Measuring Semantic Differences Between Images in the Super Resolution Domain

Josep Maria Rocafort

June 27, 2025

Abstract

Coming from the field of super-resolution and image enhancement, we encounter a significant problem. When we upscale and enhance a low-resolution image, the output may exhibit very high perceptual quality and high pixel-wise similarity to the ground truth. However, substantial semantic errors can still occur. While the general structure and colors may remain consistent, specific objects can change meaningfully. For example, a low-resolution plant on a table in the input image might become a decorative pot, or a red box of cigarettes might be transformed into a red velvet cake.

Current metrics used to evaluate super-resolution models are not equipped to detect these kinds of semantic errors; they primarily focus on overall image fidelity. In this work, we define the semantic fidelity objective, construct a dataset that emphasizes semantic errors, and propose a semantic fidelity metric.

Keywords: Super Resolution, Semantic Fidelity, Image Quality Assessment, CLIP, Computer Vision

1 Introduction

In the field of super-resolution (SR), we aim to convert low-quality, low-resolution images into clean, high-resolution versions. In this process, SR models must extract information from the low-resolution image and, combined with learned world priors, attempt to generate a high-resolution image that matches the original in terms of composition, detail, and semantics.

Typically, to evaluate these models, we use datasets composed of high-resolution images and their corresponding low-quality counterparts. We then compare the model's outputs against the ground truth images.

To evaluate these outputs, we can begin by comparing pixel-wise differences, such as mean squared error (MSE) or PSNR[1]. We can also assess textures and low-level details using kernel-based metrics like SSIM [2]. To measure perceptual errors, we use methods based on pre-trained image encoder features. For example, LPIPS [3] utilizes feature differences from a pre-trained VGG network to train a model on human preference data. These features, derived from deep convolutional layers, carry more abstract, perceptual information.

Similarly, ViTScore [4] leverages features from the last hidden layer of a Vision Transformer to compute patch-wise similarity. ClipScore [5] uses the CLIP vision-language model to assess the semantic similarity between images. However, as we will show, **CLIP can produce inconsistent results** for this task. While these methods capture various levels and types of semantic differences, they lack interpretability and alignment with human perception, largely due to the limitations and biases inherent in the models they rely on.

Although these approaches are valuable for evaluating SR outputs, there remains a gap: the lack of a dedicated **semantic fidelity** metric. The goal of this project is to explore what such a metric would entail—and to propose one.

Take, for instance, the result shown in Figure 1 from SeeSR [6], where a cigarette box is modified to a velvet cake. Existing methods struggle with these types of cases because they misinterpret the low-resolution input. While current metrics report strong perceptual and pixel-wise similarity—since the output is indeed visually similar—some objects change in meaning and context. In the example, there is a clear **semantic shift** that existing metrics fail to detect. This is precisely the issue we aim to address, and the gap we intend to fill.

The main contributions of this project are:

- We raise awareness of the **semantic fidelity problem** in super-resolution models.
- We **provide two evaluation datasets** designed to capture semantic errors: one based on controlled inpainting and another from real SR model outputs.

• Contact E-mail: josepmarocafort@gmail.com
• Supervised by: Javier Vázquez Corral (Computer Vision Center)
• Academic Year 2024/25

Fig. 1: Example of a clear semantic change between an image ground truth and a super-resolution output. In the SR-output (right side image) we can see objects in the original image such as the cigarrete box change to a red-velvet cake. We can also see different boxes and plastic wrappers also transform into cake-like objects.

- We introduce two complementary evaluation approaches:
  - A **global evaluation method** that measures overall semantic alignment.
  - A **local evaluation method** that focuses on semantic changes in specific regions.
- We find that a backbone pretrained on Imagenet transfers better to the semantic fidelity analysis than a Laion contrastive-pretrained backbone such as the image encoder from clip.

1.1 Measuring Semantics from a Human Perspective

In the context of super-resolution (SR), we can define three distinct evaluation paradigms for assessing semantic fidelity: **no-reference**, **full-reference (GT)**, and **semi-reference (LQ)**.

No-reference evaluation is of limited value in our setting. While some existing approaches—such as natural image assessment or rationale-based evaluation—could serve as proxies for semantic fidelity without a ground truth, they fall short in critical ways. These methods are prone to overestimating quality when the SR output appears visually coherent, realistic, and undistorted, even if semantic inconsistencies are present.

Full-reference (GT) evaluation offers the most comprehensive measure of semantic fidelity. By comparing the SR output directly to its corresponding ground truth, we can reliably identify semantic discrepancies. Through qualitative analysis of numerous SR samples, we have defined a taxonomy of semantic changes, including:

- Object class substitutions (e.g., *cigarette box* $\rightarrow$ *cake*),
- Appearance modifications (e.g., changes in color, texture, or material),
- Structural alterations (e.g., shape deformation or incorrect object placement).

Semi-reference (LQ) evaluation compares the SR output to the original low-resolution input. While it can help detect semantic errors, its reliability is constrained by the quality of the LQ image itself. The more degraded or ambiguous the input, the harder it is—even for humans—to infer the intended semantics. This raises important questions: *Where is the limit of human perception when interpreting LQ images? Should SR models aim to exceed that limit? And if so, does this goal even make sense from a semantic perspective?*

These questions highlight the subjective nature of semantic fidelity and the challenge of aligning SR outputs with human expectations.

2 Previous Work

A wide range of methods has been proposed to evaluate image similarity, particularly in the context of super-resolution (SR) and image enhancement. Traditional evaluation metrics such as **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, **Peak Signal-to-Noise Ratio (PSNR)** [1], and the **Structural Similarity Index Measure (SSIM)** [2] focus on pixel-level fidelity and local pattern similarity. While these metrics are computationally efficient and widely used, they often fail to align with human perception—especially for SR outputs that may appear overly smooth or lack fine texture despite achieving high PSNR or SSIM scores.

To address this gap, **feature-based metrics** have been introduced. The **Learned Perceptual Image Patch Similarity (LPIPS)** metric [7] compares deep features from pretrained convolutional networks (e.g., VGG), capturing perceptually meaningful differences. More recent approaches, such as **ViTScore** [4] and **ClipIQA**, extend this idea using Vision Transformers and CLIP-based embeddings to assess quality in a full-reference and no-reference setting respectivelly. These methods show improved correlation with human judgment and are useful for training and evaluating SR models. However, while some are sensitive to semantic errors, none are specifically designed to detect **semantic distortions** in SR, and are often biased by low-level image quality artifacts.

Unsupervised image-language models such as **CLIP** [8] have demonstrated strong performance in cross-modal tasks by learning a shared embedding space for images and text through contrastive learning on large-scale datasets. These models excel at capturing high-level semantic relationships and have been widely adopted in tasks like zero-shot classification and cross-modal retrieval While one might extend this to compare two images by computing the cosine similarity between their embeddings, this approach has fundamental limitations.

In 2022, **Liang and Ziang et al.** [9] highlighted a core issue with dual-encoder architectures: the **intra-modal misalignment** caused by modeling two disjoint **cone-shaped embedding spaces**. When two encoders (image and text) are randomly initialized, their embeddings fall into separate cones within the shared space. Contrastive learning, as used in CLIP, does not collapse these cones into a unified space—instead, it only aligns directions within each cone to maximize similarity across modalities. As a result, while CLIP organizes inter-modal relationships well, it is **subop-**

timal for intra-modal similarity, **such as comparing two images directly using CLIP's image encoder alone.**

In 2025, **Mistretta et al.** [10] further addressed this problem by proposing **modality inversion**. Their method bridges modalities by optimizing pseudo-tokens that transform an image embedding into the text domain, allowing for intra-modal comparisons via the opposite modality. For example, to compare two images, one can project one image into the text space and then use it to compare with the other image embeddings. However, this approach is **computationally expensive**, and the authors recommend instead using models trained with **stronger intra-modal objectives**, such as **SLIP** [11].

In our early experiments, we found **no clear benefit** in using SLIP over CLIP, so this work focuses exclusively on CLIP-based embeddings. However, our final approach could potentially benefit from SLIP or similar models, and we leave this direction for future investigation.

3 DATASETS

To our knowledge, **there is currently no dataset that focuses specifically on semantic changes between images**. Therefore, it is of vital importance to create one ourselves. We aim to build a dataset in which two images appear visually similar but exhibit subtle differences—such as the disappearance, appearance, or transformation of certain objects—resulting in the types of semantic errors discussed in the introduction. Additionally, we want these images to originate from super-resolution (SR) model outputs, as this is the domain we are addressing, and the images should reflect that distribution.

We explored two main approaches to construct this dataset. The first involves using a dataset like COCO, which contains annotated masks and class labels. We can then **inpaint selected regions** and condition them for semantic change. This approach is highly controllable and requires minimal additional annotation work; however, it does not produce images or semantic errors that fall within the SR distribution.

The second approach involves **running SR models** on a dataset while applying different levels of degradation to the low-quality (LQ) inputs. This method allows us to generate, for the same ground truth (GT) image, multiple SR outputs with varying degrees of semantic errors.

Ultimately, we chose to pursue both approaches in parallel, as they can complement each other. The inpainting dataset provides a controlled setting to verify whether our network can learn semantic distinctions, and later serves as a proof of concept for measuring semantic differences across object class changes.

3.1 COCO Inpainting Dataset

For this dataset, we leveraged the **COCO 2017 training set**. We selected images that were at least 512 pixels in both dimensions, as SR models typically operate on images at least this size. We further filtered the images based on the number of object masks to control scene complexity. Finally, we used the KonIQ++ model to evaluate image quality and exclude low-quality samples.

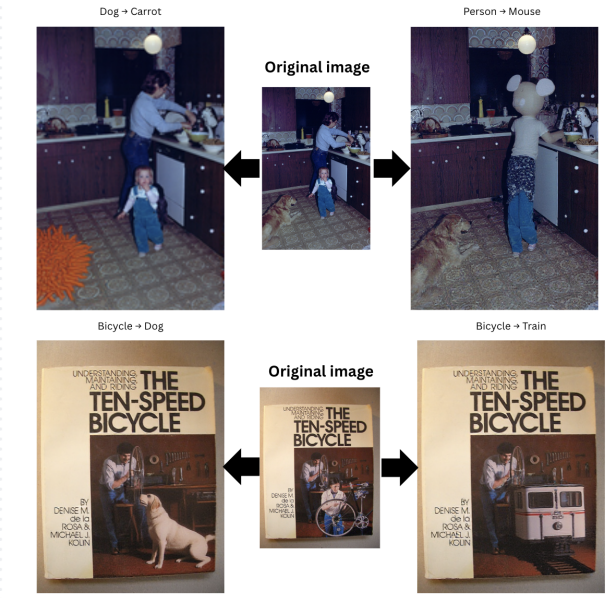

Fig. 2: Examples of inpainted images from the COCO Inpainting Dataset. In the center we have the original images and on each side two different variations of changing supercategory(left-side) and changing category within the original supercategory (right-side).

To determine which objects to inpaint, we sampled **up to three masks per image**, ensuring a **balanced distribution across COCO 2017 classes**. For each selected instance, we generated three variations:

No class change (e.g., bicycle → bicycle)

Intra-supercategory change (e.g., bicycle → car)

Inter-supercategory change (e.g., bicycle → toothbrush)

To prompt the diffusion model, we used simple strings such as "A toothbrush."

The inpainting was performed using the Flux 1.5-fill [12] model within the ComfyUI platform, which provided flexibility during development. Initially, we experimented with conditioning the generation using ControlNets (e.g., depth maps or Canny edges) to better resemble SR outputs, but results were inconsistent.

We processed 1% of the filtered dataset, yielding 476 images. From these, we sampled 963 object instances, producing 4,399 proposals (output images).

We then **filtered the images using CLIP[8]**. Specifically, we measured the cosine similarity between the **masked region's embeddings** and the **original vs. target class labels.** Images showing a large semantic discrepancy were retained. For borderline cases, manual inspection was used to determine whether a true semantic change had occurred. The resulting images were organized into **no-semantic-change** and **semantic-change categories** for each GT image.

To generate **additional examples without semantic change** but with **minor pixel-level variations** we reused the pipeline within the same class. Instead of starting the inpainting process from pure noise, we used partial denoising (starting from 10–30% noise), allowing us to reconstruct the region with minor differences.

3.2 SR Outputs Dataset

To create a dataset of SR outputs containing varying levels of semantic errors, we **applied the BSRGAN degradation pipeline[13]** at three intensity levels to the KonIQ-10k dataset[14], which was selected for its diversity and strong reputation in IQA benchmarking.

BSRGAN applies a sequence of degradations, including camera noise, compression artifacts, pixelation, and blur, to produce challenging LQ images. **We applied this pipeline at three scaled intensities: 0.3, 0.7, and 1.0.** Each LQ image was then processed through five different SR models:

- BSRGAN [15]
- SeeSR [6]
- StableSR [16]
- SwinIR [17]
- PASD [18]

This resulted in approximately **150,000 images**. To refine the dataset, we filtered for samples with moderate performance on classical metrics (SSIM and LPIPS). Specifically, we normalized the scores and selected 25,000 images within the 0.6–0.8 range. We aim to select moderate results as images with best results do not have enough pixel wise variation for semantic errors to occur as frequently.

Next, we evaluated the images using the KonIQ++ [19] no-reference IQA model to e**stimate image quality and blur**. **The 15,000 most blurred images were excluded.**

We retained a final set of 10,000 images for further processing, including pseudo-labeling and a planned user study, which will be described in the following sections.

4 Global Evaluation Methods

We explore methods to assign a **single semantic fidelity score to each image pair (SR and GT)**. Two complementary approaches were used: **(1)** conducting a **user study** to gather human semantic judgments, and **(2) leveraging pretrained models** (e.g., CLIP[8], captioning, text retrieval) to generate pseudo-labels for both training and evaluation.

4.1 Pseudo-labels

We designed two types of pseudo-labels intended to reflect semantic alignment and correlate with human judgments. Both rely on captions generated using the Qwen2.5-VL-7B-Instruct[20] model, applied across the full 10k SR image dataset.

Caption2Caption (C2C) Cosine Similarity: We computed cosine similarity between GT and SR captions using the widely adopted text embedding model all-MiniLM-L6-v2[21]. This score aims to reflect semantic similarity between the captions.

Caption2Image (C2I) CLIP Cosine Similarity: Here, we compared the text embedding of the GT caption with the image embedding of the SR image using CLIP. This attempts to directly measure semantic alignment between language and visual content.

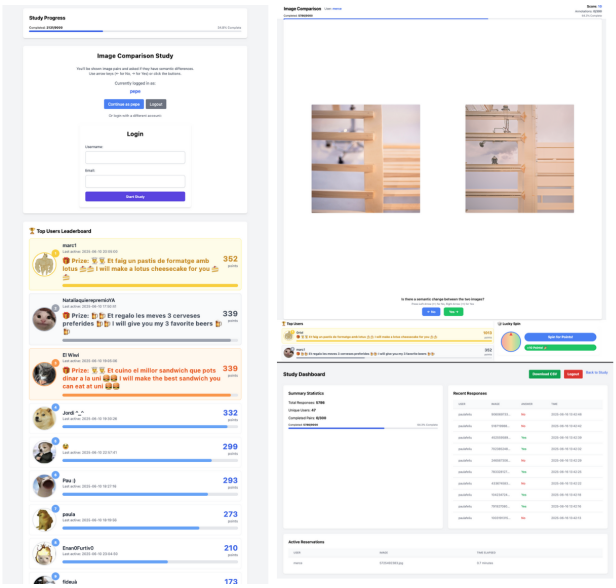

Fig. 3: The different pages of the developed user study platform. On the left side we can see the home page where users can log in and see the users in the leaderboard. On the top-right side we can see a comparaison page, two images are shown and the user is asked to click/swipe left or right depending on if there is a semantic change or not. On the bottom-right side we can see a dashboard page for the experiment administration. We can see the different user responses and also download a csv file with all the results.

4.2 User Study

We built an **interactive online platform** to c**ollect semantic fidelity assessments from human users.** Participants log in and evaluate whether a given SR image contains a semantic error compared to the GT. **Each image pair receives multiple evaluations**, and the final semantic score is computed as the **mean opinion score (MOS)**, the proportion of "Yes" responses.

To encourage participation, the platform included gamification features such as a point system, leaderboard, roulette mini-game, and prize rewards (Figure 3).

For the study, we selected **300 image pairs from the 10k SR dataset**. Selection was based on the **lowest CLIP[8] cosine scores (between the GT caption and SR image)**, followed by manual filtering to ensure diversity. Each image received 15–30 responses, and **trap images** (with 90% agreement on semantic error) were used to **filter unreliable participants.**

In a preliminary analysis in (Figure 4) we compare user responses with existing metrics. While **no strong correlation is observed**, weak alignment appears with LPIPS and our C2C pseudo-label. This suggests that our labels might capture aspects of perceptual or semantic quality, though further validation is needed.

4.3 CLIP-LPIPS regressor

We designed and trained a **semantic fidelity regressor** based on user study results and a binary-labeled inpainting dataset, where a label of `1` indicates a semantic error and `0` indicates no error. The architecture follows the structure

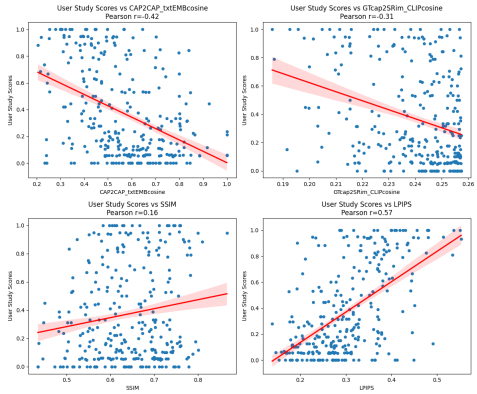

Fig. 4: Correlation plots between the User Study scores, the two proposed pseudo-labels and two classic metrics, LPIPS[7] and SSIM[2].

of **LPIPS**, but replaces the original ImageNet-trained VGG backbone with a **CLIP ResNet** backbone.

Initially, we experimented with using a **frozen backbone**, but observed significantly better results when **fine-tuning** the entire model. Therefore, all results reported from here onward use the fine-tuned backbone.

Training Details: For models trained on the user study data, we used 240 image pairs for training and left out 60 for evaluation. Training was performed over 30 epochs. For the inpainting dataset, we trained on 618 base images (with 155 held out for evaluation), each paired with **3 inpainted variations** (semantic errors) and **3 positive variations** (no errors). We evaluate all possible combinations of these variations during training, resulting in a large number of training pairs. These models were trained for 60 epochs.

Transfer Learning Experiments: We attempted to fine-tune a model pretrained on the inpainting dataset using the user study data. However, this approach resulted in **worse performance** than training from scratch using CLIP or ImageNet backbones. This highlights a key limitation: the **domain gap** between synthetic inpainting distortions and real-world semantic errors in SR images. As such, the inpainting dataset in its current form is not well-suited for transfer learning to the SR domain.

Model ablation: We conducted an ablation study to compare variations of our model by altering the number of ResNet layers (i.e., model depth) and by using two ResNet-50 backbones with different pretraining strategies. Specifically, we trained the CLIP-LPIPS regressor model on both the user study and the inpainting datasets, using either a standard ResNet-50 pretrained on ImageNet[22] or the image encoder from CLIP[8] pretrained on LAION[23], which is also based on a ResNet-50 architecture but differs in its initial layers. We will call the two pretrained resnet-50 models R50-CLIP-IMENC and R50-IMAGENET respectivelly. The results of this ablation are presented in Table 1.

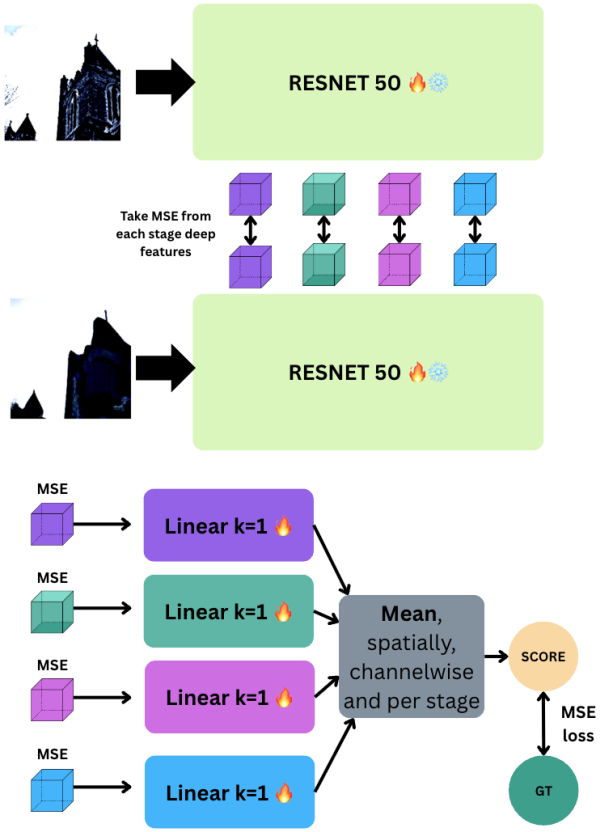

Fig. 5: Architecture of the semantic fidelity regressor based on CLIP (ResNet) and LPIPS structure. Both input images are processed through a ResNet-50 backbone, and deep features are extracted at each stage. The model computes the mean squared error (MSE) between corresponding features of the two images at each stage. These MSE values are then passed through individual linear layers, and the resulting outputs are averaged to produce the final semantic fidelity score.

4.4 Results and analysis

Table 1 presents an **ablation study** evaluating the effect of varying feature depth (from 2 to 4 stages), different pretraining ResNet-50 models (R50-CLIP-IMENC vs. R50-IMAGENET), and training on either the User Study or the Inpainting dataset. A notable observation is that the **R50-IMAGENET generally performs better** on the User Study dataset, whereas the **R50-CLIP-IMENC yields superior results** on the Inpainting dataset.

We hypothesize that this is due to the nature of the two datasets. The User Study data consists of **fine-grained annotations** that reflect medium semantic fidelity—subtle changes that often stem from small image details. These are aspects that traditional classification models, such as those pretrained on ImageNet [22], are typically **more sensitive to**. In contrast, the Inpainting dataset involves more **drastic semantic alterations**, where entire objects change shape or class, significantly affecting the overall meaning of the image.

In such cases, **R50-CLIP-IMENC**, as it is pretrained to align image features with textual descriptions, may be better suited to capturing these **global semantic shifts**. Conversely, R50-CLIP-IMENC may struggle to detect finer

Pretraining	Training	Depth	ImpDS MSE		ImpDS MSE Bin		User Study MSE		User Study SRCC	
			Test	Train	Test	Train	Test	Train	Test	Train
R50-IMAGENET	UserS.	1	0.526*	0.560*	0.636*	0.651*	**0.043**	0.040	0.729	0.941
R50-IMAGENET	UserS.	2	0.544*	0.545*	0.635*	0.638*	0.053	0.049	**0.776**	0.947
R50-IMAGENET	UserS.	3	0.550*	0.553*	0.638*	0.635*	0.061	**0.011**	0.658	**0.950**
R50-CLIP-IMENC	UserS.	1	0.639*	0.656*	0.697*	0.707*	0.063	0.014	0.720	0.943
R50-CLIP-IMENC	UserS.	2	0.603*	0.614*	0.666*	0.681*	0.060	0.021	0.707	0.916
R50-CLIP-IMENC	UserS.	3	0.644*	0.662*	0.706*	0.720*	0.065	0.015	0.712	0.932
R50-IMAGENET	ImpDS	1	0.168	0.159	**0.200**	0.184	0.330*	0.414*	-0.174*	-0.328*
R50-IMAGENET	ImpDS	2	0.179	0.166	0.222	0.212	0.287*	0.302*	-0.177*	-0.152*
R50-IMAGENET	ImpDS	3	**0.167**	0.155	0.208	0.191	0.358*	0.334*	-0.116*	-0.257*
R50-CLIP-IMENC	ImpDS	1	0.171	0.119	0.208	0.110	0.406*	0.368*	-0.612*	-0.487*
R50-CLIP-IMENC	ImpDS	2	0.178	**0.110**	0.209	**0.103**	0.411*	0.431*	-0.513*	-0.510*
R50-CLIP-IMENC	ImpDS	3	0.179	0.115	0.213	0.118	0.574*	0.556*	-0.362*	-0.409*

TABLE 1: Comparison of ResNet-based models with different encoder depths and pretraining strategies. We test on both datasets in all train and test sets. "*" indicates that a set is out of domain for a given model, thus a train set from one model is a validation to the counterpart

MSE and MSE Bin (lowest better) SRCC (highest better).

detail-level variations, as its pretraining objective encourages **abstraction over localized precision**. This might also explain why fine-tuning R50-CLIP-IMENC based models trained on the Inpainting dataset (ImpDS) yielded **poor results on the User Study task**: the domain gap between the datasets is too significant.

The fact that we were able to train models and achieve **reasonable SRCC and MSE validation values** demonstrates the **internal consistency** of our User Study dataset. While the dataset is too small to produce a deployable semantic fidelity metric at this stage, our results serve as a **promising proof of concept** for future work involving larger-scale user studies.

Subjective Analysis of Model Behavior

In this section, we qualitatively analyze both successful and failed model predictions. We consider not only the final output but also intermediate feature activations at different stages, prior to their aggregation by the regressor.

We compare the best-performing CLIP image encoder and ImageNet-pretrained backbone models for each dataset.

The activation maps reveal clear differences between the two pretraining strategies. The ImageNet-pretrained model exhibits a classic classification network pattern: early layers attend to fine details, while deeper layers progressively compress features into a focused representation of the main object. In contrast, the CLIP image encoder backbone model seems to attend to the entire object from the very first layers, consistent with its goal of encoding global image-level semantics to match textual captions.

These findings support our earlier observations: the ImageNet-pretrained model is better suited for our task, as it captures both localized detail and global structure. This also explains why increasing feature depth benefits ImageNet-based models, but not their CLIP image encoder counterparts.

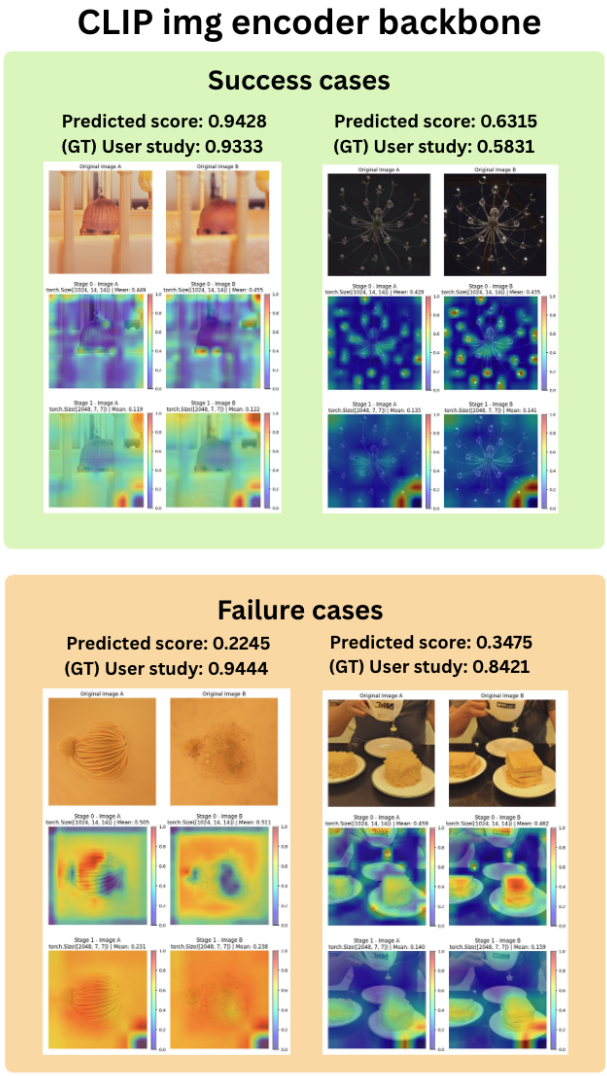

Fig. 6: Activation maps from the best CLIP-pretrained model on the User Study dataset.

4.5 Global Method Conclusions and Future Work

This study demonstrates the **feasibility of designing a semantic fidelity metric** for super-resolved (SR) images using human-annotated data. As a pilot project, our goal was to assess whether **user judgments of semantic fidelity could be meaningfully modeled**. The results are encouraging.

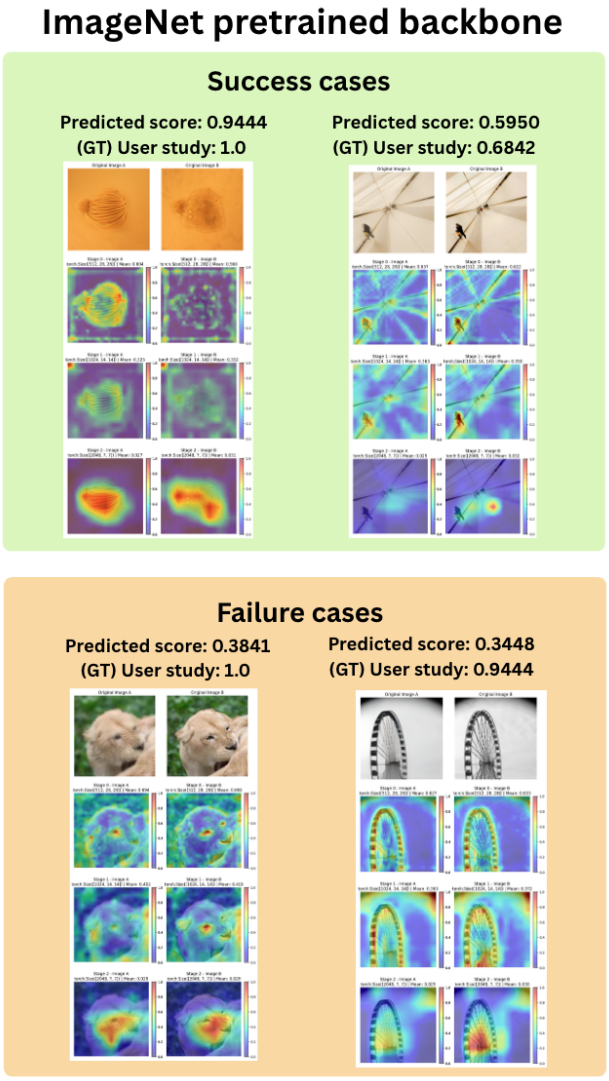

Fig. 7: Activation maps from the best ImageNet-pretrained model on the User Study dataset.

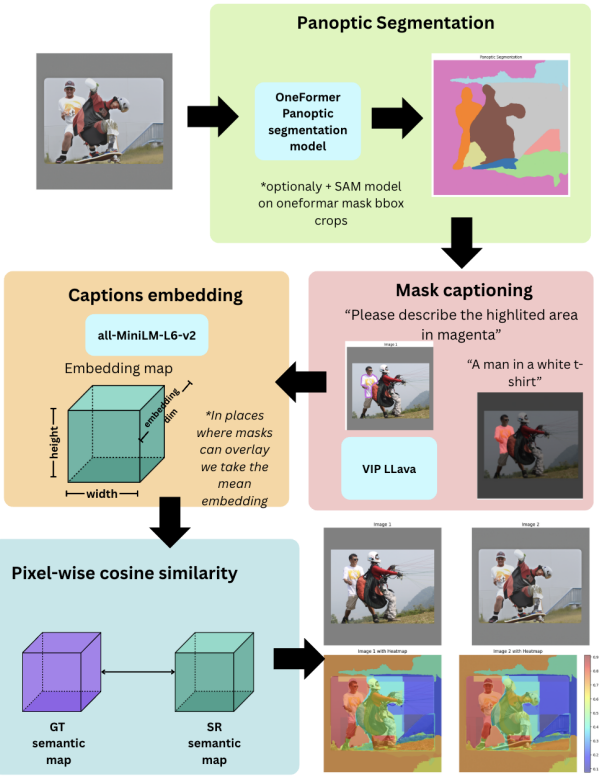

Fig. 8: Overview of the semantic map pipeline, composed of four stages. In the first three stages, both input images are processed in parallel: (1) panoptic segmentation is applied to each image, (2) each segmented mask is captioned, and (3) a pixel-wise embedding map is generated by embedding each caption and assigning it to the corresponding region. In the final stage, the pixel-wise cosine similarity is computed between the two embedding maps, yielding the semantic fidelity map.

We have also developed a **model architecture capable of capturing relevant aspects of human perception**, as reflected in the **SRCC and MSE scores**. Furthermore, we explored a **synthetic data strategy** (using the Inpainting dataset), which unfortunately did not transfer well to the User Study task due to **substantial domain differences**.

Future work will involve conducting a **more extensive and diverse user study**, incorporating a wider range of distortions and image types. Additionally, we will continue **refining our model architecture**.

5 Local Evaluation Methods

5.1 Panoptic+Captions Embedding Maps Dataset

This approach can be considered an extension of **caption embeddings pseudo-labels**, applied at a **pixel-wise level** rather than regressing a single score for an entire image. Due to the **unavailability of human-labeled data**, we developed a **synthetic dataset** for this purpose.

The pipeline to generate these pseudo-labels involves performing **panoptic segmentation** on whole images using a segmentation model: we employed **OneFormer** [24], and in some experiments, we layered it with **SAM** [25]. From each mask, we generate an **individual caption** describing that region using a **visual language model (VLM)**. Specifically, we utilized **VIP-LLava** [26], a model fine-tuned for interpreting visual prompts for captioning and question answering. This process produces many captions describing different parts of the image.

Next, we extract **embeddings** for each caption using a **text retrieval model** and overlay the masks and embeddings to create an **embedding map**. By computing the **pixel-wise cosine similarity** between two embedding maps, we can derive a **semantic similarity map** for a pair of images. This pipeline is illustrated in (Figure 8).

We developed multiple versions of this dataset and trained models on them, based both on filtering the **number of objects** appearing in the images and on pairing **ground truth (GT)** and **super-resolved (SR)** images, as well as pairing different SR images that correspond to the same GT image.

(Figure 9) illustrates examples where certain parts of the images—typically backgrounds—have **very low cosine similarity values** despite remaining unchanged. We attribute this to **hallucinations by the visual language model** when captioning those areas. This is one of the main **limi-**

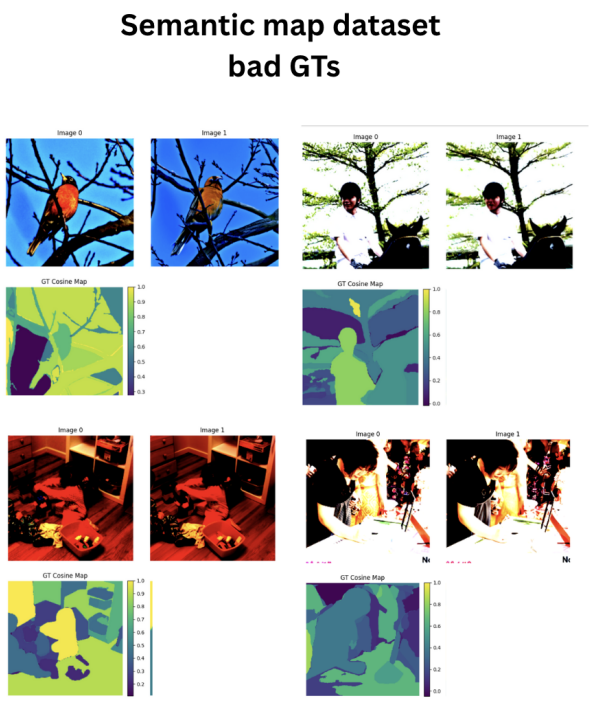

Fig. 9: Examples of poor-quality ground truth samples from the semantic maps dataset. Due to hallucinations introduced by the captioning model—particularly in background regions—some areas exhibit artificially low cosine similarity despite visual similarity. This highlights a key limitation of the current dataset, which will need to be addressed in future work.

tations of this pipeline: the synthetic data is quite **noisy**. In the current work, we trained on data containing some of this errors but acknowledge that in future work **improvements are needed**.

5.2 Inpainting dataset

We are able to harness the **inpainting dataset** described before in this local evaluation method by utilizing the **masks** that were used in the inpainting process. We compute a **semantic map** for each image pair by setting a **value of 0 in the inpainting mask region and 1 elsewhere.**

5.3 CLU model

CLU stands for **CLIP LPIPS UNet model**. This is the architecture we propose to solve the **semantic maps task**. It consists of a **R50-CLIP-IMENC or R50-IMAGENET** backbone that is fed with the images pair. Then, we take the **MSE of the deep features** from the model at each ResNet stage, as done in LPIPS, to feed a **decoder** passing each stage in a **UNet structure**.

We trained the CLU model on both the **semantic maps dataset** and the **inpainting dataset** but generally obtained **unfavorable results**. We attribute this model failure to **data noisiness**, and we believe that such an approach could work, but it requires **significantly more work on cleaning the dataset and increasing its size**.

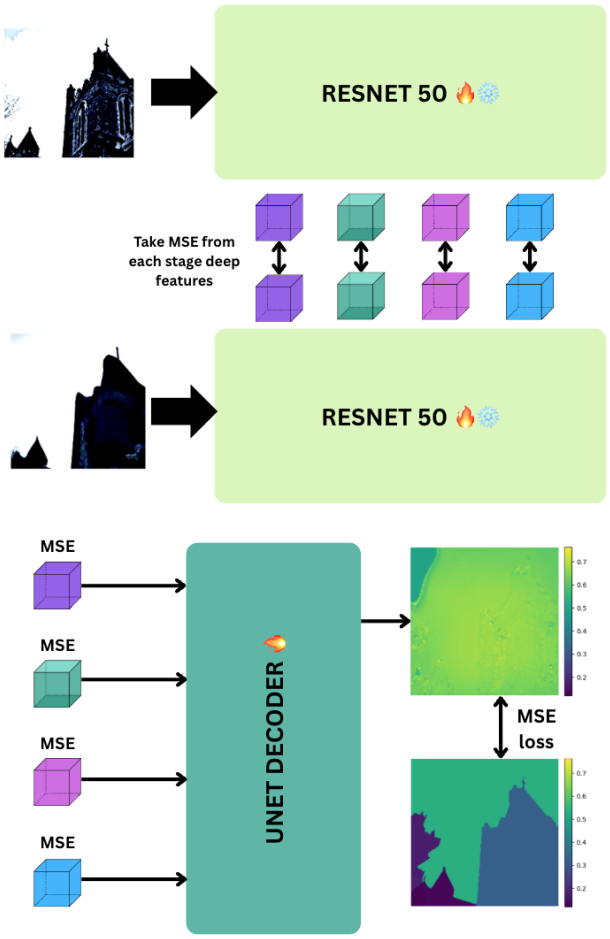

Fig. 10: CLU model architecture. The two input images are first processed through a shared pretrained ResNet-50 backbone, from which feature maps are extracted at each stage. The mean squared error (MSE) is then computed between the corresponding features of each image. These difference maps are subsequently passed through a decoder with a U-Net-like structure to generate the final output.

5.4 Results and analysis

In (Figure 11), we show some results of the CLU model trained on the **inpainting dataset**. The model does learn to identify and mask the inpainted region, but it also exhibits hallucinations in other parts of the image

In (Figure 12), we present results of the same CLU model trained on the **semantic maps dataset**. Here, we observe a **complete failure**. Although the model appears to train, and ImageNet-pretrained variant seems to capture some segmentation, **all model variants we tested produced unfavorable results**.

We attempted to **binarize the ground-truth semantic maps** using various thresholds, but this only introduced additional noise to the dataset and did not improve performance. We also **evaluated the models across datasets** for example, testing a CLU model trained on the inpainting dataset on the semantic maps task—but again, results were unfavorable.

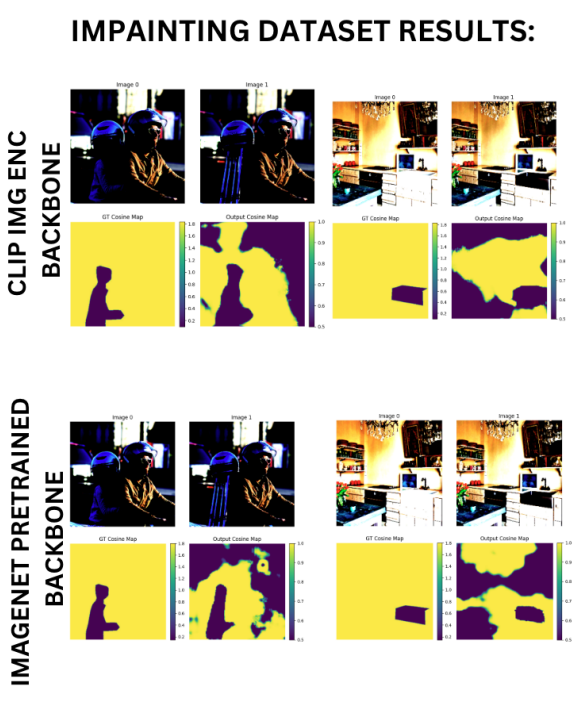

Fig. 11: Qualitative results of the CLU model trained on the Inpainting dataset. The model successfully identifies the inpainted regions; however, it also hallucinates changes in unrelated areas. Additionally, results show that using an ImageNet-pretrained backbone produces sharper and more localized predictions compared to a CLIP-pretrained image encoder.

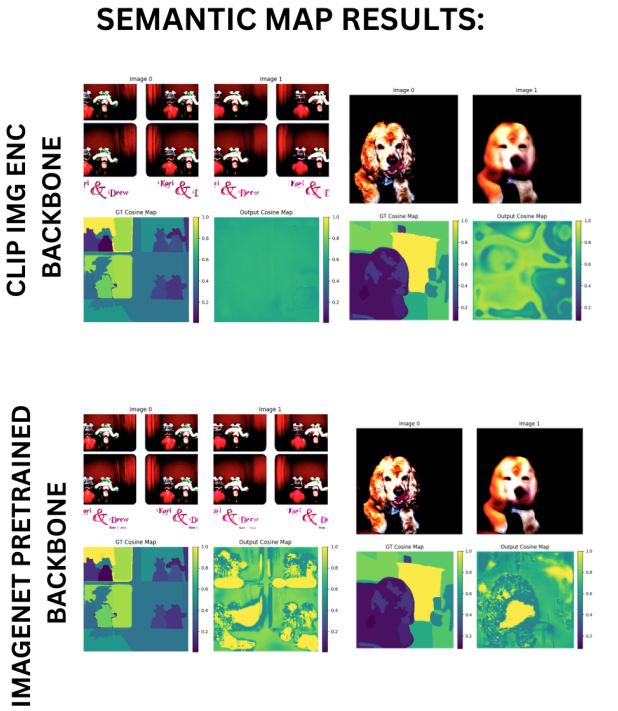

Fig. 12: Qualitative results of the CLU model trained on the semantic maps dataset. The model generally fails to learn a meaningful semantic map. However, variants using an ImageNet-pretrained backbone capture some limited structure in the segmentation, indicating partial learning despite overall poor performance.

5.5 Local Method Conclusions and Future Work

We have seen that it is indeed possible to train the CLU model on the Inpainting dataset, but it is clear that there is still significant room for improvement in the model architecture design. Given the simplicity of the inpainting task in terms of semantic changes, the model should be able to learn this more effectively. This could also be due to images not being inpainted properly in the inpainting dataset itself, which is another topic to improve further in future work.

In future work, we plan to explore initializing the model from a pretrained segmentation UNet, which we believe would provide a much stronger foundation than starting from CLIP or ImageNet pretraining.

Regarding the semantic maps, we believe that substantial improvements are needed to reduce the captioning model hallucinations and to better control the level of detail produced during panoptic segmentation. One potential strategy would be to focus segmentation only on the main scene objects, ignoring background regions. This could help reduce noise and hallucinations caused by describing irrelevant background areas.

6 Futher approaches

We initially attempted to use **CLIP embeddings** to measure semantic fidelity. The most straightforward approach involved passing both images through the image encoder and computing the **cosine similarity** between their embeddings. Since both embeddings originate from the same encoder (the image encoder), this is referred to as **intra-modal cosine similarity**. This method was somewhat effective in capturing semantic differences, but we ultimately treated it as a **baseline**.

6.1 SpLiCE for Explainability and Improving Embeddings

Our subsequent efforts focused on **increasing the explainability** of the embeddings and improving their performance relative to this intra-modal baseline.

In 2024, Bhalla et al. [27] introduced **SpLiCE**, a method designed to make CLIP image embeddings more interpretable by expressing them as a **linear combination of text concept embeddings**.

This technique enables applications such as:

- Bias detection in vision-language models
- Identification of dominant semantic concepts in images
- Modification of embeddings by filtering out specific concepts

For our task, we applied SpLiCE to analyze the **conceptual decomposition of two images** and quantify their semantic differences.

As shown in Figure 13, we explored three decomposition strategies. The **symmetric** approach constrains both decompositions to use the same set of concepts derived from

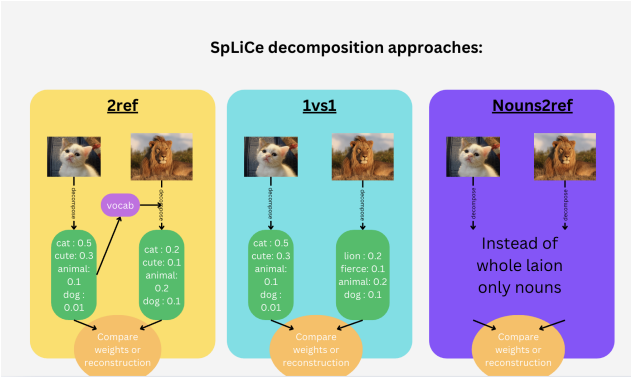

Fig. 13: Illustration of the three SpLiCE-based comparison strategies. In the **symmetric** approach, the GT image is first decomposed, and the same vocabulary is used to force the decomposition of the evaluated image. In the **non-symmetric** approach, both images are decomposed using the full concept vocabulary. The **Nouns2ref approach** is a variation of the other two restricting the vocabulary used in the decomposition

Method	ImpDS MSE Test	ImpDS MSE Train	ImpDS MSE Bin Test	ImpDS MSE Bin Train
clip_intra-modal	**0.451**	0.428	**0.551**	0.530
splice_sym	0.458	**0.420**	0.556	**0.510**
splice_nonsym	0.485	0.478	0.556	0.549

TABLE 2: Quantitative results on the inpainting dataset using training-free CLIP-based methods. The table reports MSE for both standard and binarized ground truth settings. The **clip_intra-modal** method—based on direct cosine similarity of image embeddings—outperforms both **splice_sym** and **splice_nonsym** in test scenarios, suggesting that SpLiCE-based decomposition does not improve performance in this context.

Method	User Study MSE Test	User Study MSE Train	User Study SRCC Test	User Study SRCC Train
clip_intra-modal	**0.392**	**0.389**	**-0.712**	**-0.736**
splice_sym	0.411	0.404	-0.699	-0.702
splice_nonsym	0.426	0.421	-0.669	-0.686

TABLE 3: User study dataset evaluated on the same training-free CLIP-based methods. **clip_intra-modal** produces results very correlated to the User Study scores.

the GT image. The **non-symmetric** approach uses independent decompositions for each image based on the full vocabulary. We also explored restricting the vocabulary used in the decomposition, this is the third method shown in the right side of the figure.

6.2 Clip training free results

We evaluated the previous methods on the user study and inpainting dataset, results can be found in (Table 2 and Table 3). We find that the proposed experiments do not improve clips intra-modal cosine similarity on any of the two datasets. We do find though that clip intra-modal cosine similarity has a SRCC of 0.713 on the user study test set, which we surpass with our CLIP-LPIPS regressor discussed above.

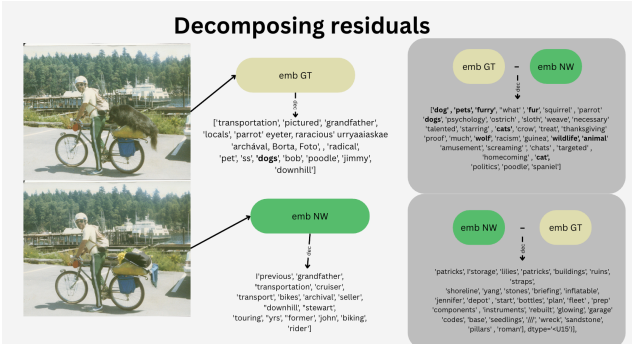

Fig. 14: Visualization of SpLiCE decomposition applied to residual embeddings. The top row shows the original images; the second row highlights concept components extracted from the residual vectors (`GT - New` and `New - GT`), which are intended to reflect the semantic differences between the images. Despite some promising visual patterns, the method proved unstable and did not yield consistent or reliable scores.

6.3 Other Results

While experimenting with SpLiCE decomposition, we explored applying it to the **residual of two image embeddings**. Specifically, we computed the residual vectors (`GT - New` and `New - GT`) and decomposed these using SpLiCE. Interestingly, the resulting concepts often corresponded to the semantic differences between the two images.

We then attempted to derive a semantic score by comparing the reconstructed embeddings of the two residuals. However, this approach also yielded **poor results**, likely due to the inherent **noisiness of the decomposition** process and instability in capturing consistent residual concepts.

7 Conclusions

Current metrics used for training and evaluating super-resolution (SR) models often fail to capture **semantic errors** that are perceptually significant to humans. While traditional methods like **PSNR**[1], **SSIM**[2], and even **LPIPS**[7] can quantify certain aspects of image fidelity, they fall short in scenarios where objects undergo a **semantic class change** while preserving overall visual similarity.

This work addresses that limitation by framing **semantic fidelity** as a primary evaluation objective and proposing methods to better capture it. Through the design of specialized datasets and the exploration of **CLIP-based approaches**, we aim to contribute more meaningful evaluation tools to the SR community.

Although we initially hypothesized that **CLIP image encoder as our backbone** would be better suited for this task due to its semantic capabilities, our results show that a **classification-pretrained backbone** (e.g., ImageNet) yielded significantly better performance — both in regression-based evaluation and in semantic map prediction.

We also developed a **pilot user study** to collect human judgments of semantic fidelity, demonstrating its viability for future large-scale studies that could support robust metric development. As part of this effort, we built and re-

leased a experimental framework, including a streamlined web interface and annotation pipeline, which enables scalable and repeatable collection of user feedback on super-resolved images. This framework lays the groundwork for future experiments involving a larger and more diverse participant pool, as well as more complex image types and distortion categories.

Additionally, we experimented with multiple **synthetic labeling strategies**, such as inpainting-based and panoptic+caption-based datasets. However, many of these proved to be **"too noisy"** for effective model training. We believe that a more **curated and controlled synthetic dataset** could still be useful, but it would require significant improvements in bad samples filtering and overall pipeline design.

8 Future Work

We plan to continue this line of research over the coming summer, with a focus on scaling and refining our methodologies. A key priority is to conduct a **larger and more carefully designed user study**, featuring more participants and a better-curated set of images. This will allow us to train a more accurate and reliable **semantic fidelity metric**, which we will then use to evaluate its impact on **super-resolution model training**.

In parallel, we will further explore the development of the **semantic fidelity map**, focusing on both **architectural improvements** and the enhancement of our **synthetic data generation** pipeline.

Through these efforts, our goal is to deliver two complementary contributions to the field:

- A robust, global **semantic fidelity metric** for image-level evaluation.
- A reliable, local **semantic fidelity map** capable of highlighting region-wise semantic discrepancies.

Together, these tools aim to fill a critical gap in the **super-resolution deep learning community**, by enabling the evaluation and training of models that better align with human perception.

Acknowledgments

This project has been a very diverse and fun learning experience, I went from very naive approaches and deep diving into a lot of papers, trying to understand why things weren't working. Learned a lot about how to make synthetic datasets, the many different models and techniques that went in to make them and even developing a user study platform and conducting it. It was also a challenging but very insightful task looking back and attempting to organize all these attempts and results into a path that makes sense.

I would like to thank both my supervisors Javier Vazquez and Shaolin Su and Alex Gómez for their invaluable guidance and support throughout this project, because of how much I've learned both through them and by trial and error but also for their interest in the project which has been very stimulating all along. I would also want to acknowledge the Computer Vision Center for providing the computational resources necessary for this research and the very enriching and fun community that constitutes it.

References

[1] O. Keleş, M. A. Yılmaz, A. M. Tekalp, C. Korkmaz, and Z. Dogan, "On the computation of psnr for a set of images or video," arXiv:2104.14868 [eess.IV], 2021, picture Coding Symposium (PCS) 2021.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[3] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[4] T. Zhu, B. Peng, J. Liang, T. Han, H. Wan, J. Fu, and J. Chen, "Vitscore: A novel vision transformer-based metric for image semantic similarity," *Conference on Computer Vision and Pattern Recognition*, 2024, proposed in GSScore comparative analysis[1][2].

[5] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: a reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 7514–7528.

[6] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, "Seesr: Towards semantics-aware real-world image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25 456–25 467.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[9] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," 2022. [Online]. Available: https://arxiv.org/abs/2203.02053

[10] M. Mistretta, A. Baldrati, L. Agnolucci, M. Bertini, and A. D. Bagdanov, "Cross the gap: Exposing the intra-modal misalignment in clip via modality

inversion," 2025. [Online]. Available: https://arxiv.org/abs/2502.04263

[11] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," 2021. [Online]. Available: https://arxiv.org/abs/2112.12750

[12] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2024.

[13] K. Zhang, J. Liang, L. V. Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," 2021. [Online]. Available: https://arxiv.org/abs/2103.14006

[14] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[15] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *IEEE International Conference on Computer Vision*, 2021, pp. 4791–4800.

[16] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *International Journal of Computer Vision*, 2024.

[17] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1833–1844, 2021.

[18] T. Yang, R. Wu, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," in *European Conference on Computer Vision (ECCV)*, 2024, arXiv:2308.14469 [cs.CV].

[19] S. Su, V. Hosu, H. Lin, Y. Zhang, and D. Saupe, "Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects," *arXiv preprint*, 2021, bMVC 2021 virtual conference. [Online]. Available: https://www.bmvc2021-virtualconference.com/assets/papers/0868.pdf

[20] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, *et al.*, "Qwen2.5 technical report," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2412.15115

[21] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5776–5788. [Online]. Available: https://arxiv.org/abs/2002.10957

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[23] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs," in *Proceedings of Neurips Data-Centric AI Workshop*, 2021.

[24] J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2211.06220

[25] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[26] M. Cai, H. Liu, D. Park, S. K. Mustikovela, G. P. Meyer, Y. Chai, and Y. J. Lee, "Vip-llava: Making large multimodal models understand arbitrary visual prompts," 2024. [Online]. Available: https://arxiv.org/abs/2312.00784

[27] U. Bhalla, A. Oesterling, S. Srinivas, F. P. Calmon, and H. Lakkaraju, "Interpreting clip with sparse linear concept embeddings (splice)," 2024. [Online]. Available: https://arxiv.org/abs/2402.10376