# Hyper-Thumbnail Super Resolution
# UROP 1100 Report

Momin Ahmed - 21017135

May 12, 2025

## 1 Abstract

Modern image resampling algorithms are focused on embedding high-resolution (HR) images within low-resolution (LR) thumbnails while preserving sufficient information for efficient HR reconstruction. Qi, Chenyang, and others [1] proposed a novel framework called HyperThumbnail for real-time rate-distortion-aware rescaling of 6K images. It employs an encoder with a quantization prediction mechanism to encode the HR image in a JPEG LR thumbnail at the expense of file size for negligible reconstruction quality. Extensive experiments reveal that this method outperforms existing image rescaling techniques in rate-distortion efficiency and makes real-time 6K image reconstruction possible.

In this work, we experiment with a different encoder architecture. We replace the proposed RDBUnet encoder with Transformer encoder and notice the change in bpp, lpips and psnr. The modified architecture tends to achieve higher degrade PSNR, higher LPIPS_Y, lower LPIPS_RGB and similar bpp suggesting potentially better pixel-level accuracy, less distortion, and potentially less perceptual similarity in the luminance channel compared to the original architecture.

# 2 Introduction

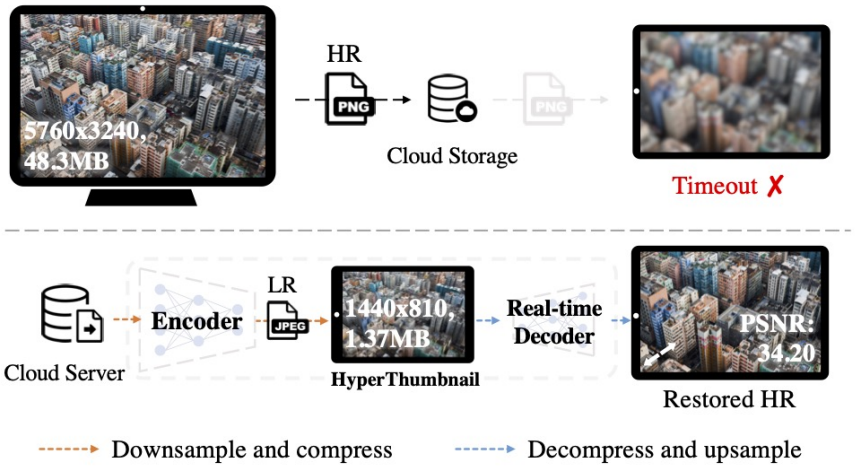



Figure 1: Task Overview

The rapid growth of high-resolution (HR) images being posted on the web has generated issues in efficient storage and transmission. While cloud storage platforms such as iCloud allow users to save space by maintaining low-resolution (LR) thumbnail copies on their devices, the process of accessing the complete HR images can lead to compromised user experience, especially over slow internet connections or high resolution images.

For the sake of enhancing user experience, real-time rescaling of images is proposed as a solution. HyperThumbnail framework compresses HR images to LR JPEG thumbnails to enable rapid preview and local reconstruction of HR images when necessary. This approach can also circumvent resolution limitations on platforms like WhatsApp.

Although advances have occurred in capturing ultra-high-resolution footage (e.g., 4K and 6K), present rescaling methods have limitations, such as being not optimally designed regarding file size and fidelity of reconstruction. Comparisons suggest that past super-resolution techniques would compromise on high-frequency details, while more recent flow-based approaches, albeit very powerful, are plagued by their complexity as well as by inefficiencies of file size to apply in practice.

On the other hand, the HyperThumbnail framework works to optimize rate-distortion performance and real-time 6K image reconstruction and corrects the disadvantages of existing methods.

## 2.1 Important Evaluation Metrics

### 2.1.1 Bits-per-pixel

It is defined as the rate as the ratio between the thumbnail file size and the number of pixels in the HR image, also known as the bits-per-pixel (bpp). The distortion consists of two parts: the perceptual quality of the thumbnail (LR distortion) and the fidelity of the restored HR image (HR distortion). The rate-distortion performance evaluates an image rescaling framework in both storage cost and visual quality [1].

### 2.1.2 Learned Perceptual Image Patch Similarity (LPIPS_)

The Learned Perceptual Image Patch Similarity (LPIPS_) calculates perceptual similarity between two images.
LPIPS essentially computes the similarity between the activations of two image patches for some pre-defined network. This measure has been shown to match human perception well. A low LPIPS score means that image patches are perceptual similar [2].

### 2.1.3 Peak Signal-to-Noise Ratio (PSNR)

Peak signal-to-noise ratio (PSNR) is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed as a logarithmic quantity using the decibel scale[3].

# 3 HyperThumbnail

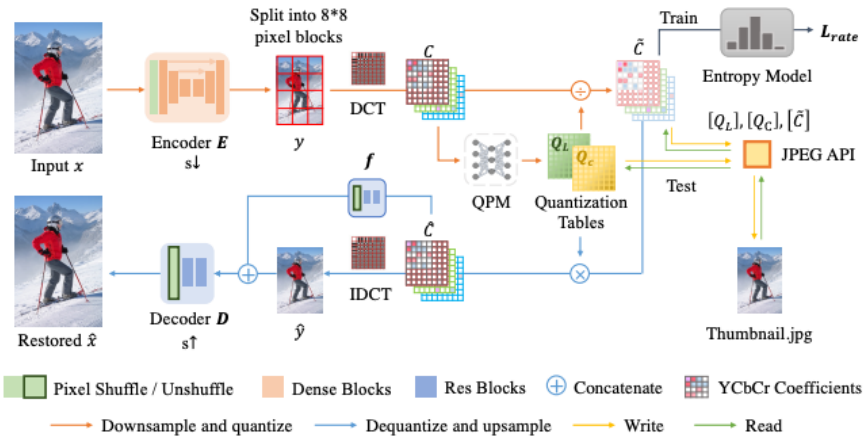



Figure 2: Architecture

- Framework Design:

  The framework puts high-resolution (HR) images into low-bitrate JPEG thumbnails using an encoder and a quantization table predictor.

  It employs an asymmetric encoder-decoder model such that the encoder handles most computation to enable efficient upscaling of thumbnails to 6K real-time, overperforming previous flow-based solutions.

- Rate-Distortion (RD) Optimization

  The document terms rate in bits-per-pixel (bpp) based on thumbnail file size relative to the number of pixels in the HR image. Distortion takes into account not only thumbnail quality but also recovered HR image quality.

  Unlike previous rescaling work, which is often biased to neglect RD performance, the approach here optimizes image quality and bpp from entropy models simultaneously.

- Quantization Prediction Module (QPM):

  The paper suggests a novel QPM for image-adaptive quantization tables to improve RD performance beyond the traditional fixed tables.

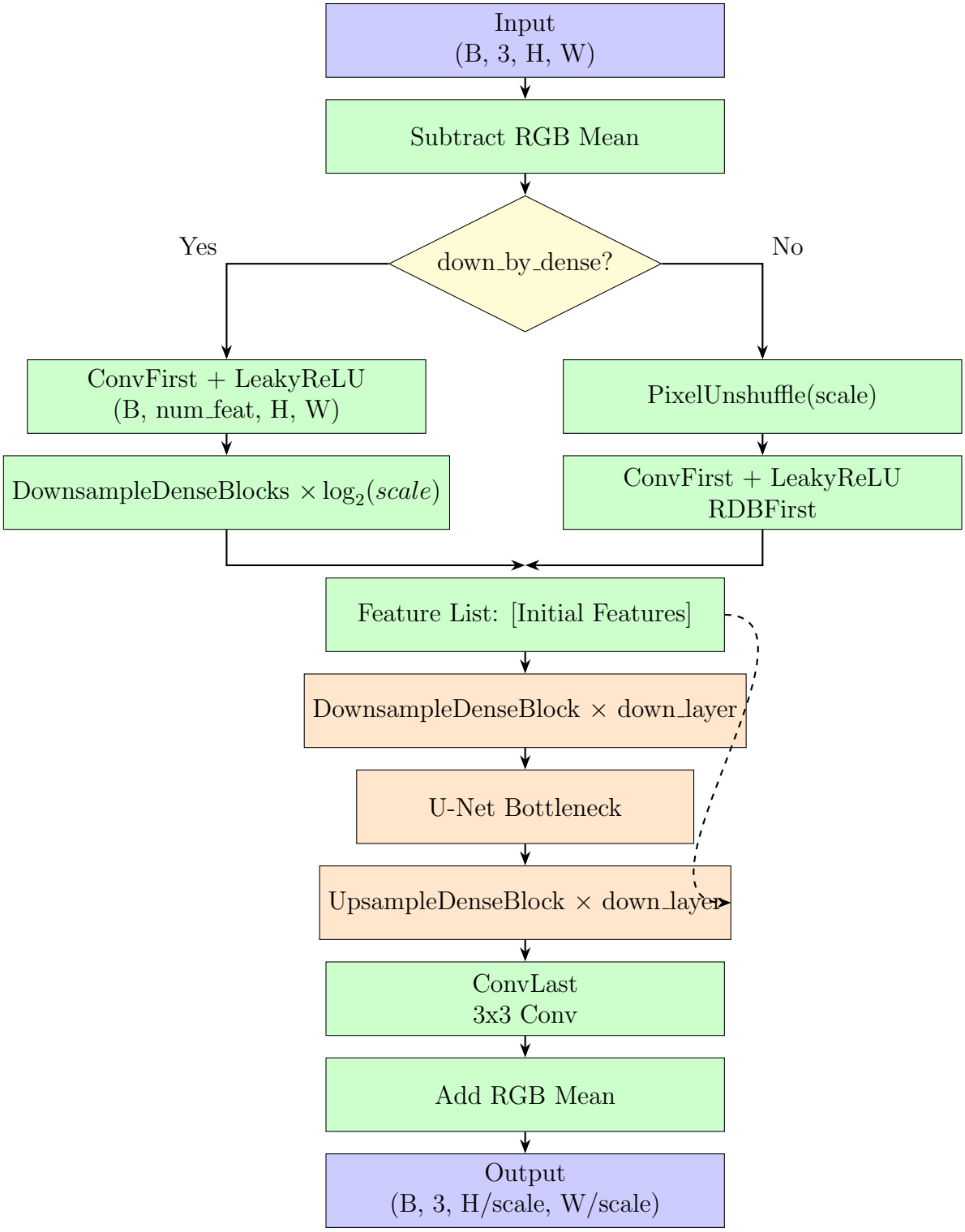
  Frequency-aware decoding is also included to minimize JPEG artifacts and improve the quality of HR reconstruction.

- Experimental Validation:

  Experiments demonstrate that HyperThumbnail performs better than the current state-of-the-art image rescaling methods in both LR and HR images, with improved quality, faster reconstruction time, and similar file sizes.

## 3.1 RDBUnet Encoder

The current HyperThumbnail framework employs a U-Net architecture with residual dense blocks (RDBs) for the encoder. The RDBUnet encoder is designed to efficiently capture and reconstruct high-frequency details in images, making it suitable for tasks like image rescaling. The architecture is summarized in the diagram below.




UNet type structures like RDBUnet are able to maintain fine details in its encoder-decoder structure, which effectively preserves spatial information through downsampling and upsampling. This is especially necessary in recovering high-quality images

from thumbnails with low bitrates. The residual connections in RDBUnet facilitate efficient gradient propagation, helping to avoid issues like vanishing gradients and potentially enhance training stability and convergence. Also, its form is inherently suitable for such processes as super-resolution and thus is a strong candidate for high-fidelity image reconstruction.

On the downside, UNets tend to have more chances of overfitting particularly for more complicated models, and if the training dataset is either too small or too homogeneous. While U-Net models can reconstruct images nicely, they may not reconstruct model long-range dependency as well as transformer-based approaches, which would limit their capacity in situations in which contextual consideration across the entire image is absolutely essential.

# 4 Transformer Encoder

Using a transformer encoder in the context of the HyperThumbnail framework for rescaling high-resolution images can offer certain advantages. First, the ability of the transformer to capture long-range dependencies makes it appropriate to represent sophisticated interactions between image regions, which is critical to recover the high-fidelity details in the rescaling process. This can enhance the quality of the upscaled image by avoiding the loss of contextual information in the rescaling process.

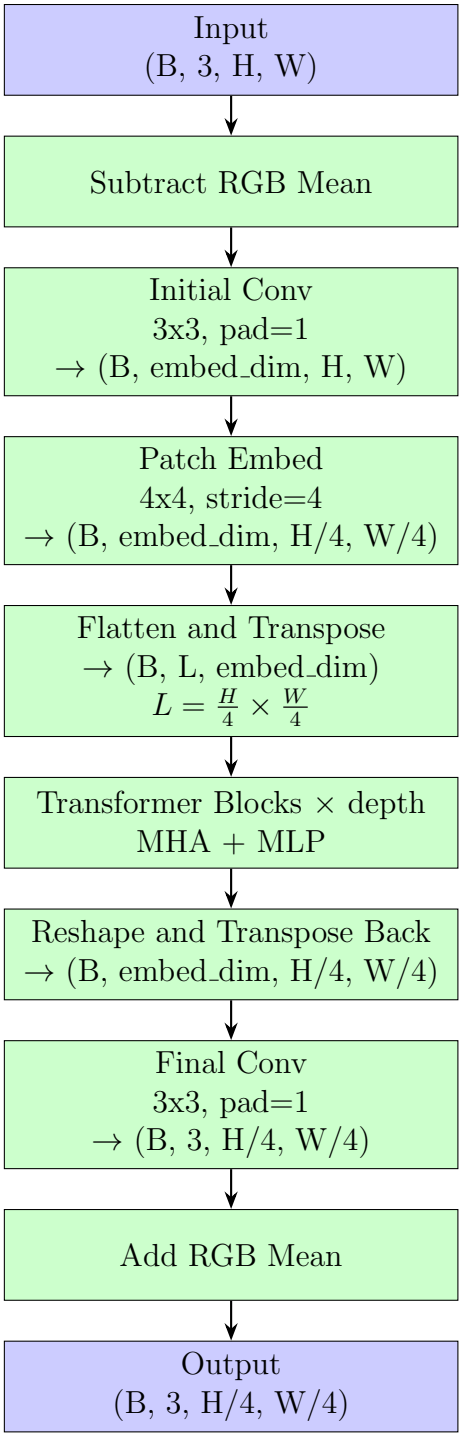
Additionally, the capacity of transformers to process in parallel could significantly speed up the encoding process, allowing real-time image reconstruction to be more efficient. This is particularly valuable when handling large high-definition images because it reduces latency and improves user experience. Although due to a more complex model, this is probably not an expectation.

The self-attention mechanism also enhances the model's performance by enabling it to dynamically attend to the most suitable areas of the image, simplifying the reconstruction process, and minimizing artifacts. The flexibility of transformer architectures also makes it easy to incorporate other features or enhancements, such as incorporating multimodal inputs or processing various types of images.

Finally, using pre-trained transformer models can accelerate development and improve performance because pre-trained models can be fine-tuned for the specific requirements of the HyperThumbnail framework, leading to better overall results in image rescaling operations.

## 4.1 Architecture

The transformer encoder architecture used is summarized in the diagram below.




## 4.2 Results

We trained both the RDBUnet and Transformer encoder architectures on the same dataset and compared their performance using the same evaluation metrics. Both

models were trained for approximately 12 hours on one RTX 5080 GPU. The results are summarized in the graphs below.


Figure 3: Progress and BPP


Figure 4: LPIPS_ and PSNR

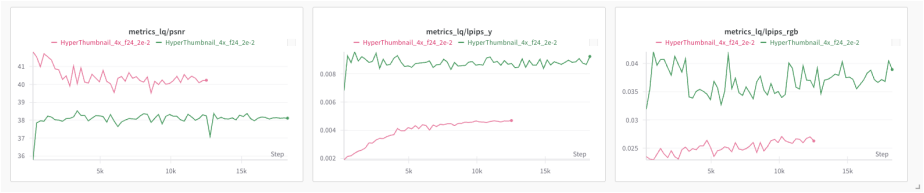



Figure 5: LPIPS_ and PSNR

The green represets the original RDBUnet architecture and the pink represents the transformer encoder architecture.
Here is a sample image from the transformer encoder architecture.

Figure 6: * scaled_lr

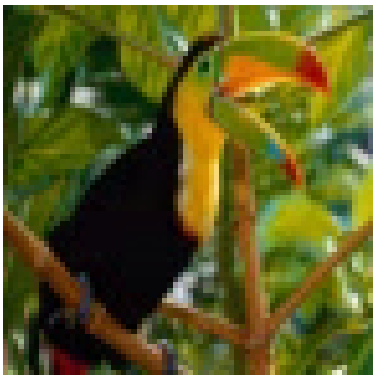


Figure 7: * scaled_degrade_lr

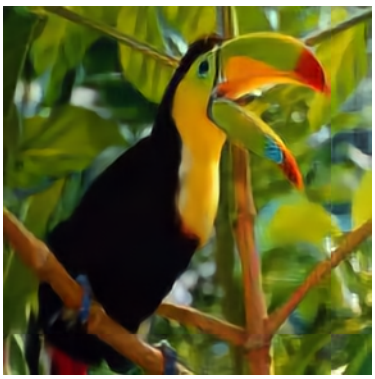


Figure 8: * reconstructed_hr

Figure 9: Comparison of scaled inputs and reconstructed output

Here is the sample image from the RDBUnet architecture.

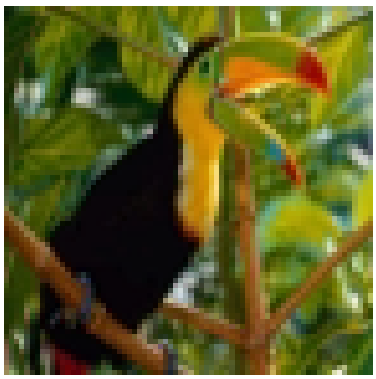


Figure 10: * scaled_lr

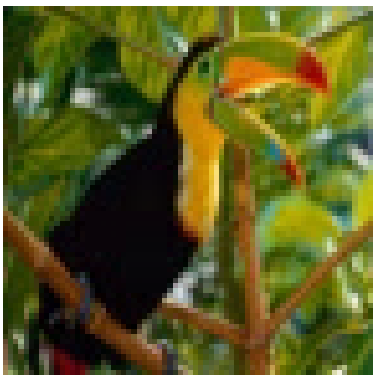


Figure 11: * scaled_degrade_lr

Figure 12: * reconstructed_hr

Figure 13: Comparison of scaled inputs and reconstructed output

## 4.3 Conclusion

The first thing to note is that the BPP for both models are similar. The Unet based model trained faster as the transofrmer was only able to progress about 75% in global_steps for the same time.
There was slight difference in the PSNR as metric/psnr is slightly lower while degrade_psnr is slightly higher for the transformer model. This suggests that the transformer model is able to achieve better pixel-level accuracy, but it produces reconstructions that are numerically less similar to the ground truth, indicating slightly worse pixel-level fidelity.

There is notable differents in LPIPS_Y and LPIPS_RGB. The transformer model has a lower LPIPS_ value which suggests that it is able to achieve less distortion in the luminance channel. Particularly, in low-quality input scenarios, the modified architecture better preserves perceptual quality than the original.

# 5 References

(1): Qi, Chenyang, et al. "HyperThumbnail: Real-time 6K Rate-Distortion-Aware Image Rescaling." arXiv preprint arXiv:2303.14123 (2023).
(2) `https://lightning.ai/docs/torchmetrics/stable/image/learned_perceptual_image_patch_similarity.html`
(3) `https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio`