KAIST RCV ROBOTICS AND COMPUTER VISION LAB

2018 ChaLearn Looking at People Challenge - Track 2. Video Decaptioning

DVDNet

Deep Blind Video Decaptioning with 3D-2D Gated Convolutions

Dahun Kim*, Sanghyun Woo*, Joonyoung Lee, In So Kweon

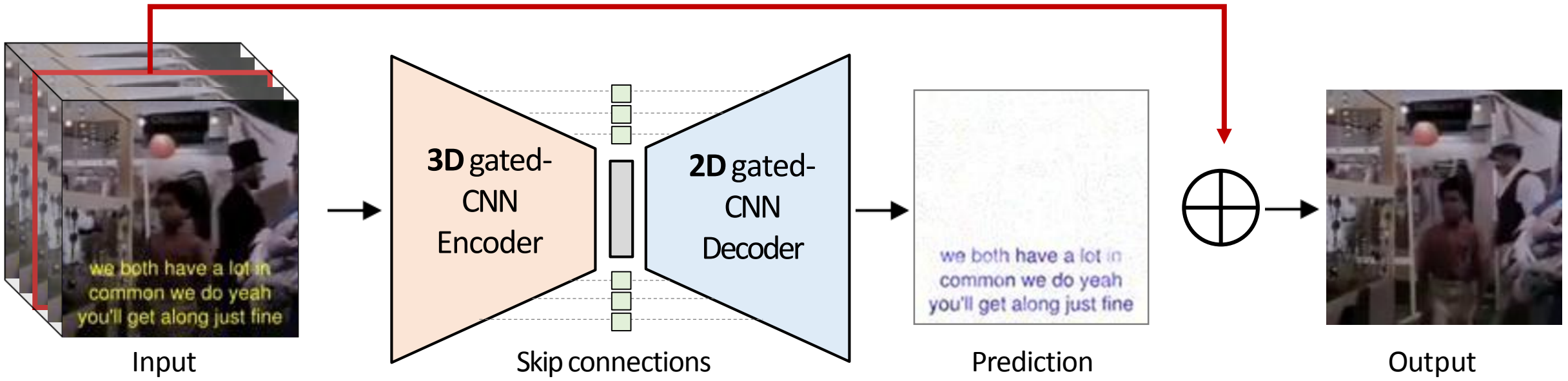
Our Problem

Remove text overlays in video

Need to consider two important points:

1. Video : Sequence of frames)
2. Blind : No inpainting mask)

Model Overview



Two important points :

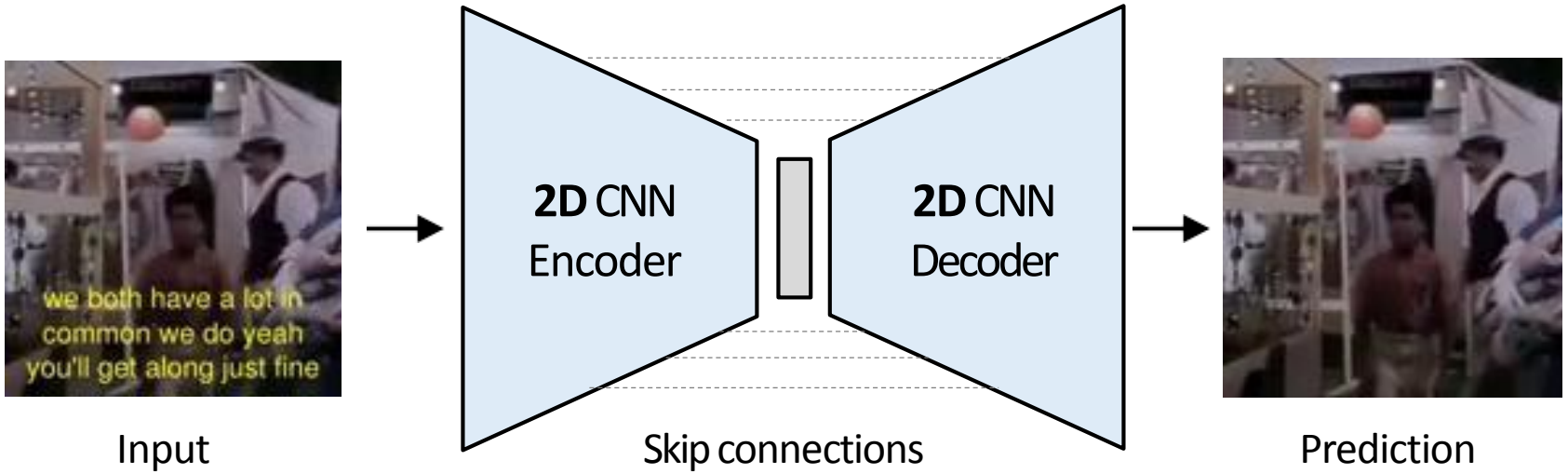
- Video : Sequence of frames
- Blind : No inpainting mask

- **3D-2D U-net**
- **Residual learning**
- **+ Gated convolution**

Vanilla 2D U-Net*

Frame-by-frame operation

- Spatial context



Two important points :

- Video : Sequence of frames
- Blind : No inpainting mask

- **Scene dynamics**

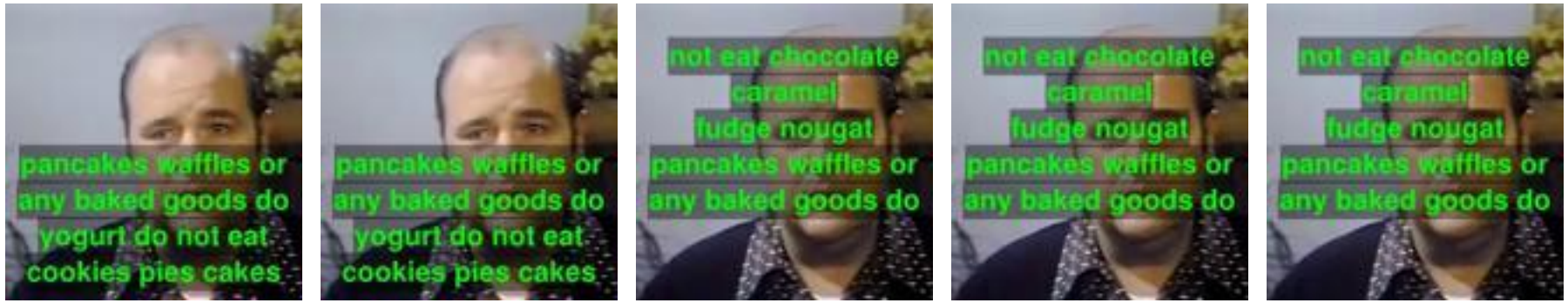
* Ronneberger, O.et al. “U-net: Convolutional networks for biomedical image segmentation.” MICCAI 2015.

Input : Multiple frames

Scene dynamics

- Aggregate hints from **spatio-temporal** neighborhoods

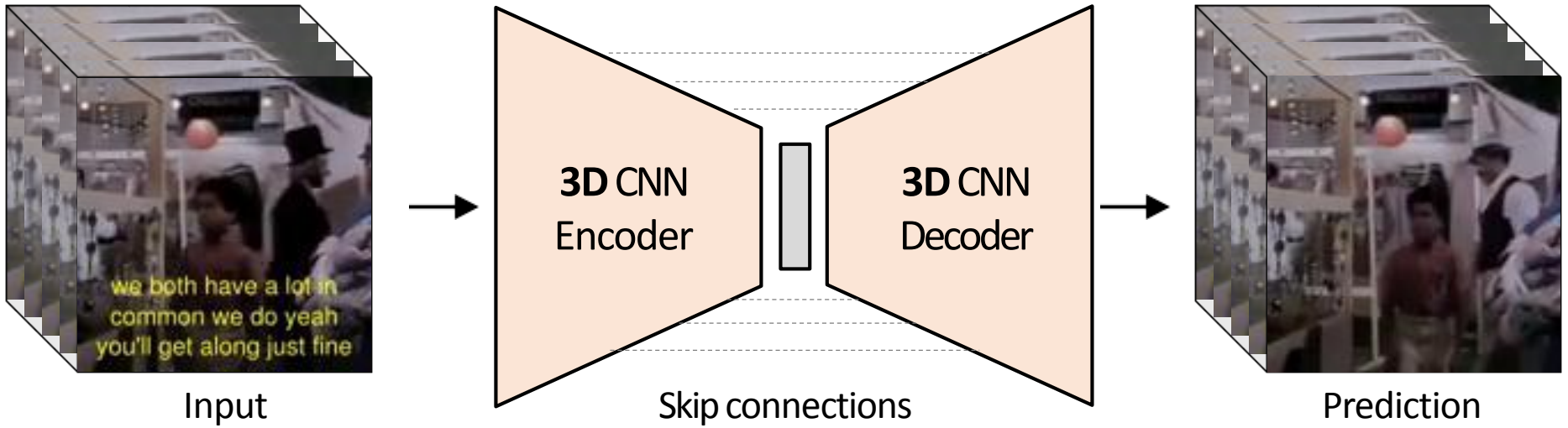
→ *Object movements*



→ *Subtitle changes*

Vanilla 3D U-Net*

Multiple frame prediction



- Hard problem
- Heavy
- Not uniform prediction

* C¸ ic¸ek, O¨ .et al. “3d u-net: learning dense volumetric segmentation from sparse annotation.” MICCAI 2016.

Output : Single frame

Focus on a single frame

- Aggregate hints from **lagging and leading frames.**

Lagging frames *Leading frames*

3D-2D U-Net

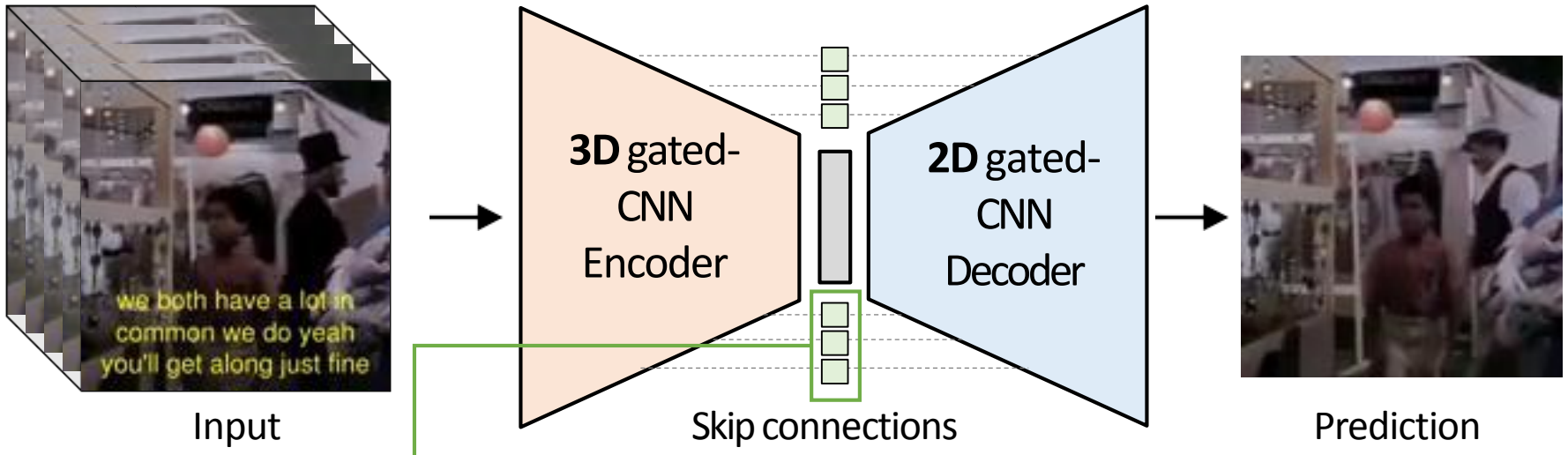
- Easy problem
- Light-weight
- Temporal view range

Output

Center frame

3D-2D U-Net architecture

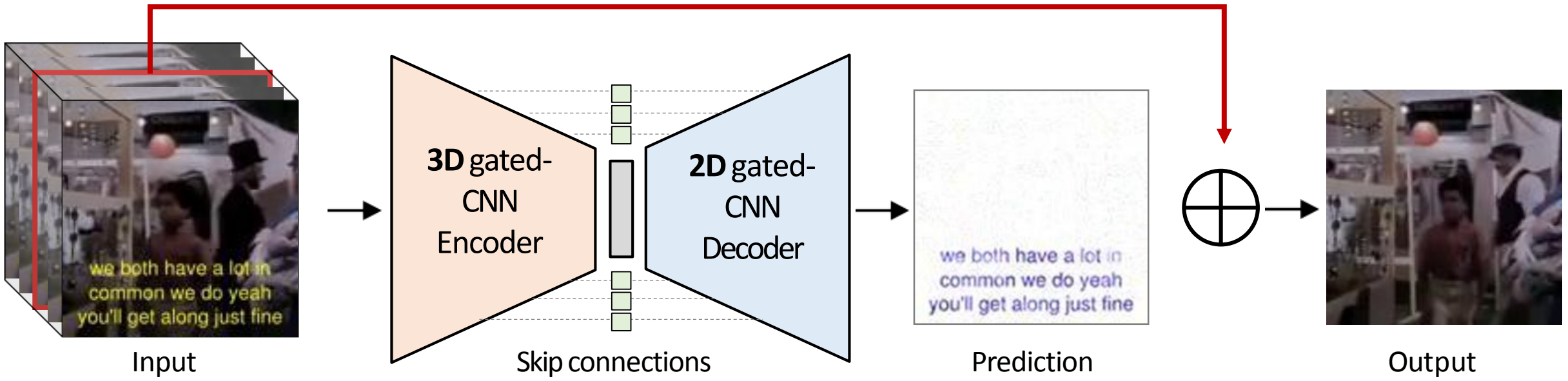
Focus on a single frame



- 3D convolutions to ***flatten*** the encoder features ***into one frame***.

→ <u>to match the shape and concatenate.</u>

Residual Learning



→ Implicitly knows the inpainting mask

Two important points :

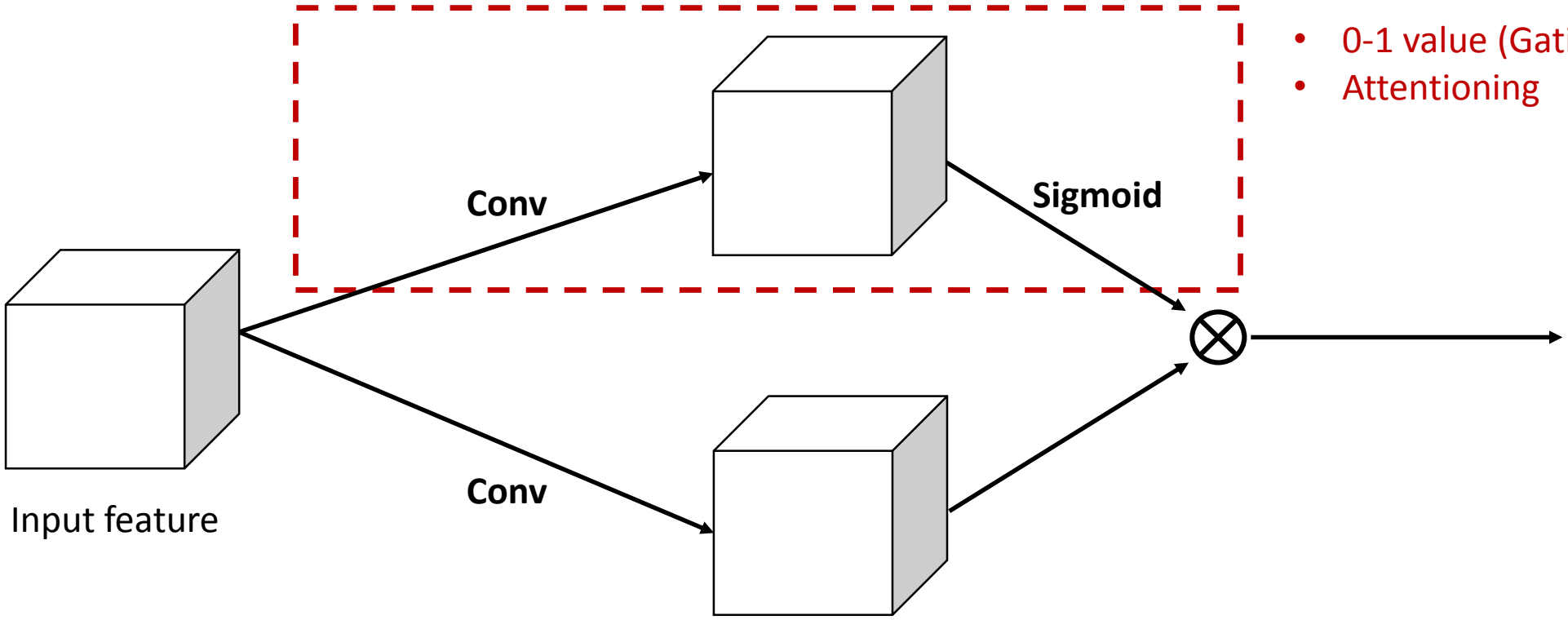
- Video : Sequence of frames
- Blind : No inpainting mask

- **Residual learning**
  - *Not touching good pixels*
  - *Focus on the corrupted regions*

+ Attention

Gated Convolution*

$$Gate = \sigma(W_g \otimes Input)$$
$$Feature = \phi(W_f \otimes Input)$$
$$Out = Feature \odot Gate$$

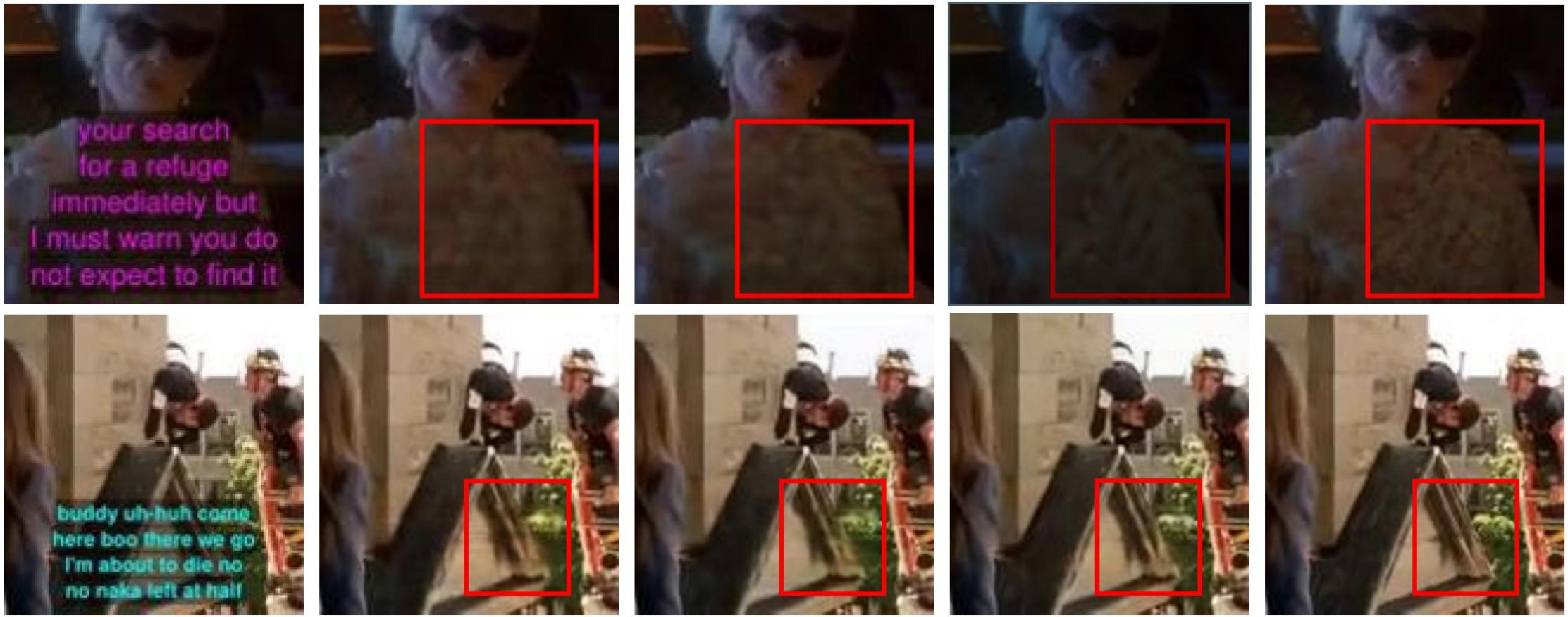


- 0-1 value (Gating)
- Attentioning

* Yu, J. et al. “Free-form image inpainting with gated convolution”. arXiv preprint arXiv:1806.03589.

Loss Function

L1 + gradient L1 + SSIM loss



(a) input (b) L1 (c) L1 + grad.L1 (d) L1+grad.L1+SSIM (e) target(GT)

Quantative Results

User	MSE	PSNR	DSSIM
KAIST-RCV	0.0011	33.3527	0.0404
ucs	0.0011	33.0052	0.041
hcilab	0.0012	33.0228	0.0424
anubhap93	0.0012	32.0021	0.0499
arnavkj95	0.0012	32.1713	0.0482
Baseline	0.0022	30.1856	0.0613

Qualitative Results

(a)

(b)

(c)

(d)

KAIST RCV ROBOTICS AND COMPUTER VISION LAB

2018 ChaLearn Looking at People Challenge - Track 2. Video Decaptioning

DVDNet

Deep Blind Video Decaptioning with 3D-2D Gated Convolutions

Dahun Kim*, Sanghyun Woo*, Joonyoung Lee, In So Kweon