RTSR: Enhancing Real-time H.264 Video Streaming using Deep Learning based Video Super Resolution

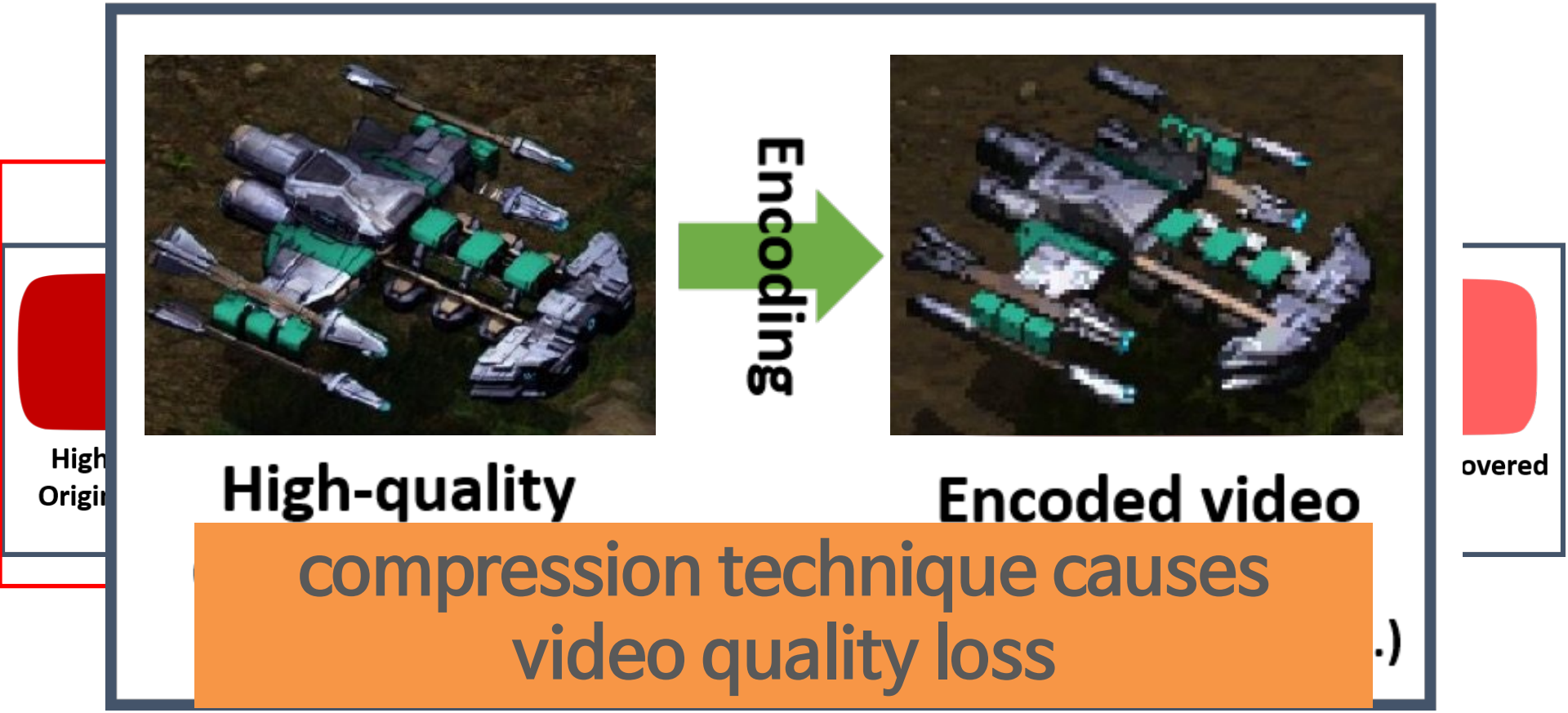
Spring 2017 CS570 Project Presentation
June 8, 2017

Team 16
Soomin Kim
Leslie Tiong
Youngki Kwon
Insu Jang
Wonyoung Lee

Index

Problem

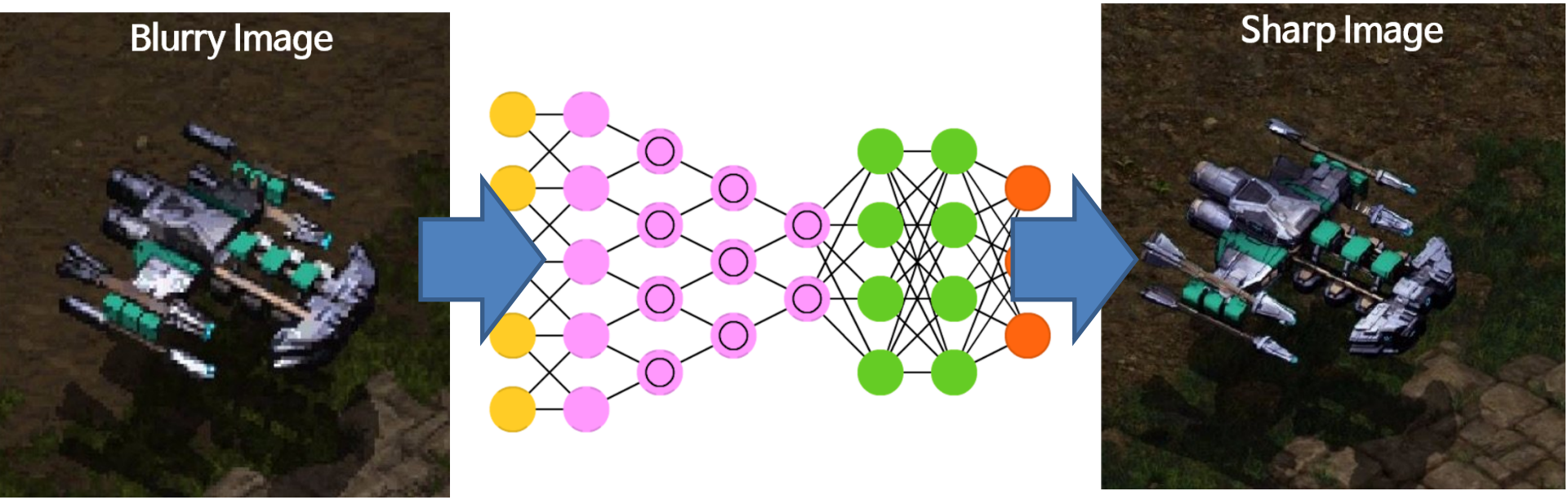
Ir
tr



compression technique causes video quality loss

Super-Resolution (SR)

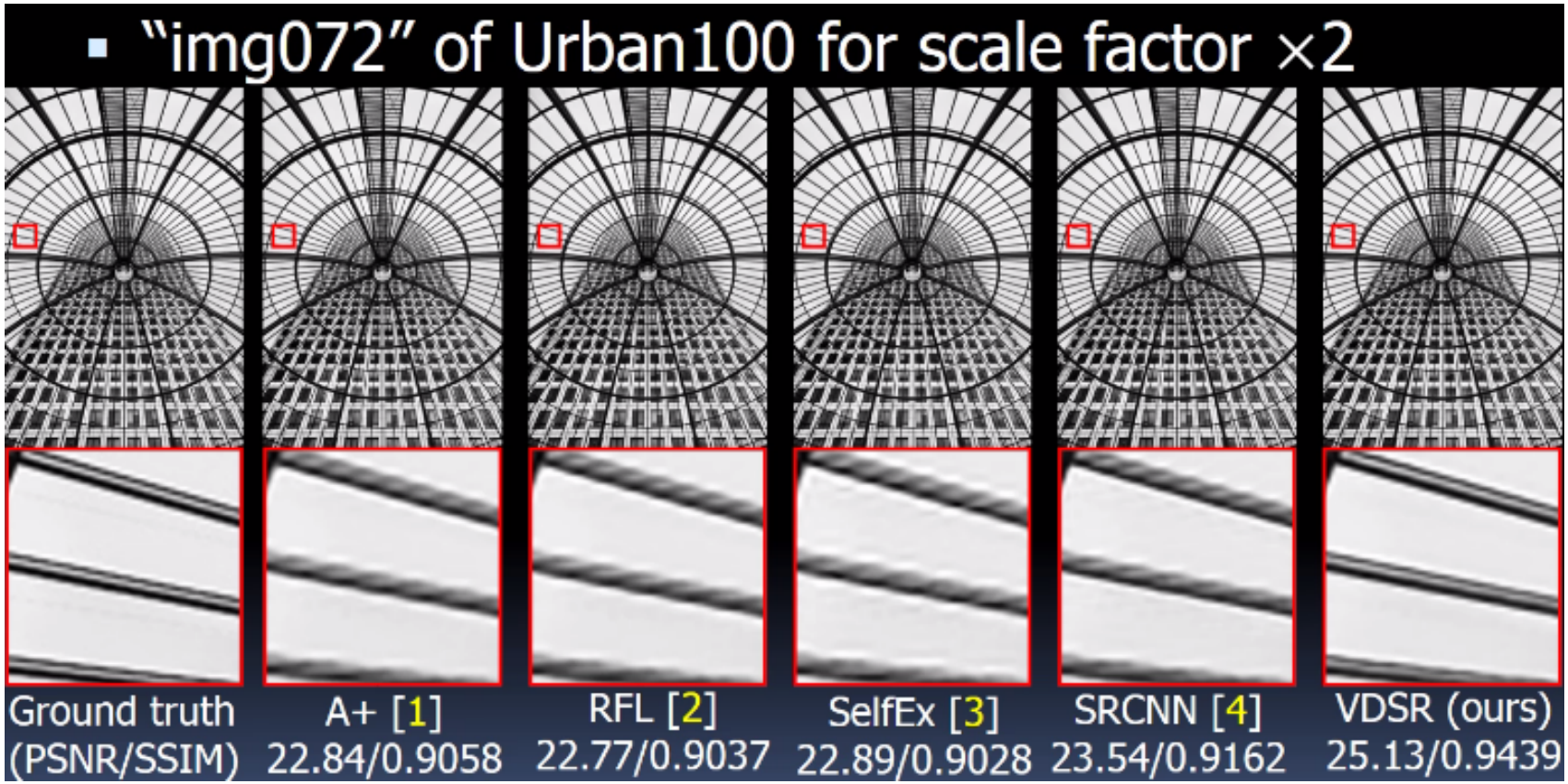
Super-Resolution technique makes image more sharper



Super-Resolution

Single Image Super-Resolution

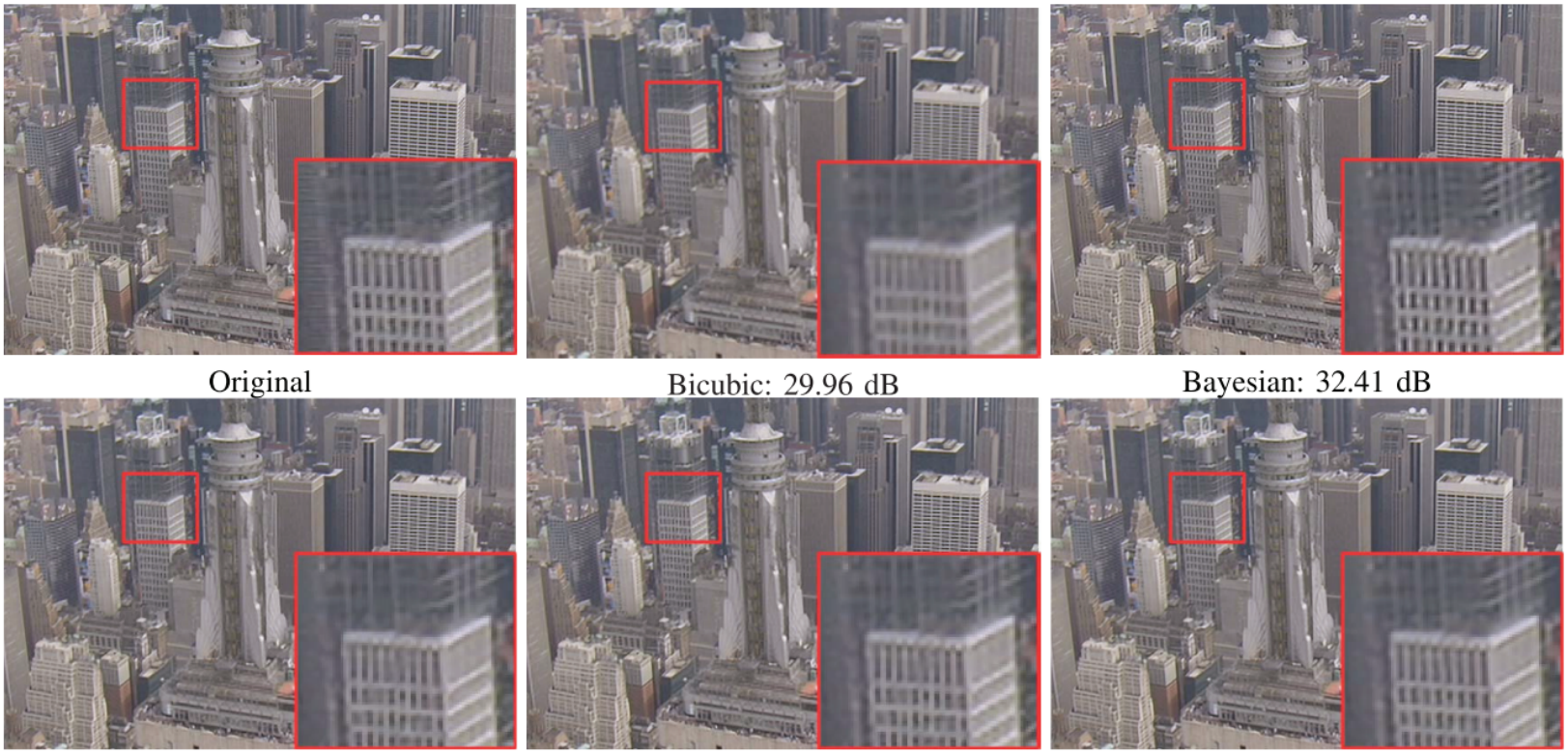
Accurate Image Super-Resolution Using Very Deep Convolutional Networks - Kim et al. (CVPR16)



Fast super-resolution technique but it's target is not video

Video Super-Resolution

Video super-resolution with convolutional neural networks - Kappeler et al. (TCI16)



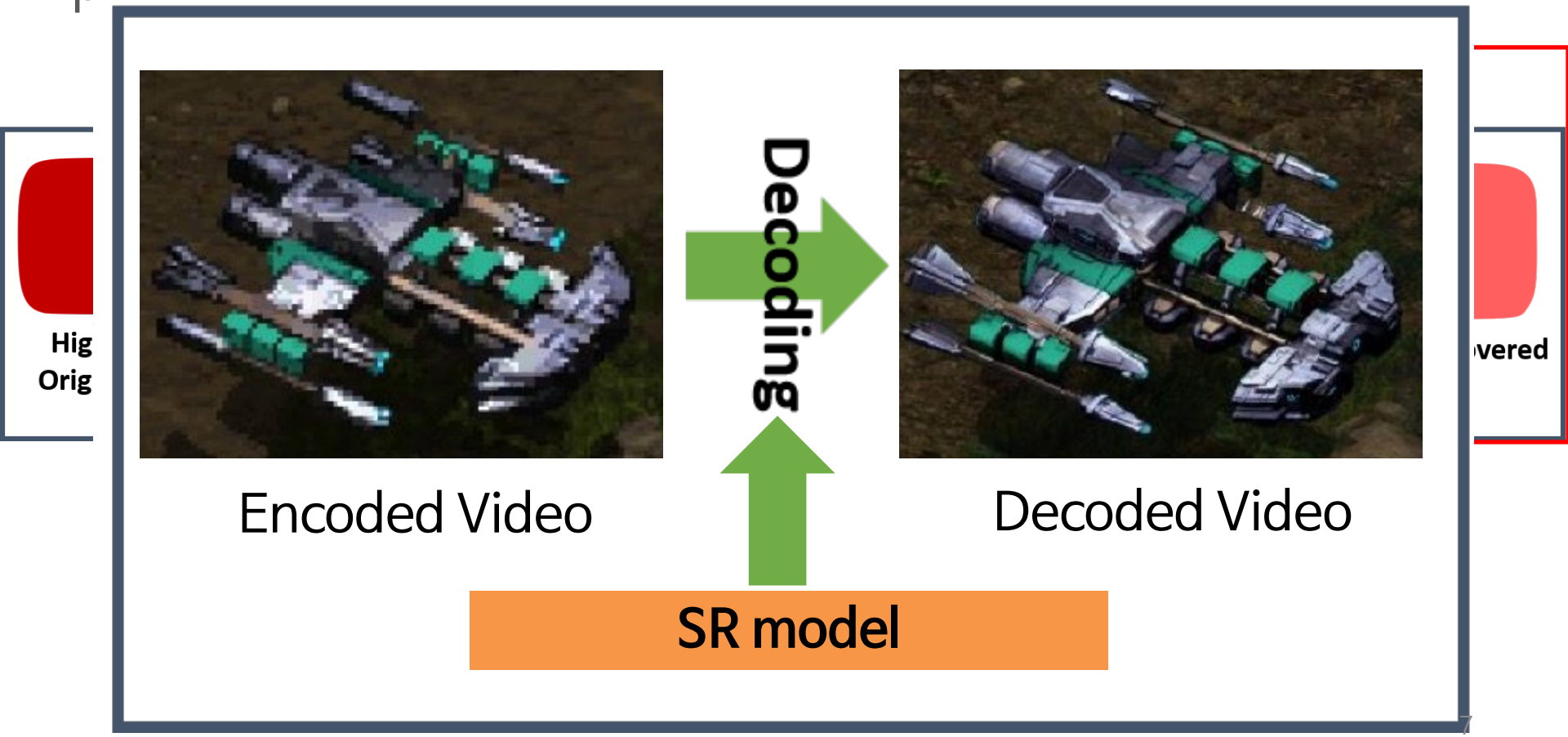
Original | Bicubic: 29.96 dB | Bayesian: 32.41 dB

A+: 32.22 dB | SRCNN: 32.43 dB | VSRnet: **34.15 dB**

The processing time is about 10 seconds per frame.
-> not suitable real-time video processing

Goal

F
c
p


End-user
Hig
Orig
Encoded Video
Decoding
Decoded Video
SR model
vered

Approach

Exploit data structure used in video compression:
Group of Picture (GOP)

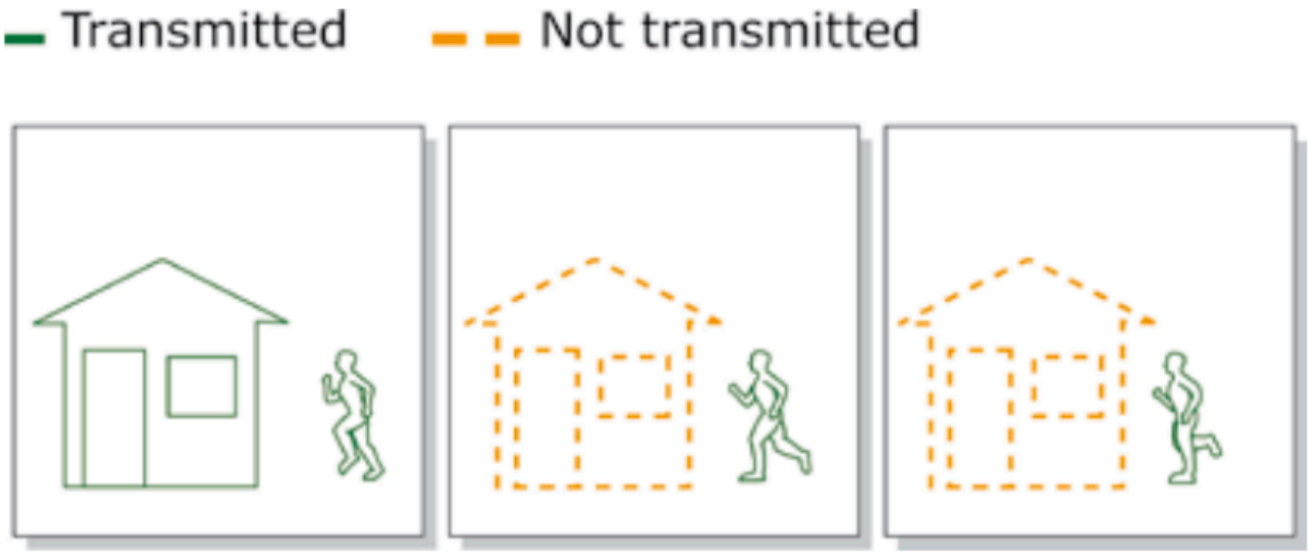
How H.264 codec reduces the size of video?

A video containing 3 frames.
Pixel information for home is duplicated

Approach

Exploit data structure used in video compression:
Group of Picture (GOP)

How H.264 codec reduces the size of video?

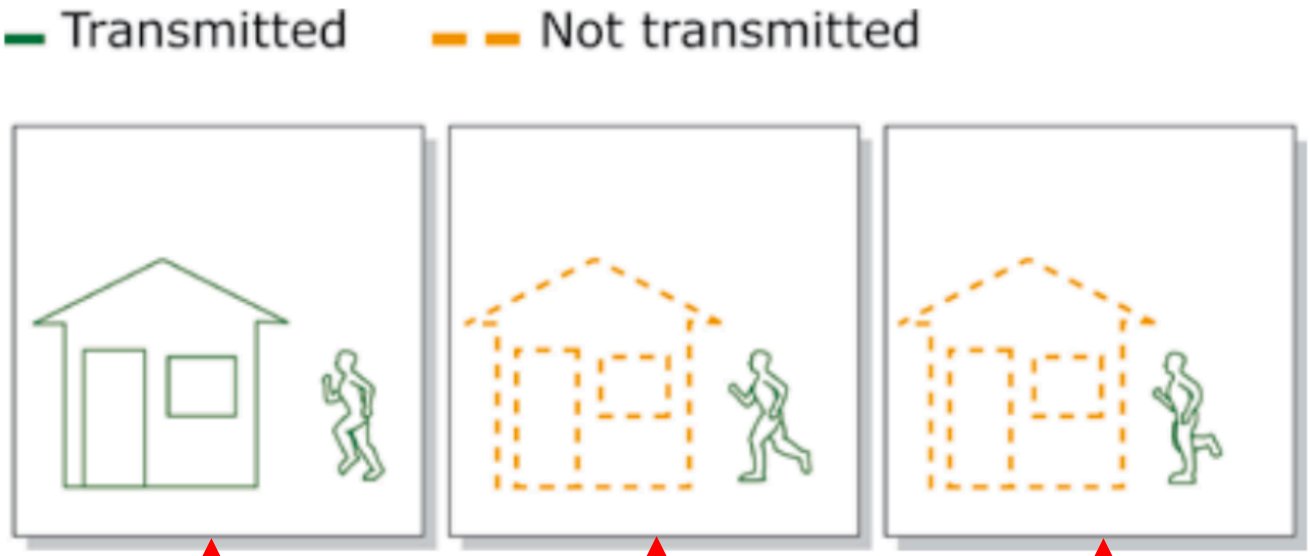


H.264 does not store home in 2nd and 3rd frame

Approach

Exploit data structure used in video compression:
Group of Picture (GOP)

How H.264 codec reduces the size of video?



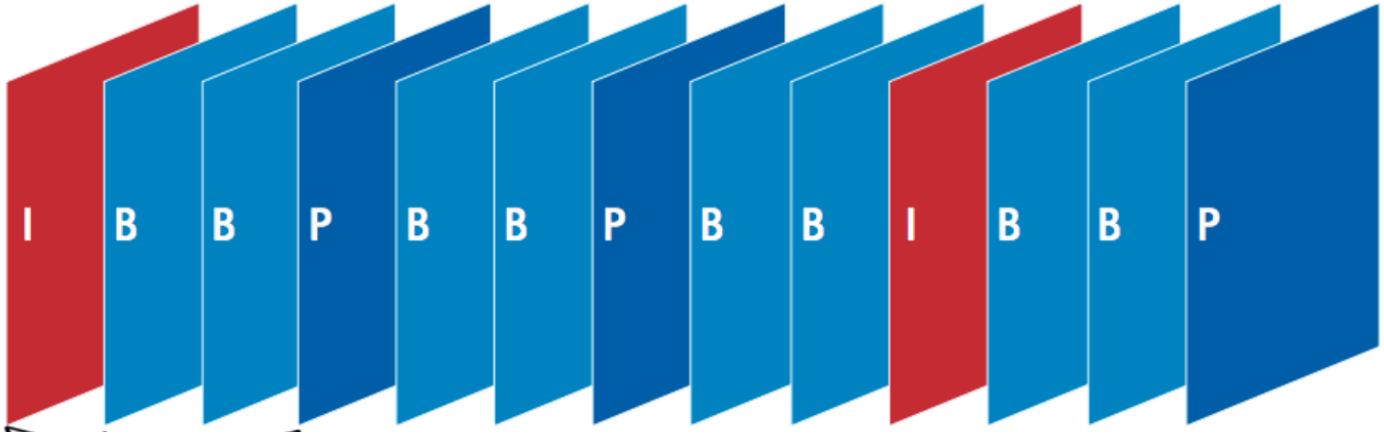
H.264 does not store home in 2nd and 3rd frame

I-frame

P-frame or B-frame

Approach

Exploit data structure used in video compression:
Group of Picture (GOP)



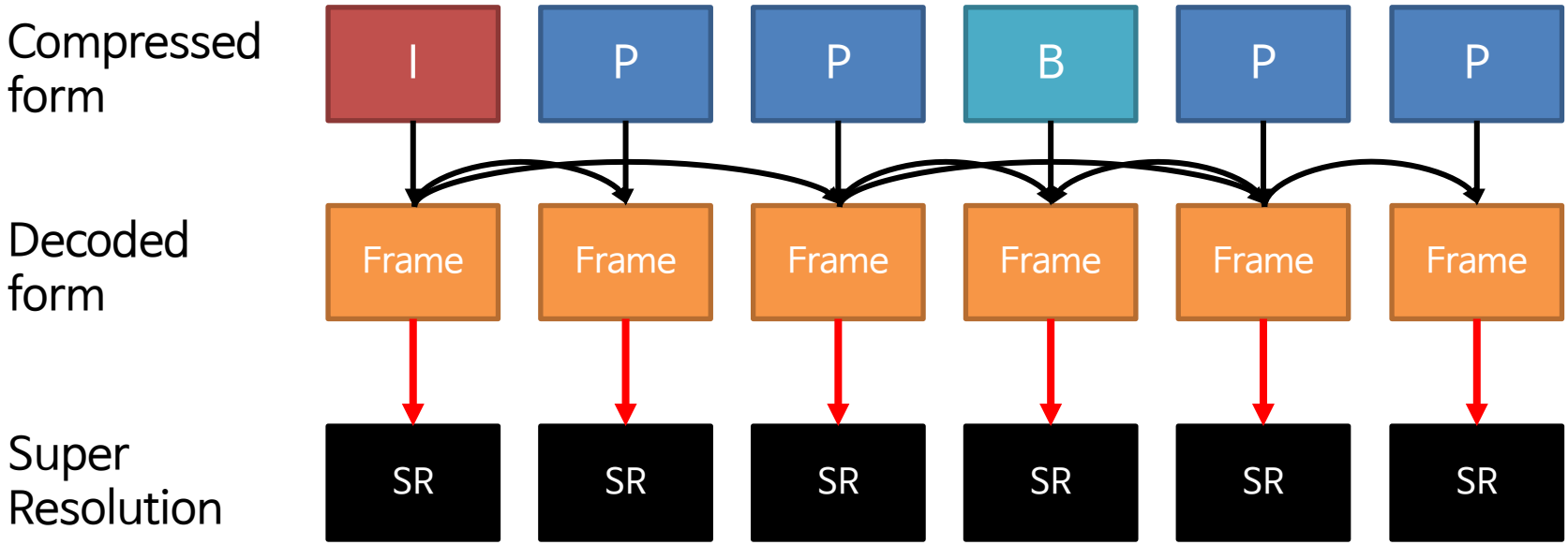
GOP: Group of frames (I-, P-, or B-)
There should be 1≥ I-frame in GOP

4~8 frames in a GOP is ideal for LTE [1]
30~60 frames in a GOP in practical [2, 3]

[1] Zulpratita, Ulil S. "GOP length effect analysis on H. 264/AVC video streaming transmission quality over LTE network." (2013).
[2] Twitch TV, Broadcast Requirements. [Online] https://help.twitch.tv/customer/portal/articles/1253460-broadcast-requirements
[3] Youtube Help, Live encoder settings, bitrates, and resolutions. [Online] https://support.google.com/youtube/answer/2853702?hl=en

Approach

Performing super resolution on video..

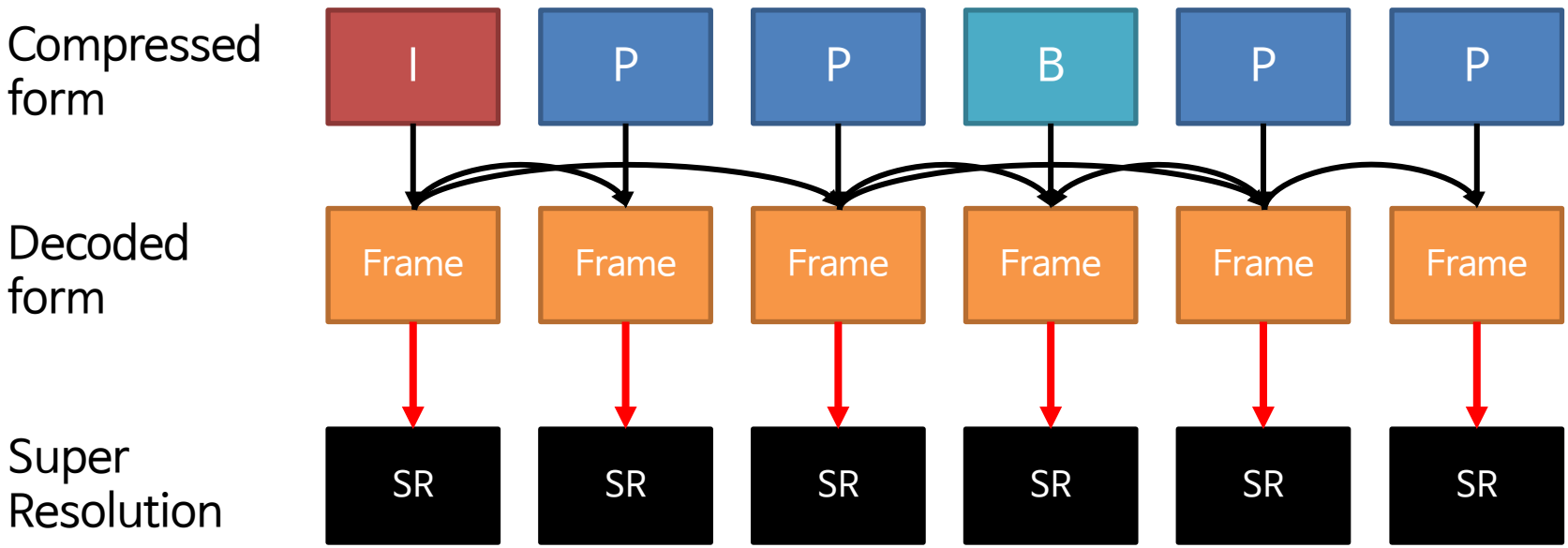


Performing super resolution for every frame

SR cannot be done in a frame interval
High computation overhead

Approach

Performing super resolution on video..



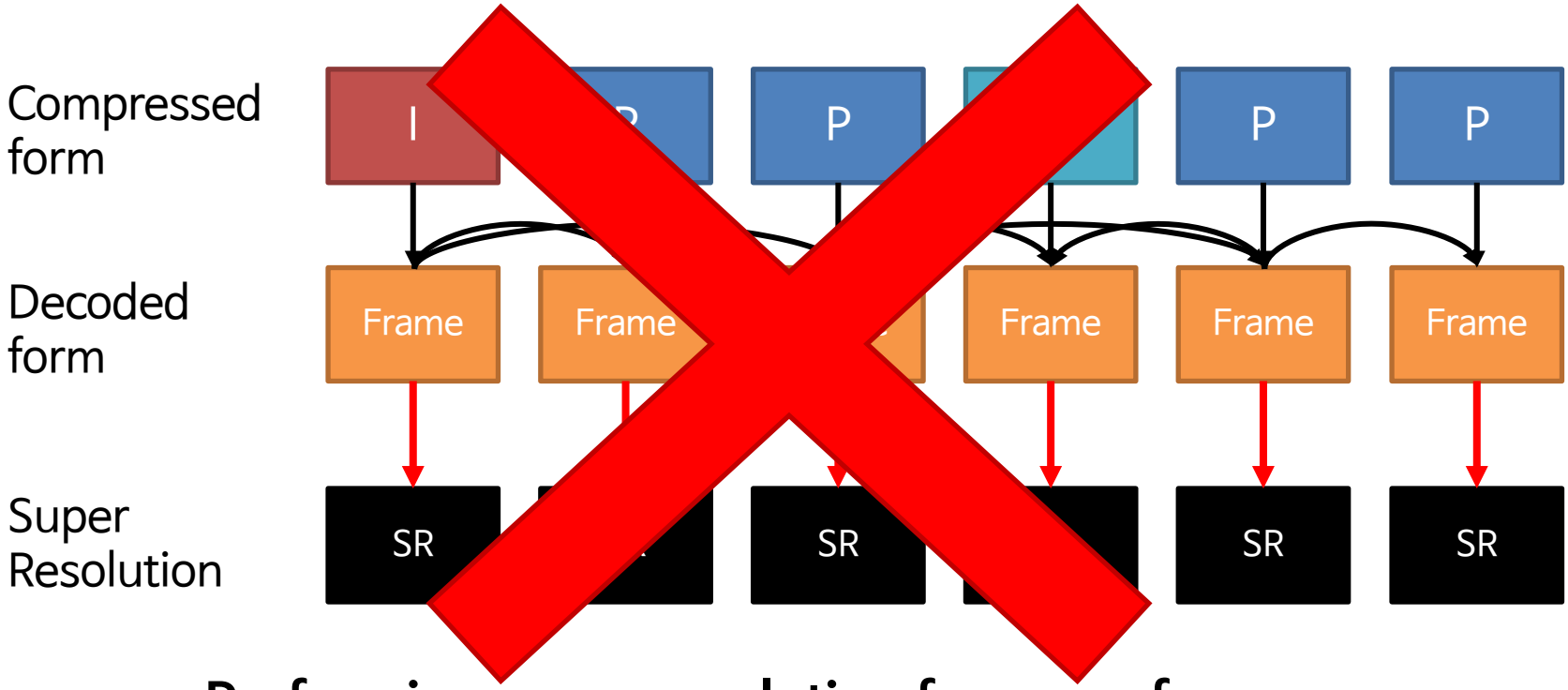
Performing super resolution for every frame

SR cannot be done in a frame interval
High computation overhead

Frame interval	0.03s
VDSR	0.1s
SRCNN	2s
A+	0.9s

Approach

Performing super resolution on video..



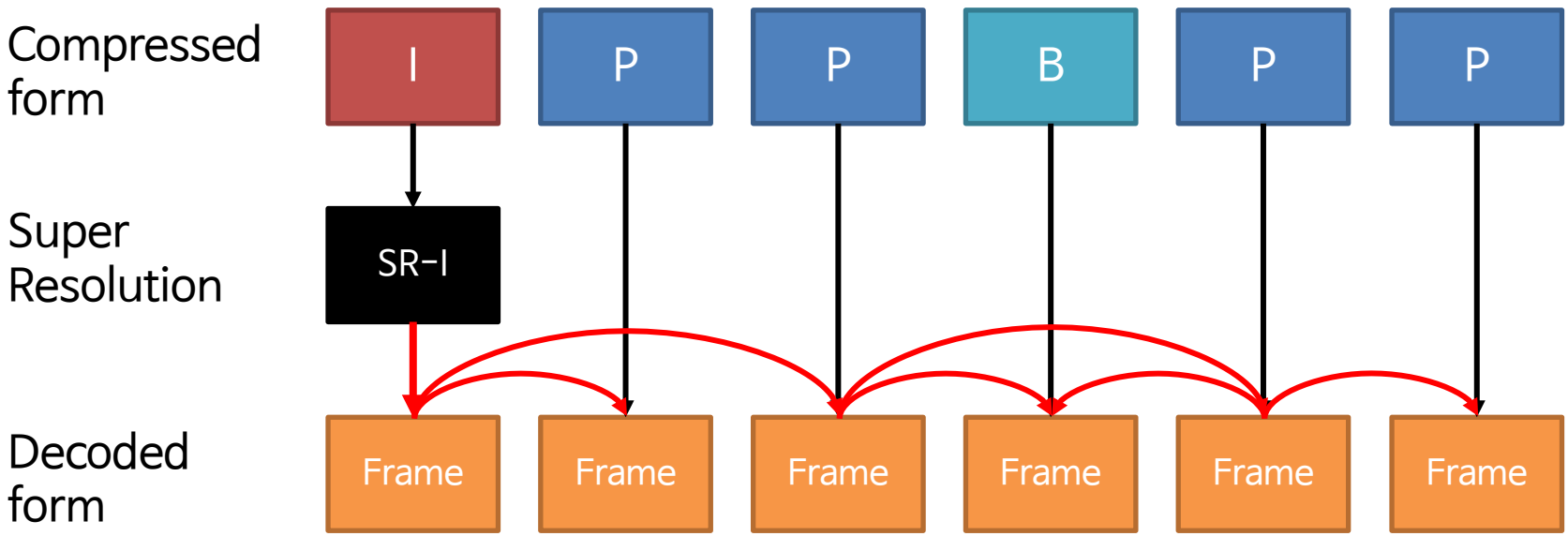
Performing super resolution for every frame

SR cannot be done in a frame interval
High computation overhead

Frame interval	0.03s
VDSR	0.1s
SRCNN	2s
A+	0.9s

Approach

Our approach: perform super resolution on **I-frame only**

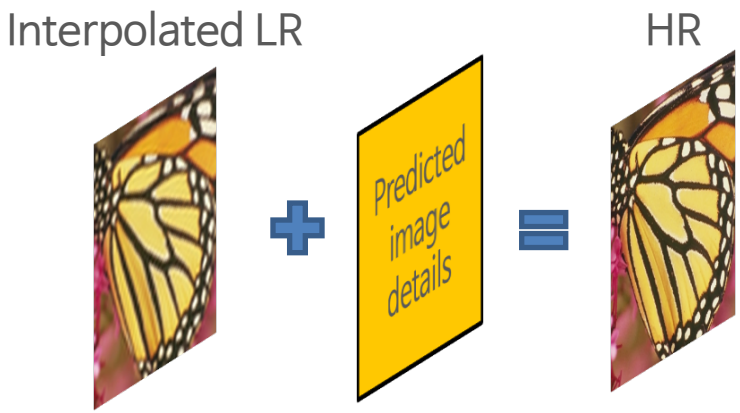


Performing super resolution for I-frame only

SR can be done in GOP interval (0.5s ~ 2s)
All frames are affected by SR I-frame (as they reference it)

Learning based SR - Algorithm (VDSR)

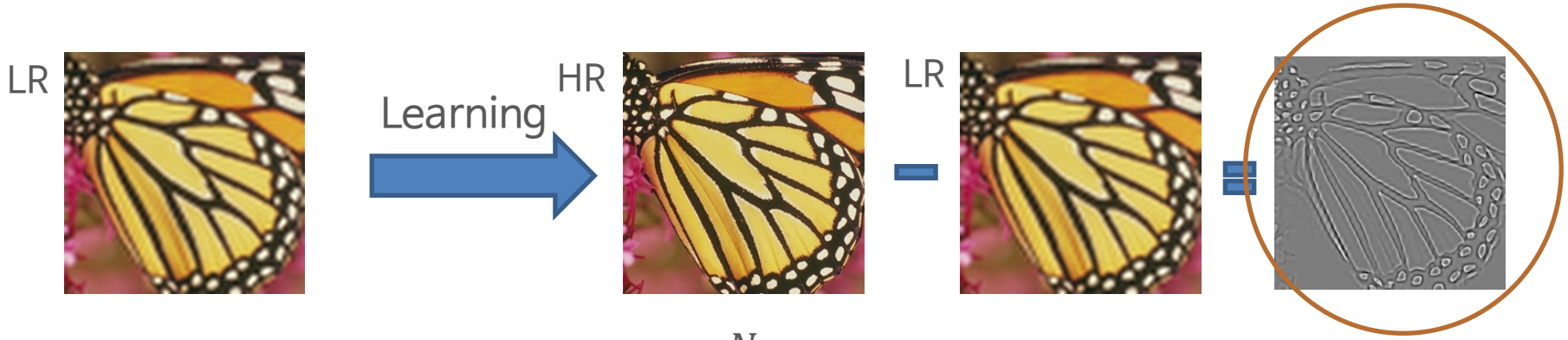
- Learn residual only for fast convergence
- Input : interpolated low resolution image



- Residual image learning

Learning based SR - Algorithm (VDSR)

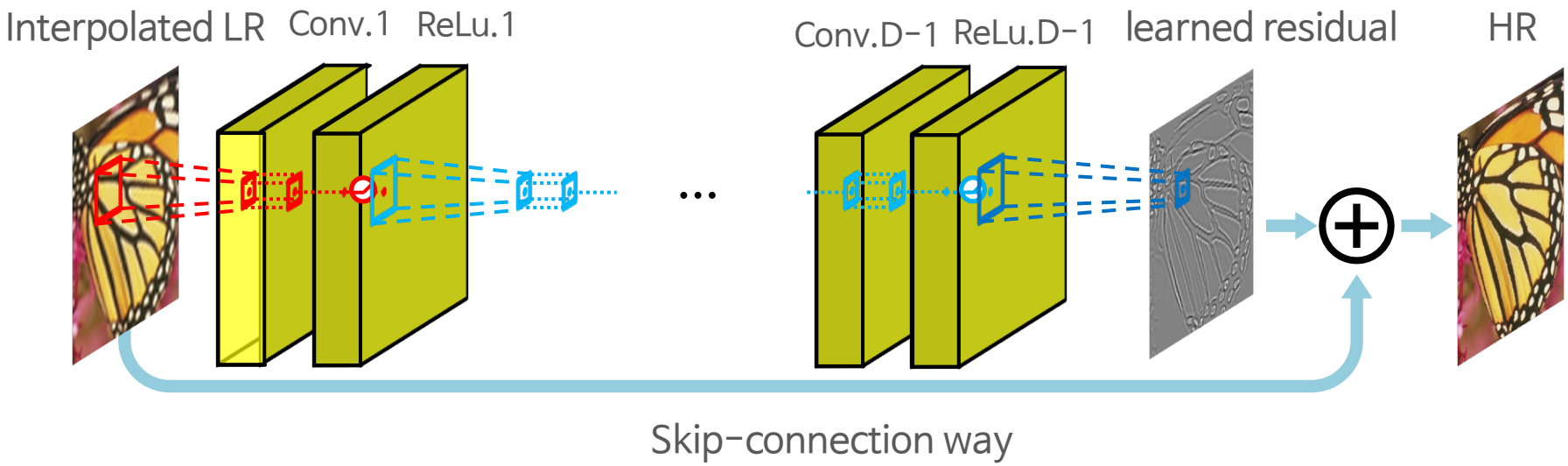
- Learn residual only for fast convergence
- Input : interpolated low resolution image
- Residual image learning



- Training dataset $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$
- Define residual image $\mathbf{r} = \mathbf{y} - \mathbf{x}$
- Goal is to minimize $\frac{1}{2}\|\mathbf{r} - f(\mathbf{x})\|^2$

x: interpolated LR
y: HR
r: residual image
f: network prediction
f(x)+x: final image

Learning based SR - Model (VDSR)



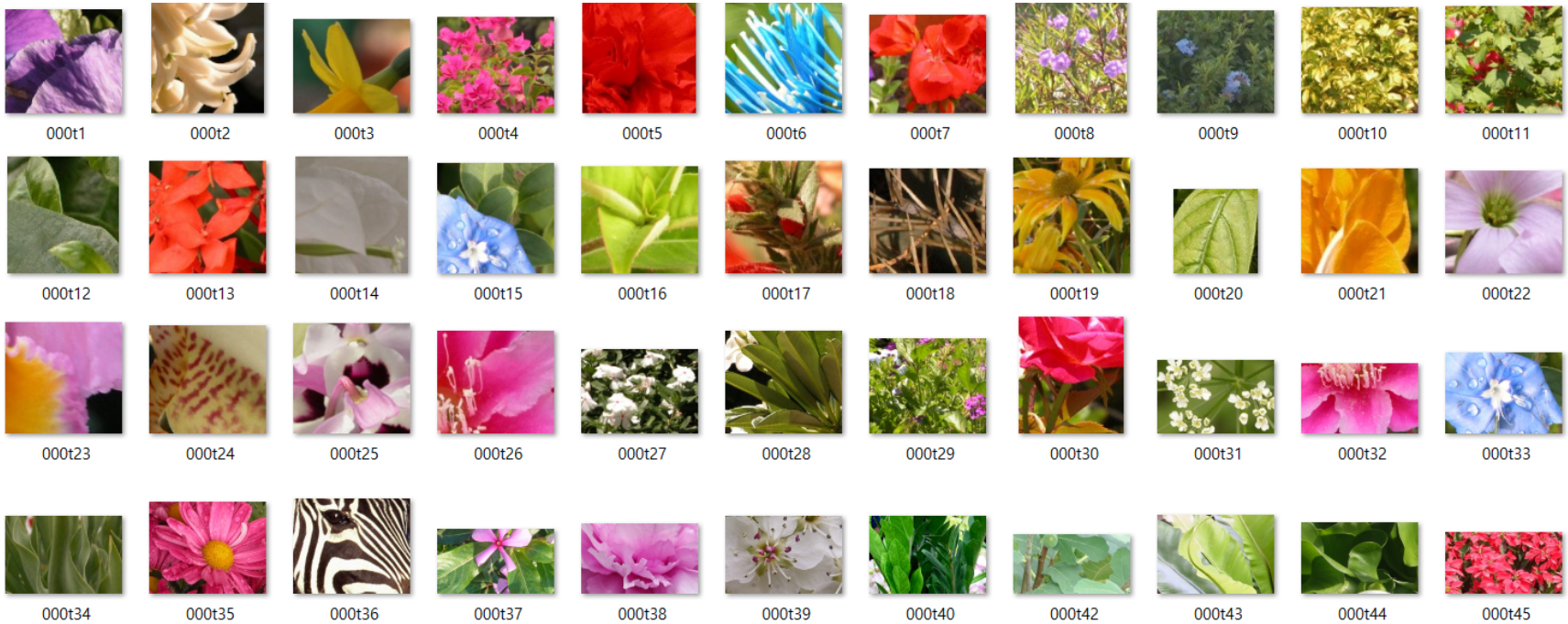
- CNN to SR
  - **20** convolutional layers
  - **Skip connection** to learn residual only
  - **No dimension reduction** such as pooling

Accurate Image Super-Resolution Using Very Deep Convolutional Networks by Kim et al. CVPR 2016

Learning based SR - Training Data (VDSR)

Training data

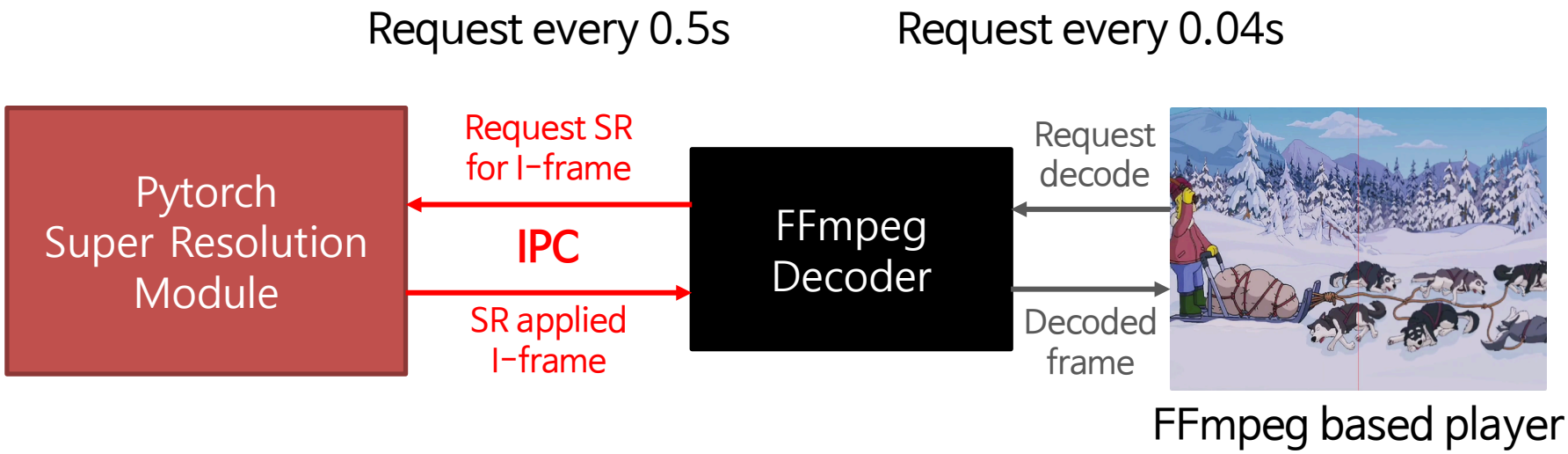
- 291 images with data augmentation (rotation or flip)
  - 91 from Yang et al. [1] + 200 from Berkeley segmentation dataset [2]



[1] Image superresolution via sparse representation by Yang et al. TIP, 2010
[2] A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. by Martin et al., ICCV, 2001

Implementation

Target movie: all videos with 24 fps, GOP length=12
Movie Player: Simple FFmpeg based player [link]
H.264 Decoder: FFmpeg [link]
SR Module: Pytorch VDSR [link] with pre-trained checkpoint
Communication: Message queue btw SR module / FFmpeg



Performance Measurement

1. Peak Signal-to-Noise Ratio (PSNR)

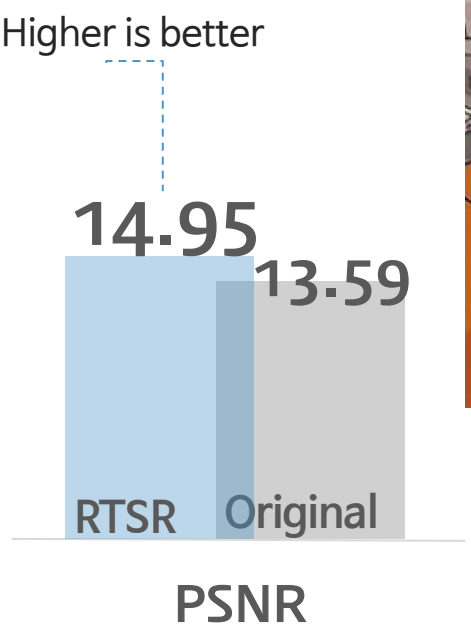
- Most commonly used to measure the quality of reconstruction of lossy compression codecs
- Commonly used metric for super resolution

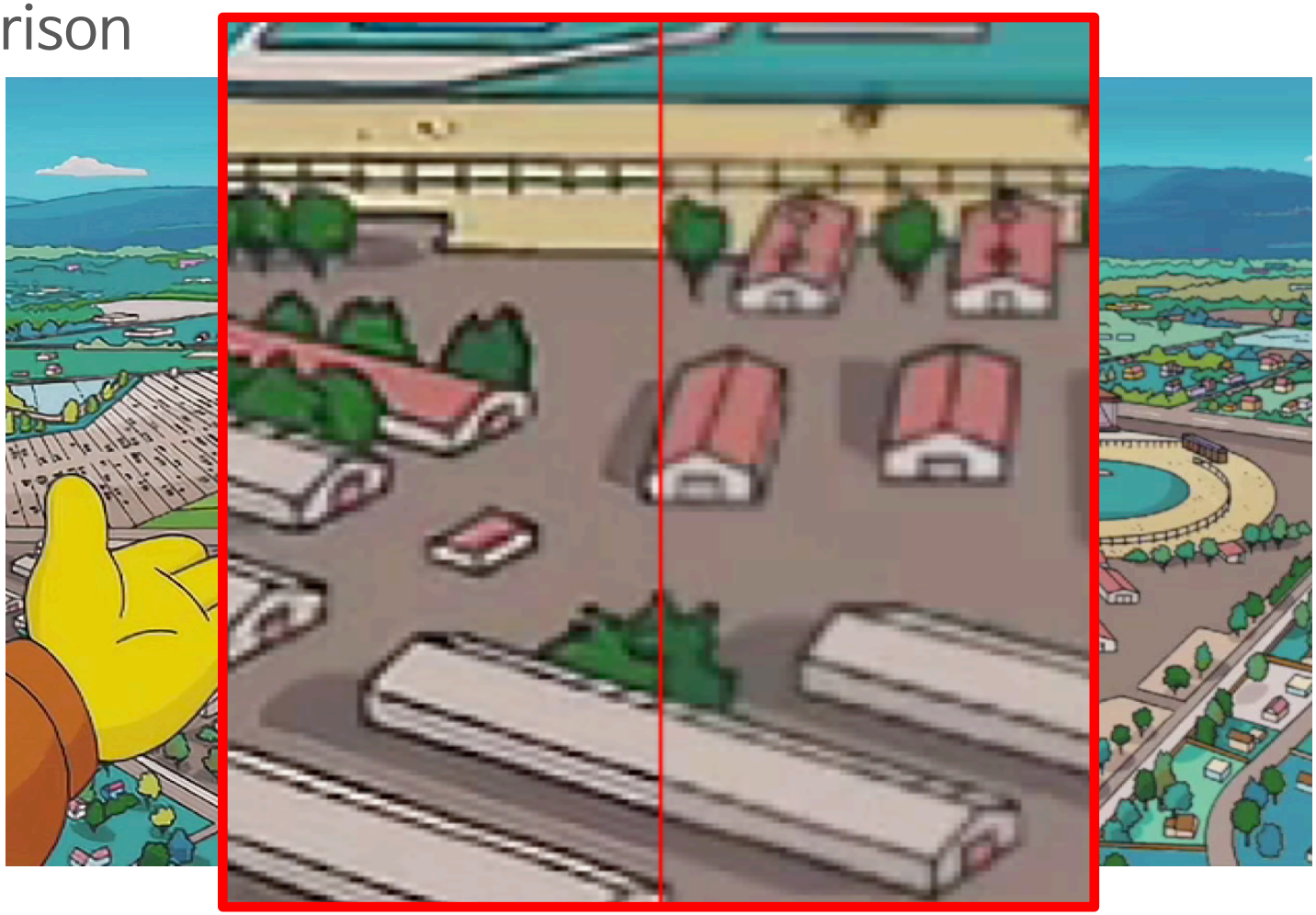
2. Time Delay

- We argued that performing super resolution for all frames cannot be done in real-time.
- Measure how the delay is in (1) SR for all frames, (2) ours

Interpret Results

PSNR Comparison





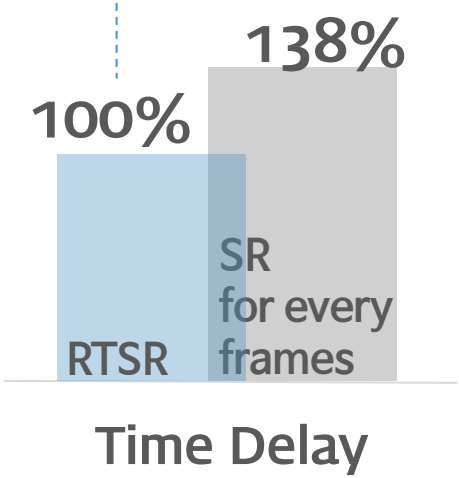
RTSR (Ours)

Original

Interpret Results

Time Delay Comparison

Playtime compared to original video
Closer to 100% is better



RTSR (Ours)

SR for every frames

Discussion

1. Contribution

- Applied SR with Deep Learning in real-time streaming to reconstruct high-resolution videos.
- Our proposed model only applied SR to i-frames, which is outperformed without time-delayed than applied SR to all frames.

2. Limitation

- The high quality of reconstructed video is depending on original video sources.

Demo

Test Video 1

Demo

Test Video 2

RTSR (ours) Original

Thank you.

Q&A