learning, and artificial intelligence, with a focus on Geometric Deep Learning. He is also interested in the applications for Trustworthy AI and Science AI. His research has appeared in several top-tier journals and conferences, such as *ACM Computing Surveys*, *IEEE TPAMI*, *IEEE TIP*, *ICCV*, and *USENIX Security*. His works offer new designs of invariant representations — from global to local and hierarchical assumptions.

Chao Wang received the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2025. She has authored or coauthored papers in top-tier venues such as *IEEE Transactions on Information Forensics and Security* and *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Her research interests include trustworthy artificial intelligence, adversarial learning, and media forensics.

Zhiqiu Huang received the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1999. He is a Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. He has authored or coauthored more than 80 journal and conference papers. His research interests include software engineering, formal methods, and knowledge engineering.



Yushu Zhang received the Ph.D. degree from the Chongqing University, Chongqing, China, in 2014. He held various research positions with the City University of Hong Kong, Southwest University, University of Macau, Deakin University, and Nanjing University of Aeronautics and Astronautics. He is a Professor with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include privacy, security, and trustworthy AI. Dr. Zhang is an Associate Editor of *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Network and Service Management*, *Signal Processing*, and *Information Sciences*.

Xiangyu Chen is currently a Research Scientist at the Institute of Artificial Intelligence (TeleAI), China Telecom. He received his Ph.D. at University of Macau in 2025. During the doctoral study, He worked as a research intern in Shanghai Artificial Intelligence Laboratory. He received his B.E. degree and M.E. degree from Northwestern Polytechnical University, Xi'an, in 2017 and 2020. He worked as research assistant in Multimedia Laboratory, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences from 2019 to 2021. His research interests includes general low-level vision and large multimodal model.

Yi Zhang received the B.S., M.S., and Ph.D. degrees in computer science and technology from the College of Computer Science, Sichuan University, Chengdu, China, in 2005, 2008, and 2012, respectively. From 2014 to 2015, he was with the Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA, as a Postdoctoral Researcher. He is currently a Full Professor with the School of Cyber Science and Engineering, Sichuan University, and is also the Director with Deep Imaging Group (DIG). His research interests include medical imaging, compressive sensing, and deep learning. He authored more than 140 papers in the field of image processing. These papers were authored or coauthored in several leading journals and conferences, including IEEE TMI, IEEE TIFS, MedIA, IJCV, and CVPR, and reported by the Institute of Physics (IOP) and during the Lindau Nobel Laureate Meeting. He was the recipient of the major funding from the National Key R&D Program of China, National Natural Science Foundation of China, and Science and Technology Support Project of Sichuan Province, China. He is also a Guest Editor of the International Journal of Biomedical Imaging, Sensing and Imaging, and an Associate Editor for IEEE TMI and IEEE TRPMS.

Tieyong Zeng is a Professor at the Department of Mathematics, The Chinese University of Hong Kong (CUHK). Together with colleagues, he has founded the Center for Mathematical Artificial Intelligence (CMAI) since 2020 and served as the director of CMAI. He received the B.S. degree from the Peking University, Beijing, China, the M.S. degree from the Ecole Polytechnique, Palaiseau, France, and the Ph.D. degree from the University of Paris XIII, Paris, France, in 2000, 2004, and 2007, respectively. His research interests include image processing, optimization, artificial intelligence, scientific computing, computer vision, machine learning, and inverse problems. He has published around 100 papers in the prestigious journals such as *SIAM Journal on Imaging Sciences*, *SIAM Journal on Scientific Computing*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *International Journal of Computer Vision*. He is laureate of the 2021 Hong Kong Mathematical Society (HKMS) Young Scholars Award.

Fenglei Fan received the B.S. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 2017, and the Ph.D. degree from Rensselaer Polytechnic Institute (RPI), Troy, NY, USA, in 2021. He is currently an Assistant Professor with the Department of Data Science at the City University of Hong Kong (CityU), and is also the Director with Frontier of Artificial Network (FAN) Group. He was a Research Assistant Professor with the Department of Mathematics, The Chinese University of Hong Kong (CUHK). Before that, he spent one year as a postdoc researcher in Weill Cornell Medicine. He has authored 26 papers in top-tier venues such as JMLR and TPAMI. His primary research interests lie in deep learning theory and methodology, neuroscience, and medical image processing. Dr. Fan served as a PC Member in many conferences such as the *International Joint Conference on Artificial Intelligence* and the *Association for the Advancement of Artificial Intelligence*. He was a recipient of OlympusMons Pioneer Award, CVPR Best Paper Award Candidates, IEEE TRPMS Best Paper Award from the IEEE Nuclear and Plasma Society, and International Neural Network Society Doctoral Dissertation Award.