PRINT-TO-CAMERA IMAGE RESTORATION USING CONDITIONAL ADVERSARIAL NETWORKS

Daniel L. Lau

Department of Electrical and Computer Engineering
University of Kentucky, Lexington, KY, USA

ABSTRACT

Machine vision systems used in print quality inspection face a fundamental challenge: camera-captured images of printed materials exhibit systematic degradations including color shifts, reduced contrast, and scanning artifacts that differ from the original digital source images. This paper presents a deep learning approach using conditional generative adversarial networks (cGANs) to transform raw camera captures back to their original digital appearance. We implement a U-Net based pix2pix architecture with perceptual loss to learn the inverse mapping from captured print images to pristine originals. Our bidirectional training framework enables both forward modeling (predicting print appearance) and reverse restoration (recovering original quality). Experimental results on paired print-capture datasets demonstrate effective artifact removal and color restoration, achieving 26.65 dB PSNR and 0.75 SSIM—an 8.6 dB improvement over the best traditional baseline. The system processes 512×512 images in real-time and requires only 35 minutes of training on consumer hardware.

Index Terms— image-to-image translation, generative adversarial networks, print quality inspection, machine vision, U-Net

1. INTRODUCTION

Industrial print quality inspection systems rely on machine vision cameras to capture images of printed materials for automated defect detection and quality assessment. A critical challenge in these systems is the systematic degradation introduced during the print-capture pipeline: even when printed images are visually near-perfect renditions of their digital sources, the captured images exhibit noticeable differences including color shifts, contrast reduction, geometric distortions, and sensor-specific artifacts.

These degradations arise from multiple sources in the imaging chain. The printing process itself introduces halftoning patterns, ink absorption variations, and substrate interactions. The capture process adds camera sensor noise, optical aberrations, and illumination non-uniformities. While individual degradations may be subtle, their cumulative effect creates a significant domain gap between digital originals and their printed-then-captured counterparts.

Traditional approaches to this problem include color calibration using reference targets, image registration with geometric correction, and histogram-based color mapping. However, these methods address individual degradations in isolation and struggle to model the complex, nonlinear interactions between printing and capture artifacts. Furthermore, they require manual tuning and domain expertise to achieve acceptable results.

We propose a data-driven approach using conditional generative adversarial networks (cGANs) to learn the complete inverse transformation from captured images to original digital sources. By training on paired examples of original images and their printed-captured counterparts, our model learns to jointly correct all degradations without explicit modeling of individual artifact sources.

Our key contributions are:

- A bidirectional pix2pix framework for print-to-camera image transformation that supports both artifact simulation (forward) and quality restoration (reverse)
- Integration of perceptual loss with adversarial training to preserve fine details and produce visually coherent results
- Demonstration of effective training on modest hardware with limited paired data

2. RELATED WORK

2.1. Image-to-Image Translation

The pix2pix framework introduced by Isola et al. [1] established conditional GANs as a general-purpose solution for paired image-to-image translation. The architecture combines a U-Net generator with skip connections and a PatchGAN discriminator that classifies local image regions. This approach has been successfully applied to tasks including semantic segmentation, colorization, and style transfer.

CycleGAN [2] extended this paradigm to unpaired training using cycle-consistency loss, enabling translation between domains without explicit correspondences. While

powerful, unpaired methods may introduce semantic changes inappropriate for quality-critical applications where pixel-accurate restoration is required.

Recent advances include attention mechanisms [3], multi-scale discriminators [4], and diffusion-based approaches [5]. However, the original pix2pix architecture remains effective for applications with well-aligned paired data.

2.2. Print Quality and Document Analysis

Document image enhancement has been extensively studied for applications including OCR preprocessing and historical document restoration. Hradis et al. [6] applied convolutional networks to document deblurring. More recent work has explored end-to-end learning for document binarization [7] and degradation removal.

Color management in print workflows traditionally relies on ICC profiles and colorimetric calibration [8]. While effective for average color accuracy, these approaches do not address spatially-varying artifacts or texture degradations introduced during capture.

2.3. Perceptual Loss Functions

Johnson et al. [9] demonstrated that optimizing in the feature space of pretrained networks produces more visually pleasing results than pixel-wise losses alone. Perceptual loss computed using VGG features has become standard in image restoration tasks, enabling networks to match high-level image structure while allowing flexibility in low-level details.

3. METHODOLOGY

3.1. Problem Formulation

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ denote an original digital image and $\mathbf{y} \in \mathbb{R}^{H \times W \times 3}$ its corresponding printed-and-captured version. We seek to learn a mapping $G : \mathbf{y} \to \mathbf{x}$ that restores the original image quality from the degraded capture.

For bidirectional modeling, we also train a forward model $G_f : \mathbf{x} \to \mathbf{y}$ that predicts how images will appear after printing and capture. Both models share identical architectures but are trained on data with swapped input-target roles.

3.2. Network Architecture

3.2.1. U-Net Generator

Our generator follows the U-Net architecture with symmetric encoder-decoder structure and skip connections. The encoder consists of four downsampling blocks that progressively extract features at resolutions 512, 256, 128, and 64 pixels. Each block applies two 3×3 convolutions with batch normalization and LeakyReLU activation (slope 0.2), followed by 2×2 max pooling.

The bottleneck operates at 32×32 resolution with 1024 channels. The decoder mirrors the encoder with transposed convolutions for upsampling. Skip connections concatenate encoder features with decoder activations at each resolution level, preserving spatial details that would otherwise be lost through the bottleneck.

The final layer is a 1×1 convolution producing 3-channel RGB output. The complete generator contains approximately 31 million parameters.

3.2.2. PatchGAN Discriminator

The discriminator follows the PatchGAN design that classifies 70×70 overlapping patches rather than the entire image. This approach provides dense gradient signal and captures high-frequency structure effectively.

The discriminator receives concatenated input and target images (6 channels) and applies four convolutional blocks with stride 2, expanding channels from 64 to 512. Batch normalization is omitted in the first layer. The final layer outputs a spatial map of patch classifications.

3.3. Loss Functions

The total generator loss combines three terms:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{L1} + \lambda_p \mathcal{L}_{perceptual} + \lambda_{adv} \mathcal{L}_{adv} \tag{1}$$

3.3.1. L1 Reconstruction Loss

The L1 loss encourages pixel-wise accuracy:

$$\mathcal{L}_{L1} = \mathbb{E}[\|\mathbf{x} - G(\mathbf{y})\|_1] \tag{2}$$

We weight this term heavily ($\lambda_1 = 100$) to ensure faithful reconstruction.

3.3.2. Perceptual Loss

Perceptual loss measures similarity in VGG-19 feature space:

$$\mathcal{L}_{perceptual} = \sum_l \|\phi_l(\mathbf{x}) - \phi_l(G(\mathbf{y}))\|_1 \tag{3}$$

where ϕ_l extracts features from VGG layers relu1_2, relu2_2, and relu3_4. We use $\lambda_p = 10$.

3.3.3. Adversarial Loss

The adversarial loss follows the standard GAN formulation:

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(\mathbf{y}, G(\mathbf{y})))] \tag{4}$$

with $\lambda_{adv} = 1$. The discriminator is trained to maximize classification accuracy on real versus generated pairs.

Table 1. Training Configuration

Parameter	Value
Image Resolution	512×512
Batch Size	4
Training Steps	10,000
Learning Rate	2×10^{-4}
L1 Weight (λ_1)	100
Perceptual Weight (λ_p)	10
Adversarial Weight (λ_{adv})	1
Optimizer	AdamW
Precision	FP16 (mixed)

3.4. Training Procedure

Training alternates between discriminator and generator updates. For each iteration, we first update the discriminator on a batch of real pairs (label 1) and generated pairs (label 0). The generator is then updated to minimize the combined loss while keeping discriminator weights frozen.

We use AdamW optimization with learning rate 2×10^{-4} and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Cosine annealing reduces the learning rate to 1% of initial value over training. Mixed-precision training (FP16) accelerates computation and reduces memory usage.

4. EXPERIMENTS

4.1. Dataset

Our dataset consists of paired images where each original digital image has a corresponding version that was printed and recaptured using an industrial machine vision camera. The source images are drawn from the Kaggle Dogs vs. Cats dataset [10], providing diverse content with varying colors, textures, and lighting conditions suitable for evaluating image restoration quality. Images are stored as TIFF files to preserve quality, with pairing established through numeric identifiers in filenames.

The complete dataset contains 16,193 paired images. All images are resized and center-cropped to 512×512 pixels. Pixel values are normalized to $[-1, 1]$. We use a 90/10 train/validation split with fixed random seed for reproducibility, yielding 14,574 training images and 1,619 validation images.

4.2. Implementation Details

Training was performed on an NVIDIA RTX 4070 Ti GPU with 12GB memory. Batch size of 4 images fits comfortably in memory with mixed-precision training. Complete training of 10,000 steps requires approximately 35 minutes. Table 1 summarizes key parameters.

Table 2. Quantitative Comparison on Validation Set

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Pix2Pix (Ours)	**26.65**	**0.7454**	**0.2483**
Channel Regression	18.07	0.4244	0.5445
Histogram Matching	16.79	0.3210	0.6099
Reinhard Color Transfer	16.43	0.3998	0.5856
Identity	10.59	0.1962	0.5899

(a) Original (b) Captured (c) Restored

Fig. 1. Image restoration results. (a) Original digital image. (b) Camera-captured image after printing, showing reduced contrast and color degradation. (c) Restored image produced by our model.

Checkpoints are saved every 2,000 steps, enabling selection of optimal model based on validation performance. Visual results on held-out samples are generated at each checkpoint for qualitative assessment.

4.3. Quantitative Results

We evaluate our model on the validation set (1,619 images) using three standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [11], and Learned Perceptual Image Patch Similarity (LPIPS) [12]. We compare against four baseline methods: Identity (no transformation), Reinhard Color Transfer [13], per-channel Linear Regression, and Histogram Matching.

Table 2 shows that our method significantly outperforms all baselines across all metrics. Compared to the best traditional baseline (Channel Regression), our model achieves an 8.6 dB improvement in PSNR, a 0.32 improvement in SSIM, and a 0.30 reduction in LPIPS. The large gap between Identity and other methods confirms the substantial degradation introduced by the print-capture process. Traditional color transfer methods partially address color shifts but fail to restore fine details and texture, as evidenced by their high LPIPS scores.

4.4. Qualitative Results

Visual inspection of results demonstrates effective artifact removal across diverse image content (Fig. 1). The reverse model (captured → original) successfully corrects:

- **Color shifts**: Restores accurate color reproduction

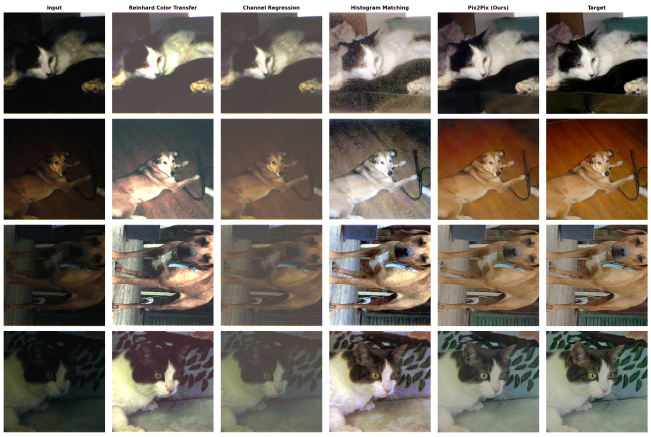

Fig. 2. Visual comparison of restoration methods. Left to right: Input (camera capture), Reinhard Color Transfer, Channel Regression, Histogram Matching, Pix2Pix (Ours), and Ground Truth.

Table 3. Effect of Loss Components

Loss Configuration	Sharpness	Color	Artifacts
L1 only	Medium	Good	Some
L1 + Perceptual	High	Good	Few
L1 + GAN	High	Medium	Few
Full (L1 + Perc + GAN)	High	Good	Minimal

from captures exhibiting warm or cool bias

- **Contrast reduction**: Recovers dynamic range compressed during printing
- **Dark artifacts**: Removes vignetting and uneven illumination patterns
- **Detail loss**: Sharpens edges softened by the print-capture process

The forward model (original $\rightarrow$ captured) learns to simulate print degradations, which serves as a useful augmentation tool and validates that the learned transformations are meaningful.

Fig. 2 compares restoration methods on sample validation images. Our method produces results most visually similar to the original images, recovering accurate colors and contrast that traditional methods cannot achieve.

4.5. Ablation Study

Table 3 summarizes qualitative effects of different loss configurations. L1 loss alone produces blurry results. Adding perceptual loss improves sharpness and detail preservation. The adversarial component further enhances high-frequency content and produces more natural-looking textures. The combination of all three losses yields the best overall results.

4.6. Computational Efficiency

Inference time for a single 512×512 image is approximately 15ms on the RTX 4070 Ti, enabling real-time processing at over 60 frames per second. The model's 31M parameters result in a checkpoint size of approximately 125MB, suitable for deployment in resource-constrained environments.

Memory usage during training peaks at approximately 6–8GB with batch size 4, making the approach accessible on consumer-grade GPUs. The mixed-precision training reduces memory footprint by roughly 40% compared to FP32.

5. CONCLUSION

We presented a conditional GAN approach for restoring machine vision camera captures of printed images to their original digital quality. The pix2pix architecture with U-Net generator and PatchGAN discriminator, combined with perceptual and adversarial losses, effectively learns the inverse print-capture transformation from paired training data. Quantitative evaluation demonstrates significant improvements over traditional baselines, achieving 26.65 dB PSNR and 0.75 SSIM compared to 18.07 dB and 0.42 for the best traditional method.

Our bidirectional framework supports both quality restoration and degradation simulation, enabling applications in print inspection, color management, and data augmentation. The system trains efficiently on consumer hardware and processes images in real-time, making it practical for industrial deployment.

Future work includes extending the approach to higher resolutions, investigating unpaired training with cycle-consistency for scenarios where paired data is unavailable, and incorporating explicit modeling of known degradation sources such as halftone patterns and illumination gradients.

6. REFERENCES

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[2] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[3] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao, "Attention-gan for object transfiguration in wild images," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 164–180.

[4] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8798–8807.

[5] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH*, 2022, pp. 1–10.

[6] Michal Hradis, Jan Kotera, Pavel Zemcik, and Filip Sroubek, "Convolutional neural networks for direct text deblurring," in *British Machine Vision Conference (BMVC)*, 2015.

[7] Chris Tensmeyer and Tony Martinez, "Document image binarization with fully convolutional neural networks," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 99–104.

[8] Abhay Sharma, *Understanding Color Management*, Wiley, Hoboken, NJ, USA, 2nd edition, 2018.

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 694–711.

[10] Kaggle, "Dogs vs. cats," `https://www.kaggle.com/c/dogs-vs-cats`, 2013.

[11] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[13] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.