

# UniShield: An Adaptive Multi-Agent Framework for Unified Forgery Image Detection and Localization

## Supplementary Material

### 001 A. Training Details

002 In this section, we present further details on the training  
003 of UniShield and the comparative methods adopted in our  
004 experiments.

005 Specifically, we retrain the IML-ViT [22]. The original  
006 pretrained weights are trained solely on forged images  
007 from the CASIA2 [9] dataset, lacking the ability to recognize  
008 authentic images. To address this, we incorporate authentic  
009 images from the CASIA2 dataset into the training process to  
010 alleviate distribution imbalance between real and fake samples.  
011 This adjustment significantly improves the model’s ability to  
012 discriminate real content and enhances its generalization  
013 performance. To further improve UniShield’s capability in  
014 DeepFake detection, we train both the CLIP [25] and DFD-R1  
015 models on the FS subset of the DF40 dataset [36]. This allows  
016 the models to better capture semantic inconsistencies and subtle  
017 facial forgery details. Additionally, our DMDL-R1 model is  
018 trained on the RTM-train dataset [20], specifically optimized  
019 for tampering detection in document images. All other expert  
020 models integrated into UniShield are used with their official  
021 released pretrained weights without further modification.

023 In Section 4.2, we train two representative models, CLIP  
024 and ResNet [13], on a mixed-domain dataset to evaluate their  
025 ability to generalize across different forgery types. We then  
026 compare their performance with that of our proposed UniShield  
027 system. The mixed-domain dataset covers data from four major  
028 forgery domains, including IMDL (CASIA2 [9]), AIGCD (DiffusionForensics [29]),  
029 DFD (DF40 [36]’s FS subset), and DMDL (RTM-train [20]).  
030 By training CLIP and ResNet on this dataset, we aim to  
031 examine whether these models can perform cross-domain  
032 detection. It is also worth noting that our task router is  
033 trained on the same mixed-domain dataset.

### 035 B. Task Router Performance

036 In Section 3.2.1, we provide a detailed description of the  
037 design of the task router module. We fine-tune Qwen2.5-VL-7B  
038 [1] on a mixed dataset consisting of IMDL, DMDL, DFD, and  
039 AIGCD data using GRPO [11]. Since the task router’s task  
040 decision directly determines whether the image is routed to the  
041 correct downstream detector, its reliability is crucial. Therefore,  
042 in this section, we evaluate the performance of the task router.

044 To establish a fair comparison, we train a ResNet-50 [13]  
045 on the same dataset used for the task router as a

Table 1. Task domain detection performance.

Method	IMDL	AIGCD	DFD	DMDL
ResNet-50 [13]	0.912	0.262	0.939	1.000
Task Router (Ours)	0.983	0.947	0.992	1.000

046 comparison method. Then we evaluate them on domain-  
047 selection tasks across four benchmarks: CASIA1+ (IMDL),  
048 RTM (DMDL), DF40 (DFD), and AIGCDetectBenchmark  
049 (AIGCD). Noted that our objective in this setting is solely  
050 domain identification; we do not distinguish between real  
051 and fake images within each domain. Consequently, both  
052 the training and testing data for every domain contain a  
053 mixture of real and fake samples.

054 The accuracy results are summarized in Table 1. The  
055 DMDL domain is highly distinguishable due to the unique  
056 characteristics of document-style images, which differ  
057 significantly from the content of the other domains. As a  
058 result, both ResNet and the task router achieve a perfect  
059 accuracy of 1.000. In the DFD domain, characterized by  
060 close-up facial images, both models perform well, although  
061 the task router achieves a higher accuracy of 0.992, compared  
062 to 0.939 for ResNet. In contrast, the IMDL and AIGCD  
063 domains exhibit more complex and diverse content, making  
064 domain separation based on visual cues more challenging.  
065 ResNet performs particularly poorly in the AIGCD domain,  
066 with an accuracy of only 0.262, frequently misclassifying  
067 AIGCD images as belonging to IMDL. The task router,  
068 however, maintains strong and stable performance, achieving  
069 accuracies of 0.983 on IMDL and 0.947 on AIGCD.

070 As emphasized in Section 3.2.1, domain classification in  
071 this context is not a simple image classification task: it  
072 requires an understanding of manipulation techniques. Simple  
073 models like ResNet are not well-suited for this challenge,  
074 while task router’s consistently high accuracy ensures a  
075 reliable basis for downstream detection tasks.

### C. Prompts Design

076 We carefully designed a vision-language prompt to guide  
077 the tool scheduler, enabling the powerful yet training-free  
078 Qwen2.5-VL [1] model to analyze both low-level visual  
079 cues and high-level semantic inconsistencies. This prompt  
080 allows the system to make informed decisions when  
081 selecting between LLM-based and non-LLM-based detection  
082 tools. The full prompt design is illustrated in Figure 1 and  
083 Figure 2.

You are a professional image forensics analyst. Your task is to evaluate images that may have been digitally manipulated, and decide which type of forgery detection method is most appropriate based on the visual characteristics of each image.

You have access to two categories of forgery detection tools:

(1) LLM-based methods: These tools are specialized in identifying high-level semantic or contextual inconsistencies. Use this category when the image likely contains:

- Logical contradictions (e.g., a person casting a shadow in the wrong direction)
- Improbable object relationships (e.g., a cat walking a dog)
- Implausible human behavior, physics violations, or common sense errors (e.g., floating objects, duplicate people)
- Incoherent visual storytelling or scene setup

(2) Non LLM-based methods (e.g., CNNs or Transformer-based vision models) These are designed to detect low-level visual anomalies.

Use this category when the image may include:

- Texture discontinuities
- Color or lighting inconsistencies at the pixel level
- Splicing, copy-move forgeries, or editing traces
- Compression artifacts or irregular noise patterns

Instructions:

- For each image, inspect the content carefully.
- Determine the likely nature of the forgery (semantic-level or pixel-level).
- Provide your reasoning briefly within a <think> tag.
- Then, output your final decision inside an <answer> tag.
- Use one of the following labels for <answer>:
  - LLM Based Tool
  - Non LLM Based Tool

Important:

- Focus strictly on the type of potential anomaly, not on the relative performance, speed, or popularity of the tools.
- Output must follow this format:
 

```
<think>Your brief reasoning...</think>
<answer>Your decision...</answer>
```

Figure 1. Tool Scheduler Prompt.

085 In addition, we designed a structured prompt for the re-  
086 port summarizer, which standardizes the output format of  
087 detection reports. This ensures that the results are presented  
088 in a clear, organized, and interpretable manner, significantly  
089 enhancing user understanding.

## 090 D. Compared Methods

### 091 D.1. IMDL Compared Methods

092 To evaluate the **IMDL performance** of UniShield, we  
093 compare it against SPAN [14], MantraNet [31], HiFi-  
094 Net [12], PSCC-Net [19], CAT-Net [15], MVSS-Net [8],  
095 FakeShield [34], and IML-ViT [22].

096 **SPAN [14]**. SPAN builds a spatial pyramid of local self-  
097 attention blocks with positional projection to model multi-

scale patch relations. Trained on synthetic data and option-  
ally fine-tuned, it generalizes well across diverse manipula-  
tions. 098 099 100

**MantraNet [31]**. ManTra-Net is an end-to-end FCN that  
detects and localizes forgeries without preprocessing. It  
learns manipulation traces via a large self-supervised task  
and performs localization through Z-score-based anomaly  
detection with an LSTM module. 101 102 103 104 105

**HiFi-Net [12]**. HiFi-Net introduces hierarchical fine-  
grained attribute learning using multi-level labels and a  
multi-branch feature extractor. Its unified classification and  
localization modules yield strong performance across di-  
verse IFDL benchmarks. 106 107 108 109 110

**PSCC-Net [19]**. PSCC-Net uses a progressive two-path de-  
sign with dense cross-scale connections to refine manipu-  
111 112

113	lation masks. Its Spatio-Channel Correlation Module enhances robustness, achieving real-time performance with high accuracy.	
114		
115		
116	<b>CAT-Net [15]</b> . CAT-Net leverages DCT-domain artifacts for manipulation localization, using a specialized architecture that preserves spatial alignment of DCT coefficients and jointly models acquisition and compression cues.	164
117		165
118		166
119		167
120	<b>MVSS-Net [8]</b> . MVSS-Net fuses boundary artifacts and noise-view features, two semantic-agnostic cues, and applies multi-scale supervision, enabling strong generalization and robustness to compression, blur, and recapturing.	168
121		169
122		170
123		171
124	<b>FakeShield [34]</b> . FakeShield provides explainable IFDL via multi-modal analysis, combining authenticity assessment, mask prediction, and textual justifications. Trained with GPT-4o-enhanced data, it achieves superior cross-domain tamper detection and localization.	172
125		173
126		174
127		175
128		176
129	<b>IML-ViT [22]</b> . IML-ViT is a Transformer-based IML model with high-resolution capacity, multi-scale features, and edge supervision, enabling effective artifact learning with limited data and outperforming prior localization methods.	177
130		178
131		179
132		180
133		181
134	<b>D.2. DFD Compared Methods</b>	182
135	To evaluate the <b>DFD performance</b> of UniShield, we compare it against a range of DeepFake detectors, including Xception [7], CLIP [25], SRM [21], SPSL [18], RECCE [3], RFM [26], and DFD-R1 on the DF40 benchmark.	183
136		184
137		185
138		186
139		187
140	<b>Xception [7]</b> . Xception replaces Inception modules with depthwise separable convolutions, achieving more efficient parameter usage and outperforming Inception-V3 on ImageNet and large-scale classification tasks.	188
141		189
142		190
143		191
144	<b>CLIP [25]</b> . CLIP learns from 400M image-text pairs using contrastive training, enabling natural language-driven zero-shot transfer that performs competitively across diverse vision tasks without task-specific supervision.	192
145		193
146		194
147		195
148	<b>SPSL [18]</b> . SPSL combines spatial information and phase spectra to capture up-sampling artifacts in forged faces, using a shallow network to focus on local textures and achieving strong cross-dataset generalization.	196
149		197
150		198
151		199
152	<b>RECCE [3]</b> . RECCE learns generalizable real-face representations via joint reconstruction and classification, enhanced by multi-scale encoder-decoder graphs and reconstruction-difference cues for robust forgery detection.	200
153		201
154		202
155	<b>RFM [26]</b> . RFM occludes a detector's most sensitive facial regions to force broader attention, providing a simple augmentation strategy that boosts CNN-based fake-face detectors to state-of-the-art performance.	203
156		204
157		205
158		206
159		207
160	<b>D.3. DMDL Compared Methods</b>	208
161	To evaluate the <b>DMDL performance</b> of UniShield, we compare it with various document forgery detection methods, including UperNet [32], DeepLabV3+ [4],	209
162		210
163		211
	HRNet-OCR [27], SegFormer [33], MaskFormer [5], Mask2Former [6], RRU-Net [2], PSCC-Net [19], MVSS-Net++ [8], CAT-Net v2 [16], Liang et al. [17], DTD [24], ASC-Former [20], and DMDL-R1.	212
	<b>UperNet [32]</b> . UperNet is a multi-task framework for Unified Perceptual Parsing that learns scenes, objects, textures, and parts from heterogeneous annotations, enabling broad and effective visual concept segmentation.	213
	<b>DeepLabV3+ [4]</b> . DeepLabv3+ adds a simple decoder to DeepLabv3 for sharper boundaries and applies Xception-style depthwise separable convolutions, forming a fast, strong encoder-decoder model with state-of-the-art segmentation results.	214
	<b>HRNet-OCR [27]</b> . HRNet-OCR keeps high-resolution features via parallel multi-resolution branches and enhances them with region-aware context, yielding precise and robust representations for various vision tasks.	215
	<b>SegFormer [33]</b> . SegFormer pairs a hierarchical Transformer encoder with a lightweight MLP decoder to produce strong multi-scale representations without positional encoding, achieving highly efficient and accurate semantic segmentation across a range of model sizes.	
	<b>MaskFormer [5]</b> . MaskFormer unifies semantic and panoptic segmentation by predicting a set of masks with class labels via a single mask-classification framework, replacing per-pixel classification and achieving state-of-the-art results across both tasks.	
	<b>Mask2Former [6]</b> . Mask2Former unifies panoptic, instance, and semantic segmentation with a masked-attention Transformer that extracts region-specific features, achieving strong generality and state-of-the-art performance across multiple segmentation benchmarks.	
	<b>RRU-Net [2]</b> . RRU-Net is an end-to-end U-Net-style splicing forgery detector that strengthens CNN learning through residual propagation and feedback, enhancing attribute differences between tampered and untampered regions and achieving strong performance without preprocessing.	
	<b>DTD [24]</b> . DTD introduces a frequency-aware head and a multi-view iterative decoder to detect subtle tampered text in document images, combined with a curriculum learning strategy for improved robustness and generalization, achieving large gains on the DocTamper benchmarks.	
	<b>ASC-Former [20]</b> . ASC-Former targets real-world text tampering with the RTM dataset and introduces a dual-stream framework equipped with consistency-aware aggregation, gated cross-neighborhood attention, and contrastive learning, yielding substantially improved localization performance on challenging real tampering cases.	
	<b>D.4. AIGCD Compared Methods</b>	
	To evaluate the <b>AIGCD performance</b> of UniShield, we compare it against a set of expert detectors from the AIGCDetectBenchmark [37], including CNNSpot [28],	

216 FreDect [10], UnivFD [23], DIRE [29], PatchCraft [37],  
217 AIDE [35], and FakeVLM [30].

218 **CNNSpot [28]**. CNNSpot shows that a classifier trained  
219 only on ProGAN fakes, paired with careful preprocessing  
220 and augmentation, can generalize to images from many un-  
221 seen generators, revealing common artifacts across CNN-  
222 based synthesis models and enabling universal fake-image  
223 detection.

224 **FreDect [10]**. This work reveals that GAN-generated im-  
225 ages contain consistent, architecture-agnostic artifacts in  
226 the frequency domain, caused by universal upsampling op-  
227 erations, and shows that these signatures enable highly ef-  
228 fective, automated deep-fake detection that surpasses prior  
229 methods.

230 **UnivFD [23]**. UnivFD shows that traditional real vs fake  
231 classifiers overfit to specific generative patterns, and instead  
232 proposes detection in an untrained feature space using sim-  
233 ple nearest-neighbor or linear probes, achieving markedly  
234 stronger generalization to unseen GAN, diffusion, and au-  
235 toregressive models.

236 **DIRE [29]**. DIRE detects diffusion-generated images by  
237 measuring the reconstruction error from a pretrained diffu-  
238 sion model: synthetic images reconstruct well, whereas real  
239 images do not. This results in a simple, general, and robust  
240 representation that outperforms previous detectors across a  
241 wide range of diffusion models.

242 **PatchCraft [37]**. This method improves generalization in  
243 AI-generated image detection by discarding global seman-  
244 tics via Smash&Reconstruction to amplify texture clues,  
245 and by exploiting inter-pixel correlation differences be-  
246 tween rich and poor texture regions, where generative mod-  
247 els struggle most, yielding robust detection across diverse  
248 generators.

249 **AIDE [35]**. AIDE combines CLIP-based semantic embed-  
250 dings with high- and low-frequency patch features to cap-  
251 ture both contextual cues and fine-grained artifacts, achiev-  
252 ing improved generalization on challenging AI-generated  
253 image benchmarks such as Chameleon and advancing  
254 progress in robust fake-image detection.

255 **FakeVLM [30]**. FakeVLM is a multimodal large model for  
256 detecting synthetic and DeepFake images that also provides  
257 natural-language artifact explanations, trained with the fine-  
258 grained FakeClue dataset and achieving strong authenticity  
259 classification and interpretable detection without auxiliary  
260 classifiers.

## 261 E. Evaluation Metrics

262 Since our work involves performance evaluation across four  
263 distinct sub-tasks, we follow standard experimental settings  
264 in each domain and adopt appropriate evaluation metrics ac-  
265 cordingly.

266 In the IMDL domain, following FakeShield [34], we use  
267 accuracy and image-level F1 score to evaluate forgery de-

268 tection performance, and adopt IoU and pixel-level F1 score  
269 to assess localization performance. In the DFD domain, fol-  
270 lowing the DF40 [36] benchmark, we use the AUC metric  
271 to evaluate DeepFake detection performance. In the DMDL  
272 domain, based on the RTM [20] benchmark, we use pixel-  
273 level IoU, precision, recall, and F1 score to evaluate local-  
274 ization quality, and image-level F1 score for detection per-  
275 formance. In the AIGCD domain, following the AIGCDe-  
276 tectBenchmark [37], we use image-level accuracy as the  
277 evaluation metric for forgery detection.

278 For the cross-domain FIDL task, we exclusively assess  
279 forgery detection ability, using accuracy and image-level F1  
280 score as evaluation metrics.

## 281 F. More Examples

282 To better illustrate UniShield’s interactive detection pro-  
283 cess and response capabilities, we present representative  
284 dialogue samples from the test sets of four major forgery  
285 domains: IMDL, DMDL, DFD, and AIGCD. For each do-  
286 main, we selected two sample dialogues. Additionally, we  
287 included dialogues based on authentic images to highlight  
288 UniShield’s ability to distinguish between real and forged  
289 content. These examples are shown in Figure 3, Figure 4,  
290 Figure 5 and Figure 6.

291 In addition, we attach the output results of the task router  
292 and tool scheduler, as shown in Figure 8 and Figure 9.

## 293 G. Limitation and Future Works

294 Despite UniShield’s strong performance in multi-domain  
295 image forgery detection, several limitations remain. Firstly,  
296 although UniShield currently outperforms existing expert  
297 models across various forgery detection tasks and demon-  
298 strates strong performance, we cannot guarantee its robust-  
299 ness against future advances in generative techniques. Sec-  
300 ondly, UniShield is primarily designed for static images and  
301 lacks sufficient adaptation to multimodal forgeries, such as  
302 those involving video, audio, or image-text combinations.  
303 To address these limitations, we plan to expand the ex-  
304 pert model pool by incorporating more specialized detectors  
305 that can effectively handle new and diverse manipulation  
306 types. Furthermore, we aim to develop a unified multimodal  
307 forgery detection that extends UniShield’s capabilities to  
308 video, audio, and textual modalities, enabling broader ap-  
309 plicability in content verification and digital forensics sce-  
310 narios.

You are a professional image forgery detection report assistant working within the UniShield multi-agent system. Your task is to generate a structured, interpretable, and user-friendly detection report based on the outputs from the perception agent and detection agent. The report must strictly follow the format below and maintain clear, precise, and logical descriptions.

### 1. Input Overview

Before generating the detection report, you will receive the following inputs:

- Image: The raw image to be analyzed.
- Task Router Output: The predicted forgery type category:
  - IMDL: Image Manipulation Detection and Localization
  - DMDL: Document Manipulation Detection and Localization
  - DFD: DeepFake Detection
  - AIGCD: AI-Generated Content Detection
- Tool Scheduler Output: The selected expert model and whether it is LLM-based or non-LLM-based.
- Detector Output:
  - For non-LLM-based models: Includes detection result, confidence score, and optional tampering mask
  - For LLM-based models: Includes detection result, explanation (textual rationale), and optional tampering mask

### 2. Agent Workflow Overview

The UniShield system handles image forgery detection via two collaborative agents:

#### Perception Agent

- Analyzes the image and assigns it to the most suitable forgery detection track using a task router.
- The tool scheduler then selects the best expert detector from a toolbox based on the forgery category.

#### Detection Agent

- Runs the selected detection model to identify forgeries and optionally localize them.
- Based on model type:
  - LLM-based: Outputs both the detection result and a human-readable explanation
  - Non-LLM-based: Outputs structured detection result, confidence score, and mask (if available)

### 3. Detection Report Format

Description: Briefly describe the main content or scene of the image.

- For documents, mention the document type, language, key contents, and layout format.
- For portraits or natural scenes, describe visual details such as character appearance, clothing, posture, and environment.

Detection (Result):

- For authentic images, write: "The image is identified as real with no visible signs of tampering."
- For forged images (forgery methods include IMDL, DMDL, DFD and AIGCD), write directly: "the image is forged by the (IMDL/DMDL/DFD/AIGCD) method, and write whether it is the detection of the LLM Based Tool or Non LLM Based Tool."

Localization:

- If a tampering mask is available, list the tampered areas (describe the altered content and its location).

Judgment Basis:

- Low-Level Visual Clues: Edge artifacts, unnatural textures, inconsistent noise patterns, abnormal colors or mismatched lighting.
- High-Level Semantic Clues: Incoherent context, unnatural facial expressions, or illogical scene composition.
- For authentic images, rationale can also be included, focusing on natural and internally consistent features.

Figure 2. Report Summarizer Prompt.

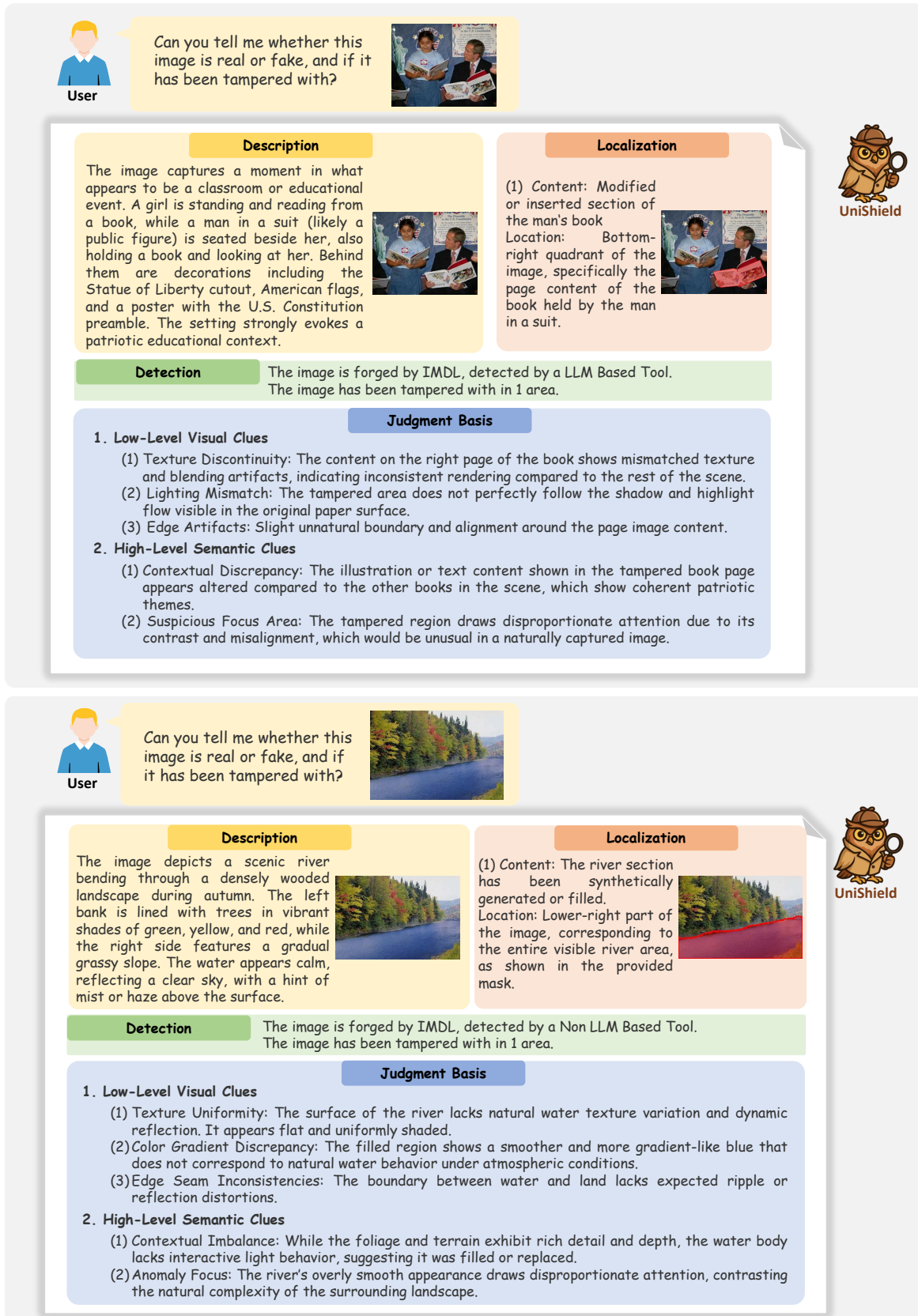




Figure 3. UniShield's detection reports on IMDL domain.



**User**

Can you tell me whether this image is real or fake, and if it has been tampered with?




**Description**

The image shows a printed receipt from a retail transaction. It lists payment details including GST, total amount, cash tendered, and change returned. There is a section summarizing item count, GST summary, and tax calculation, followed by a promotional message encouraging membership. The receipt includes URLs and a slogan about Chinese books.

**Localization**

(1) Content: The number of 'Total Savings' Location: To the right of the text "Total Savings"



**UniShield**

**Detection** The image is forged by DMDL, detected by a Non LLM Based Tool. The image has been tampered with in 1 area.


**Judgment Basis**

**1. Low-Level Visual Clues**

- (1) Edge Artifacts: The tampered text area exhibits smoother and less pixelated edges than surrounding text, suggesting post-processing or overlay.
- (2) Font Irregularity: The altered portion has inconsistencies in character alignment and stroke thickness compared to the rest of the line.
- (3) Texture Disturbance: Slight local brightness or texture shift around the edited text area.


**2. High-Level Semantic Clues**

- (1) Contextual Disruption: The suspected modified section interferes with the regular structure of the URL and printed footer layout.
- (2) Linguistic Anomaly: The phrase near the mask ("Buy Chinese books online") appears partially misaligned and stylistically different, which may suggest manipulation of promotional content or branding.



**User**

Can you tell me whether this image is real or fake, and if it has been tampered with?




**Description**

The image shows a printed cash receipt issued by BEMED (SP) SDN. BHD., a medical or supplement-related business. The receipt lists a transaction on 14/Jun/2017 for one item: "ALPHA LIPID LIFE LINE" priced at MYR 165.00, paid via VISA. It includes GST summary information, totals, rounding, and a thank-you message. Handwritten annotations like "CP", "OK", and "165/-" appear on the receipt.

**Localization**

(1) Content: Number '03-7845 8155' Location: Middle of the receipt, on the line listing the purchased item  
 (2) Content: Location name 'Shah Alam' Location: In the business address section, mid-upper part of the receipt.  
 (3) Content: Word 'Rounding' Location: Near the bottom-left of the receipt, in the line "Points : Rounding"



**UniShield**

**Detection** The image is forged by DMDL, detected by a Non LLM Based Tool. The image has been tampered with in 3 area.

**Judgment Basis**

**1. Low-Level Visual Clues**

- (1) Texture Disruption: The altered fields show smoother regions with different noise patterns compared to surrounding printed text.
- (2) Edge Sharpness Difference: The tampered text appears either sharper or more smudged than original thermal print.
- (3) Background Brightness Anomaly: There are subtle but detectable tonal changes beneath the modified regions.

**2. High-Level Semantic Clues**

- (1) Amount Discrepancy Risk: Modifications involve numerically significant fields such as item price and tax summary, which are critical for financial records.
- (2) Visual Irregularity: Handwriting overlaps and spatial inconsistency around modified digits suggest human-induced editing, possibly to align or conceal the change.

Figure 4. UniShield’s detection reports on DMDL domain.

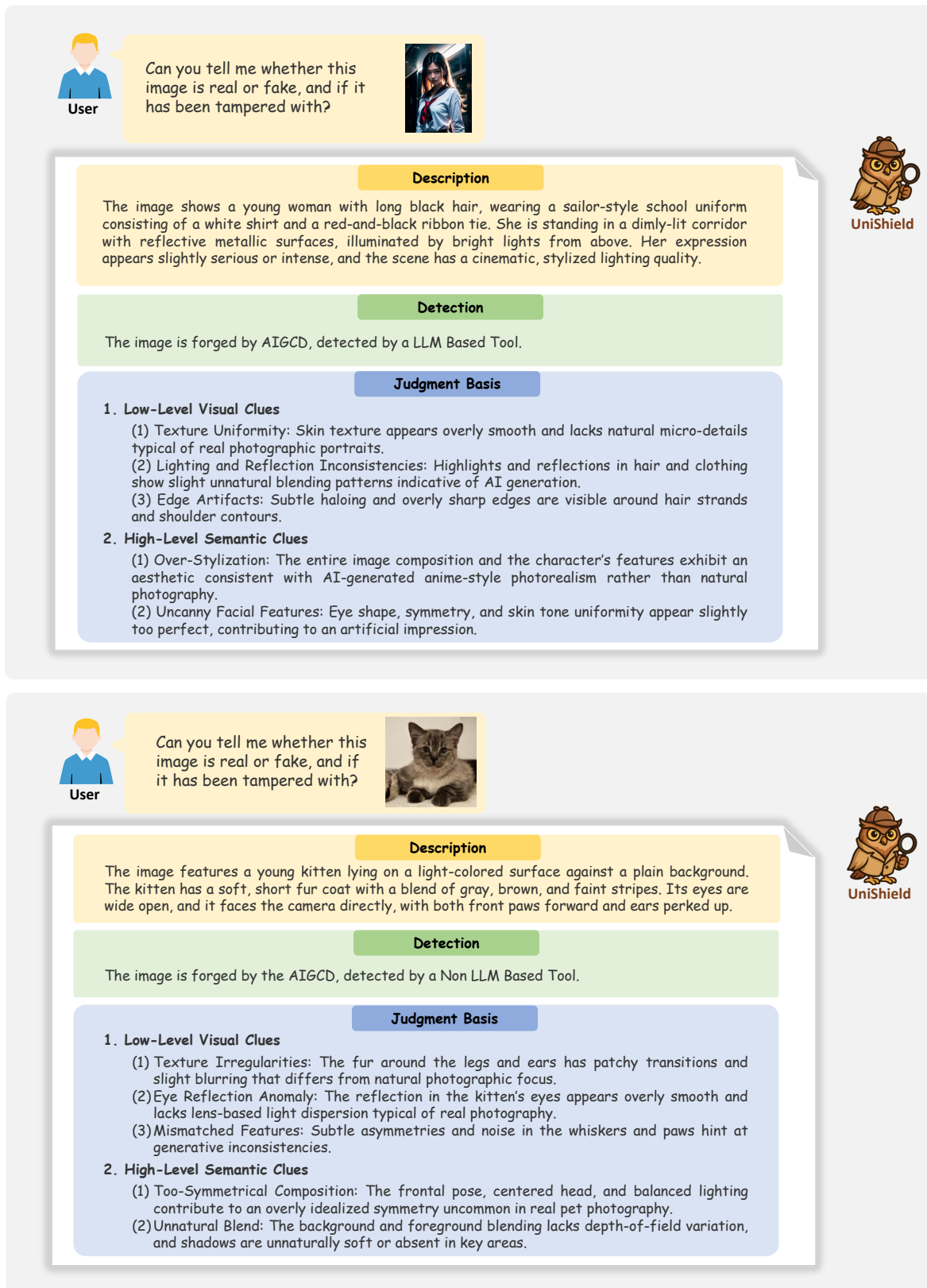




Figure 5. UniShield's detection reports on AIGCD domain.



Can you tell me whether this image is real or fake, and if it has been tampered with?



**Description**

The image shows a man wearing a dark flat cap, with light facial hair and a faint smile. The background is dim, focusing attention on the face, which is slightly angled to the left. His skin tone is even, and the overall composition resembles a video frame or an interview shot.

**Detection**

The image is forged by DFD, detected by a Non LLM Based Tool.


**Judgment Basis**

**1. Low-Level Visual Clues**


- (1) Facial Blending Issues: The transition between the facial features (especially the eyes, mouth, and chin) and surrounding skin shows mild blending inconsistencies.
- (2) Edge Smoothing: Unnaturally smooth edges around the jawline and cap border, lacking real-world camera sharpness.
- (3) Compression Artifacts: Localized blocky patterns around the cheeks and lips suggest generative rendering or post-synthesis compression.

**2. High-Level Semantic Clues**


- (1) Unnatural Expression Sync: The facial expression and mouth posture appear contextually inconsistent, particularly around lip motion and eye engagement.
- (2) Identity Conflict: While the face resembles a known individual, slight uncanny discrepancies exist in facial symmetry and detail realism—common signs of identity-swapping DeepFakes.



UniShield



Can you tell me whether this image is real or fake, and if it has been tampered with?



**Description**

The image depicts a bald man with a mustache, dressed in a dark jacket over a collared shirt, speaking or appearing in front of a blue studio background. The facial expression appears neutral, with slightly parted lips as if talking.

**Detection**

The image is forged by the DFD, detected by a LLM Based Tool.


**Judgment Basis**

**1. Low-Level Visual Clues**

- (1) Skin Texture Artifacts: Visible patchy texture on the skin, particularly on the forehead and cheeks, inconsistent with natural human skin.
- (2) Blending Issues: The contour between the face and background shows subtle blending artifacts and uneven edges.
- (3) Compression Anomalies: Portions of the face exhibit blocky compression patterns that differ from surrounding image areas.

**2. High-Level Semantic Clues**

- (1) Facial Dynamics Mismatch: Expression, mouth shape, and eye focus seem slightly detached from the implied speech motion, typical in DeepFake videos.
- (2) Contextual Inconsistency: The person's facial features look slightly exaggerated and lack natural asymmetry found in authentic human faces.




UniShield

Figure 6. UniShield’s detection reports on DFD domain.



Can you tell me whether this image is real or fake, and if it has been tampered with?



**Description**

The image depicts a man wearing a dark shirt and a maroon scarf, playing an electronic keyboard with his eyes closed, suggesting deep concentration or emotion. In the background, another musician plays a drum set, visible through a transparent partition. The setting appears to be an indoor music studio or performance space with musical instruments and equipment around.




**Detection**


The image is identified as an authentic photograph.

**Judgment Basis**

1. **Low-Level Visual Clues**
  - (1) **Natural Textures:** Skin textures, clothing fabrics, and instrument surfaces exhibit natural variation and fine details without signs of artificial smoothing or rendering artifacts.
  - (2) **Lighting Consistency:** Light sources, reflections, and shadows are coherent throughout the scene, consistent with a real photographic capture.
  - (3) **Clarity:** Object boundaries and edges appear naturally integrated, with no visible compositing lines or blending errors.
2. **High-Level Semantic Clues**
  - (1) **Realistic Scene Composition:** The arrangement of people, instruments, and background is natural and contextually plausible.
  - (2) **Authentic Expression:** The emotional expression and body language of the keyboard player convey genuine human presence, lacking the subtle stiffness often found in AI-generated imagery.




Can you tell me whether this image is real or fake, and if it has been tampered with?



**Description**

The image shows a man wearing a traditional white keffiyeh and agal, along with a white thobe. He is facing the camera directly with a neutral facial expression. A black headset is visible on his right ear. The background appears to be part of a broadcast or video call setting with soft blue and gray tones.



**Detection**

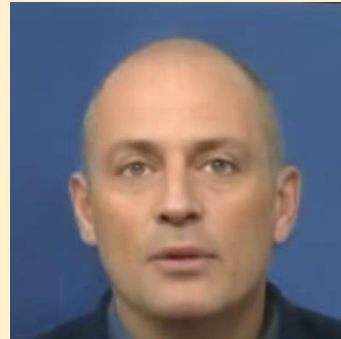
The image is identified as an authentic photograph.

**Judgment Basis**

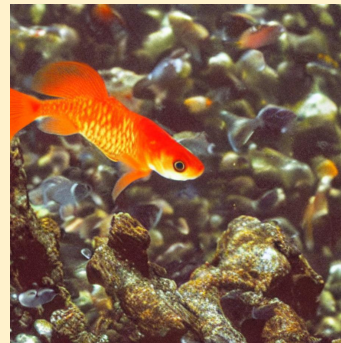
1. **Low-Level Visual Clues**
  - (1) **Natural Skin Texture:** The skin displays realistic pore structure, lighting reflections, and natural tonal gradients.
  - (2) **Edge Sharpness and Noise Consistency:** The edges around the face, headwear, and headset are well-defined and consistent with the resolution of the entire image.
  - (3) **Lighting Coherence:** Facial highlights and shadows match the direction and quality of the light source across the whole scene.
2. **High-Level Semantic Clues**
  - (1) **Contextual Naturalness:** The clothing, facial features, expression, and headset placement are all contextually appropriate and consistent.
  - (2) **Realistic Imperfections:** The slight asymmetry in facial features and natural eye movement reflect genuine human presence.

Figure 7. UniShield's detection reports on authentic images.

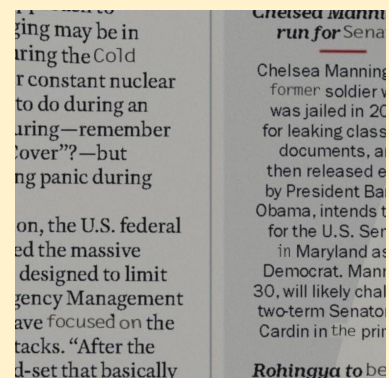
<think>Let me think... The image shows a human face against a plain background, with neutral expression. However, facial details appear slightly rigid, and the skin texture and lighting look somewhat unnatural. These are indicative of potential face swapping or synthesis, suggesting typical DeepFake characteristics.</think>  
<answer>DFD</answer>



<think>Let me think... The image appears to contain synthetic elements. The fish in the foreground has an unusually smooth texture, overly saturated colors, and lacks natural blending with the surrounding water and lighting environment. The background also shows inconsistencies in depth and structure, suggesting generation artifacts. These characteristics align with fully generated content.</think>  
<answer>AIGCD</answer>



<think>Let me think... The image shows a printed document with typographic and layout characteristics. Upon inspection, there are slight distortions around some characters, uneven kerning, and possible blending artifacts—especially noticeable in areas with mixed font weights and colors. These low-level inconsistencies suggest pixel-level tampering often seen in document image manipulation.</think>  
<answer>DMDL</answer>



<think>Let me think... The image shows children playing on a wooden dock by a lake under natural lighting. All shadows, poses, reflections, and environmental details appear coherent and consistent. I don't observe any semantic contradictions or pixel-level anomalies.</think>  
<answer>IMDL</answer>



Figure 8. Task Router Output.

<think>There is a clear visual inconsistency in the image: the seated woman in the foreground does not blend naturally with the background in terms of lighting, resolution, and shadowing. Her presence seems artificially inserted—there's a lack of consistent shadows or blending at the edges. These suggest splicing or digital cut-paste manipulation, which are low-level pixel artifacts rather than high-level logical inconsistencies. Thus, non-LLM-based methods like CNNs or Vision Transformers are more suitable for detecting such forgeries.</think>

<answer>Non LLM Based Tool</answer>



<think>The receipt lists the second item with a quantity of 1 and unit price of RM10.00, but the total is listed as RM200.00, which is a clear violation of arithmetic logic. Additionally, the formatting of "200.00" appears visually inconsistent, suggesting potential tampering. However, the primary issue is the high-level semantic inconsistency in the pricing logic, making LLM-based tools the appropriate choice.</think>

<answer>LLM Based Tool</answer>

ENW Hardware Centre (M) Sdn. Bhd.			
Co. Reg. No. : 795225-A			
GST Reg. No. : 000155453440			
No. G3, Blk G, Jln PJU 1A/3,			
Ara Damansara, 47301			
Petaling Jaya, Selangor Darul Ehsan.			
TEL: 03-7832 6929, 012-659 9829			
10/03/2018	Tax Invoice		No. CS00230322
Cash Sale (Ara D'sara)			
Tel: 03-7832 6929			
H/P: 012-659 9829			
Item	Qty	U.Price	RM
001633	1.00	19.00	19.00 SR
2kgs. Wood Lacquer Thinner			
000920	1.00	10.00	200.00SR
1.5ka. Ufixx Putty Filler.			

<think>While the image looks natural at first glance, closer inspection of the hands—especially the left hand (right side of image)—shows potential anatomical abnormalities such as finger spacing, joint alignment, and count, which may signal AI-generated inconsistencies. These types of errors relate to human anatomy and semantic expectations, thus are best detected by LLM-based tools that understand logical coherence and physical realism.</think>

<answer>LLM Based Tool</answer>



<think>The image shows signs of low-level visual inconsistencies: the subject's facial texture is overly smooth and lacks realistic detail, while features like the ears appear oddly shaped or blurry. Additionally, the sharpness mismatch between the subject and the blurred background introduces suspicion of either AI generation or synthetic composition. These artifacts are best detected through non-LLM-based tools that specialize in identifying texture anomalies, edge blending issues, and generative traces.</think>

<answer>Non LLM Based Tool</answer>



Figure 9. Tool Scheduler Output.

311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1
- [2] Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. Rru-net: The ringed residual u-net for image splicing forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3
- [3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 3
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3
- [5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 3
- [6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3
- [7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3
- [8] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022. 2, 3
- [9] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013. 1
- [10] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 4
- [11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1
- [12] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1
- [14] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [15] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2, 3
- [16] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 3
- [17] Weipeng Liang, Li Dong, Rangding Wang, Diquan Yan, and Yuanman Li. Robust document image forgery localization against image blending. In *2022 IEEE international conference on trust, security and privacy in computing and communications (TrustCom)*, pages 810–817. IEEE, 2022. 3
- [18] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021. 3
- [19] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscn-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. 2, 3
- [20] Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, page 110828, 2024. 1, 3, 4
- [21] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021. 3
- [22] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023. 1, 2, 3
- [23] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023. 4
- [24] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust

- 424 tampered text detection in document image: New dataset and  
425 new solution. In *Proceedings of the IEEE/CVF Conference*  
426 *on Computer Vision and Pattern Recognition*, pages 5937–  
427 5946, 2023. 3
- 428 [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya  
429 Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,  
430 Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning  
431 transferable visual models from natural language supervi-  
432 sion. In *International conference on machine learning*, pages  
433 8748–8763. PmLR, 2021. 1, 3
- 434 [26] Chengrui Wang and Weihong Deng. Representative forgery  
435 mining for fake face detection. In *Proceedings of the*  
436 *IEEE/CVF conference on computer vision and pattern*  
437 *recognition*, pages 14923–14932, 2021. 3
- 438 [27] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang,  
439 Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui  
440 Tan, Xinggang Wang, et al. Deep high-resolution repre-  
441 sentation learning for visual recognition. *IEEE transactions*  
442 *on pattern analysis and machine intelligence*, 43(10):3349–  
443 3364, 2020. 3
- 444 [28] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew  
445 Owens, and Alexei A Efros. Cnn-generated images are sur-  
446 prisingly easy to spot...for now. In *CVPR*, 2020. 3, 4
- 447 [29] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun  
448 Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for  
449 diffusion-generated image detection. In *Proceedings of the*  
450 *IEEE/CVF International Conference on Computer Vision*,  
451 pages 22445–22455, 2023. 1, 4
- 452 [30] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen  
453 Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and  
454 Weijia Li. Spot the fake: Large multimodal model-based  
455 synthetic image detection with artifact explanation. *arXiv*  
456 *preprint arXiv:2503.14905*, 2025. 4
- 457 [31] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan.  
458 Mantra-net: Manipulation tracing network for detection and  
459 localization of image forgeries with anomalous features. In  
460 *Proceedings of the IEEE/CVF Conference on Computer Vi-*  
461 *sion and Pattern Recognition (CVPR)*, 2019. 2
- 462 [32] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and  
463 Jian Sun. Unified perceptual parsing for scene understand-  
464 ing. In *Proceedings of the European conference on computer*  
465 *vision (ECCV)*, pages 418–434, 2018. 3
- 466 [33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar,  
467 Jose M Alvarez, and Ping Luo. Segformer: Simple and  
468 efficient design for semantic segmentation with transform-  
469 ers. *Advances in neural information processing systems*, 34:  
470 12077–12090, 2021. 3
- 471 [34] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing  
472 Huang, and Jian Zhang. Fakeshield: Explainable image  
473 forgery detection and localization via multi-modal large lan-  
474 guage models. In *International Conference on Learning*  
475 *Representations*, 2025. 2, 3, 4
- 476 [35] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xi-  
477 aolong Jiang, Yao Hu, and Weidi Xie. A sanity  
478 check for ai-generated image detection. *arXiv preprint*  
479 *arXiv:2406.19435*, 2024. 4
- 480 [36] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao,  
481 Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang,  
Shouhong Ding, Yunsheng Wu, et al. Df40: Toward  
next-generation deepfake detection. *arXiv preprint*  
*arXiv:2406.13495*, 2024. 1, 4
- [37] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and  
Xinpeng Zhang. Patchcraft: Exploring texture patch for  
efficient ai-generated image detection. *arXiv preprint*  
*arXiv:2311.12397*, 2023. 3, 4