Diffusion Model for Super Resolution of Ship Images

Student
Andrea Mazzitelli - 1835022

Advisor
Prof. **Danilo Comminiello**

Co-Advisor
Ing. **Alessandro Nicolosi**

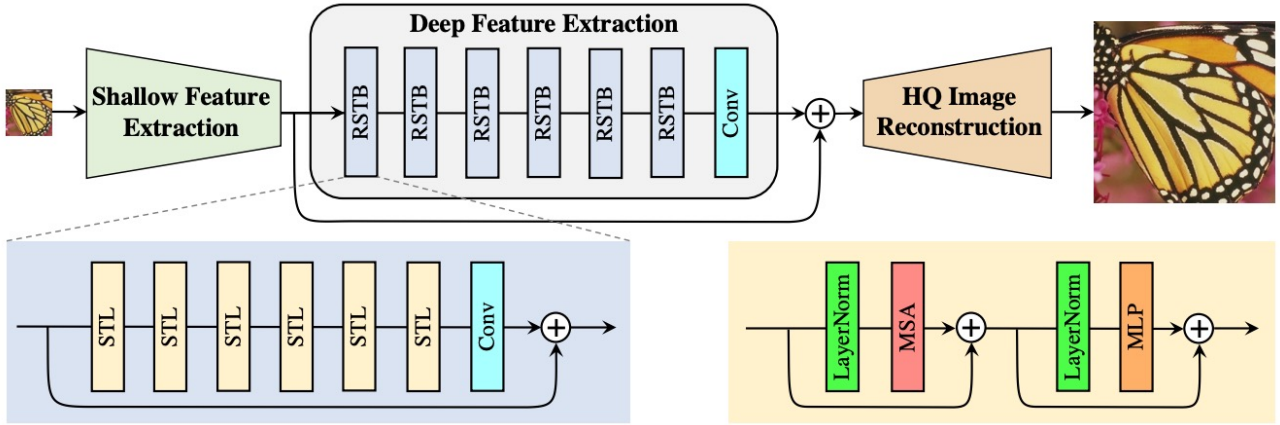
Engineering in Computer Science
Dipartimento di Ingegneria informatica, automatica e gestionale
Facoltà di Ingegneria dell'informazione, informatica e statistica
Anno Accademico 2022-2023
Sapienza University of Rome

In Collaboration with

Table of contents

Background - Image Super Resolution

Baseline - SwinIR



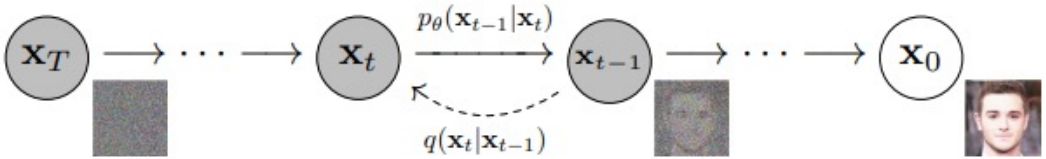
(a) Residual Swin Transformer Block (RSTB)

(b) Swin Transformer Layer (STL)

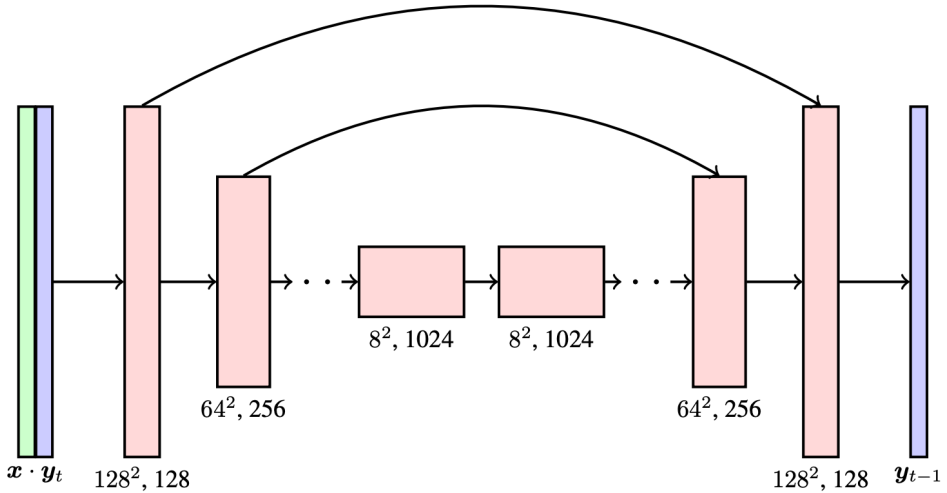
Chosen Approach - Diffusion Model

Diffusion Models leverage a diffusion process to iteratively transform an initial distribution into a target one. This is achieved using a **Forward Process** and a **Reverse Process.**

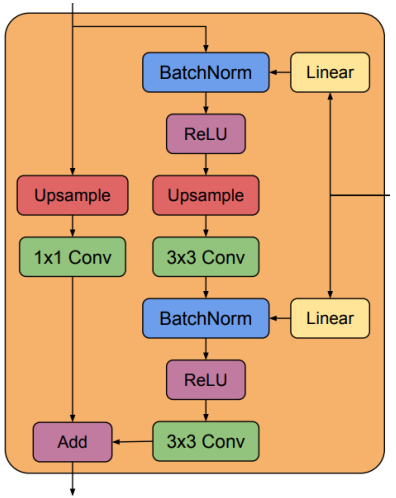
By including a **Conditioning Element** in the process is possible to guide the diffusion process in generating an output that have specific characteristics.



Super Resolution via Repeated Refinement - SR3



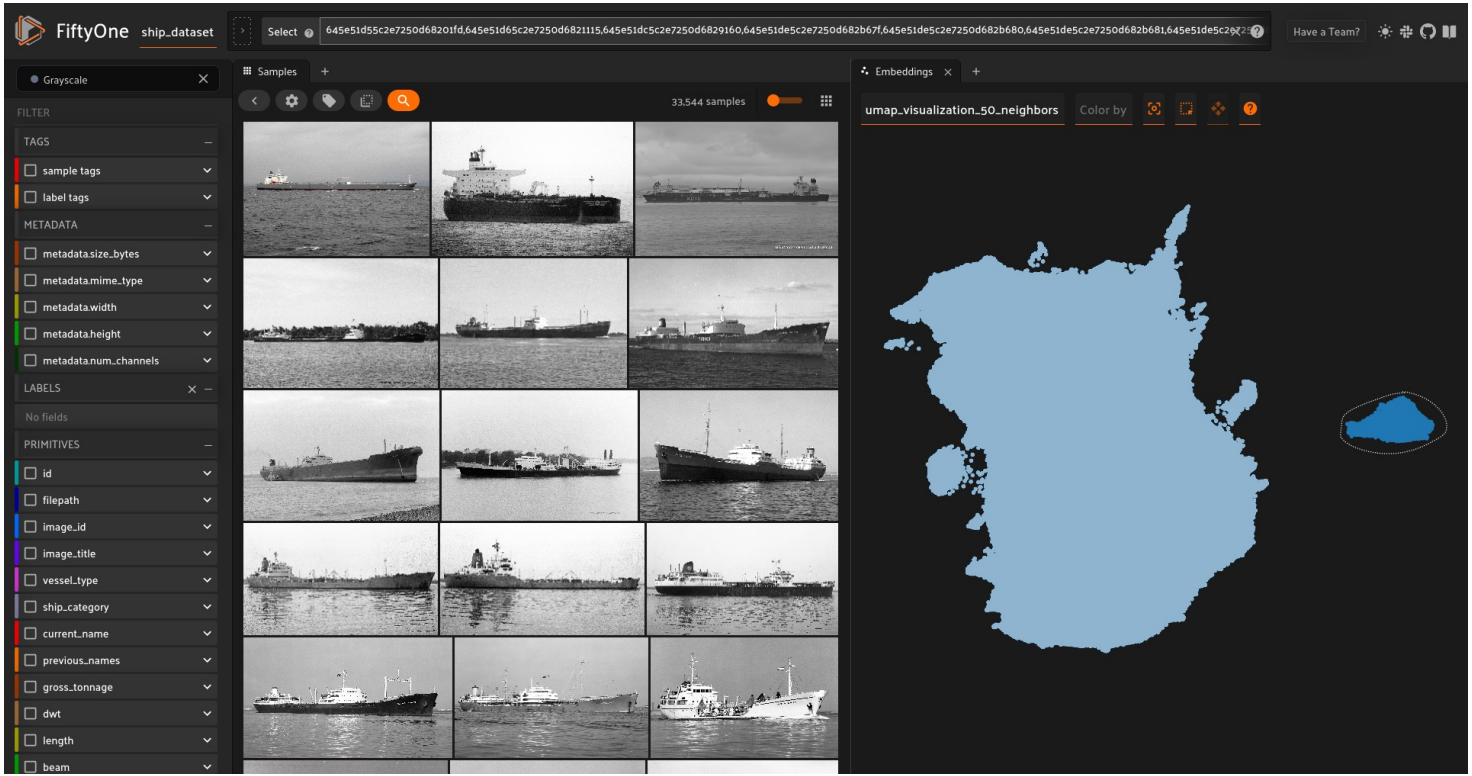
SR3 Architecture



G-Block

Dataset

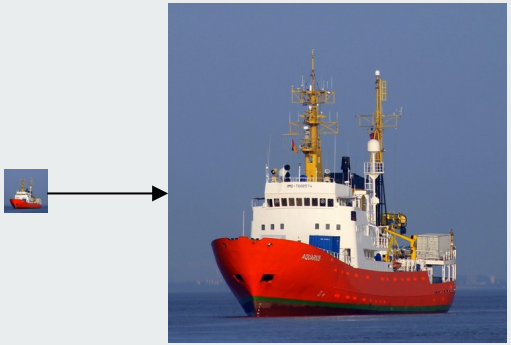

SHIPSPOTTING.COM
scraped ≃ 1.5 images
CleanVision
Open-Source
FiftyOne
FINAL DATASET
1.5 Millions Images


FiftyOne
ship_dataset
Select
Have a Team?
Grayscale
FILTER
TAGS
sample tags
label tags
METADATA
metadata.size_bytes
metadata.mime_type
metadata.width
metadata.height
metadata.num_channels
LABELS
No fields
PRIMITIVES
id
filepath
image_id
image_title
vessel_type
ship_category
current_name
previous_names
gross_tonnage
dwt
length
beam
Samples
33,544 samples
Embeddings
umap_visualization_50_neighbors
Color by

Training Details

TASK : 8x Super Resolution
64x64 → 512x512

Devices : 2 NVIDIA RTX A6000, 48GB



- **SwinIR:**
  - 12 million parameters
  - 500k iterations
  - batch size :32
  - 10 days
  - inference time: few seconds
- **SR3:**
  - 155 million parameters
  - 1M iterations
  - batch size: 16
  - 17 days
  - inference time: 2:14

Results - Objective Metrics

	PSNR (db)	SSIM	FID
SwinIR	**23.80**	**0.6805**	50.79
SR3	21,07	0.57093	**16.59**

PSNR - Peak Signa to Noise Ration, SSIM - Structural Similarity Index Measures, FID - Fréchet Inception Distance

SwinIR outperforms **SR3** in both PSNR and SSIM but not in FID. This may appear to be somewhat counterintuitive but there is a reason for it.

Results - Subjective Metric

2AFC - 2 Alternative Force Choice

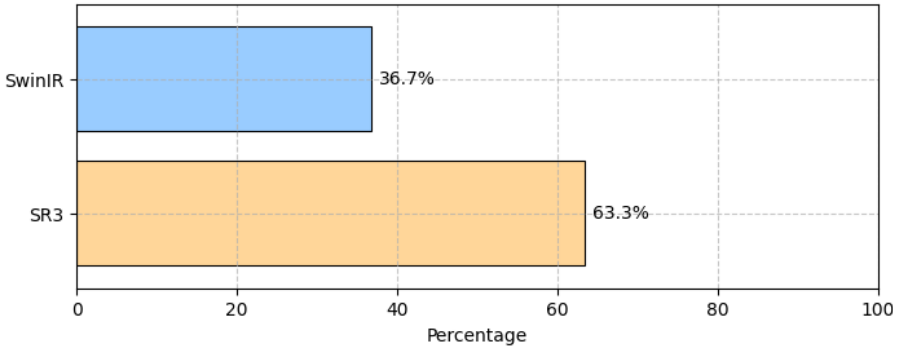
INPUT

SwinIR

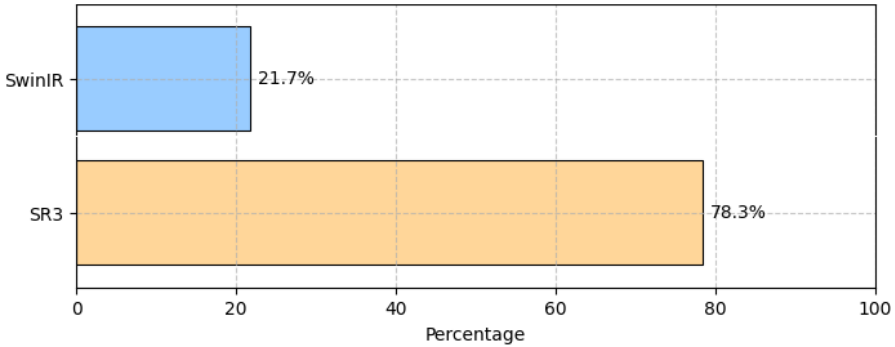
SR3

Results - Subjective Metrics contd.

Which is the best high-resolution version of the given low-resolution image?



Which is the most realistic version of the given low-resolution image?



Results - Visual Analysis

INPUT

SwinIR

SR3

Results - Visual Analysis

INPUT

SwinIR

SR3

Results - Visual Analysis

INPUT

SwinIR

SR3

Results - Visual Analysis

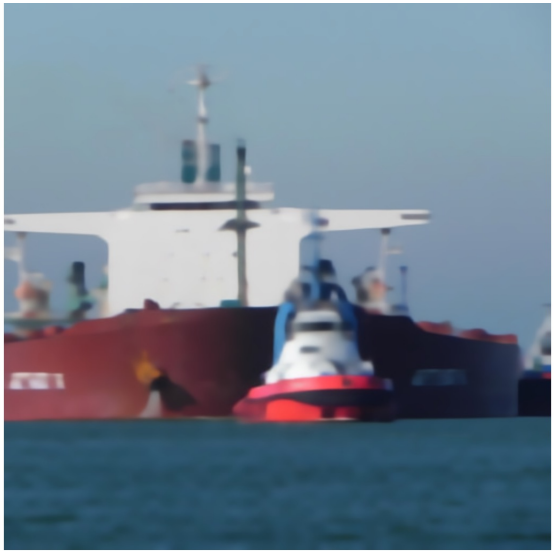
INPUT

SwinIR

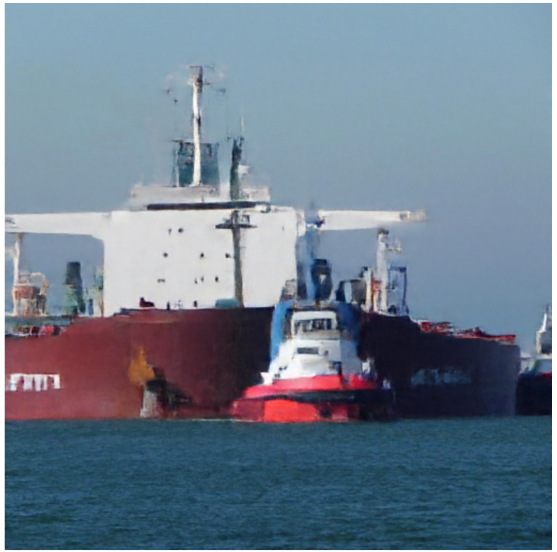
SR3

Results - Visual Analysis

INPUT



SwinIR



SR3

Results - Visual Analysis

INPUT

SwinIR

SR3

Results - Visual Analysis

INPUT

SwinIR

SR3

Results - Visual Analysis

INPUT

SwinIR

SR3

Conclusions and Future Works

- **SwinIR :** generates results that are perceived as worse by the human eye but inference is fast
- **SR3 :** produces higher quality images and more realistic to the human eye but inference is long

TRADE OFF : Image Quality vs Inference Speed

Future Works : reduce inference time of the SR3 model

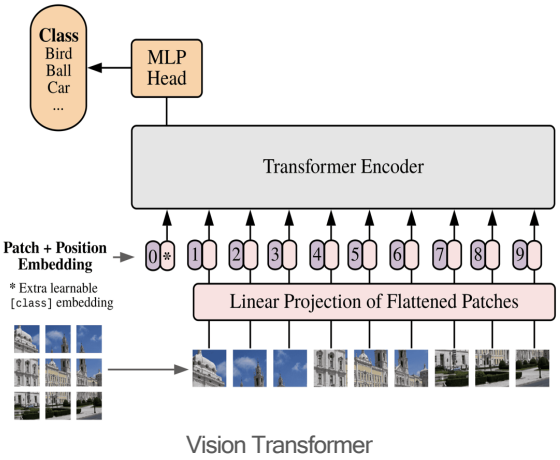
Thank you for the attention!

Transformers for Computer Vision

Transformers are encoder-decoder architecture built to process sequence of tokens. So by default they are impractical for computer vision task due to the high number of tokens needed to represent an image.

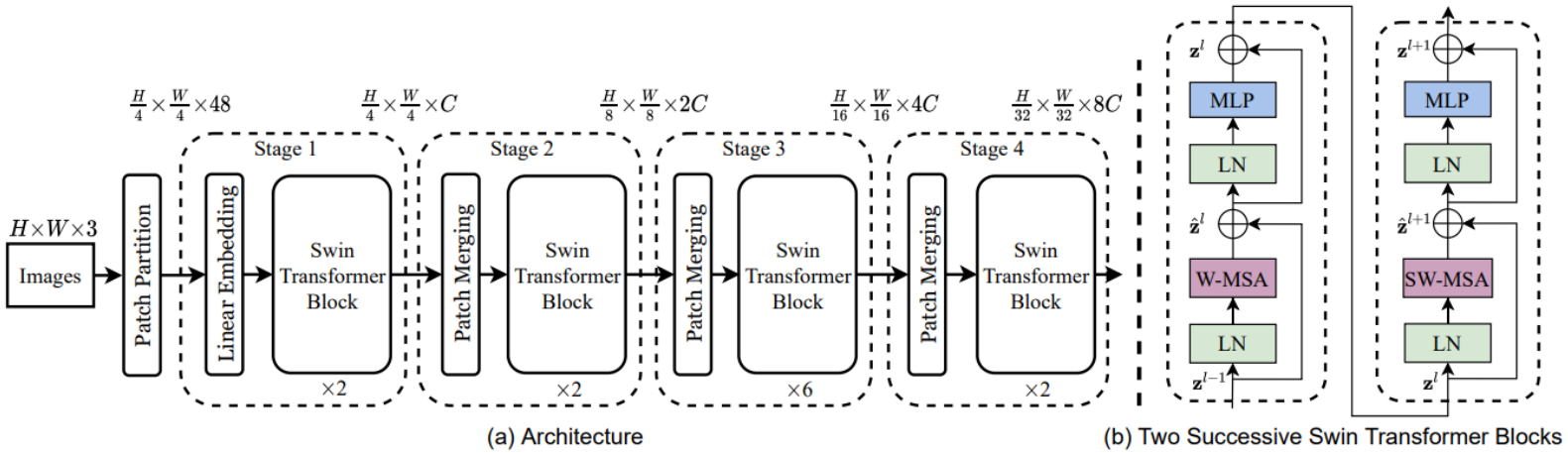
The **Vision Transformer** solved this issue by considering patches of the image as tokens, reducing the length of the sequence representing the image. Still, ViT is limited to process the image on a single scale

So, the **Swin Transformer** introduce a hierarchical shift-based window mechanism with local self-attention to process images at multiple scales.



Vision Transformer

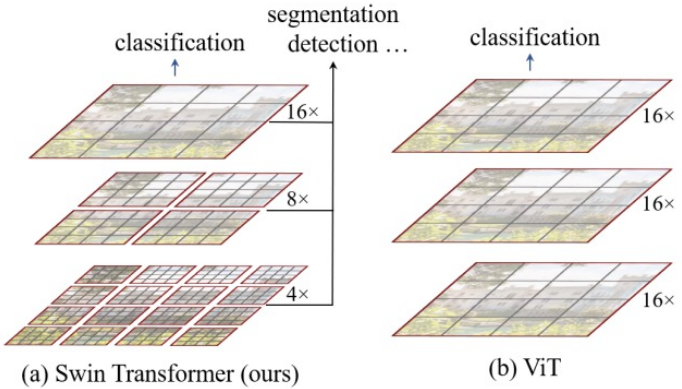
Swin Transformer

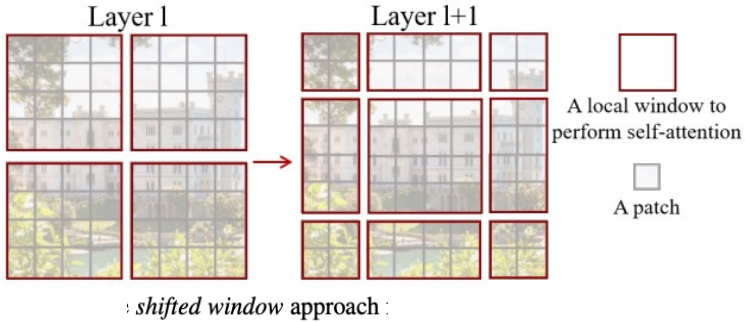


(a) Architecture

(b) Two Successive Swin Transformer Blocks

Swin Transformer





SR3 Generation Process