Document image shadow removal based on Vision Transformers

21521604
Nguyen Quoc Truong
Supervisor: Ph.D Nguyen Duy Khanh

Table of contents

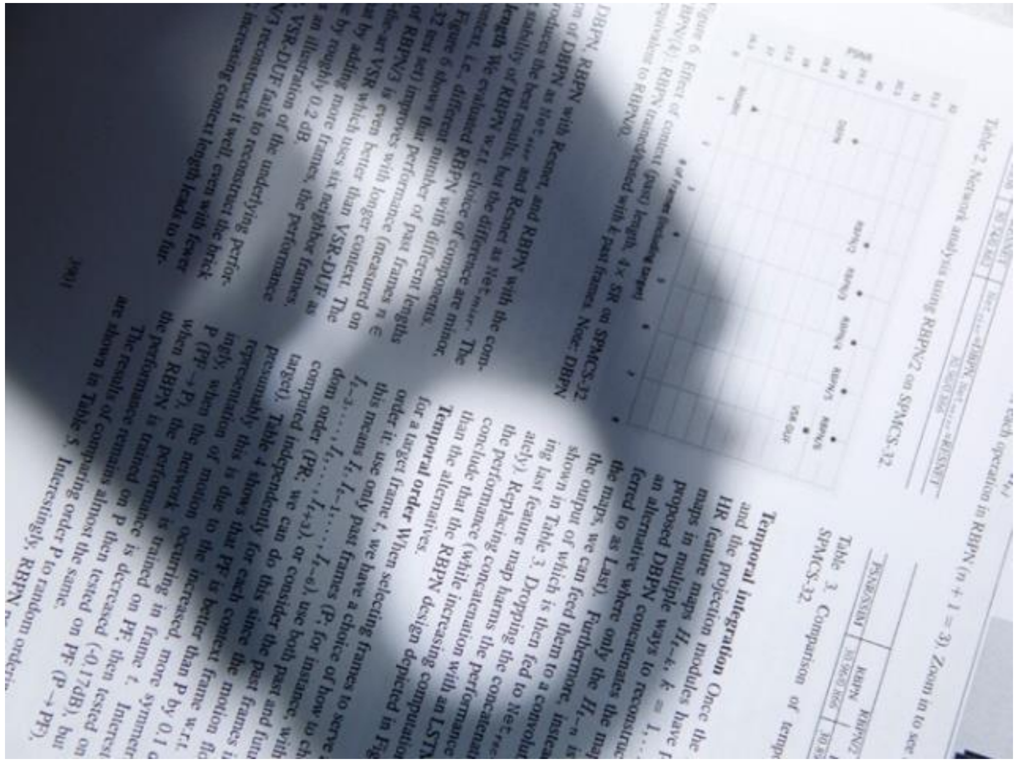
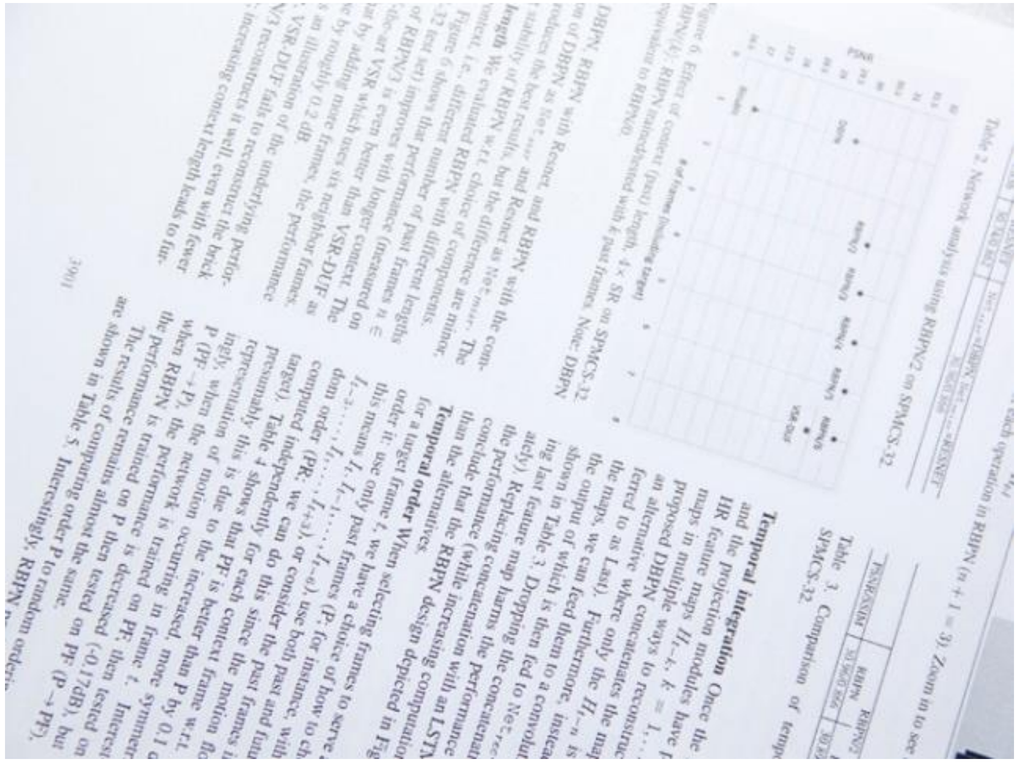
Introduction

Document image shadow removal

- **Document image shadow removal** is the process of image processing aimed at eliminating unwanted shadow regions that appear on photographs of paper documents.

- These shadows make it difficult to read and automatically process the text on the documents.

Input - Output

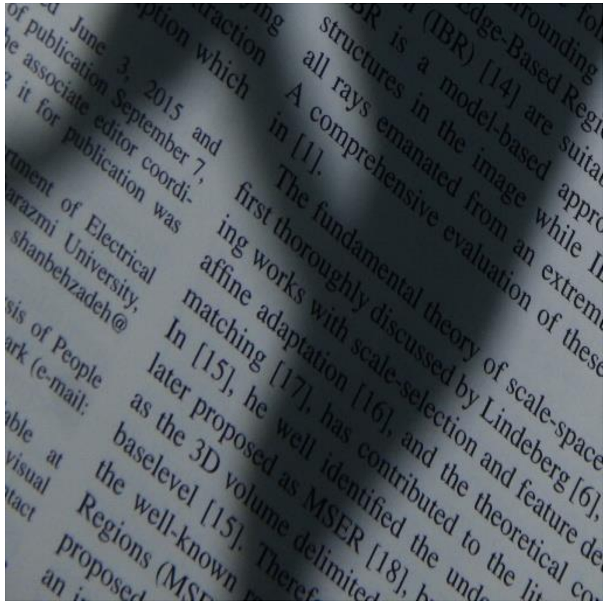
 

Motivation

- Improving Readability
- Enhancing OCR Accuracy
- Improving Image Quality

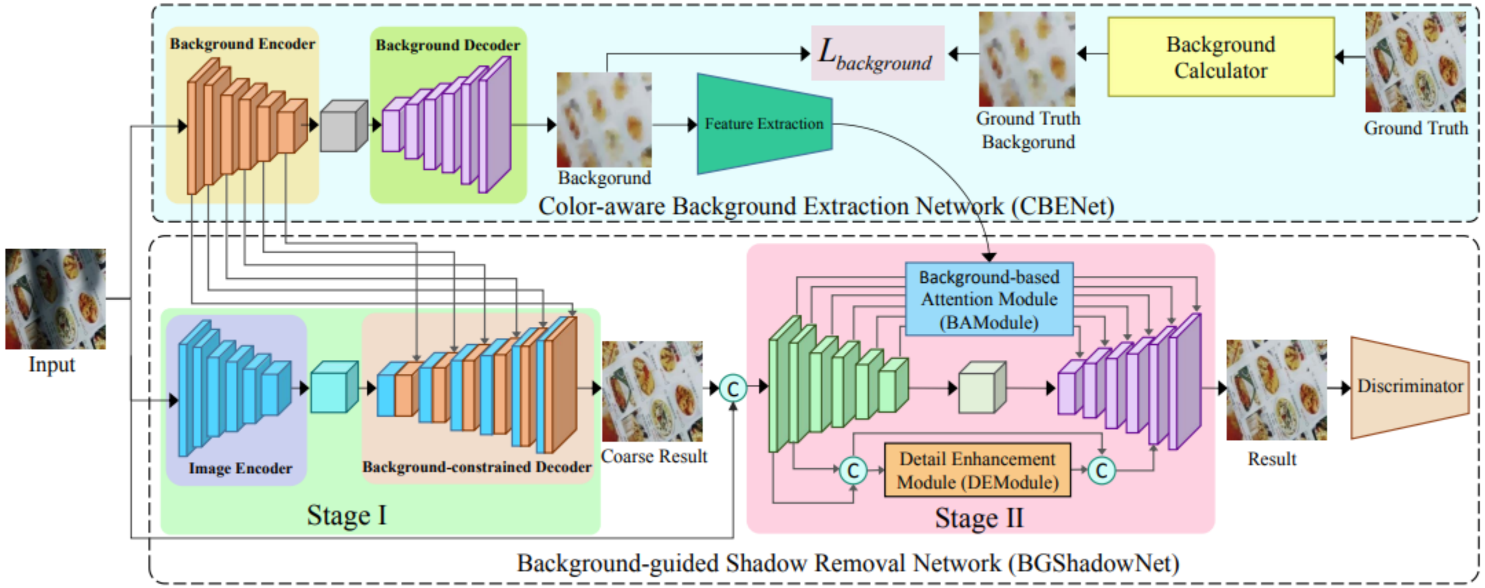
Challenges

- High Contrast
- Detail Preservation
- Complexity of Images



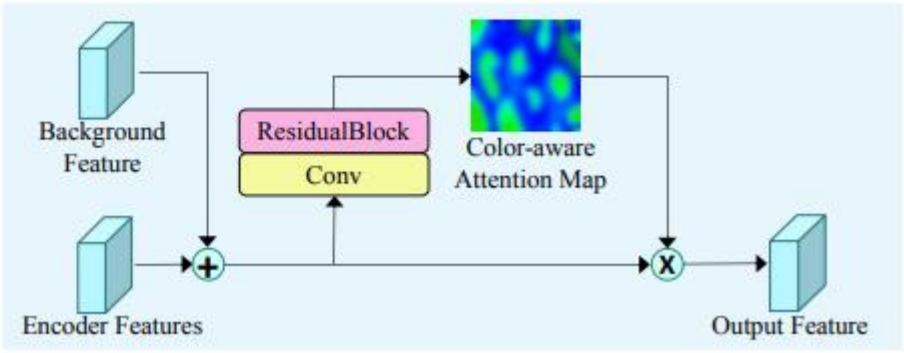
Methods

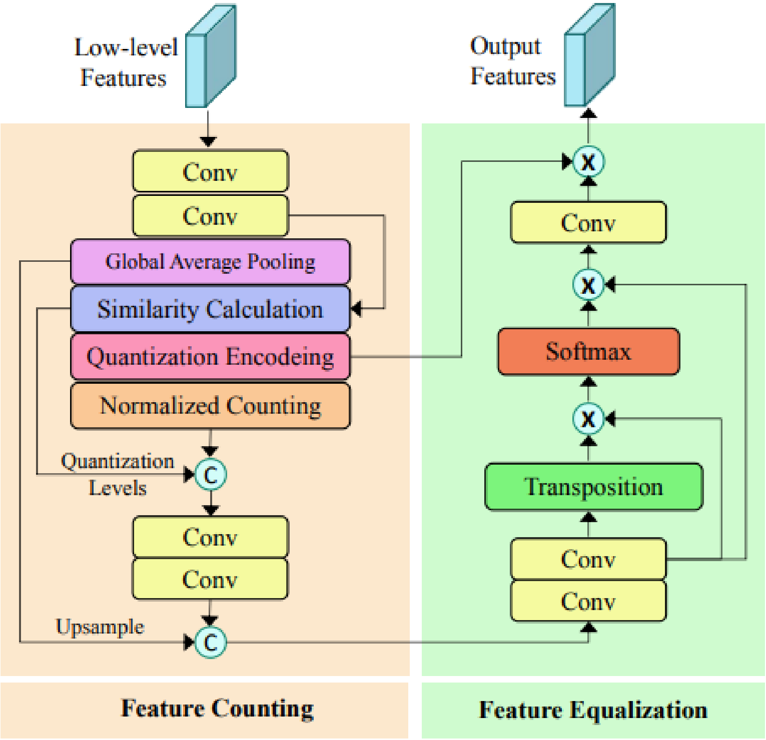
Document image shadow removal guided by Color-aware Background



- CBENet
- BGShadowNet
  - Stage I
  - Stage II
    - BAModule
    - DEModule

BGShadowNet – Stage II



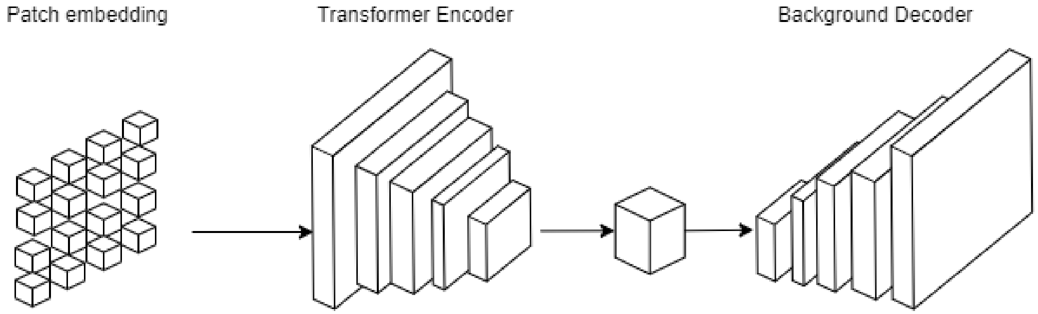


Vision Transformers

- ViTs is developed based on the Transformer architecture.
- Divide the input image into fixed-size patches then flatten each into a vector.
- Patch embedding: transform the patch vectors into fixed-size embedding vectors.
- Transformer Encoder: process and learn complex features.

CBENet based on Vision Transformers

- Idea: divide images into 16x16 patches ⟶ calculations and training
- Replace conv layers in U-Net with ViTs Encoder layers

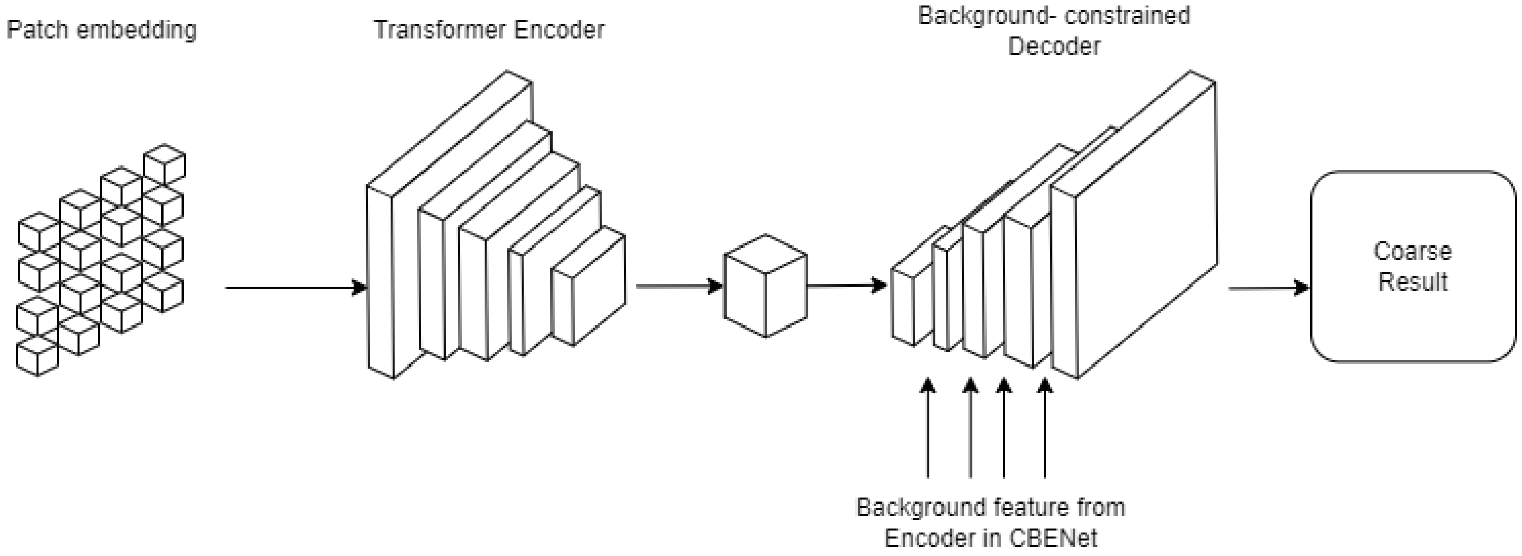


Proposed methods

BGShadowNet based on Vision Transformers

Stage I

- Similar to the CBENet
- Use the background features from ViTs encoder in CBENet as additional features to train Decoder.



Proposed methods

Proposed Methods

BGShadowNet based on Vision Transformers

Stage II (BAM & DEM)

Adjustments at this stage have shown ineffective results → No changes

Loss function

Background reconstruction loss

$$\mathcal{L}_{background} = \|B - \hat{B}\|$$

Structure consistency loss

$$\mathcal{L}_{structure} = \lambda_3 \|VGG(I_{gt}) - VGG(I_{free})\|^2$$

Appearance consistency loss

$$\mathcal{L}_{appearance} = \lambda_1 \mathcal{L}_{coarse} + \lambda_2 \mathcal{L}_{final}$$
$$= \lambda_1 \|I_{gt} - I_{coarse}\| + \lambda_2 \|I_{gt} - I_{free}\|$$

Adversarial loss

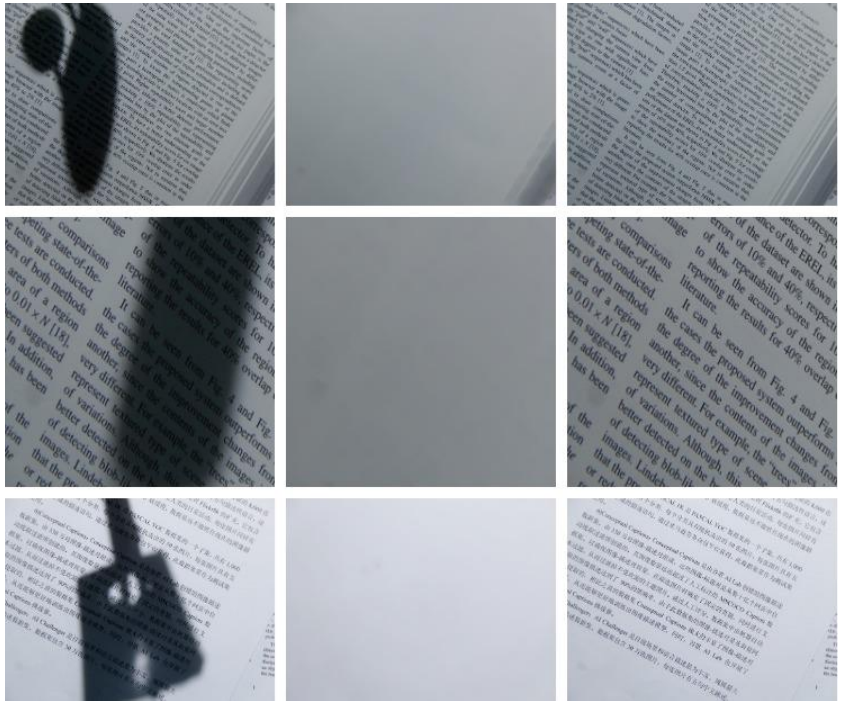
$$\mathcal{L}_{adv} = \lambda_4 \mathbb{E}_{(I, I_{free}, I_{gt})} \left[\log\left(D(I_{gt})\right) + \log(1 - D(I))\right]$$

Experiments

Dataset

RDD dataset

- 4916 pairs of shadow, background and shadow_free images
- Training set: 4371
- Test set: 545



Setting

- CBE and BGS are trained separately.
- Epochs: 200
- Optimizer: Adam
- Learning rate: 0.0004
- Weight parameters: λ_1, λ_2, λ_3 and λ_4 set to 1, 1, 0.05 and 0.01

Metrics

RMSE

$$RMSE = \sqrt{MSE}$$

PSNR

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(I_{pred}(i) - I_{gt}(i))^2$$

SSIM

$$SSIM(x, y) \\ = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma}$$

Evaluation

Models	RMSE	PSNR	SSIM
CBENet + BGShadowNet	2.572	36.301	0.975
CBETransformer + BGSTransformer	2.477	36.624	0.977
CBETransformer + BGShadowNet	2.583	36.263	0.972
CBENet + BGSTransformer	2.639	36.077	0.968

Evaluation

Methods	RMSE	PSNR	SSIM
My method(*)	2.477	36.624	0.977
BGShadowNet(*)	2.572	36.301	0.975
BGShadowNet	2.219	37.585	0.983
BEDSR-Net	2.937	34.928	0.973
Bako	14.648	20.741	0.894
Jung	30.190	14.364	0.861

Conclusion

Limitation

- High Resource Demand
- Training Time
- Data Diversity
- Language and Handwriting
- Generalization

Conclusion

- Effective Shadow Removal: Significantly improves image quality over traditional methods.
- Enhanced Image Accuracy and Clarity.
- Reduces recognition errors caused by shadows.
- Global Context Handling: Efficiently captures and processes complex shadow variations.
- Automation: Minimizes human intervention, increasing efficiency and practical application.

Thank you