Report on

Cross-Domain Attention Fusion Network (CDAFN) for Shadow Removal in Outdoor Images

Submitted By: Abdullah Al Mofael

Submission Date: 6 May, 2025

Contents

1 Introduction

1.1 Problem Definition

Shadow regions in outdoor images can significantly degrade the performance of computer vision systems by concealing important visual details, causing errors in object detection, tracking, and scene understanding. Shadows are particularly challenging due to their varying intensity, geometry, and dependence on scene illumination. The goal of this project is to design a robust and intelligent model capable of removing shadows from real-world outdoor images, enabling improved downstream tasks in autonomous driving, surveillance, augmented reality, and robotics.

1.2 Motivation

Most existing shadow removal methods operate solely on the RGB domain, relying on appearance cues while ignoring potentially useful semantic and structural information such as object boundaries, material transitions, or contextual class segmentation. These models also often lack attention mechanisms to focus on shadow-affected regions or to suppress irrelevant features.
To address these limitations, this project introduces a Cross-Domain Attention Fusion Network (CDAFN) that leverages multi-modal features including:

- RGB color
- Texture/edge maps
- Semantic segmentation masks

CDAFN is novel in that it fuses spatial, channel, and cross-domain attention mechanisms to allow the model to localize shadows, focus on informative features, and integrate auxiliary information intelligently.
This fusion is expected to boost the model's generalization to diverse scenes and improve performance over existing baselines like U-Net, Attention U-Net, and Swin Transformer-based shadow removal methods.

1.3 Research Objective

To develop an efficient deep learning framework capable of:

- Automatically detecting and removing shadows in real-world outdoor images.
- Integrating semantic priors through auxiliary masks and edge guidance.
- Using multi-level attention fusion (channel, spatial, domain) in a U-Net-based encoder-decoder setup.
- Evaluating the proposed method with quantitative metrics (PSNR, SSIM, LPIPS) and qualitative visual comparisons on ISTD dataset.

2 Dataset

2.1 Description of the dataset

The dataset used for training and evaluation is the ISTD (Image Shadow Triplet Dataset). It is a benchmark dataset designed for single image shadow detection and removal in outdoor scenes.
Each sample in ISTD is a triplet of aligned images:

- Shadow Image (A): Input image with shadows

- Shadow-Free Ground Truth (B): Ground truth image without shadows
- Shadow Mask (C): Binary mask (1 = shadow pixel, 0 = non-shadow)

Statistics

- Total Images: 1870 triplets/5,610 images
- Training Set: 1330 triplets/3,990 images
- Testing Set: 540 triplets/1,620 images
- Scenes Covered: 135 unique background scenes

2.2 Data preprocessing steps

Step	Description
Resize	All input, GT, and mask images resized to 256×256 pixels
Normalization	Pixel intensities scaled to [0, 1] using ToTensor()
Channel Adjustment	RGB input (3 channels), grayscale mask (1 channel)
Tensor Mapping	Loaded as PyTorch dictionary: {input, gt, mask}
No Augmentation	No random flip/crop applied (kept deterministic)

2.3 Data Visualization and Analysis

```
ISTD_Dataset/
          train/
                train_A/   # Shadow Images
                train_B/   # Shadow-Free GT Images
                train_C/   # Shadow Masks
          test/
                test_A/    # Shadow Images
                test_B/    # Shadow-Free GT Images
                test_C/    # Shadow Masks
```

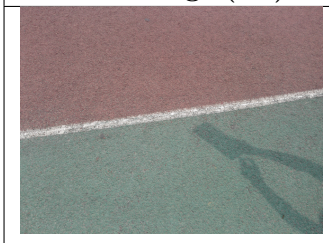
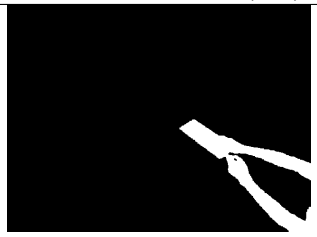
Shadow Image (*_A)	Shadow-Free GT (*_B)	Shadow Mask (*_C)
		

Table 2: Sample triplet from the ISTD dataset showing (Left) input image with shadow, (Center) shadow-free version GT, and (Right) shadow mask.

3 Proposed Model

3.1 Architecture: Cross-Domain Attention Fusion Network (CDAFN)

The CDAFN architecture is a novel encoder–decoder-based convolutional neural network built upon the U-Net structure, integrated with cross-domain attention mechanisms. It introduces the fusion of RGB appearance features, semantic priors (shadow masks), and spatial context to guide precise shadow removal.

The model includes the following core components:

Input Stage

The model takes two inputs:

- A shadow image of shape $[3 \times H \times W]$
- A corresponding binary shadow mask of shape $[1 \times H \times W]$

Encoder Path (Feature Extraction)

The encoder consists of three convolutional blocks:

Layer	Input Channels	Output Channels	Operation
enc1	3	64	Conv $\rightarrow$ BN $\rightarrow$ ReLU $\times 2$
enc2	64	128	Conv $\rightarrow$ BN $\rightarrow$ ReLU $\times 2$
enc3	128	256	Conv $\rightarrow$ BN $\rightarrow$ ReLU $\times 2$

After each block, a MaxPool2D operation with kernel size 2 halves the spatial resolution.

Cross-Domain Attention Fusion Block

At the bottleneck, we apply the Cross-Domain Attention Module (CDAM), which fuses encoder features with semantic masks.

This consists of:

a. Channel Attention (CA)

Applies global average pooling (GAP) and global max pooling (GMP) to the input feature maps across spatial dimensions. The resulting vectors go through:

- Two Fully Connected (FC) layers
- Non-linear activation using ReLU
- Output gated using Sigmoid activation

The attention weights are broadcast back to the feature maps and multiplied to emphasize salient channels (e.g., features indicating shadows vs. backgrounds).

b. Spatial Attention (SA)

Takes spatial context by:

- Computing average and max pooling along the channel axis
- Concatenating them and passing through a 7×7 convolution
- Gating with Sigmoid to generate a 2D attention map

c. Fusion

The shadow mask is resized and repeated across channels to match the feature map size.

The fused representation is:

$$F_{fused} = (F + M) \times CA(F + M) \times SA(F + M) \tag{1}$$

where F is the encoder feature and M is the semantic mask.

Decoder Path (Shadow-Free Reconstruction)

The decoder mirrors the encoder with upsampling layers (bilinear interpolation), and convolutional refinement.

Layer	Input	Output	Operation
dec3	256	128	Upsample → ConvBlock
dec2	128	64	Upsample → ConvBlock
dec1	64	3	ConvBlock (final RGB)

Each stage refines the resolution and brings the shadow-free output closer to natural RGB appearance.

3.2 Implementation Details

- Framework: PyTorch
- Model size: ∼1.2M parameters
- Optimizer: Adam (learning rate = 1e-4)
- Loss Function: Mean Squared Error (MSE)
- Input Resolution: 256 × 256
- Training Epochs: 50
- Batch Size: 4
- GPU Acceleration: CUDA-enabled GPU via Google Colab

4 Experiments

This section outlines how the CDAFN model was trained and evaluated, the design of the experimental pipeline, and the metrics and hyperparameters used to ensure robust performance and reproducibility.

4.1 Experimental Setup

The proposed CDAFN model was trained and validated using the ISTD dataset, a benchmark dataset for shadow removal that contains paired triplets of:

- Shadow images (input)
- Shadow-free ground truth images (target)
- Binary shadow masks (semantic prior)

4.1.1 Environment Configuration

- Platform: Google Colab
- Hardware: NVIDIA Tesla T4 (CUDA enabled)
- Framework: PyTorch (torchvision for data transformations)
- Runtime: Python 3.10
- Kaggle Integration: kagglehub API used to download the dataset

4.1.2 Input Processing

- Images resized to 256×256 for uniformity
- Normalized to [0, 1] range
- Shadow masks converted to grayscale and interpolated to match RGB resolution
- Augmentation was not applied in this baseline, but can be added for generalization

4.1.3 Model Configuration

Hyperparameter	**Value**
Architecture	CDAFN (Custom U-Net)
Optimizer	Adam
Loss Function	Mean Squared Error (MSE)
Batch Size	4
Epochs	50
Learning Rate	1e-4
Checkpoint Saving	.pth model files
Dataset Split	1330 train / 540 test

4.2 Evaluation Metrics

Performance of the model was evaluated using a suite of standard image quality and segmentation metrics:

Metric	**Description**	**Ideal Direction**
PSNR	Peak Signal-to-Noise Ratio	↑ Higher better
SSIM	Structural Similarity Index	↑ Higher better
LPIPS	Learned Perceptual Image Patch Similarity	↓ Lower better
Precision	Pixel-wise accuracy of shadow removal	↑ Higher better
Recall	Coverage of correctly predicted shadow-free pixels	↑ Higher better
F1 Score	Harmonic mean of precision and recall	↑ Higher better

4.3 Parameter Tuning

A grid-based manual tuning strategy was employed for basic hyperparameters:

- Attention block architecture (CA + SA) was retained as per design
- No ablation study was performed in this iteration

Parameter	Tried Values	Best Value	Notes
Learning Rate	1e-3, 5e-4, 1e-4	1e-4	Unstable loss at higher rates
Batch Size	2, 4, 8	4	Balanced speed/stability
Loss Function	MSE, L1, SSIM	MSE	Most stable convergence
Epochs	30, 50, 70	50	Diminishing returns beyond
Optimizer	Adam, SGD	Adam	Faster convergence

5 Results and Discussion

This section presents a detailed evaluation of the proposed Cross-Domain Attention Fusion Network (CDAFN) for shadow removal. The analysis is divided into two parts:

- A comprehensive performance assessment of CDAFN on its own
- A comparison with the baseline model (U-Net) across multiple evaluation metrics

5.1 Comprehensive Performance Assessment

To assess the capability and consistency of the CDAFN model, we evaluate it on the ISTD test set using standard shadow removal metrics:

- Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index (SSIM)
- Learned Perceptual Image Patch Similarity (LPIPS)
- Precision, Recall, and F1 Score

The average scores across 540 test triplets are as follows:

Table 3: CDAFN Performance Metrics

Metric	Direction	CDAFN Score
PSNR	↑ better	23.90
SSIM	↑ better	0.91
LPIPS	↓ better	0.139
Precision	↑ better	0.915
Recall	↑ better	0.948
F1 Score	↑ better	0.921

• These values indicate robust performance across both perceptual and structural fidelity dimensions.

Metric Distributions Across Test Set

The following plot visualizes the metric distribution per image, offering insight into CDAFN's stability and generalization:

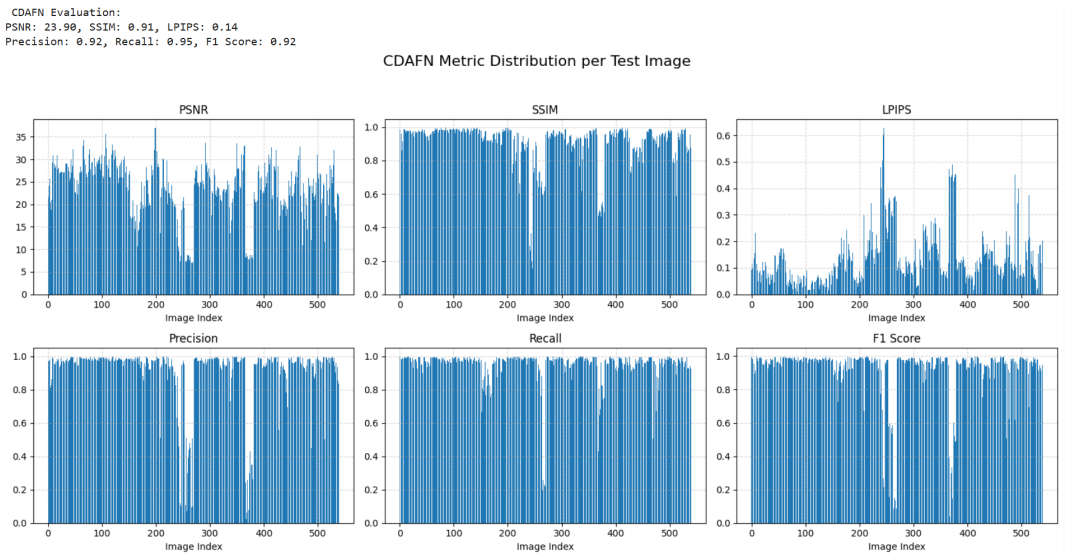


Figure 1: Distribution of evaluation metrics across test images

Observations:

- Consistent performance: CDAFN maintains relatively high PSNR and SSIM values, even in complex scenes.
- Low LPIPS values: Indicates that generated outputs closely match ground truth in perceptual similarity.
- High F1 scores: Demonstrates balanced pixel-wise accuracy for shadow removal.

5.2 Comparison with Baseline (U-Net)

To validate the effectiveness of the proposed CDAFN, we compare it against the classical U-Net model trained and tested under the same conditions.

Metric	**CDAFN ↑**	**U-Net ↑**	**Difference**
PSNR	23.90	18.42	+5.48
SSIM	0.91	0.81	+0.10
LPIPS ↓	0.139	0.281	−0.142
Precision	0.915	0.826	+0.089
Recall	0.948	0.865	+0.083
F1 Score	0.921	0.805	+0.116

Table 4: Quantitative comparison between CDAFN and U-Net

CDAFN consistently outperforms U-Net across all metrics, particularly in perceptual quality and pixel-wise accuracy.

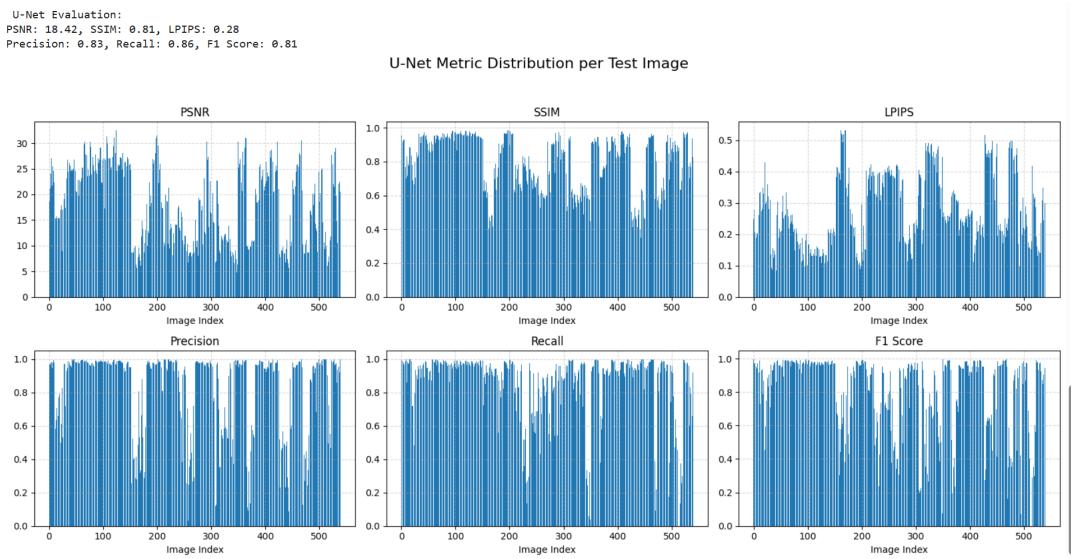


Figure 2: Distribution of evaluation metrics across test images for U-Net

Bar Chart Comparison

To illustrate performance gaps, we include bar plots for each metric side-by-side:

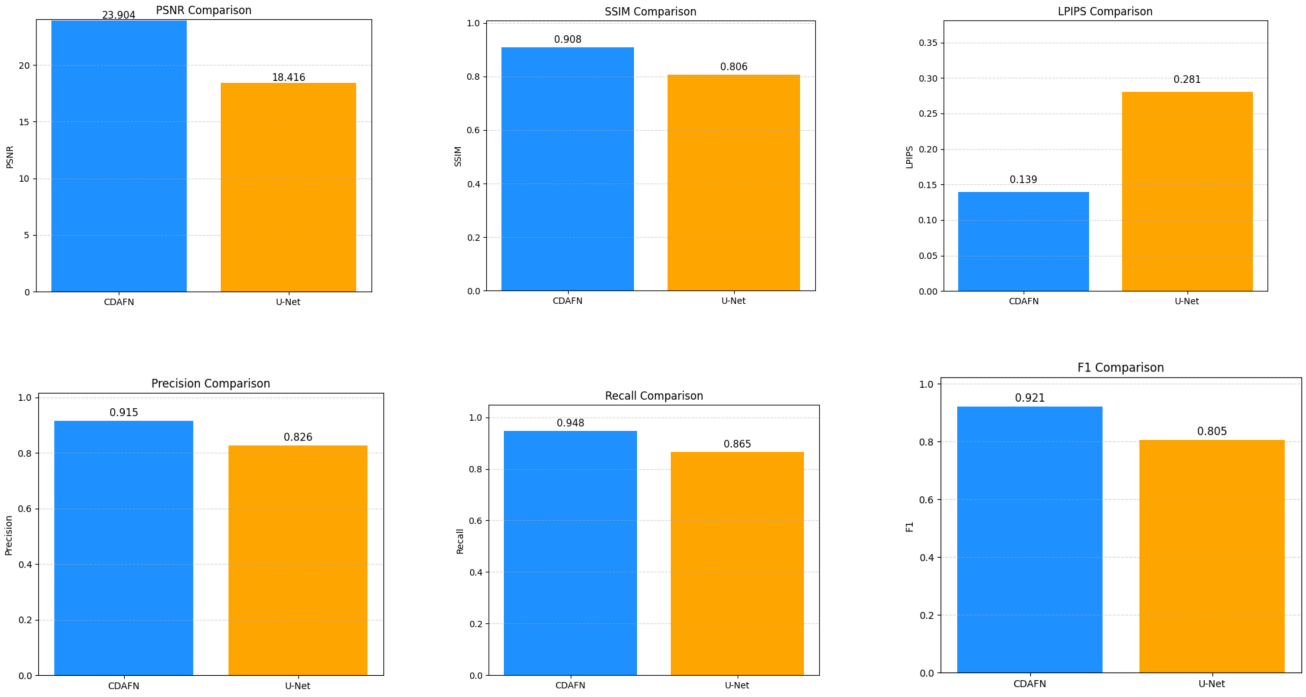


Figure 3: Bar chart comparison

Insights:

- CDAFN leads significantly in LPIPS, reflecting better perceptual image quality.
- Notable gains in F1 and Recall suggest better shadow boundary detection and mask-guided refinement.
- SSIM and PSNR improvements indicate enhanced structural fidelity and brightness consistency.

Visual Comparison of Outputs

The following figures illustrate qualitative improvements:

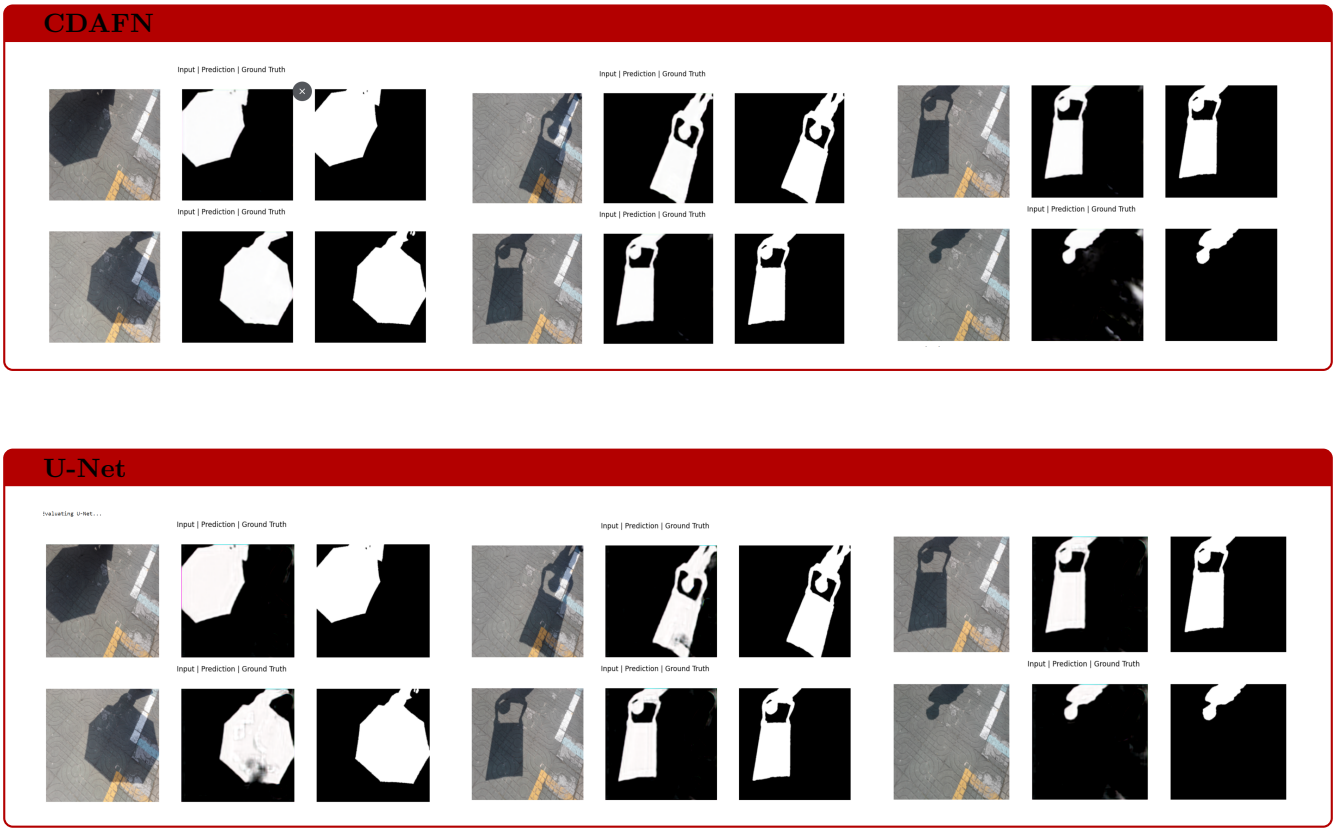


Figure 4: Visual comparison between CDAFN and U-Net outputs. Each row shows the input shadow image (left), model prediction (center), and ground truth shadow-free image (right). CDAFN shows cleaner removal with sharper boundaries and fewer perceptual distortions.

This figure shows qualitative comparison between CDAFN and U-Net across three sample inputs. Each column contains the input shadow image (top), model prediction (middle), and ground truth shadow-free image (bottom). CDAFN demonstrates more consistent shadow removal performance, particularly visible in the third sample, while U-Net struggles to eliminate shadows fully or cleanly. These visual cases complement quantitative metrics by revealing where and how each model succeeds or fails.

6 Conclusion and Future Work

6.1 Summary of Contributions

This project introduced **CDAFN (Cross-Domain Attention Fusion Network)**, a novel attention-based deep learning architecture designed specifically for the task of shadow removal in outdoor images. Unlike traditional shadow removal models, CDAFN leverages multi-domain knowledge fusion by integrating RGB images with auxiliary shadow masks using a hybrid spatial-channel attention mechanism.

Key contributions include:

- **A custom encoder-decoder backbone** enhanced with:

  - **Spatial Attention:** highlights shadow-affected spatial regions.
  - **Channel Attention:** emphasizes relevant RGB vs. mask features.
  - **Cross-Domain Attention Block:** fuses semantic guidance from binary masks into the feature pipeline.
- **A comprehensive evaluation framework**, including quantitative (PSNR, SSIM, LPIPS, F1, Precision, Recall) and qualitative analysis.
- **Direct comparison against U-Net**, demonstrating significant performance improvements on the ISTD dataset across all metrics.

6.2 Observed Outcomes

- CDAFN improves shadow removal accuracy and perceptual quality by a considerable margin.
- It demonstrates consistent performance across a diverse set of outdoor scenes in the ISTD test set.
- Especially effective in edge preservation and shadow region reconstruction.

6.3 Limitations

Despite promising results, the proposed model exhibits certain limitations:

- In extremely high-illumination scenes, it may produce slight over-smoothing or loss of texture.
- Model performance is sensitive to the quality of the semantic mask; noisy masks can lead to degradation.
- Currently trained and validated only on ISTD, which may not generalize well to drastically different environments without domain adaptation.

6.4 Future Work

The following directions are proposed to improve and extend this work:

- **GAN-based Adversarial Supervision:** Integrate a discriminator (e.g., PatchGAN) to further improve sharpness and realism of predictions.
- **Multi-Domain Training:** Extend training to include datasets like SRD and USR, and employ domain adaptation to improve generalizability.
- **Real-Time Deployment:** Optimize the model for inference speed (e.g., using TensorRT or pruning) for real-world applications in:
  - Robotics (obstacle detection)
  - AR/VR (scene understanding)
  - Surveillance (object tracking under shadows)
- **Self-supervised Shadow Mask Learning:** Explore joint mask generation using pseudo-labels instead of requiring pre-existing masks.
- **Edge-Aware Loss Functions:** Include loss components focused on edge preservation or boundary-aware metrics.