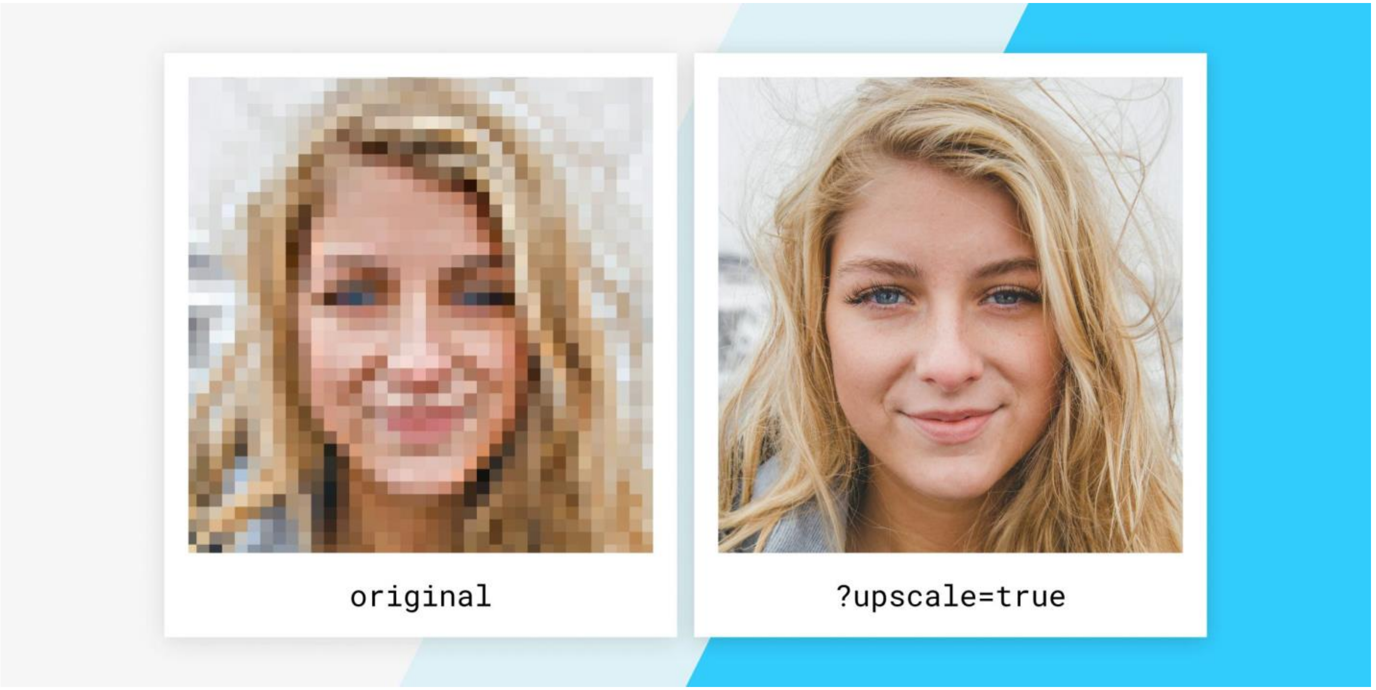
Aligning Subjective and Objective Assessments in Super-Resolution Models

Muhammad Hamza Zafar · Jon Y. Hardeberg

COLOURLAB NTNU, Norway

Introduction

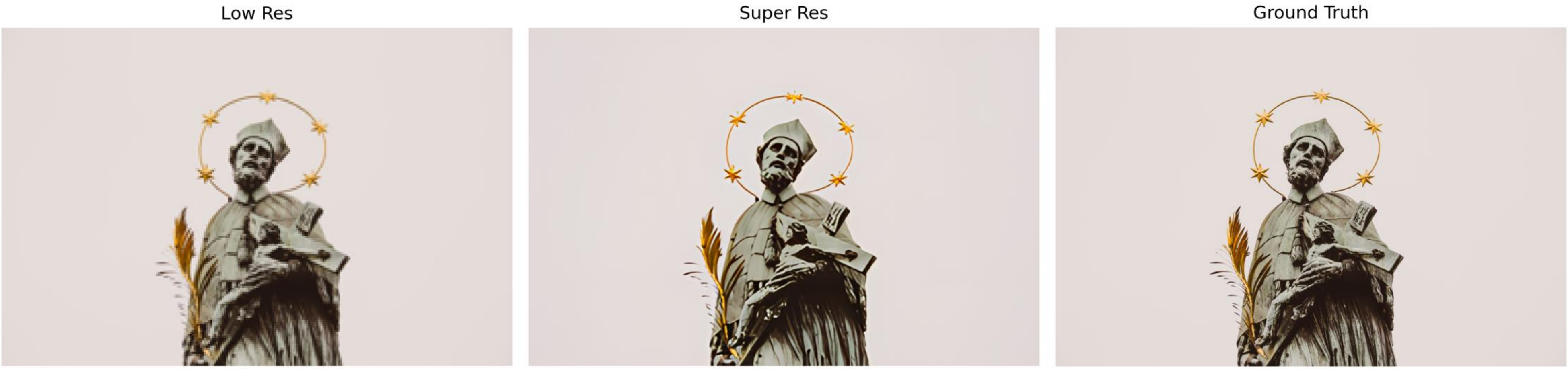


- Super-resolution (SR) enhances image details
- Objective metrics dominate (e.g., PSNR, SSIM)
- Need: Incorporate human perceptual assessments

Ref: https://www.imgix.com/blog/ai-powered-image-super-resolution

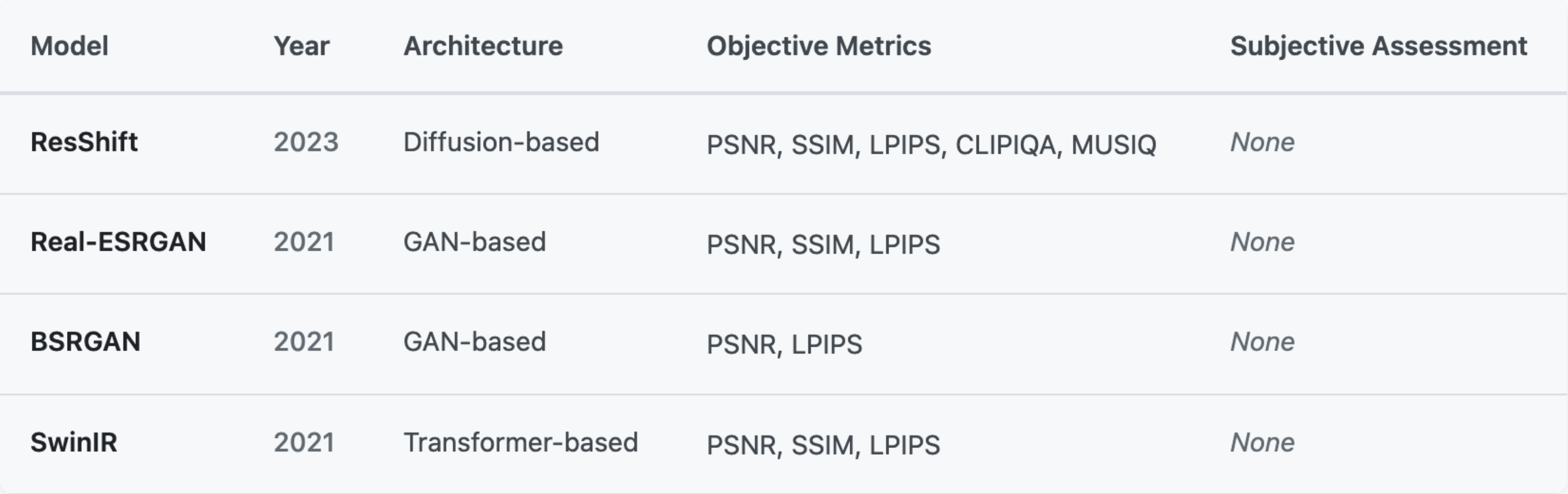
Motivation

- Why evaluate SR models beyond PSNR/SSIM?
- Problem: Objective metrics ≠ Human perception
- Visual: A high-PSNR image with poor perceptual quality
- Despite high **PSNR of 27.92** and **SSIM of 0.940**, the reconstructed image (center) appears **lacks perceptual sharpness**.
- **LPIPS = 0.047** also indicates high perceptual similarity
- Traditional metrics like PSNR and SSIM can be misleading when evaluating perceptual quality in super-resolution.



Evaluated Models

- 4 SR models:
  - **ResShift, Real-ESRGAN, BSRGAN, SwinIR**
- Some models report only a subset of objective metrics, which can mislead comparisons.
  - For example, BSRGAN omits SSIM, while ResShift includes LPIPS, SSIM, PSNR, and even CLIPIQA.
  - Fair evaluation requires consistent reporting across metrics.
- **ResShift** (Diffusion-based) - Claims SOTA
- **Critical Missing Piece**
  - **No subjective evaluation** for these models
  - **Limited understanding** of human preferences
  - **Objective-subjective alignment** unclear



Model	Year	Architecture	Objective Metrics	Subjective Assessment
ResShift	2023	Diffusion-based	PSNR, SSIM, LPIPS, CLIPIQA, MUSIQ	*None*
Real-ESRGAN	2021	GAN-based	PSNR, SSIM, LPIPS	*None*
BSRGAN	2021	GAN-based	PSNR, LPIPS	*None*
SwinIR	2021	Transformer-based	PSNR, SSIM, LPIPS	*None*

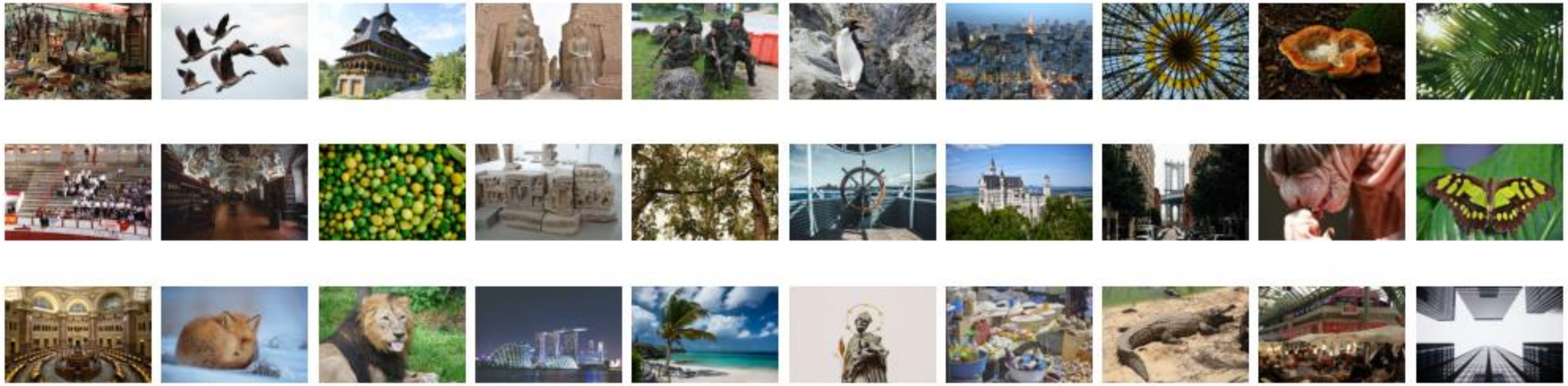
Dataset: DIV2K

- Standard benchmark used in super-resolution challenges
- High-quality, diverse 2K resolution images across various content types
- Low-resolution and high-resolution image pairs created using bicubic x8 and x2 downscaling
- Resolution details:
  - Low-resolution (LR): 255 × 169
  - High-resolution (HR): 1020 × 676 (4× upsampled)

Dataset Visualization: Low Res vs High Res

Setup

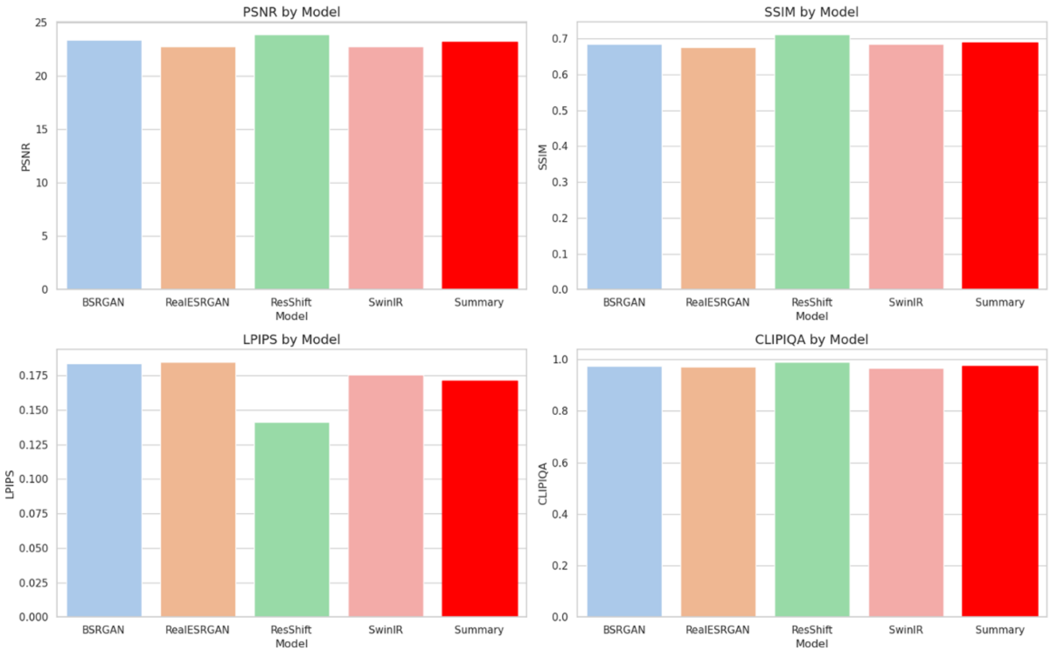
- No additional preprocessing.
- Setup follows official benchmark protocol
- Used in both objective and subjective evaluations:
  - 30 images for the online single-choice study
  - 10 images for the controlled lab pairwise comparison
    - Subset of the above 30 images.



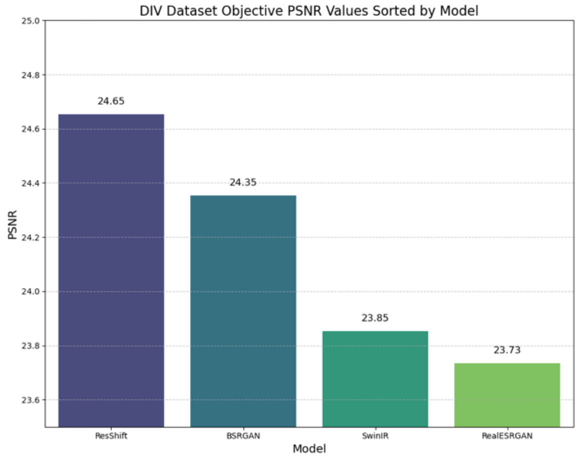
30 images for the online single-choice study

10 images for the controlled lab pairwise comparison

Objective Evaluation



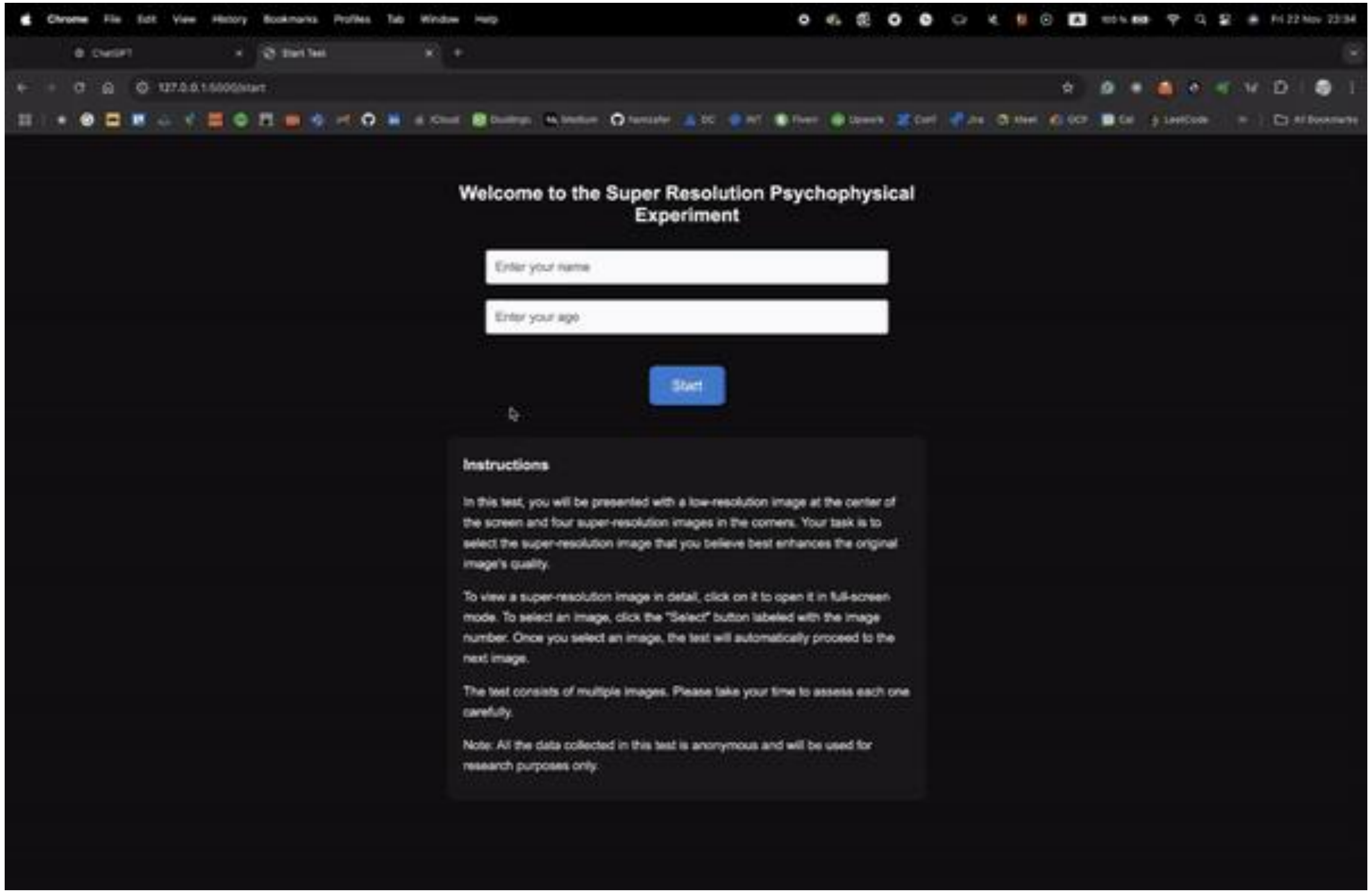
- Metrics: PSNR, SSIM, LPIPS, CLIPIQA on 30 DIV2K Images
- **ResShift leads in 4/4 metrics**
- **BSRGAN shows strong pixel-level performance** with second-highest PSNR
- **RealESRGAN underperforms** across most metrics despite being widely used
- **SwinIR surprisingly lags** in this comparison



Performance Metrics (DIV2K)

Model	PSNR ↑	SSIM ↑	LPIPS ↓	CLIPIQA ↑
ResShift	**24.65**	**0.723**	**0.136**	**0.986**
BSRGAN	24.35	0.703	0.172	0.977
SwinIR	23.85	0.705	0.164	0.968
Real-ESRGAN	23.73	0.697	0.169	0.972

Experiment 1 – Online Setup



Welcome to the Super Resolution Psychophysical Experiment

Enter your name

Enter your age

Instructions

In this test, you will be presented with a low-resolution image at the center of the screen and four super-resolution images in the corners. Your task is to select the super-resolution image that you believe best enhances the original image's quality.

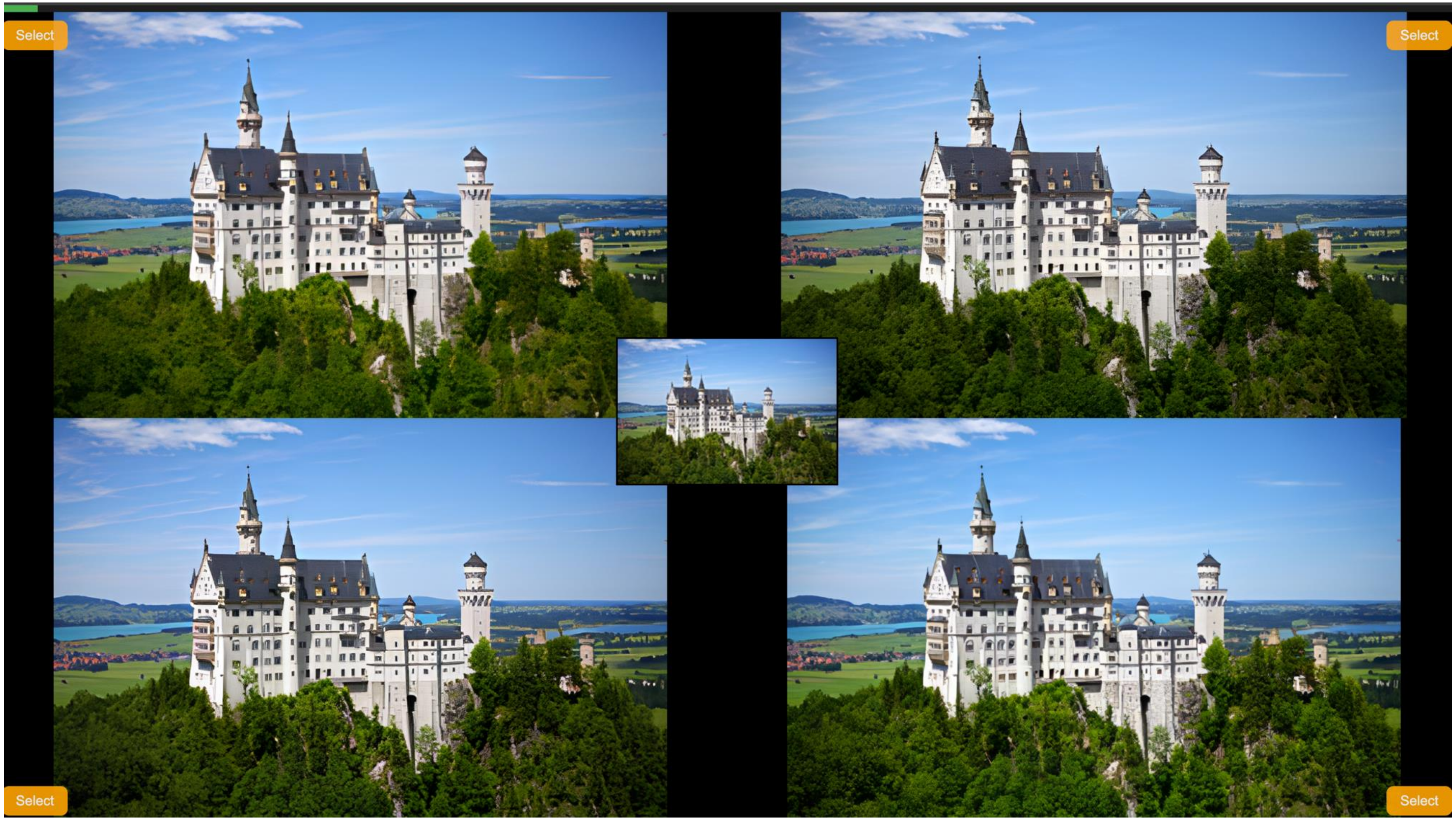
To view a super-resolution image in detail, click on it to open it in full-screen mode. To select an image, click the "Select" button labeled with the image number. Once you select an image, the test will automatically proceed to the next image.

The test consists of multiple images. Please take your time to assess each one carefully.

Note: All the data collected in this test is anonymous and will be used for research purposes only.

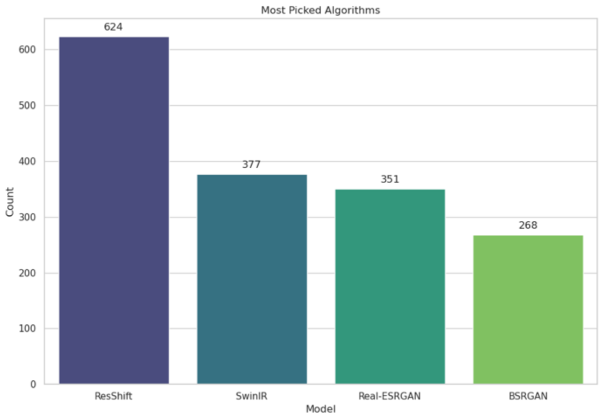
- **Total Images**: 30
- **Estimated Time**: 15 minutes
- **Low-Resolution Image**:
  - Size: 255×169 pixels
  - Position: Center of the screen
- **High-Resolution Images**:
  - Quantity: 4
  - Size: 1020×676 pixels each
  - Display: Surrounding the low-resolution image
- Choose the best HR image from randomized comparisons.

EXPERIMENT 1
SETUP



Experiment 1 – Results

- **Subjective Results:**
  - **What Humans Actually Prefer**
- 134 participants
- 54 completed
- ResShift dominant
  - ResShift: 624 selections ⭐
  - SwinIR: 377 selections
  - Real-ESRGAN: 351 selections
  - BSRGAN: 268 selections



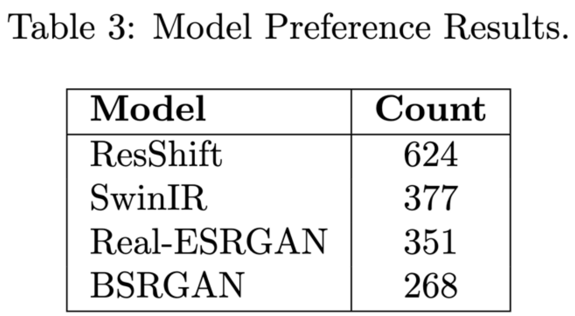
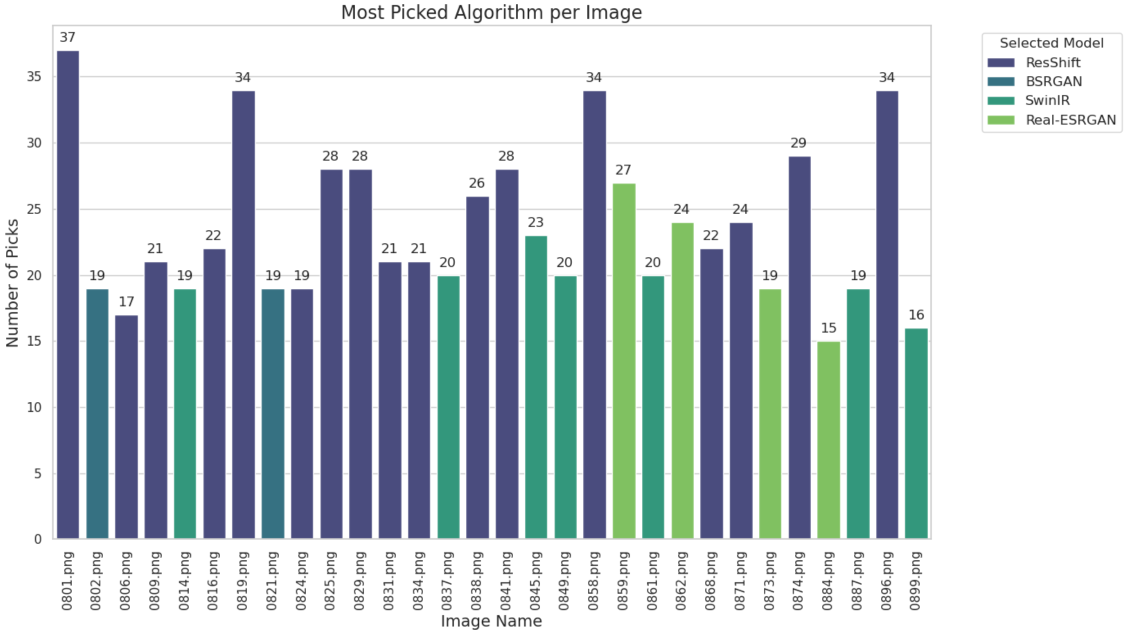


Table 3: Model Preference Results.

Model	Count
ResShift	624
SwinIR	377
Real-ESRGAN	351
BSRGAN	268



Experiment 1 – Results

Image 1: Age Distribution (18-50)

- **Participant demographics**: Heavily skewed toward younger adults
- **Peak at 24-25 years**: 29 and 16 participants respectively (majority of sample)
- **Limited older representation**: Only 3 participants over 35, which could be a limitation

Image 2: Age vs. Algorithm Preference (Age < 30)

- **ResShift dominance**: Consistently highest preference across ALL age groups from 18-28
- **Interesting insight**: ResShift is not the highest picked among 19 & 21 year old
- **SwinIR as second choice**: Generally second-most preferred across age groups

Image 3: Word Cloud - Reasons for Selection

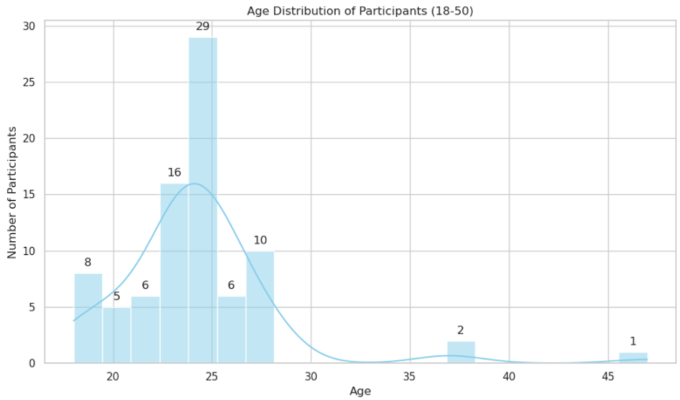
- **Most prominent terms (by size):**
- **"detail"** - Largest term, indicates primary selection criterion
- **"artifact"** - Second largest, shows participants actively avoided visual artifacts

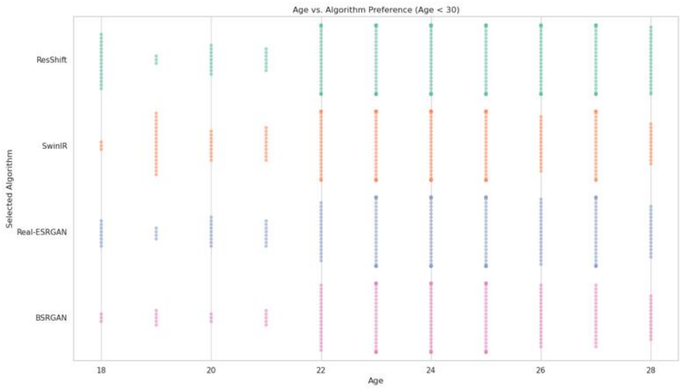
Image 4: Word Cloud - General Feedback

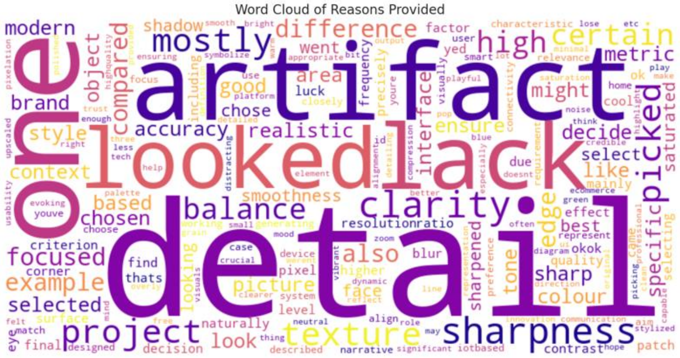
- **"generated", "detail"** - Focus on output quality
- **"quality", "resolution"** - Technical assessment terms

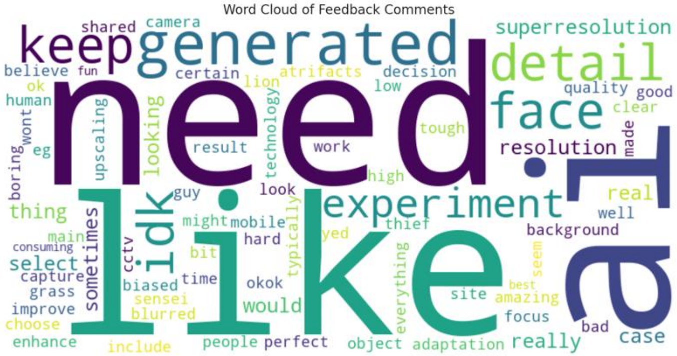
Overall Analysis:

- **Consistent preferences**: ResShift dominance holds across all demographics
- **Quality-focused selection**: Participants prioritized detail preservation and artifact avoidance



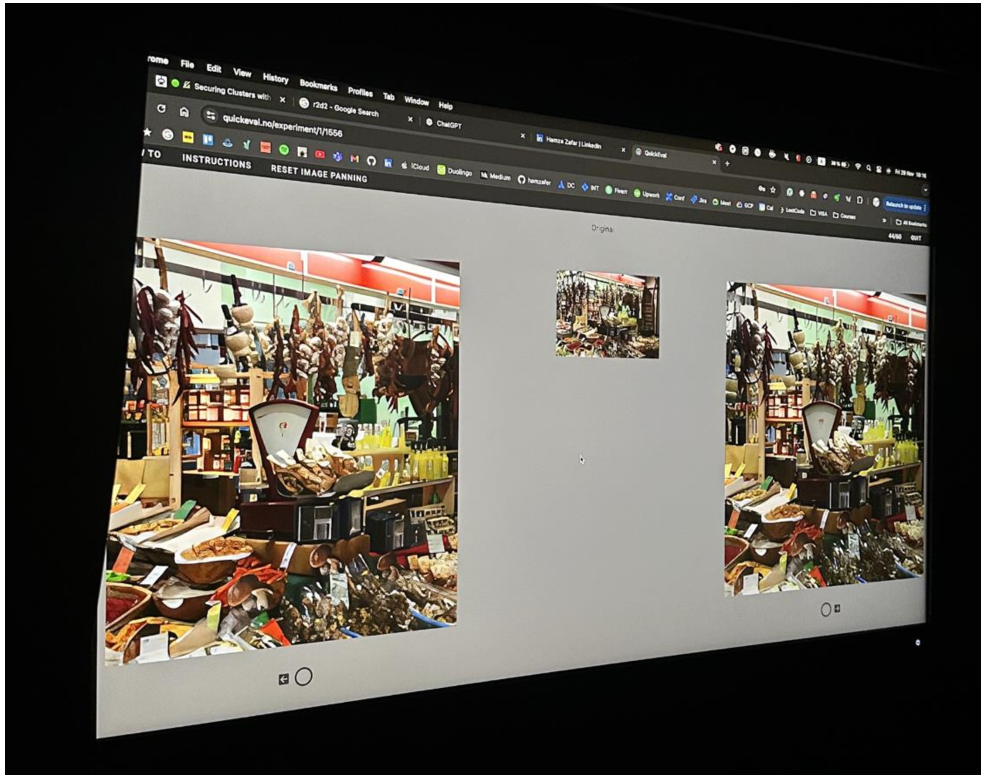


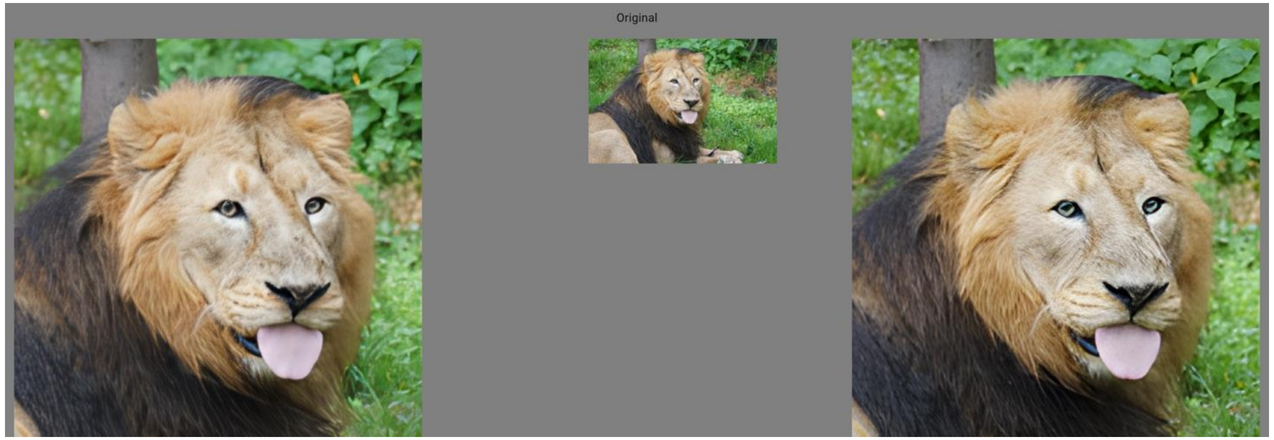




Experiment 2 – Lab Study

- Pairwise comparisons, 10 images, 60 pairs per person
- BenQ calibrated monitor, sRGB, D65, 80 cd/m^2
- Estimated Time: 15 minutes





Experiment 2 – Results

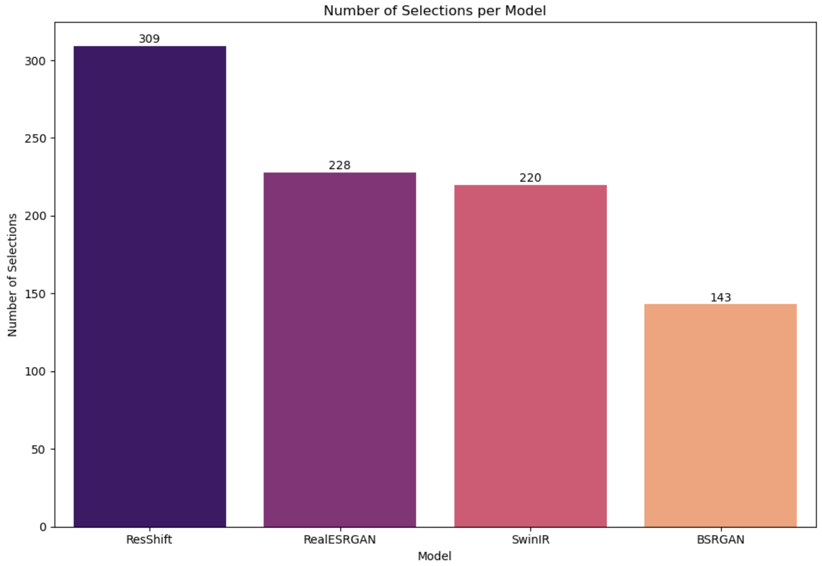
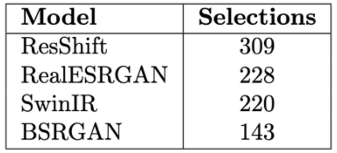
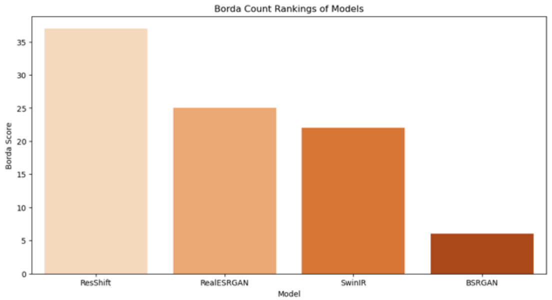


Table 4: Model selections during Experiment 2.



Model	Selections
ResShift	309
RealESRGAN	228
SwinIR	220
BSRGAN	143



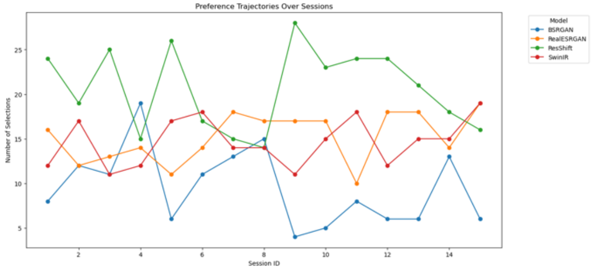
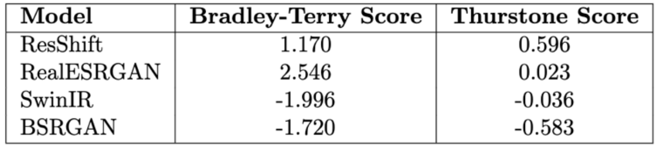


Table 5: Ability estimates for SR models using Bradley-Terry and Thurstone methods.

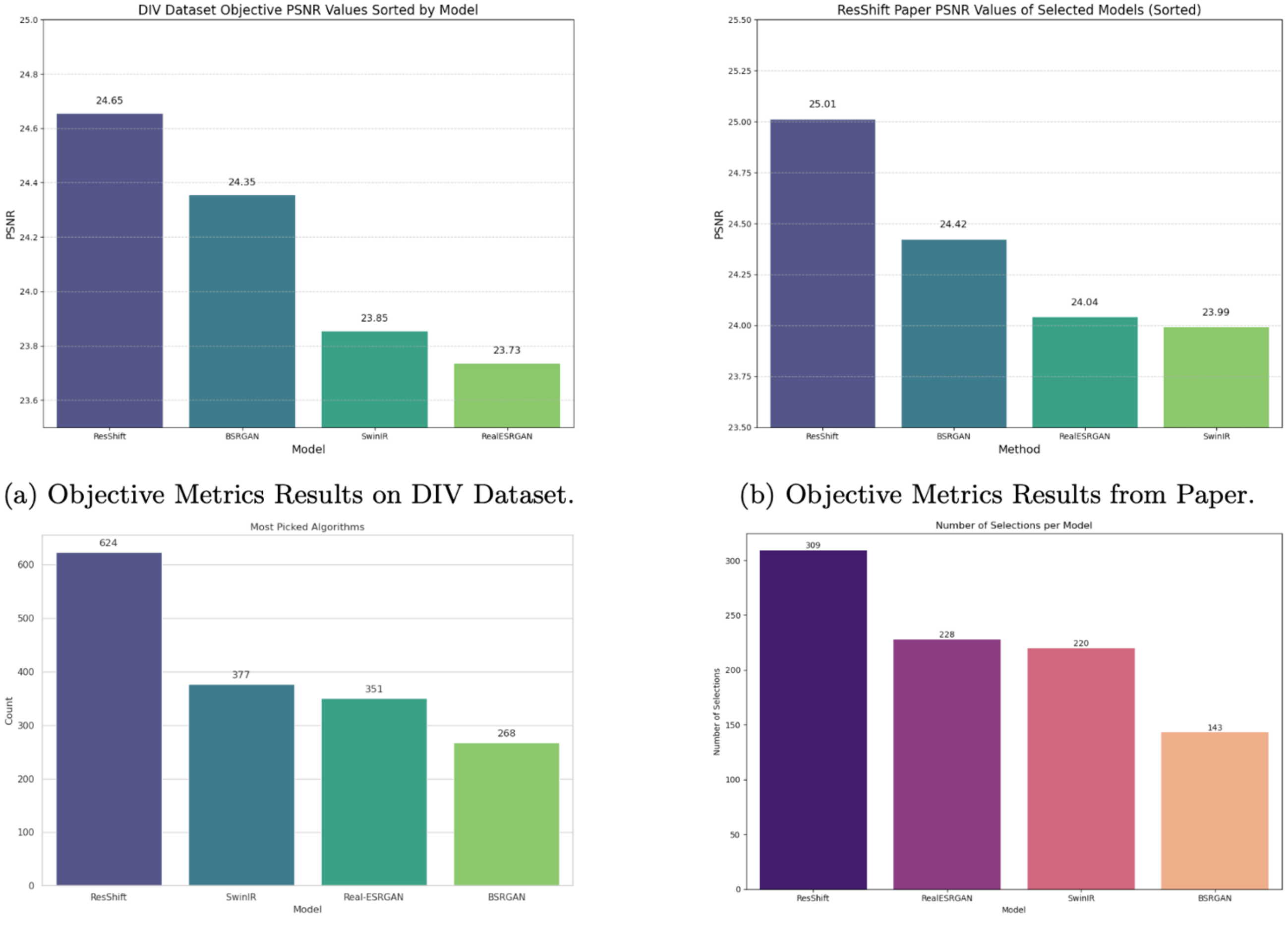


Model	Bradley-Terry Score	Thurstone Score
ResShift	1.170	0.596
RealESRGAN	2.546	0.023
SwinIR	-1.996	-0.036
BSRGAN	-1.720	-0.583

- 900 pairwise votes, 15 observers
- ResShift dominant again
- Chi-square test (p < 0.0001) Results are not due to random chance
- **Bradley-Terry Scores**: ResShift 1.170 (highest ability to be preferred)
- Multiple statistical methods confirm ResShift's superiority

Unified Comparison

- **ResShift claims validated**: Both objective metrics AND human preferences confirm SOTA status
- **Cross-study consistency**: Online (54 participants) and lab (15 observers) show identical model rankings
- BSRGAN:
  - **Objective-subjective disconnect**: 2nd best PSNR but worst human preference
- **Traditional metrics misleading**: High PSNR ≠ Visual quality



(a) Objective Metrics Results on DIV Dataset.

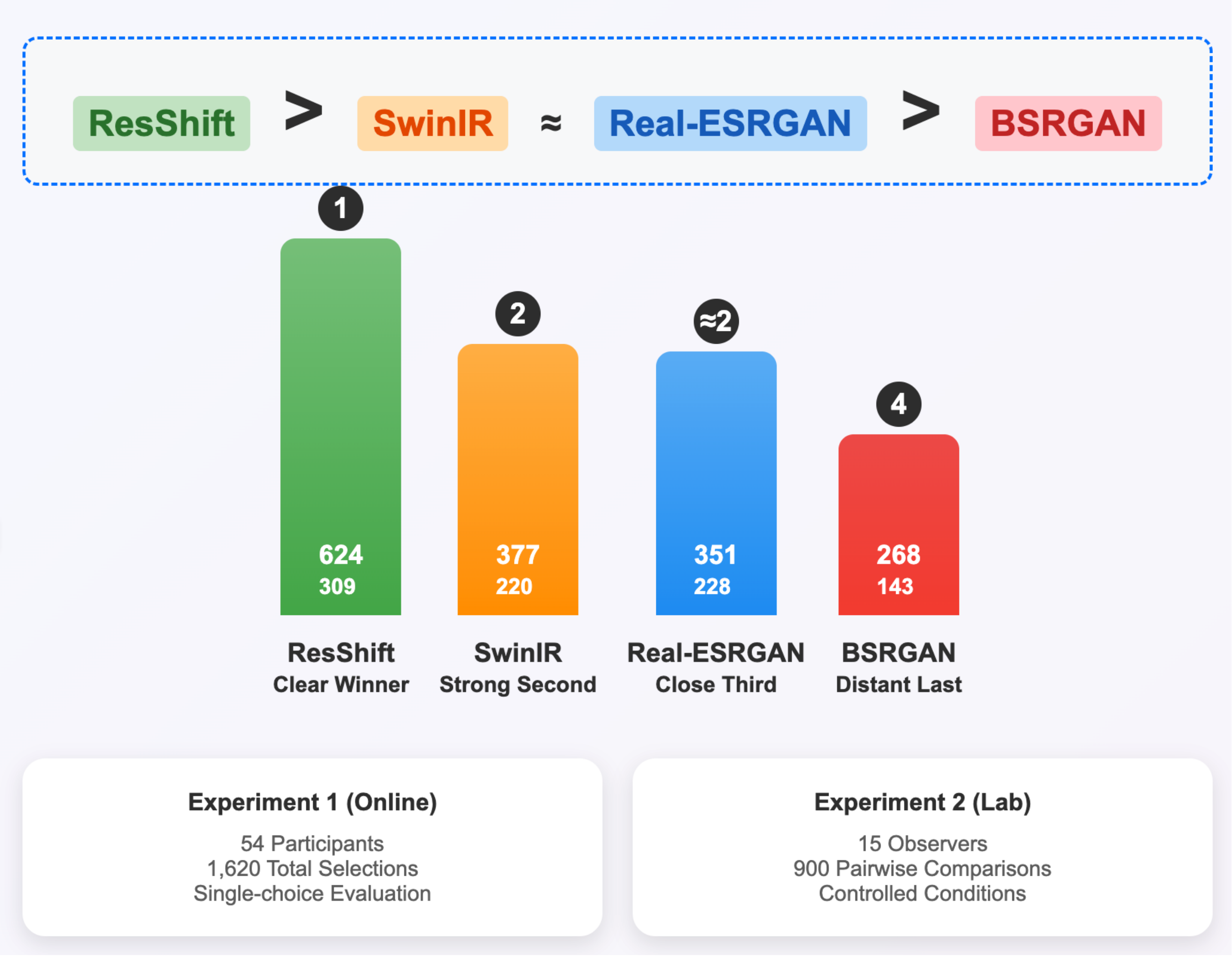
(b) Objective Metrics Results from Paper.

(c) Experiment 1 Results.

(d) Experiment 2 Results.

Key Insights & Discussion

- Objective ≠ Perceptual always
- ResShift = technically & perceptually strong
- BSRGAN = objectively good, subjectively bad
- Hybrid evaluation is necessary for real-world quality
- **Both studies show identical model rankings**, validating the robustness of human preferences for ResShift's superior perceptual quality.
- The numbers above bars show selections from Exp1 (top) and Exp2 (bottom).



ResShift > SwinIR ≈ Real-ESRGAN > BSRGAN

Experiment 1 (Online)
54 Participants
1,620 Total Selections
Single-choice Evaluation

Experiment 2 (Lab)
15 Observers
900 Pairwise Comparisons
Controlled Conditions

Limitations

- Only 4 models, DIV2K only
- No novel metric proposed
- Small lab (Exp 2) dataset (10 images)
- Still, robust reproducible framework

Conclusion & Future Work

- ResShift sets a new standard in perceptual SR
- Need larger and more diverse datasets
- Explore joint metric-subjective learning for SR
- Push toward **human-centric model evaluation**

Project Page

Contact

Thank You

- Repo: https://github.com/hamzafer/super-resolution-color
- muhamhz@ntnu.no
- https://www.linkedin.com/in/ihamzafer/
- Open for questions ☺

References

1.ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting
Authors: Zongsheng Yue, Jianyi Wang, Chen Change Loy
URL: https://github.com/zsyOAOA/ResShift

2.Real-ESRGAN: Real-Enhanced Super-Resolution Generative Adversarial Networks
Authors: Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy
URL: https://github.com/xinntao/Real-ESRGAN

3.BSRGAN: Blind Super-Resolution Generative Adversarial Networks
Authors: Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy
URL: https://github.com/cszn/BSRGAN

4.SwinIR: Image Restoration Using Swin Transformer
Authors: Jingyun Liang, Yiming Ma, Chengjie Wang, Bineng Zhong, Dong Wang
URL: https://github.com/JingyunLiang/SwinIR

5.Agustsson, E., & Timofte, R. (2017). *NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study*. In CVPR Workshops.