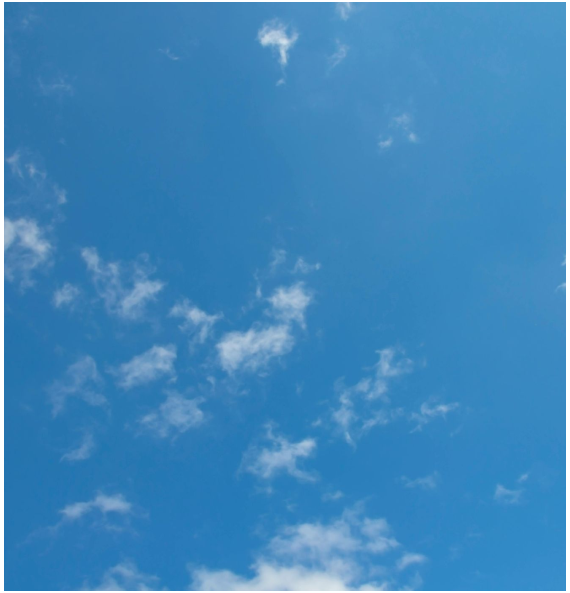
Image-Difficulty-Aware Evaluation of Super-Resolution Models

Atakan Topaloğlu*, Ahmet Bilican*, Cansu Korkmaz, A. Murat Tekalp

*** Denotes equal contribution**

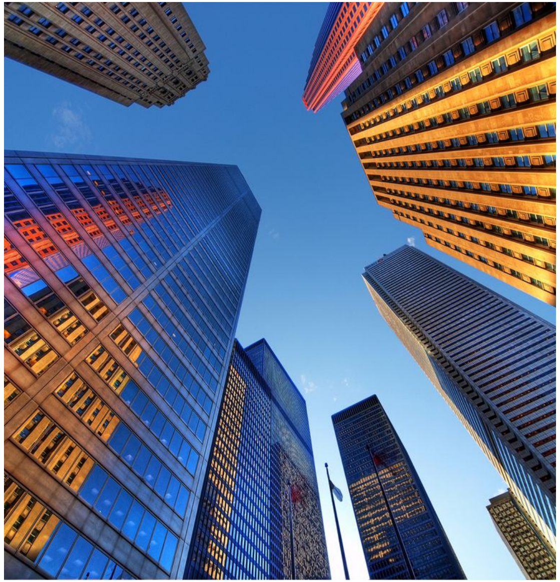
This work was supported in part by an AI Fellowship to Cansu Korkmaz by the KUIS AI Center. A. Murat Tekalp acknowledges support from Turkish Academy of Sciences (TUBA).

Problem 1: Not All Images Are Created Equal



An open sky doesn't have much high frequency content, making it easy to super-resolve.

Animal fur is notoriously difficult to super resolve, because it has irregular, high frequency, **texture**d content.



Buildings are also hard to super resolve, because they have sharp **edge**s prone to artifacts.

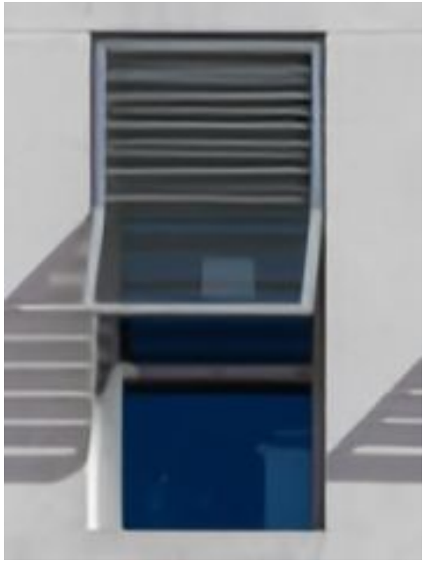
Problem 2: Localization - Average Scores Can Be Deceptive

HR Crop

Crop of Image 65 from LSDIR Validation Set

SR Crop from Model A

Average PSNR
26.1842 dB



SR Crop from Model B

Average PSNR:
26.1825 dB

Our Solution Part 1: Quantify per-Image Characteristics

Observation: Categorize images based on **Amount** and **Type** of high frequency

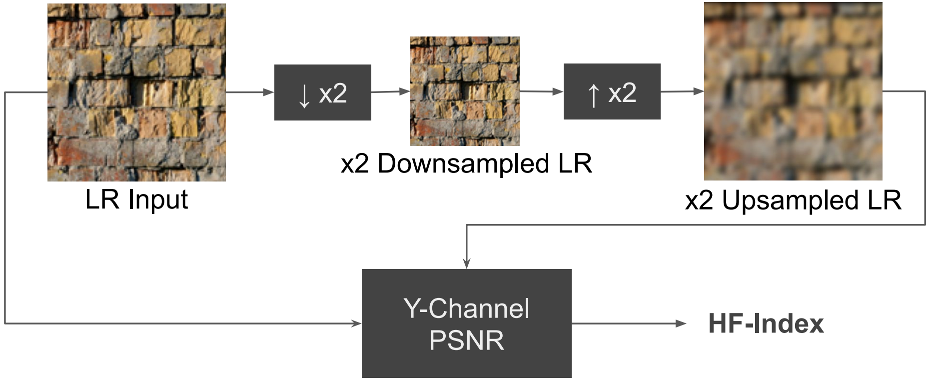
Amount of High Frequency Content

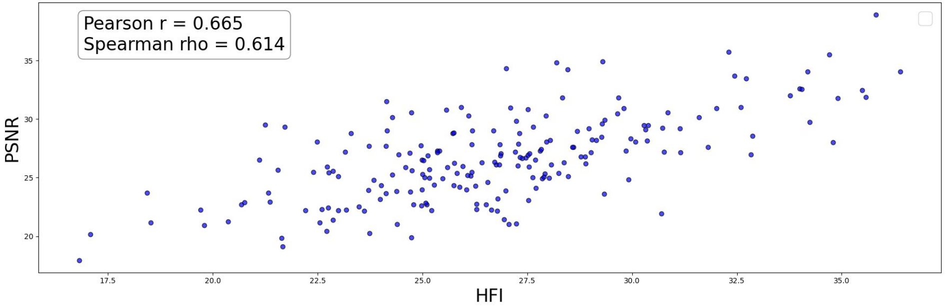
- High Frequency Index (HFI)

Type of High Frequency Content

- Rotation Invariance Edge Index (RIEI)

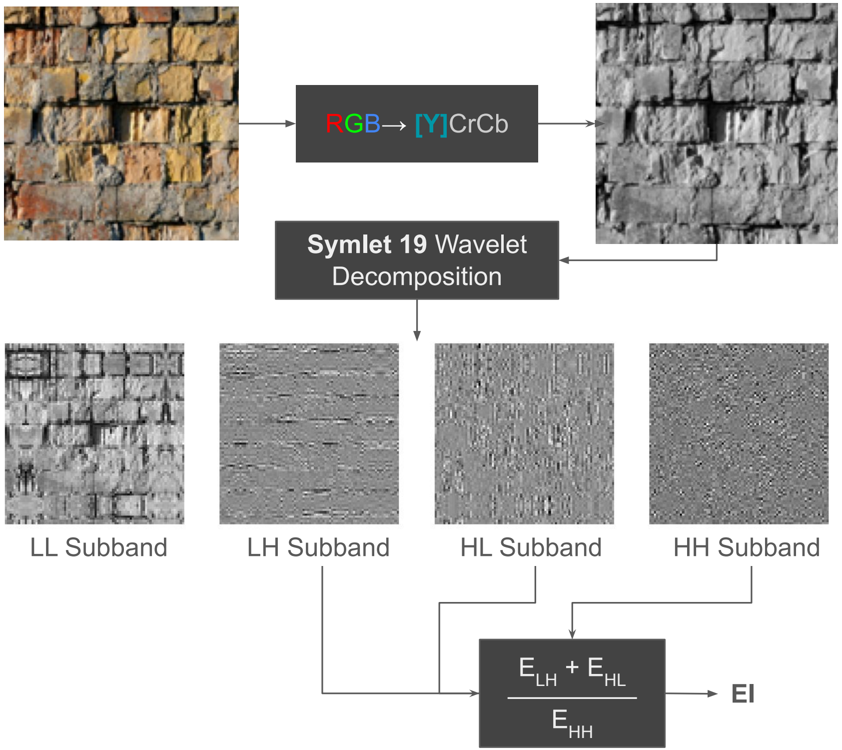
Image Characterization: **HFI Computation**





- High Frequency Index (HFI) provides an a-priori estimate of super-resolution difficulty.
- HFI is highly correlated with PSNR between HR and SR image on LSDIR Validation set.

Image Characterization: **RIEI Computation**



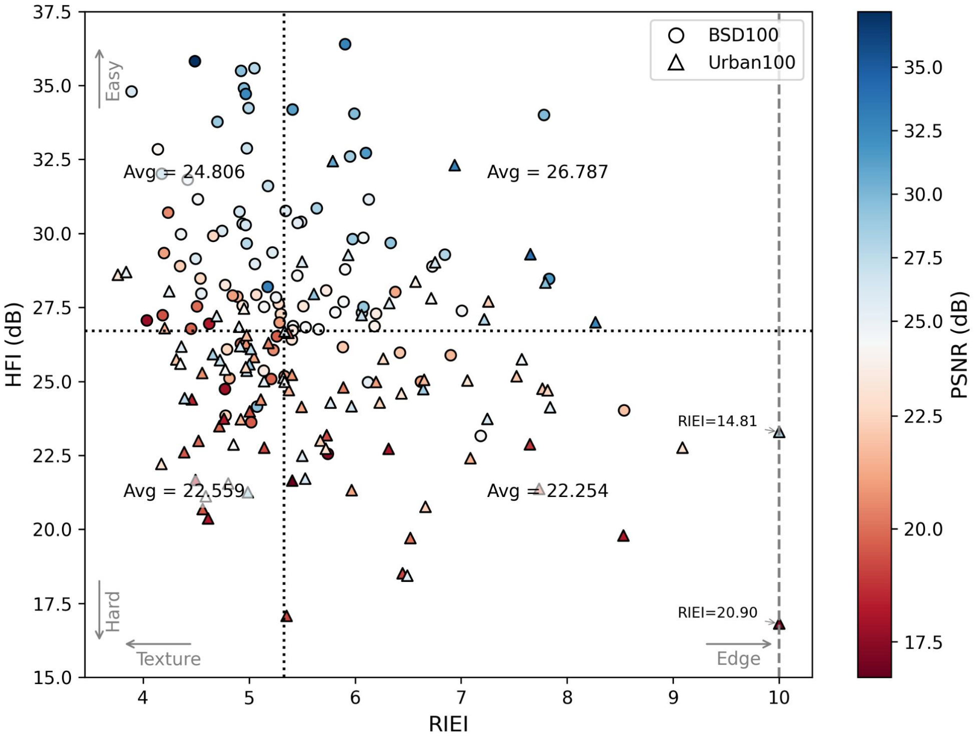
- Edge Index (EI) is not invariant to the orientation of dominant edges in the image.

- Rotationally Invariant Edge Index (RIEI) computes EI on LR images rotated with fixed angular increments, and taking the maximum.

$$\mathrm{RIEI} = max(\mathrm{EI}_\theta)$$

The Difficulty-Aware Evaluation Plane

- Each data point represents the characteristics of an image in terms of difficulty and high frequency content in the HFI-RIEI plane.
- Rather than using a single metric across all images for evaluation, use image characteristics to partition the images into semantically meaningful clusters, and analyze them in conjunction.



Quadrant-based PSNR analysis of ESRGAN+ results via HFI vs. RIEI scatter plot, where PSNR values are color-coded, on combined BSD100 and Urban100.

Our Solution Part 2: Localized Artifact Evaluation

- To capture severe but localized artifacts, we propose **PSNR99**. It calculates the PSNR on only the worst 1% of pixel errors.

Algorithm 1 Top 1% Error PSNR (PSNR99)

1: **Input:** Ground-truth image HR, SR image SR

2: Compute squared error per pixel on Y-channel:

$$E \leftarrow (HR - SR)^2$$

3: Rank pixel-wise errors and select the highest 1%:

$$E_{\text{top}} \leftarrow \text{R1\%}(E)$$

4: Compute the mean of selected errors:

$$\text{MSE}_{\text{top}} \leftarrow M(E_{\text{top}}) = \frac{1}{K} \sum E_{\text{top}}$$

where K is the number of pixels in the top 1%.

5: Compute

$$\text{PSNR99} \leftarrow 20 \log_{10}\left(\frac{255}{\sqrt{\text{MSE}_{\text{top}}}} \right)$$

6: **Return:** PSNR99

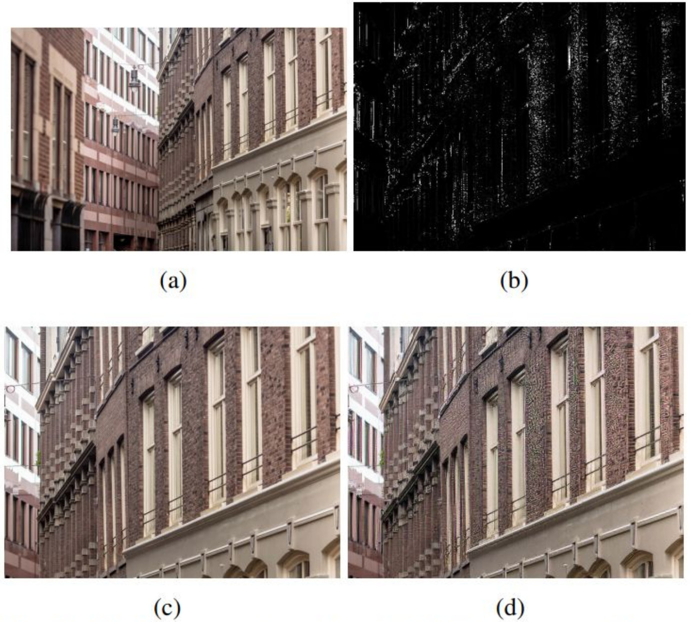


Fig. 7: Artifact map based on PSNR99 accurately captures visually disturbing areas of the SR image (e.g., hallucinations on the bricks). (a) HR Image (b) Zoomed in PSNR99 error map (c) Zoomed in HR crop (d) Zoomed in SR crop (PSNR = 21.19 dB PSNR99 = 8.38 dB)

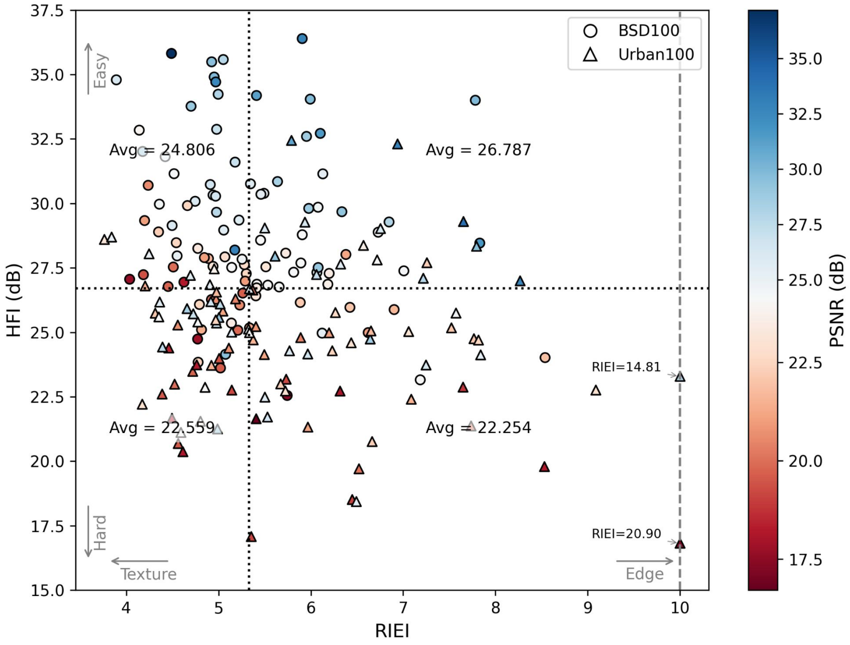
Case Study 1: Analyzing a Single Model (ESRGAN+)

Main Finding: The global average PSNR hides significant performance variations.

Global Average PSNR: 24.08 dB

Quadrant-Based PSNR:
🟢 Easy-Edge: 26.79 dB (+2.7 dB vs. average)
🟡 Easy-Texture: 24.81 dB (≈ average)
🔴 Hard-Texture: 22.56 dB (-1.5 dB vs. average)
🔴 Hard-Edge: 22.25 dB (-1.8 dB vs. average)

Key Takeaway: The model struggles with hard content, especially images with complex edges, a fact completely missed by the single average score.



Quadrant-based PSNR analysis of ESRGAN+ results via HFI vs. RIEI scatter plot, where PSNR values are color-coded, on combined BSD100 and Urban100.

Case Study 2: Explaining Why One Model is Better (WGSR vs. ESRGAN+)

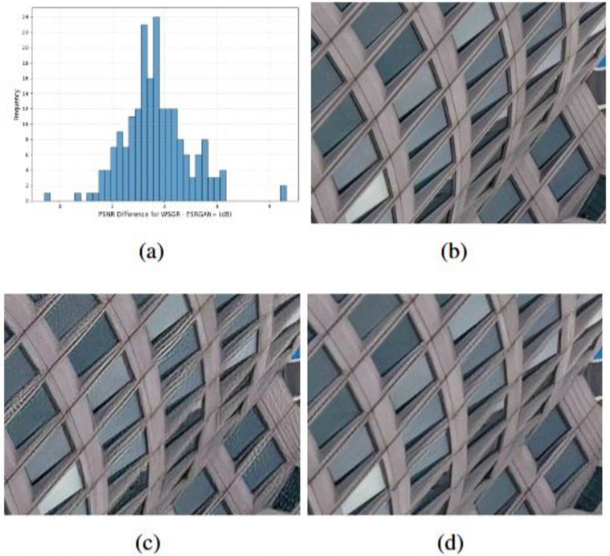


Fig. 9: Per-image comparison of ESRGAN+ vs. WGSR on image 68 from Urban100. (a) Histogram of PSNR differences, (b) Zoomed in HR crop (c) Zoomed in SR crop via ESRGAN+ (d) Zoomed in SR crop via WGSR

	Easy		Hard		Global Average
	Texture	Edge	Texture	Edge	
WGSR [15] vs. ESRGAN+ [9]					
PSNR	1.904	1.823	1.776	1.868	1.845
PSNR99	1.625	1.750	1.529	1.628	1.633
CLIPIQA [24]	-0.083	-0.063	-0.024	-0.017	-0.047

- **Simple Story:** WGSR outperforms ESRGAN+ by 1.845 dB on average.
- **Deeper Insight:** Where does this improvement come from?
- **Visual Proof:** WGSR effectively suppresses hallucinations on structured, "edge" content where ESRGAN+ fails.
- **Quantitative Proof:** Quadrant analysis confirms, largest gains in worst-case performance are on edge-heavy images.

Case Study 3: Revealing Architectural Differences (GAN vs. Diffusion)

- **Observation:** The WGSR shows a significant advantage on 'Easy-Edge' images.
    - PSNR99 Gain (WGSR vs ResShift) on Easy-Edge: +1.52 dB
    - PSNR Gain on other quadrants: ~1 dB
- **Nuanced Finding:** While overall performance is close, our method reveals a critical difference.
- **Why this Matters:** "Easy-Edge" images (e.g., graphics, text, simple architecture) are highly prone to the exact kind of hallucination artifacts that wavelet loss is designed to prevent.

	Easy		Hard		Global Average
	Texture	Edge	Texture	Edge	
WGSR [15] vs. ESRGAN+ [9]					
PSNR	1.904	1.823	1.776	1.868	1.845
PSNR99	1.625	1.750	1.529	1.628	1.633
CLIPIQA [24]	-0.083	-0.063	-0.024	-0.017	-0.047
WGSR [15] vs. ResShift [25]					
PSNR	0.918	1.190	0.961	0.989	1.012
PSNR99	0.966	1.515	1.015	1.003	1.119
CLIPIQA [24]	-0.005	0.015	0.022	0.020	0.013