An On-Paper Data Augmentation Algorithm for Dataset Generation

André Neto[1,2][0009−0001−0398−5859], Nuno Gonçalves[1,2][0000−0000−0000−0000], and João Marcos[1,2][0000−0000−0000−0000]

[1] University of Coimbra
[2] Institute of Systems and Robotics

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

The success of deep learning models relies heavily on large and diverse datasets. However, collecting and annotating such datasets is often time-consuming and expensive, especially for specialized tasks. To address this challenge, data augmentation [] has become a fundamental technique, artificially expanding datasets to improve model generalization.

Traditional data augmentation techniques apply transformations such as rotation, scaling, flipping, and noise addition. These methods are effective in many domains but often fail to capture the complex variations found in real-world data. For instance, printed materials such as documents, icons, and symbols undergo physical degradation, including smudging, ink fading, and paper wear, which current digital augmentations algorithms struggle to replicate accurately.

In this paper, we introduce a novel 2D on-paper data augmentation algorithm that simulates real-world degradation using physical sheets of paper with varying states of abrasion. Our method involves preparing multiple sheets with different levels of wear, randomly selecting a region, and imprinting the symbol or icon onto the paper. This process generates thousands of naturally degraded samples, enriching the dataset with realistic variations that improve model performance.

The key contributions of our work are:

- Novel Augmentation Technique: We propose a unique approach that leverages physical paper textures and degradation patterns, offering a realistic alternative to purely digital augmentation.
- Cost-Effective and Accessible: Our method requires minimal computational resources and can be implemented using common materials, making it practical for researchers with limited resources.
- Versatile Applications: The algorithm benefits tasks such as icon recognition, optical character recognition (OCR) [], historical document analysis, and artistic style transfer.

– Improved Model Robustness: By simulating real-world degradation, our augmentation method enhances neural networks' ability to handle variations encountered in practical scenarios.

The remainder of this paper is organized as follows: Section 2 reviews related work in data augmentation and printed symbol recognition. Section 3 details our methodology, including the preparation of paper sheets and the augmentation process. Section 4 presents our experimental setup and results. Section 5 discusses potential applications, and Section 6 concludes the paper with future research directions.

2 Related Work

Data augmentation is a widely used strategy to improve the generalization ability of deep learning models. Traditional augmentation techniques have been extensively used to artificially expand datasets. These methods have proven effective for general computer vision tasks but may not fully capture the physical distortions that appear in printed materials.

For icon recognition, augmentation is particularly crucial due to the often limited availability of labeled data. Existing work has explored digital transformations to generate additional training samples, with methods including geometric distortions, occlusions, and synthetic noise (Krizhevsky et al., 2012). Additionally, some studies have leveraged generative adversarial networks (GANs) to synthesize new icons with slight variations, allowing models to learn more robust representations (Zhu et al., 2017).

However, digitally generated augmentations fail to replicate real-world imperfections that arise from the physical printing and aging process. Physical augmentation techniques have been less explored in the context of icon datasets, with most studies focusing on handwritten character augmentation (Simard et al., 2003). Some approaches have attempted to replicate real-world distortions by capturing printed symbols under different lighting conditions or paper textures, but these remain limited in scope.

Our approach builds upon these ideas by introducing a physically-grounded augmentation technique specifically designed for printed 2D icons. By leveraging sheets of paper with varying levels of abrasion and degradation, our method generates augmented datasets that better reflect the real-world conditions of printed icons. This technique not only enhances model robustness but also provides a simple, low-cost alternative to synthetic augmentation methods.

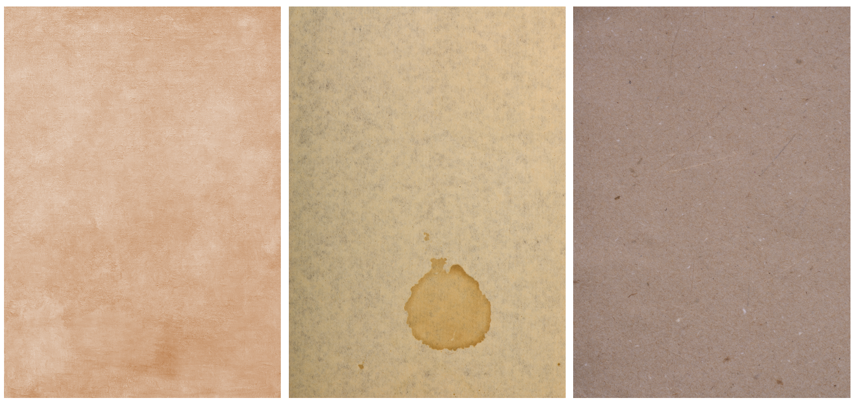
3 Methodology

The proposed method simulates the natural degradation of printed icons on textured surfaces by blending an icon with a randomly selected paper texture. The approach ensures that the icon's appearance is influenced by the physical characteristics of the paper, such as darkness, creases, and smudges, to create realistic augmentations.

3.1 Paper Texture Analysis and Degradation

To replicate realistic degradation, we collected various textured paper sheets with different levels of wear. These textures exhibit characteristics such as darkness, creasing, and smudging, which influence how the icon blends with the background.

Darker paper reduces contrast and absorbs more light, affecting the visibility of the printed icon. As shown in Figure 1, varying levels of darkness impact the overall perception of detail, requiring adjustments to the blending process to ensure proper integration.



(a) Lightly darkened. (b) Moderately darkened. (c) Heavily darkened.

Fig. 1: Examples of paper with increasing levels of darkness.

Creased and scratched paper introduces irregular surface distortions, altering the way printed elements interact with the texture. As shown in Figure 2, creases and scratches disrupt the uniformity of the surface, affecting how elements blend with the background.

Smudges and blurring effects introduce soft, unpredictable distortions, altering the paper's texture. As shown in Figure 3, these variations create areas of differing opacity and contrast, mimicking the imperfections found in real-world printed materials.

3.2 Icon Integration and Blending Process

To integrate the icon into the selected texture, a random texture T is first selected from the `texture_pack` directory. The algorithm then computes a height map H from the chosen texture. This map represents variations in the surface texture and is derived by first converting the texture to grayscale and then normalizing

(a) Slight creases. (b) Moderate crumpling. (c) Severe wrinkling.

Fig. 2: Examples of crumpled and scratched paper with increasing intensity.

(a) Minor smudge. (b) Moderate blurring. (c) Severe distortion.

Fig. 3: Examples of blurring and distortion with increasing intensity.

its pixel values:

$$H(x,y) = \frac{T_{gray}(x,y) - \min(T_{gray})}{\max(T_{gray}) - \min(T_{gray})} \times 255. \tag{1}$$

This ensures that the values range from 0 (darkest regions) to 255 (brightest regions), allowing for consistent degradation modeling.

Next, the icon is randomly positioned within the texture to simulate natural placement:

$$x_0 \sim U(0, W_T - W_I), \quad y_0 \sim U(0, H_T - H_I), \tag{2}$$

where W_T, H_T are the dimensions of the texture and W_I, H_I are the dimensions of the icon.

The degradation effect is applied using a visibility factor α, which is randomly sampled within a predefined range:

$$\alpha \sim U(\alpha_{min}, \alpha_{max}). \tag{3}$$

Using this factor, a degradation map D is constructed from the corresponding height map region:

$$D(x,y) = \left(1 - \frac{H(x_0 + x, y_0 + y)}{255}\right) \cdot \alpha. \tag{4}$$

This equation ensures that higher regions in the height map contribute to stronger degradation effects.

The icon itself is then adjusted according to the computed degradation:

$$I' = I_{gray} \cdot D. \tag{5}$$

where I_{gray} represents the grayscale version of the icon, ensuring that degradation uniformly affects its intensity.

If the icon contains an alpha channel A_I, it is modified accordingly:

$$A'_I = A_I \cdot D. \tag{6}$$

This step ensures that transparency levels are also affected by the degradation process.

Finally, the blended image B is constructed by merging the degraded icon with the texture background using alpha compositing:

$$B(x,y) = I'(x,y) \cdot A'_I(x,y) + T(x_0 + x, y_0 + y) \cdot (1 - A'_I(x,y)). \tag{7}$$

After blending, a windowed extraction is performed to crop a region around the icon, incorporating additional padding to maintain a realistic blend with the background. The cropping boundaries are determined as:

$$x_{\min} = \max(0, x_0 - B_W), \quad x_{\max} = \min(W_T, x_0 + W_I + B_W) \tag{8}$$

$$y_{\min} = \max(0, y_0 - B_H), \quad y_{\max} = \min(H_T, y_0 + H_I + B_H) \tag{9}$$

where the final extracted image contains both the icon and additional background, preserving the natural degradation effect.

3.3 Implementation Details

The method is implemented using OpenCV and NumPy to efficiently perform image transformations and blending operations. The algorithm dynamically selects textures from the `texture_pack` directory, ensuring a variety of degradation patterns. The randomized positioning and degradation factors ensure that each augmented image exhibits unique characteristics, enhancing dataset diversity for robust training applications.

4 Experimental Setup and Evaluation

To assess the effectiveness of the proposed method, we conduct qualitative and quantitative evaluations of the generated images.

4.1 Experimental Setup

We generate a dataset of augmented icons using a diverse set of textures from `texture_pack`. The visibility factor α is varied within the predefined range, and random border expansions are applied. The output images are analyzed for their blending quality and degradation realism.

To ensure a fair evaluation, we consider multiple test cases with different texture types, including smooth, rough, and highly creased surfaces. Additionally, we vary the positioning of the icons to test the robustness of the algorithm against different spatial placements.

4.2 Qualitative Evaluation

Sample images generated using the proposed algorithm are visually examined to assess their consistency with real-world degradation effects. As shown in Figure , the blended icons exhibit natural variations in contrast, smudging, and background integration. The images are compared with real-world printed and degraded icons to validate their visual authenticity.

Furthermore, we analyze how different textures impact the final blended output. Icons placed on highly wrinkled textures show more severe degradation, whereas smoother textures yield more subtle blending effects. These qualitative insights help refine the augmentation process.

4.3 Quantitative Evaluation

To quantify the effectiveness of the blending, we compute similarity metrics such as the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) between the augmented icons and real-world degraded icons. Higher SSIM values indicate a closer resemblance to naturally degraded images, while PSNR provides an objective measure of signal distortion.

Additionally, we perform a perceptual evaluation by conducting a user study where participants rate the realism of the generated images on a Likert scale. The user study includes a diverse group of individuals who assess whether the augmented images convincingly mimic real-world degradation. Statistical analysis of the responses provides insights into the subjective quality of the augmentation technique.

4.4 Results and Discussion

The evaluation results demonstrate that the proposed method produces highly realistic degraded icons with minimal artifacts. The combination of texture-based blending and height-map-driven degradation ensures that the output images closely resemble real-world printed icons exposed to wear and environmental effects.

Table presents a summary of the SSIM and PSNR values for different texture categories. The results indicate that textures with higher roughness lead to lower SSIM scores, which aligns with the expectation that increased degradation reduces structural similarity. However, user study results suggest that these variations contribute positively to the perceived authenticity of the degradation effects.

The findings highlight the strengths of the proposed method in generating realistic augmentations while also identifying areas for further improvement, such as fine-tuning degradation intensity based on specific use cases.