

EIB-Net: Entropy-Guided Information Bottleneck for Generalizable AI-Generated Image Detection

Zhida Zhang

School of Artificial Intelligence,
UCAS & NLPR, CASIA
Beijing, China
zhida.zhang@cripac.ia.ac.cn

Xinlei Ma

School of Advanced Interdisciplinary Sciences,
UCAS & NLPR, CASIA
Beijing, China
xinlei.ma@nlpr.ia.ac.cn

Jie Cao *

NLPR, Institute of Automation,
Chinese Academy of Sciences
Beijing, China
jie.cao@cripac.ia.ac.cn

Abstract—The proliferation of photorealistic AI-generated images demands robust detection methods that generalize across diverse generative models. While existing approaches target manipulation-based forgeries with local artifacts, generation-based images (e.g., from diffusion models) lack such traces, posing a fundamental challenge. We observe that generative models prioritize global semantics at the expense of local texture fidelity, making low-texture regions key indicators of synthetic origin. To exploit this, we propose EIB-Net, an Entropy-guided Information Bottleneck Network. EIB-Net introduces a novel Image Entropy (IE) metric to automatically select the most informative (lowest-entropy) patch, then processes it with a Variational Information Bottleneck (VIB) to learn compact, generalizable features. Extensive experiments on DIFF, DiffusionForensics, and GenImage benchmarks demonstrate state-of-the-art performance: EIB-Net achieves 85.7% accuracy using only 2% of training data, outperforming full-image baselines by over 15%, and maintains robust cross-generator generalization (83.5% average accuracy on GenImage). Furthermore, our entropy-guided patch selection (EGPL) consistently enhances diverse backbones (CNNs and Transformers), proving its practical value for data-efficient detection.

Index Terms—AI-generated Image Detection, Patch-based Analysis, Information Bottleneck, Generalization

I. INTRODUCTION

In recent years, the rapid advancement of generative models, notably Generative Adversarial Networks (GANs) [1] and the latest diffusion models (e.g., Stable Diffusion [2], Midjourney [3]), has pushed synthetic image realism to unprecedented levels. These text-to-image systems now produce content often indistinguishable from real photographs. While beneficial for creative applications, this capability simultaneously raises severe concerns regarding the misuse of synthetic media for disinformation, underscoring the urgent need for robust image authenticity verification techniques.

Most detection frameworks were originally developed for manipulation-based forgeries (e.g., deepfakes [4]), which typically leave clear, localized artifacts like boundary inconsistencies. However, detecting pure **generation-based images** is fundamentally more challenging. Unlike tampered images, generated images lack explicit traces; their subtle statistical

flaws are often distributed across the image, making traditional localization-based methods ineffective.

We identify the root cause of this failure: a mismatch in focus between generation and detection. Generative models are optimized to produce globally coherent semantics, often at the expense of local texture fidelity. In particular, **low-texture regions** (e.g., clear skies, smooth walls) receive weak supervision during training, leading to unnatural smoothness and a lack of characteristic high-frequency noise. This observation leads to our core insight: **the smoother, the faker**. Ironically, the most revealing artifacts of synthetic origin are found not in complex textures but in the simplest, most homogeneous areas.

This counter-intuitive insight is supported by the Gestalt principle of perception [5], which states that both humans and neural networks are biased toward global structure, often overlooking local details. By shifting the detection focus from global semantics to local statistical anomalies, we can build more robust and generalizable detectors.

To operationalize this insight, we propose the **Entropy-guided Information Bottleneck Network (EIB-Net)**. EIB-Net implements a “**select-then-compress**” strategy through two novel components:

- 1) **Entropy-Guided Patch Learning (EGPL)**: We design a novel Image Entropy metric to quantify local smoothness, automatically selecting the single most suspicious (lowest-entropy) patch as input.
- 2) **Variational Information Bottleneck (VIB)**: This module compresses the patch representation, discarding irrelevant semantic content and forcing the model to learn compact, authenticity-relevant features.

By focusing exclusively on the most artifact-prone region and suppressing content redundancy, EIB-Net learns to detect based on intrinsic texture fidelity rather than dataset-specific semantics.

Extensive experiments on DIFF, DiffusionForensics, and GenImage benchmarks validate EIB-Net’s superiority. It achieves state-of-the-art accuracy in both intra- and cross-generator settings while exhibiting exceptional data efficiency—maintaining over 85% accuracy with only 2% of training data. Furthermore, the EGPL component proves

* is the corresponding author. This work was supported by the Beijing Natural Science Foundation (Grant No. L257008 & L252145) and National Natural Science Foundation of China (Grant No. 62576338 & 62550062 & 32341009).

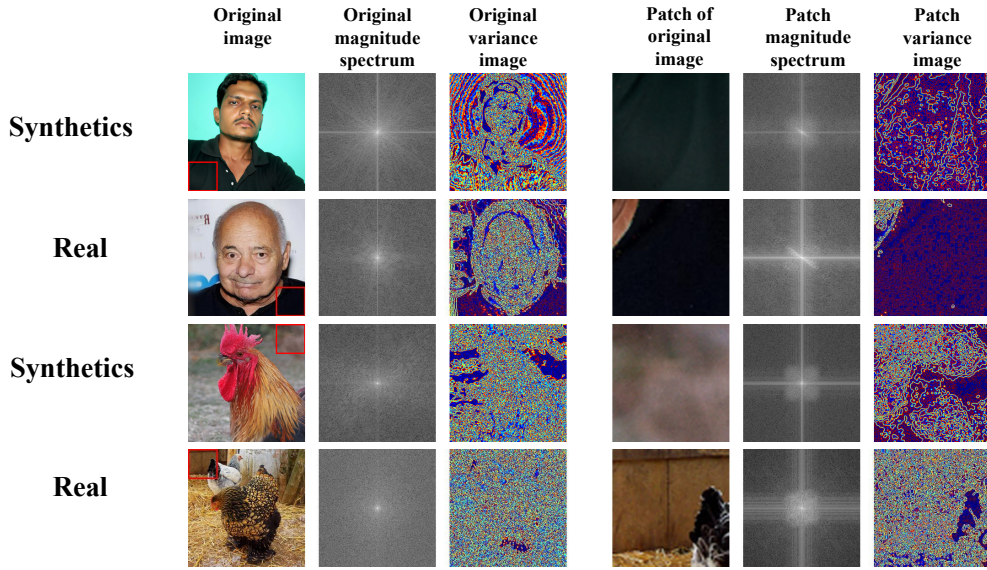


Fig. 1: Visual comparison of real and AI-generated images. The red box marks the lowest-entropy patch. The magnitude spectrum (middle) and variance image (right) show that real images contain richer high-frequency components and more complex local variance patterns.

model-agnostic, consistently enhancing diverse backbone architectures (CNNs and Transformers).

Our main contributions are summarized as follows:

- We propose **EIB-Net**, a novel patch-based detection framework combining **EGPL** and **VIB** for **robust, generalizable, and data-efficient** AI-generated image detection.
- We introduce a new **Image Entropy (IE)** metric to guide patch selection toward smooth, artifact-prone regions, which significantly enhances convergence speed and generalization, especially under low-data regimes.
- We demonstrate that the **EGPL** selection strategy is broadly applicable across diverse CNN and Transformer architectures, and that the full **EIB-Net** consistently achieves superior performance on challenging intra- and cross-generator benchmarks.

II. RELATED WORK

A. Generated Image Detection

Early research on AI-generated image detection primarily focused on manipulation-based forgeries (e.g., deepfakes), typically employing specialized CNNs and attention mechanisms [6]–[9]. However, these methods, which often rely on boundary or lighting artifacts, prove less effective against the highly photorealistic images generated by modern Latent Diffusion Models (LDMs), which lack explicit tampering traces.

To address the LDM challenge, several model-specific approaches emerged. Wang et al. [10] introduced Diffusion Reconstruction Error (DIRE), which measures discrepancies between an image and its reconstruction by a pre-trained diffusion model. Methods like AEROBLADE [11] and LA-TENTTRACER [12] perform detection by analyzing LDM’s autoencoder representations without additional training. Other

work leverages intrinsic statistical or structural flaws: Tan et al. [13] propose Neighboring Pixel Relationships (NPR) to capture upsampling artifacts, while more recent generalizable approaches like **GenDet** [14] learn discriminative features from full images, which may still inadvertently encode semantic biases.

However, most approaches still rely on **model-specific reconstruction** (e.g., DIRE), target particular **geometric inconsistencies** [15], or over-rely on **global semantic priors**. In stark contrast, our work is distinguished by its focus on **low-texture, low-entropy regions**. This deliberate departure from the established practice in camera-based forensic tasks (e.g., camera model identification) which *high-entropy* (texture-rich) patches are typically selected to capture device-specific noise patterns. We posit that for *synthesis* artifacts, the weakness of generative models lies in the *under-supervised*, smooth areas—a novel hypothesis that guides our entropy-based patch selection.

B. Mutual Information and Information Bottleneck

The **Information Bottleneck (IB) principle** [16] seeks to compress input representations X into a compact latent variable Z while preserving task-relevant information about the target variable Y . This principle provides an effective means for enhancing generalization and robustness. Variational formulations of IB, such as the Variational Information Bottleneck (VIB) proposed in [17], [18], adapt the IB principle to deep neural networks. These works reveal profound connections between IB, disentangled representation learning [19], and rate-distortion theory.

Recent research extends IB’s utility across domains including reinforcement learning, adversarial robustness, and generative modeling [20]–[22]. In our work, the VIB module

is not merely an add-on for regularization; it is integral to our strategy for overcoming semantic overfitting. By enforcing a compressed latent representation, the VIB actively **discards content-related information** from the selected image patch, compelling the network to retain only the minimal, task-relevant statistical cues for authenticity detection. This aligns with the recent direction of explicitly learning content-agnostic representations for media forensic tasks [23], and together with our low-entropy patch selection, forms a dual strategy operating at both the data and feature levels to achieve robust generalization.

III. METHOD

A. Overview of EIB-Net

Our proposed **EIB-Net** is designed to detect AI-generated images by directly countering the prevalent issue of *semantic overfitting*. Instead of analyzing the full image—which often leads models to exploit dataset-specific content biases—EIB-Net forces the detector to base its decision on subtle, local statistical anomalies. As illustrated in fig. 2, this is achieved through a synergistic two-stage strategy:

- **Entropy-Guided Patch Learning (EGPL)**: This front-end module implements our “low-entropy hypothesis”. It computes a novel Image Entropy metric to automatically identify and select the single, smoothest image patch where generative artifacts are most likely to be exposed.
- **Feature Learning with Variational Bottleneck**: The selected patch is processed by a feature encoder (ResNet) followed by a **Variational Information Bottleneck (VIB)** layer. The VIB actively compresses the feature representation, discarding content-related information and preserving only the minimal statistical cues essential for authenticity classification.

This “**select-then-compress**” paradigm ensures the model learns from artifact-prone regions while being invariant to distracting global semantics, thereby achieving robust generalization.

B. Entropy-Guided Patch Learning (EGPL)

1) *Image Entropy Computation*: To locate the most vulnerable regions, we introduce an **Image Entropy (IE)** metric that quantifies local texture complexity across color channels. Given an RGB image $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$ where $C = 3$, we first divide it into N non-overlapping patches $\{P_1, P_2, \dots, P_N\}$.

The IE for each patch is computed by aggregating multi-directional intensity gradients within each color channel. We employ two Laplacian-style convolutional kernels:

$$K_{\text{neighbor}} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \quad K_{\text{diagonal}} = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 4 & 0 \\ -1 & 0 & -1 \end{bmatrix}. \quad (1)$$

These kernels highlight intensity variations in adjacent and diagonal directions, respectively. The convolution is applied **independently to each of the three RGB channels**. For the

entire image, the total variation map $\mathbf{D}_{\text{total}} \in \mathbb{R}^{C \times H \times W}$ is obtained by a weighted sum of the absolute responses:

$$\mathbf{D}_{\text{total}} = |\mathbf{I} * K_{\text{neighbor}}| + \frac{1}{\sqrt{2}} |\mathbf{I} * K_{\text{diagonal}}|, \quad (2)$$

where $*$ denotes the 2D convolution performed per channel.

The scalar entropy value $\text{IE}(P_i)$ for a patch P_i is then defined as the sum of $\mathbf{D}_{\text{total}}$ over **all three color channels and all spatial locations** within that patch:

$$\text{IE}(P_i) = \sum_{c=1}^C \sum_{(h,w) \in P_i} \mathbf{D}_{\text{total}}^{(c,h,w)}. \quad (3)$$

A lower IE value indicates a smoother, less textured region across color channels.

2) *Patch Selection*: Following our hypothesis that low-entropy areas are weak points for generators, the EGPL module selects the patch with the minimum entropy as the sole input to the subsequent detector:

$$P_{\min} = \arg \min_{P_i} \text{IE}(P_i). \quad (4)$$

This operation focuses the model’s attention precisely on the region where generative supervision is likely weakest and statistical anomalies are most pronounced

C. Network Architecture with VIB

The selected patch $P_{\min}$ is resized to a standard resolution and fed into a ResNet-50 backbone (with the final fully-connected layer removed) to produce a feature vector $\mathbf{x} \in \mathbb{R}^d$.

To further purify this feature by stripping away residual semantic information, we employ a **Variational Information Bottleneck (VIB)** layer. The VIB models the distribution of a latent code $\mathbf{z}$ given $\mathbf{x}$. It first projects $\mathbf{x}$ into the parameters of a Gaussian distribution:

$$\boldsymbol{\mu} = \mathbf{W}_{\mu} \mathbf{x} + \mathbf{b}_{\mu}, \quad (5)$$

$$\log \sigma^2 = \mathbf{W}_{\sigma} \mathbf{x} + \mathbf{b}_{\sigma} - 5. \quad (6)$$

Here, the constant offset -5 initializes σ near zero, ensuring stable training in the early phases. The latent code $\mathbf{z}$ is then sampled using the reparameterization trick: $\mathbf{z} = \boldsymbol{\mu} + \sigma \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. This stochastic bottleneck compels the model to form a compact representation $\mathbf{z}$ that is maximally informative about the authenticity label while being minimally informative about the input $\mathbf{x}$, thereby enforcing the information bottleneck principle.

D. Loss Function

The model is optimized end-to-end with a composite loss function $\mathcal{L}$ that balances classification accuracy, prediction consistency, and feature compression:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{LS-CE}} + \alpha \cdot \mathcal{L}_{\text{JSD}}}_{\mathcal{L}_{\text{cls}}} + \beta \cdot \mathcal{L}_{\text{KL}}. \quad (7)$$

Classification Loss with Self-Regularization. To promote generalization and calibrated predictions, the classification

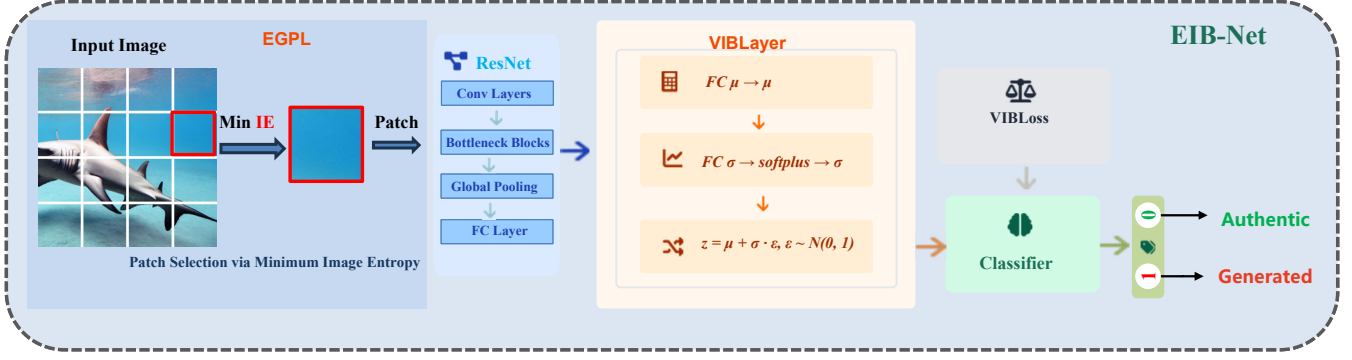


Fig. 2: The pipeline of EIB-Net. Given an input image, the Entropy-Guided Patch Learning (EGPL) module computes the Image Entropy map and selects the patch $P_{\min}$ with the minimum entropy. This patch is encoded by a ResNet backbone, and its features are compressed by the Variational Information Bottleneck (VIB) layer into a latent representation $\mathbf{z}$ for the final real/synthetic classification.

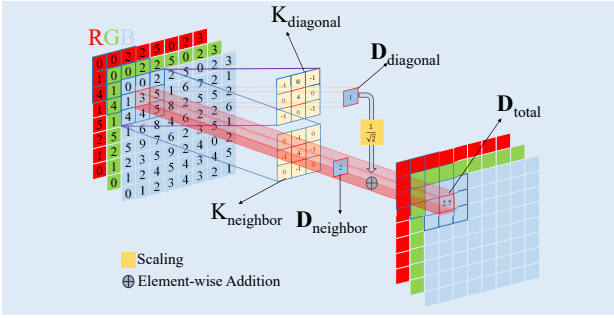


Fig. 3: Illustration of Image Entropy (IE) computation. For each pixel in the patch, differences with neighboring and diagonal pixels are computed using K_{neighbor} and K_{diagonal} . These differences are weighted, summed, and aggregated to form the total Image Entropy D_{total} .

term $\mathcal{L}_{\text{cls}}$ combines a Label Smoothing Cross-Entropy loss ($\mathcal{L}_{\text{LS-CE}}$) with a Jensen-Shannon Divergence (JSD) regularization ($\mathcal{L}_{\text{JSD}}$) weighted by $\alpha = 12$. $\mathcal{L}_{\text{LS-CE}}$ with smoothing factor $\epsilon = 0.1$ prevents over-confidence. $\mathcal{L}_{\text{JSD}}$ encourages consistent output distributions and acts as an additional regularizer. The overall classification loss is:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{LS-CE}} + \alpha \cdot \mathcal{L}_{\text{JSD}}. \quad (8)$$

Information Bottleneck Regularization. The VIB module is regularized by the KL divergence between the learned latent distribution $q(\mathbf{z}|\mathbf{x})$ and a standard Gaussian prior $p(\mathbf{z})$:

$$\mathcal{L}_{\text{KL}} = \frac{1}{N} \sum_{i=1}^N \text{KL} [q(\mathbf{z}_i|\mathbf{x}_i) \| p(\mathbf{z}_i)]. \quad (9)$$

Hyperparameters. The trade-off between task performance and latent compression is controlled by β , set to 10^{-3} in all experiments.

IV. EXPERIMENTS

A. Experimental Setup

We utilize three benchmarks: **DIFF** [24] (faces), **DiffusionForensics** [10] (LSUN/ImageNet), and **GenImage** [25]

TABLE I: Accuracy (%) on DIFF-Intra under varying training data ratios. EIB-Net achieves superior performance across all data regimes, especially under low-data conditions. * indicates reproduced results.

Model	100%	20%	5%	2%
ResNet50*	98.66	90.86	86.69	82.67
RMT-S*	98.81	93.27	84.66	70.66
ViT-B16*	97.02	86.53	71.46	62.09
Xception*	98.85	97.09	76.28	70.12
EIB-Net (Ours)	98.85	97.17	92.85	85.69

(ImageNet-base). We evaluate on two settings: (1) **Intra-model:** Train/Test on the same generator distribution; (2) **Cross-model:** Train on one source (e.g., SDv1.4) and test on unseen generators. We report Accuracy (Acc) and provide full training details in the supplementary material.

B. Intra-Model Generalization

We evaluate in-domain robustness using a curated face-image subset, DIFF-Intra (see supplementary), varying training data ratios from 100

Analysis. The superiority of EIB-Net stems from its targeted learning strategy. While full-image models, especially ViT-B16, overfit to limited semantic cues, EGPL provides a strong *inductive bias*. By focusing on low-entropy, artifact-prone regions, EGPL acts as an efficient data filter, enabling stable and sample-efficient learning from limited supervision.

C. Cross-model generalization

1) **DiffusionForensics:** Following [10], models are trained on LSUN-Bedroom and tested on ImageNet samples from ADM and SDv1, a challenging cross-scene and cross-model setup. We report results at different training ratios to assess data efficiency (table II).

Analysis. EIB-Net maintains near-perfect validation accuracy (>99% with 1% data) and robust test accuracy (>80%

TABLE II: Cross-model generalization on DiffusionForensics. Models are trained on LSUN-Bedroom and tested on ImageNet. Each row shows performance under a specific training ratio. Best results per row are in **bold**, EIB-Net column is shaded.

Train Ratio	Res.50*	RMT-S*	ViT*	Xcep.*	(Ours)
<i>Validation Accuracy (ADM_VAL)</i>					
20%	100.00	100.00	96.10	92.65	100.00
5%	99.45	99.85	92.10	96.45	99.85
1%	91.80	98.75	84.15	90.55	99.00
<i>Test Accuracy (ADM)</i>					
20%	64.73	72.17	60.26	74.54	86.70
5%	71.29	81.20	57.31	76.55	87.50
1%	70.90	72.29	57.67	68.40	85.80
<i>Test Accuracy (SDv1)</i>					
20%	46.18	56.55	34.03	52.97	87.47
5%	43.27	57.67	34.91	52.88	87.87
1%	39.98	53.91	34.70	38.87	81.93

across both ADM and SDv1). In stark contrast, all baselines suffer a catastrophic drop in test performance, despite strong validation results. This confirms our hypothesis that baselines overfit to *global, distribution-specific semantics* in the LSUN training set. EIB-Net, by leveraging *local statistical anomalies* (via EGPL) and compressing semantics (via VIB), learns fundamentally more transferable features.

2) *GenImage*: We evaluate cross-generator generalization on the GenImage benchmark, training all methods exclusively on images synthesized by SD v1.4 and testing on samples from eight different generators.

As shown in table III, **EIB-Net** consistently outperforms both classical and state-of-the-art detectors across most generators. Notably, it achieves near-perfect accuracy on SD v1.4 (**99.89%**) and the best average accuracy (**83.51%**) across unseen generators.

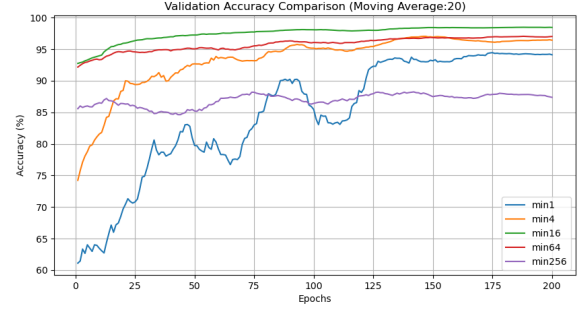
D. Ablation Study

We conduct ablation studies to validate the design of EIB-Net, with a focus on our core innovations: the Entropy-Guided Patch Learning (EGPL) module and the Variational Information Bottleneck (VIB).

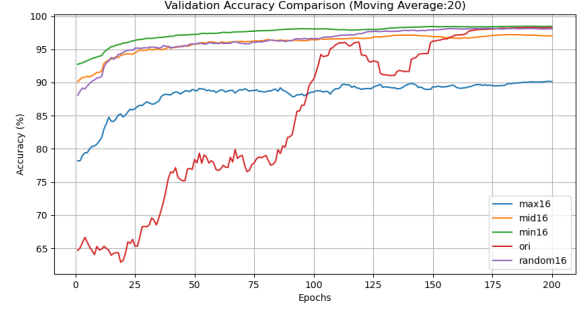
1) Analysis of EGPL Component:

a) *Why Low-Entropy Patches?*: We first ablate the patch selection strategy using a ResNet50 backbone. As shown in fig. 4(b), selecting the **minimum entropy patch (min16)** achieves accuracy comparable to using the full image (ori) but with faster convergence. In contrast, selecting the **maximum entropy patch (max16)** results in the worst performance. This critical comparison provides direct empirical support for our “low-entropy hypothesis”: high-entropy, texture-rich regions act as noisy distractors, while low-entropy, smooth regions are the most reliable indicators of generative artifacts.

b) *How Many Patches?*: We then analyze the granularity of patch splitting. Fig. 4(a) shows that dividing the image into 16 patches (**min16**) offers the optimal trade-off. Too few patches (e.g., **min4**) may contain excessive irrelevant content, while too many (**min64**) may excessively fracture the



(a) Accuracy vs. Patch Number



(b) Accuracy vs. Patch Type

Fig. 4: Validation accuracy of ResNet50 under varying patch configurations. (a) shows performance under different patch quantities, using the lowest-entropy patch from each split. (b) compares different patch selection strategies on a 16-patch split. Accuracy is smoothed with a moving average of 20 epochs.

artifact signal. The single full-image patch (**min1**) suffers from information loss due to rescaling.

c) *Is EGPL Model-Agnostic?*: To demonstrate the generality of our entropy guidance, we integrate EGPL into diverse backbone architectures. Table IV shows that EGPL provides consistent and significant performance gains across CNNs (ResNet50) and Transformers (ViT-B16, RMT-S) on both intra- and cross-model benchmarks. This confirms that the benefit of focusing on low-entropy regions is a general principle, not an architecture-specific trick.

2) Analysis of VIB and Integrated Framework:

a) *Role of the Information Bottleneck.*: The VIB module is designed to compress features and discard semantic noise. Ablating the VIB module (i.e., using only EGPL) leads to a noticeable drop in cross-model generalization performance, validating its role in improving feature robustness.

b) *Synergy of EGPL and VIB.*: The final row of table IV shows that the complete **EIB-Net (EGPL + VIB)** achieves the best overall performance, surpassing EGPL used alone on any single backbone. This demonstrates a clear synergy: **EGPL selects the most informative data (low-entropy patch), and VIB extracts the most informative features from it**, together forming a powerful “select-then-compress” pipeline that is highly efficient and robust.

TABLE III: The cross-generator results on GenImage. All methods are trained only on the SD v1.4 subset. Accuracy (%) is reported. The * symbol indicates the result of reproduction.

Method	Year	Mid	SDv1.4	SDv1.5	adms	glide	wukong	vqdm	BigGAN	Average
ResNet18*	2016	62.33	99.75	99.31	56.00	60.75	96.17	55.67	51.08	72.63
ResNet50*	2016	61.25	99.92	99.81	57.83	69.58	98.67	59.67	53.75	75.06
Swin-T*	2021	62.67	99.92	99.56	61.92	78.25	98.00	62.67	58.08	77.84
Xception*	2017	60.58	99.92	99.62	65.75	84.92	98.58	67.67	58.33	79.42
DIRE	2023	60.20	99.90	99.80	50.90	55.00	99.20	50.10	50.20	70.66
GenDet	2024	89.60	96.10	96.10	58.00	78.40	92.80	66.50	75.00	81.56
PatchCraft	2024	79.00	89.50	89.30	77.30	78.40	89.30	83.70	72.40	82.30
EIB-Net	-	76.92	99.89	99.89	72.17	80.25	99.88	64.25	74.83	83.51

TABLE IV: Impact of EGPL on different backbone architectures across datasets. DIFF(20%) and DIFF(5%) refer to intra-model validation accuracy with limited data. GenImage(Avg) is the average cross-model performance. EGPL improves generalization across all models.

Model	DIFF (20%)	DIFF (5%)	GenImage (Avg)
ResNet50	90.86	88.10	75.06
ResNet50 + EGPL	95.94	92.00	80.62
RMT-S	93.27	84.66	78.55
RMT-S + EGPL	96.39	91.81	80.78
ViT-B16	86.53	71.46	77.13
ViT-B16 + EGPL	92.16	85.92	80.46
EIB-Net (EGPL + VIB)	97.17	92.85	83.51

V. CONCLUSION

This work tackles the critical issue of *semantic overfitting* in AI-generated image detection. We propose a novel framework, **EIB-Net**, which introduces a “**select-then-compress**” strategy for robust and data-efficient detection. It first uses a novel **Image Entropy** metric to select the most artifact-prone (lowest-entropy) image patch, then applies a **Variational Information Bottleneck** to compress features and discard content semantics.

Extensive experiments show that EIB-Net achieves state-of-the-art cross-generator generalization while maintaining exceptional efficiency (e.g., >85% accuracy with only 2% training data). The entropy-guided selection is **model-agnostic**, consistently boosting diverse backbones. Future work will extend this paradigm to video forgery detection and explore multi-patch aggregation for even broader robustness.

REFERENCES

- [1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022, pp. 10684–10695.
- [3] Midjourney, “Midjourney,” <https://www.midjourney.com/home/>, 2022.
- [4] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *ICCV*, 2019, pp. 1–11.
- [5] Wolfgang Köhler, *Gestalt psychology: An introduction to new concepts in modern psychology*, vol. 18, WW Norton & Company, 1970.
- [6] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in *BTAS*, 2019, pp. 1–8.

- [7] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu, “Fakespotter: a simple yet robust baseline for spotting ai-synthesized fake faces,” in *IJCAI*, 2021, pp. 3444–3451.
- [8] Jian Wang, Yunlian Sun, and Jinhui Tang, “Lisiam: Localization invariance siamese network for deepfake detection,” *IEEE TIFS*, vol. 17, pp. 2425–2436, 2022.
- [9] Junxian Duan, Siyu Liu, Yiming Hao, Huaibo Huang, and Ran He, “Dual frequency-guided spatiotemporal feature learning for face forgery detection,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.
- [10] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li, “Dire for diffusion-generated image detection,” in *ICCV*, 2023, pp. 22445–22455.
- [11] Jonas Ricker, Denis Lukovnikov, and Asja Fischer, “Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error,” in *CVPR*, 2024, pp. 9130–9140.
- [12] Zhenting Wang, Vikash Sehwal, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma, “How to trace latent generative model generated images without artificial watermark?,” in *ICML*, 2024.
- [13] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei, “Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection,” in *CVPR*, 2024, pp. 28130–28139.
- [14] Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang, “Gendet: Towards good generalizations for ai-generated image detection,” *arXiv preprint arXiv:2312.08880*, 2023.
- [15] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad, “Shadows don’t lie and lines can’t bend! generative models don’t know projective geometry... for now,” in *CVPR*, 2024, pp. 28140–28149.
- [16] Naftali Tishby, Fernando C Pereira, and William Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.
- [17] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, “Deep variational information bottleneck,” *arXiv preprint arXiv:1612.00410*, 2016.
- [18] Alessandro Achille and Stefano Soatto, “Information dropout: Learning optimal representations through noisy computation,” *TPAMI*, vol. 40, no. 12, pp. 2897–2905, 2018.
- [19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” in *ICLR*, 2017.
- [20] Greg Ver Steeg and Aram Galstyan, “The bottleneck principle: Enhancing interpretability in deep neural networks,” in *AAAI*, 2019, vol. 33, pp. 3656–3663.
- [21] Xinwei Sun, Wenhui Chen, and William Yang Wang, “Information bottleneck learning for out-of-distribution generalization,” in *ICML*, PMLR, 2021, pp. 9940–9950.
- [22] Binghui Wang and Neil Zhenqiang Gong, “Information bottleneck approach to adversarial robustness,” in *ICLR*, 2021.
- [23] Veronika Solopova, Lucas Schmidt, and Dorothea Kolossa, “Extending information bottleneck attribution to video sequences,” 2025.
- [24] Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli, “Diffusion facial forgery detection,” *arXiv preprint arXiv:2401.15859*, 2024.
- [25] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang, “Genimage: A million-scale benchmark for detecting ai-generated image,” *NIPS*, vol. 36, 2024.

Supplementary Material for *EIB-Net: Entropy-Guided Information Bottleneck for Generalizable AI-Generated Image Detection*

I. DATASET DETAILS

A. Overview of Evaluation Datasets

We employ three public benchmarks and construct one customized subset for a comprehensive evaluation:

- **DiffusionForensics** [1]: Contains real images from LSUN-Bedroom [2], ImageNet, and CelebA-HQ [3], alongside synthetic samples generated by pretrained diffusion models.
- **GenImage** [4]: Includes over one million synthetic images generated by eight diffusion methods [5]–[10], with real images from ImageNet. The BigGAN subset is included in our cross-generator evaluation despite its non-diffusion origin.
- **DIFF** [11]: A face-focused dataset with over 500K images synthesized by 13 diffusion models [7], [12]–[21], under four generative conditions. Real faces are sourced from VoxCeleb2 and other celebrity datasets [22]–[24].

B. Construction of the DIFF-Intra Subset

For the controlled intra-model generalization study in Sec. 4.2 of the main paper, we construct a balanced subset from the DIFF dataset, denoted as **DIFF-Intra**. It is constructed by randomly sampling an equal number of real and synthetic face images from the full DIFF dataset, ensuring a balanced and representative subset for intra-domain evaluation. This setup provides a clean and controlled distribution, specifically designed for evaluating in-domain generalization and data efficiency (few-shot learning) capabilities.

C. Data Splits for Cross-Model Evaluation

- **DiffusionForensics** [1]: We follow the standard split. Training/Validation: 80,000 real and 80,000 ADM-generated images from LSUN-Bedroom. Test: 10,000 ImageNet-based images each from ADM and SDv1 generators.
- **GenImage** [4]: We use the official subset. Training is exclusively on the SD v1.4 subset (200k images). Evaluation covers all eight generators (including BigGAN), each with 20k images.

II. COMPLETE EXPERIMENTAL CONFIGURATION

A. Backbone Architectures and Baseline Training

To fairly evaluate the general applicability of our EGPL module, we integrate it into four diverse, off-the-shelf backbone networks with minimal modification. The baseline training configurations for each are as follows:

- **ResNet50** [25]: Initialized with ImageNet-pretrained weights. Trained with SGD (momentum=0.9, initial

lr=1e-2, cosine annealing schedule). Uses JsdCrossEntropy loss (num_splits=1, $\alpha = 12$, label smoothing=0.1).

- **RMT-S** [26]: Uses the official implementation. Configured with AdamW (lr=1e-4, betas=(0.9,0.99), weight decay=1e-4) and exponential LR decay (rate=0.95 every 5 epochs). Uses BCEWithLogitsLoss.
- **ViT-B16** [27]: Modified for binary classification. Trained with SGD (momentum=0.9, lr=1e-3, cosine annealing). Uses JsdCrossEntropy loss ($\alpha = 12$).
- **Xception** [28]: Optimized with AdamW (lr=1e-4) and exponential decay. Uses BCEWithLogitsLoss.

For all baselines, we only replace the final classification layer. This controlled setup ensures that any performance gain from EGPL is attributable to our method rather than extensive backbone-specific tuning.

B. Configuration for EIB-Net (EGPL + VIB)

The full EIB-Net framework uses ResNet50 as its feature encoder backbone, extended with our proposed modules. The specific configuration, which aligns exactly with the provided training code, is as follows:

- **Optimizer**: SGD with momentum (0.9). No explicit weight decay is applied.
- **Learning Rate**: Initialized at 1×10^{-2} (via `args.lr`), scheduled with a Cosine Annealing scheduler where `T_max` equals the total number of epochs.
- **Batch Size**: 64 (via `args.BATCH_SIZE`).
- **Training Epochs**: 200 (via `args.max_epoch`).
- **Data Augmentation**: RandomResizedCrop(224), RandomHorizontalFlip(p=0.5), RandomErasing (p=0.6).
- **Loss Function**: $\mathcal{L} = \mathcal{L}_{\text{JSD}} + \beta \mathcal{L}_{\text{KL}}$, where $\mathcal{L}_{\text{JSD}}$ is the Jensen-Shannon Divergence enhanced cross-entropy with parameters $\alpha = 12$, num_splits = 1, and label smoothing factor $\epsilon = 0.1$. The VIB regularization weight is $\beta = 10^{-3}$.
- **VIB Latent Dimension**: 256.
- **Reproducibility**: A fixed random seed (42) is used for all experiments.

III. ALGORITHMIC AND IMPLEMENTATION DETAILS

A. Training Pipeline

B. Implementation of Image Entropy

The Image Entropy (IE) metric is computed efficiently on GPU to enable batch processing. Given an input batch of RGB images $\mathbf{I} \in \mathbb{R}^{B \times H \times W \times 3}$, the core steps are as follows:

Convolution with Directional Kernels. We employ two predefined 3×3 convolutional kernels, K_{neighbor} and K_{diagonal} , which are mathematical Laplacian filters designed to highlight intensity variations in adjacent and diagonal directions,

Algorithm 1: Training Pipeline of EIB-Net

Input: Training image set $\mathcal{D}$, number of patches N , backbone network f_θ , VIB layer g_ϕ , classifier h_ψ

Output: Optimized parameters $\Theta = \{\theta, \phi, \psi\}$

```

1 Initialize model parameters  $\Theta$ ;
2 foreach  $epoch = 1 \rightarrow \text{MaxEpochs}$  do
3   foreach  $\text{mini-batch } \{\mathbf{I}_b, y_b\} \subset \mathcal{D}$  do
4     Step 1: Entropy-Guided Patch Selection (EGPL) ;
5     foreach  $\text{image } \mathbf{I}$  in  $\mathbf{I}_b$  do
6       Divide  $\mathbf{I}$  into  $N$  non-overlapping patches  $\{P_1, \dots, P_N\}$ ;
7       foreach  $\text{patch } P_i$  do
8         Compute per-channel variation via Eq. (1) in main paper;
9         Aggregate to get Image Entropy  $\text{IE}(P_i)$  via Eq. (2);
10      Select patch  $P_{\min} = \arg \min_i \text{IE}(P_i)$ ;
11      Resize  $P_{\min}$  to standard input size;
12      Form selected-patch batch  $\mathbf{P}_{\text{batch}}$ ;
13      Step 2: Feature Extraction & Compression ;
14       $\mathbf{x} \leftarrow f_\theta(\mathbf{P}_{\text{batch}})$ ; // Backbone feature extraction
15       $\boldsymbol{\mu}, \boldsymbol{\sigma} \leftarrow g_\phi(\mathbf{x})$ ; // VIB: Eq. (3,4) in main paper
16       $\mathbf{z} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I})$ ; // Reparameterization
17       $\hat{\mathbf{y}} \leftarrow h_\psi(\mathbf{z})$ ; // Classification
18      Step 3: Optimization ;
19      Compute  $\mathcal{L}_{\text{cls}}$  (JSD loss) and  $\mathcal{L}_{\text{KL}}$  (KL divergence);
20       $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{cls}} + \beta \cdot \mathcal{L}_{\text{KL}}$ ; // Eq. (5) in main paper
21      Update  $\Theta$  by descending  $\nabla_{\Theta} \mathcal{L}_{\text{total}}$ ;

```

respectively (their explicit forms are given in Sec. 3.2 of the main paper). To process the RGB channels correctly, each kernel is repeated across the channel dimension (‘repeat(3, 1, 1, 1)’). The 2D convolution is then applied **independently to each color channel** using PyTorch’s `conv2d` function with the argument `groups=3`. This is equivalent to having three separate convolutional filters for the R, G, and B channels, ensuring that color information is processed independently without cross-channel mixing.

Aggregation into Scalar Entropy. The absolute responses of the two convolutions are summed to obtain a per-pixel, per-channel variation map $\mathbf{D}_{\text{total}}$, following Eq. (2) in the main paper. The final scalar entropy value $\text{IE}(P)$ for a patch P is computed by summing $\mathbf{D}_{\text{total}}$ over all spatial locations (h, w) and all three color channels c within that patch: $\text{IE}(P) = \sum_{c=1}^3 \sum_{(h,w) \in P} \mathbf{D}_{\text{total}}^{(c,h,w)}$. A lower IE value indicates a smoother, less textured region. This operation

is implemented as a summation over the designated tensor dimensions (‘dim=[1, 2, 3]’).

This design allows the IE metric to capture fine-grained, channel-wise texture anomalies that are critical for distinguishing subtle generative artifacts.

IV. ADDITIONAL ABLATION STUDIES AND RESULTS

A. Comprehensive Analysis of Patch Selection

TABLE I: Performance of Different Patch Selection Strategies on DIFF-Intra. Only the best achieved accuracy (**Best Acc(%)**) is retained for brevity.

Type of Patches	Number of Patches	Dataset Size	Acc(%)
max	16	all	94.91
max	16	20%	89.94
max	64	all	94.26
max	64	20%	83.78
mid	4	all	98.66
mid	4	20%	89.56
mid	16	all	97.93
mid	16	20%	86.99
mid	64	all	96.94
mid	64	20%	86.11
min	1	all	98.13
min	4	all	97.86
min	4	20%	90.17
min	16	all	98.85
min	16	20%	97.17
min	64	all	97.67
min	256	all	92.65
ori	-	all	98.66
ori	-	20%	90.86
random	16	all	97.62

To validate our patch selection strategy, we experiment with different criteria based on **Image Entropy**, including patches with minimum, maximum, and central entropy, as well as the entire image and randomly selected patches. All tests are conducted using ResNet50 with varying numbers of patches and dataset sizes.

As shown in table I, our findings reveal the following:

1. **Patch-based detection is highly effective:** Even random patch selection achieves competitive accuracy (97.62%), confirming that localized regions contain sufficient forensic cues to distinguish AI-generated content.

2. **Minimum-entropy patches consistently outperform others:** The best configuration—16-patch division with minimum-entropy selection—yields a peak accuracy of **98.85%**, and maintains high performance even with only 20% of the training data (97.17%).

3. **High-entropy patches are suboptimal:** Despite their complexity, maximum-entropy regions likely introduce irrelevant noise, reducing generalization.

These results support the design rationale of EGPL: low-entropy regions are more robust and artifact-prone, making them ideal candidates for patch-based detection. While random or center patches also perform well, our entropy-guided approach offers greater consistency and data efficiency.

B. Sensitivity Analysis of the VIB Weight β

The variational information bottleneck (VIB) module is regulated by the hyperparameter β , which controls the trade-off between preserving task-relevant information and compressing the latent representation. While $\beta = 10^{-3}$ is a commonly adopted default, we conduct a focused sensitivity analysis to verify its optimality for our specific task.

Experimental Setup: We trained the full EIB-Net (with EGPL selecting the minimum entropy patch from a 16-patch split) on the DIFF-Intra subset. All other hyperparameters (learning rate, batch size, data augmentation) were held constant. We evaluated three values: $\beta \in \{10^{-4}, 10^{-3}, 10^{-2}\}$.

Results and Analysis: The validation accuracy for each β is reported in table II. The results confirm that $\beta = 10^{-3}$ yields the best performance (97.17%). A weaker regularization ($\beta = 10^{-4}$) leads to under-compression and lower accuracy (96.75%), while a stronger regularization ($\beta = 10^{-2}$) likely causes excessive information loss, also harming performance (96.86%). This analysis justifies our selection of $\beta = 10^{-3}$ for all main experiments.

TABLE II: Sensitivity analysis of the VIB regularization parameter β on the DIFF-Intra subset.

β	Validation Accuracy (%)	Relative Performance
1×10^{-4}	96.86	-0.31%
1×10^{-3} (<i>Ours</i>)	97.17	-
1×10^{-2}	96.75	-0.42%

C. Ablation on Classification Loss Components

To decouple the contributions of different components within our chosen Jensen-Shannon Divergence (JSD) loss, we conduct a progressive ablation study on the DIFF-Intra subset. The JSD loss can be decomposed into a Label Smoothing Cross-Entropy (LSCE) base and an additional JSD consistency regularization term. We compare three settings:

- **CE:** Standard Cross-Entropy loss.
- **LSCE only:** Only the Label Smoothing component ($\epsilon = 0.1$).
- **JSD (Ours):** The full loss combining LSCE and JSD regularization ($\alpha = 12$).

TABLE III: Progressive ablation on the classification loss. The full JSD loss yields the best performance.

Loss Function	Validation Accuracy (%)	Relative Performance
CE	95.60	-
LSCE only	96.25	+0.65%
JSD (Ours)	97.17	+1.57%

Analysis: The results demonstrate that both components contribute to the final performance. Label smoothing provides a stable gain of +0.65% over the CE baseline by mitigating over-confidence. On top of that, the JSD consistency regularization further improves accuracy by +0.88%, leading to a total gain of +1.57%. This confirms that the full JSD

loss is not merely effective due to label smoothing, but also because of its inherent self-consistency regularization, which aligns well with our goal of learning robust and generalizable features.

REFERENCES

- [1] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li, "Dire for diffusion-generated image detection," in *ICCV*, 2023, pp. 22445–22455.
- [2] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.
- [3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *ICLR*, 2018.
- [4] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang, "Genimage: A million-scale benchmark for detecting ai-generated image," *NIPS*, vol. 36, 2024.
- [5] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*. PMLR, 2022, pp. 16784–16804.
- [6] Wukong, "Wukong," <https://xihe.mindspore.cn/modelzoo/wukong>, 2022.
- [7] Midjourney, "Midjourney," <https://www.midjourney.com/home/>, 2022.
- [8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.
- [9] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo, "Vector quantized diffusion model for text-to-image synthesis," in *CVPR*, 2022, pp. 10696–10706.
- [10] Prafulla Dhariwal and Alexander Nichol, "Diffusion models beat gans on image synthesis," *NIPS*, vol. 34, pp. 8780–8794, 2021.
- [11] Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli, "Diffusion facial forgery detection," *arXiv preprint arXiv:2401.15859*, 2024.
- [12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *ICLR*, 2024.
- [13] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang, "Freedom: Training-free energy-guided conditional diffusion model," in *ICCV*, 2023, pp. 23174–23184.
- [14] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li, "Better aligning text-to-image models with human preference," *arXiv preprint arXiv:2303.14420*, vol. 1, no. 3, 2023.
- [15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [16] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023, pp. 22500–22510.
- [17] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee, "Diffface: Diffusion-based face swapping with facial guidance," *arXiv e-prints*, pp. arXiv-2212, 2022.
- [18] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu, "Dcfac: Synthetic face generation with dual condition diffusion model," in *CVPR*, 2023, pp. 12715–12725.
- [19] Bahjat Kavar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani, "Imagic: Text-based real image editing with diffusion models," in *CVPR*, 2023, pp. 6007–6017.
- [20] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu, "Collaborative diffusion for multi-modal face generation and editing," in *CVPR*, 2023, pp. 6080–6090.
- [21] Chen Henry Wu and Fernando De la Torre, "A latent space of stochastic diffusion models for zero-shot image editing and guidance," in *ICCV*, 2023, pp. 7378–7387.

- [22] Yeqi Bai, Tao Ma, Lipo Wang, and Zhenjie Zhang, “Speech fusion to face: Bridging the gap between human’s vocal characteristics and facial imaging,” in *ACM MM*, 2022, pp. 2042–2050.
- [23] Joon Son Chung, Arsha Nagrani, and Andrew Senior, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.
- [24] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *CVPR*, 2020, pp. 5549–5558.
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [26] Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He, “Rmt: Retentive networks meet vision transformers,” in *CVPR*, 2024, pp. 5641–5651.
- [27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ICLR*, 2021.
- [28] François Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *CVPR*, 2017, pp. 1251–1258.