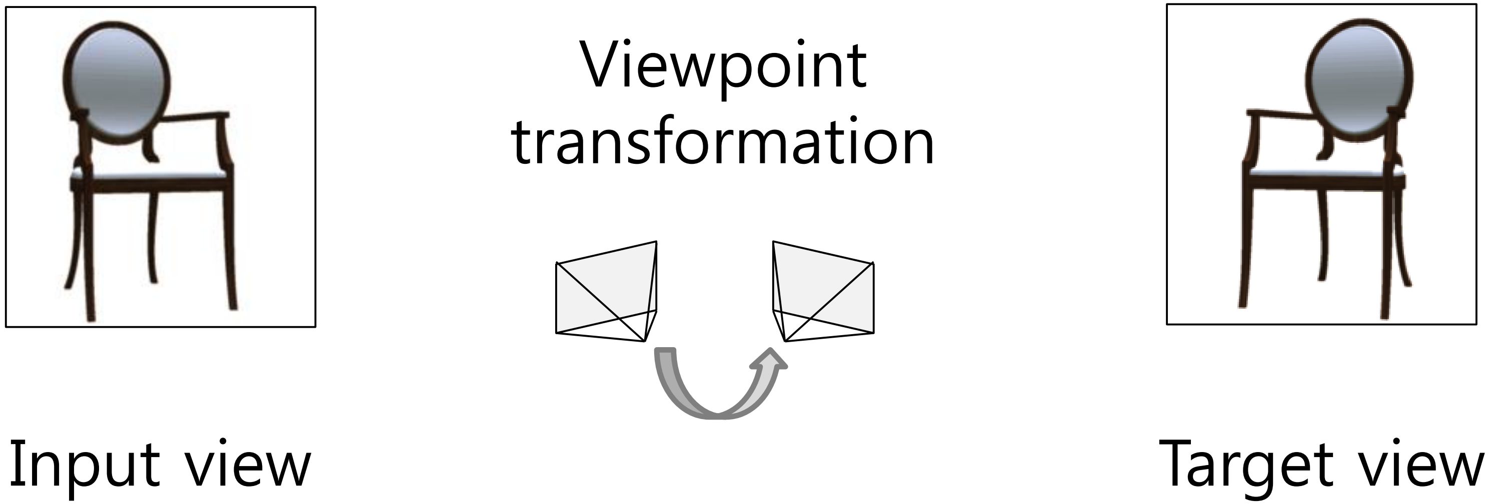
NOVEL VIEW SYNTHESIS WITH SKIP CONNECTIONS

Juhyeon Kim and Young Min Kim
Seoul National University 3D Vision Lab

INTRODUCTION

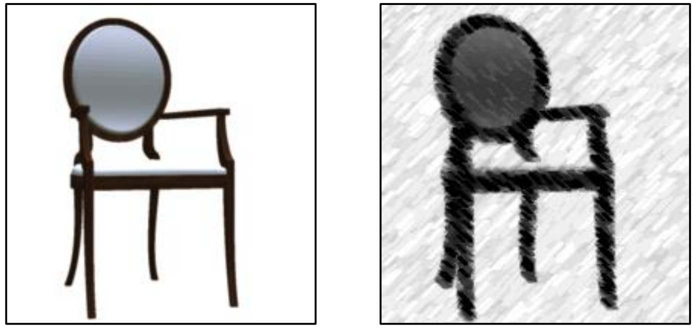
Novel View Synthesis



Considering the camera pose as a domain, novel view synthesis can be thought as an **image-to-image** translation task

Novel View Synthesis

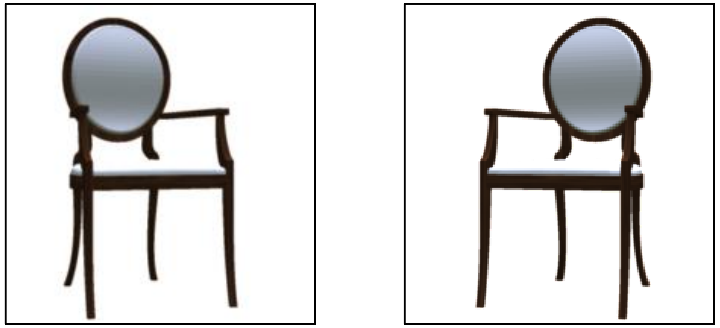
- One of the most successful in image-to-image translation is the **skip connections**, also known as **U-Net**.



Traditional case

- style/texture transfer

- no global structure change



Novel view synthesis

- dramatic geometrical transformation

→ skip connections cannot be simply applied

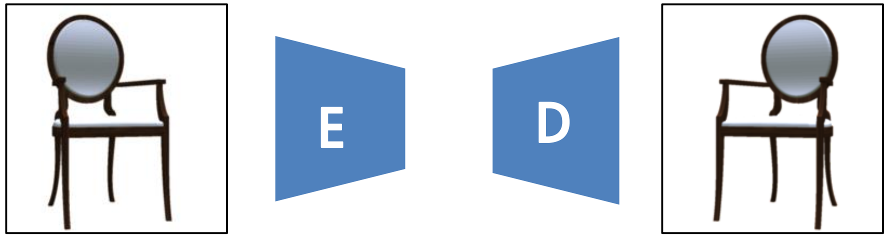
(Park, 2017)

Previous Works

- Encoder-decoder structure without any skip connections.
- The works can be divided into two groups:

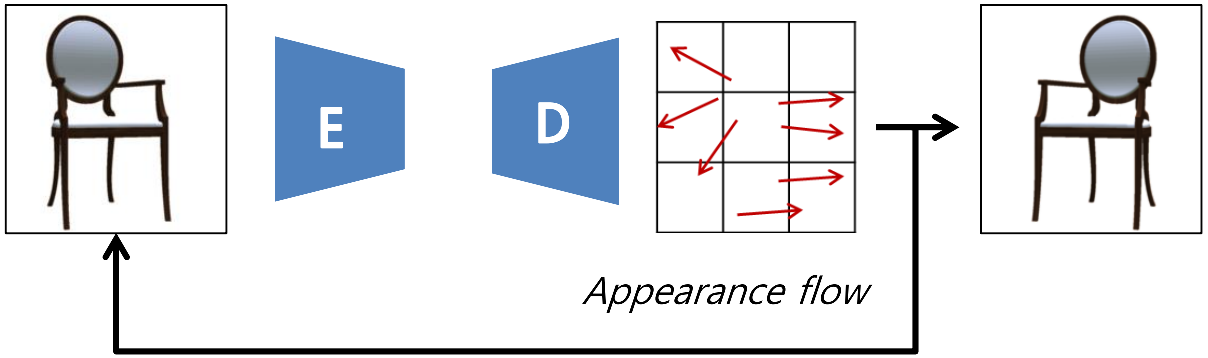
Pixel Generation

(Tatarchenko, 2015)



Flow Prediction

(Zhou, 2016)



- Both modules used in sequential or parallel way.

(Park, 2017) (Sun, 2018)

Previous Works

- These works imply standard **skip connections** are not directly applicable to novel view synthesis due to significant shape change. However, none of them have thoroughly investigated the possibilities.

Our Contribution

- We study the effect of various types of **skip connections** on pixel generation and flow prediction module and find the appropriate way to apply residual connections.

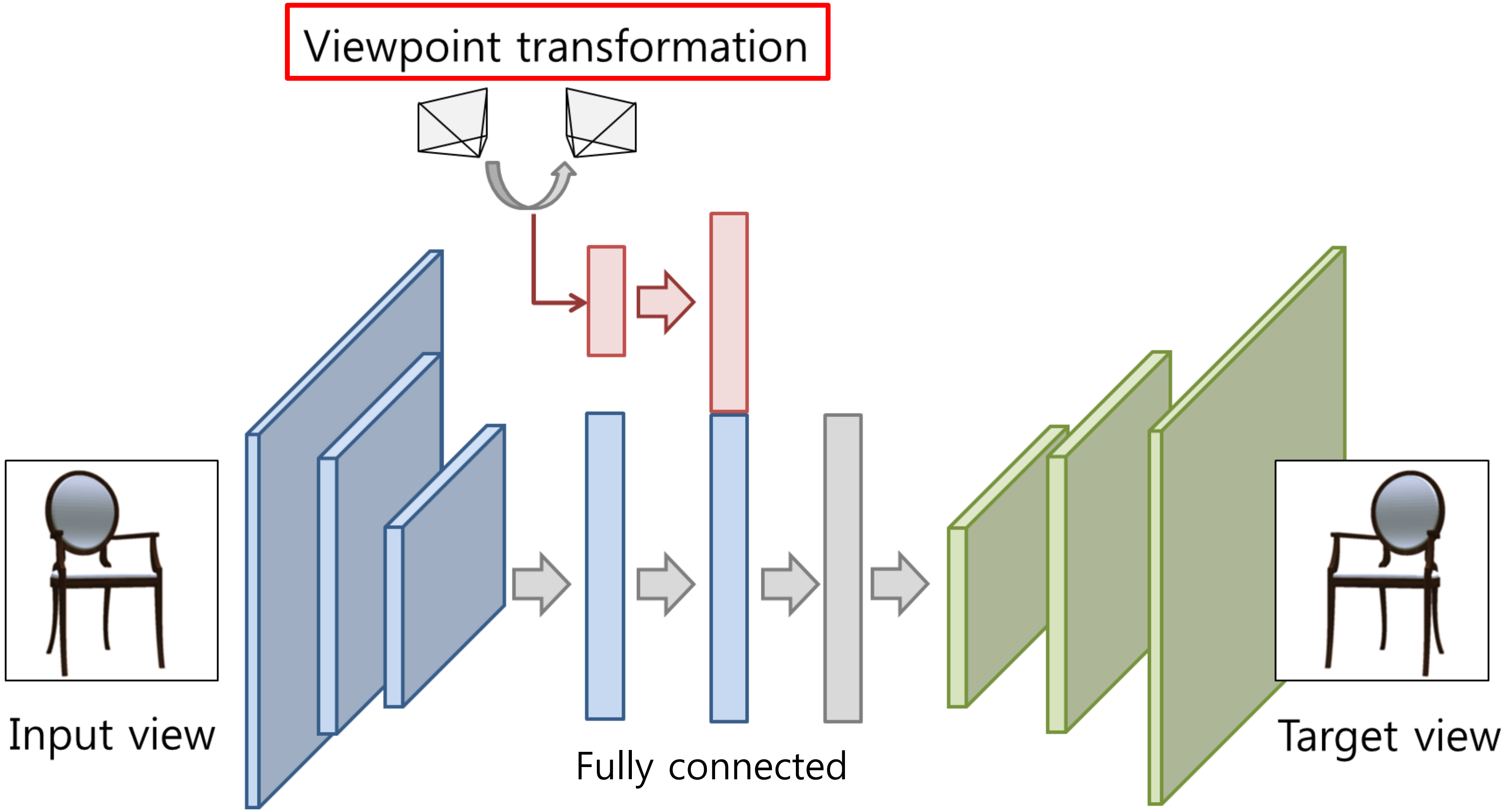
- For pixel generation, we find that using ***flow-based hard attention*** mechanism to skip connection is effective. Flow prediction enjoys marginal benefit from skip connections in deeper layers.

PROPOSED METHOD

Encoder-Decoder Architecture

Encoder layers: $\{x_e^1, x_e^2, \cdots x_e^L\}$

Decoder layers: $\{x_d^1, x_d^2, \cdots x_d^L\}$



U-Net Architecture

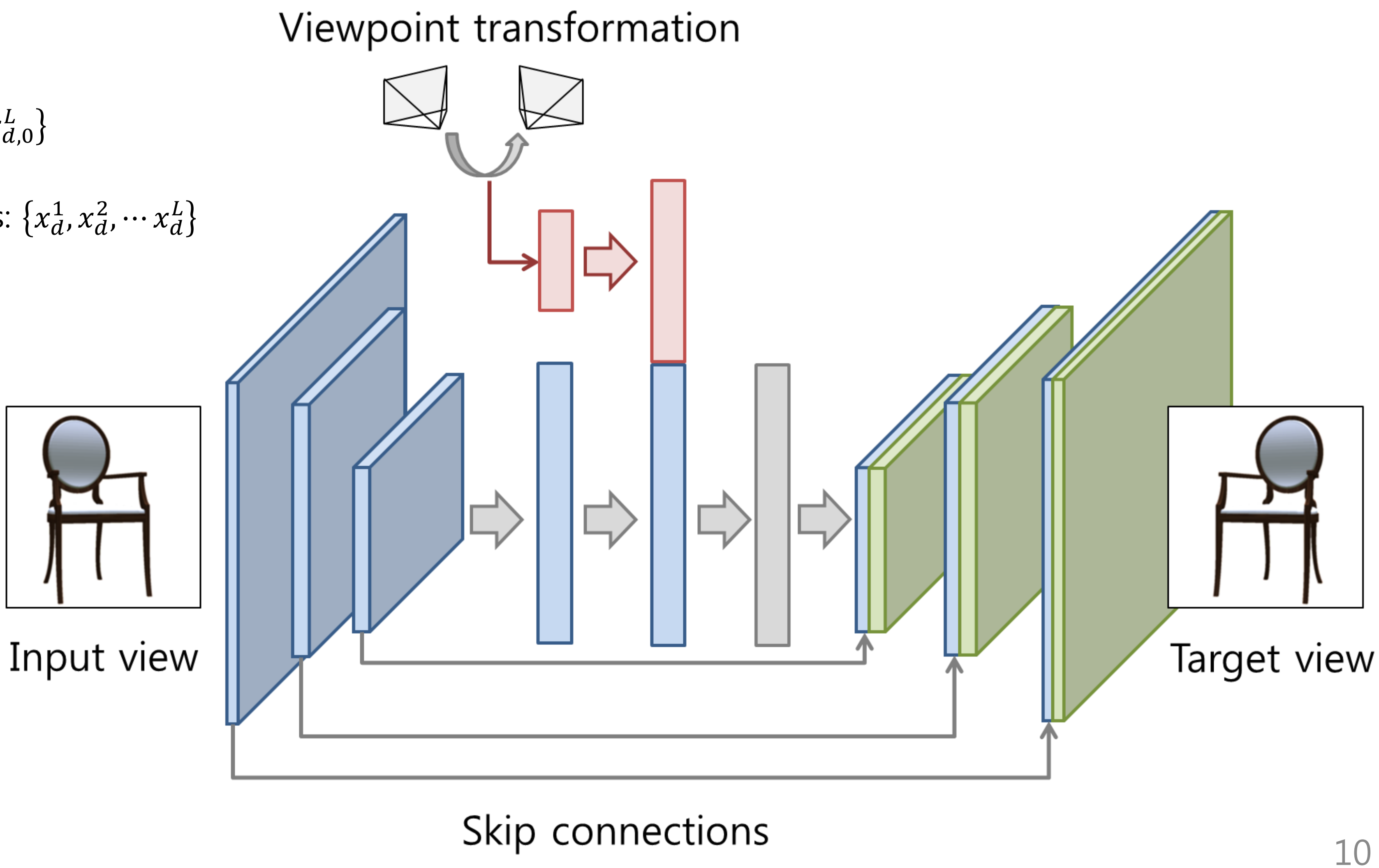
Encoder layers: $\{x_e^1, x_e^2, \cdots x_e^L\}$

Decoder layers: $\{x_{d,0}^1, x_{d,0}^2, \cdots x_{d,0}^L\}$

Concatenated Decoder layers: $\{x_d^1, x_d^2, \cdots x_d^L\}$

$$x_d^l = x_e^l \oplus x_{d,0}^l$$

$\oplus$: concatenation



U-Net with Attention Mechanism

Encoder layers: $\{x_e^1, x_e^2, \cdots x_e^L\}$

Decoder layers: $\{x_{d,0}^1, x_{d,0}^2, \cdots x_{d,0}^L\}$

Concatenated Decoder layers: $\{x_d^1, x_d^2, \cdots x_d^L\}$

Attention mechanism

Attn U-Net (Oktay, 2018) :

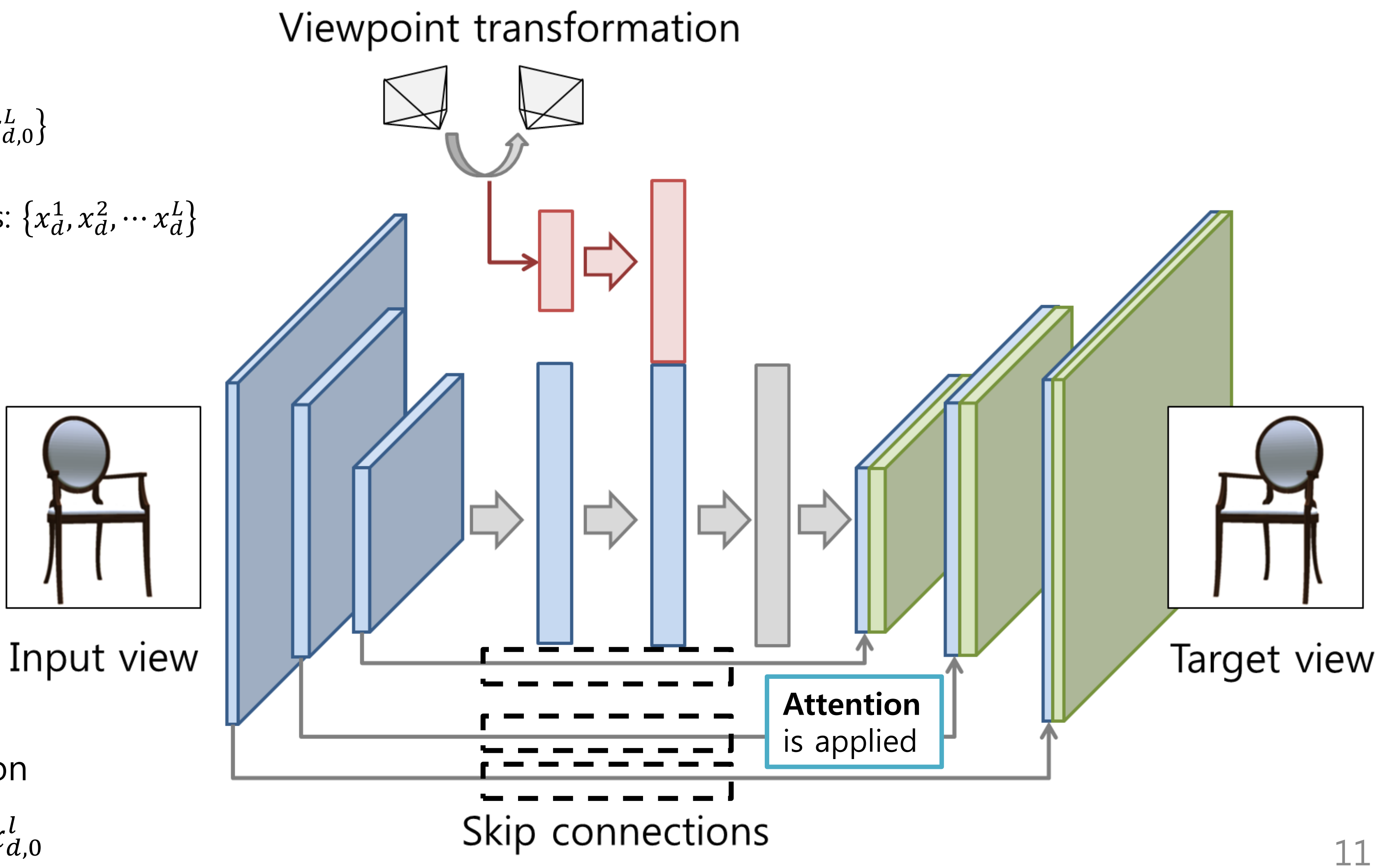
$$x_d^l = (\psi^l \otimes x_e^l) \oplus x_{d,0}^l$$

Cross Attn :

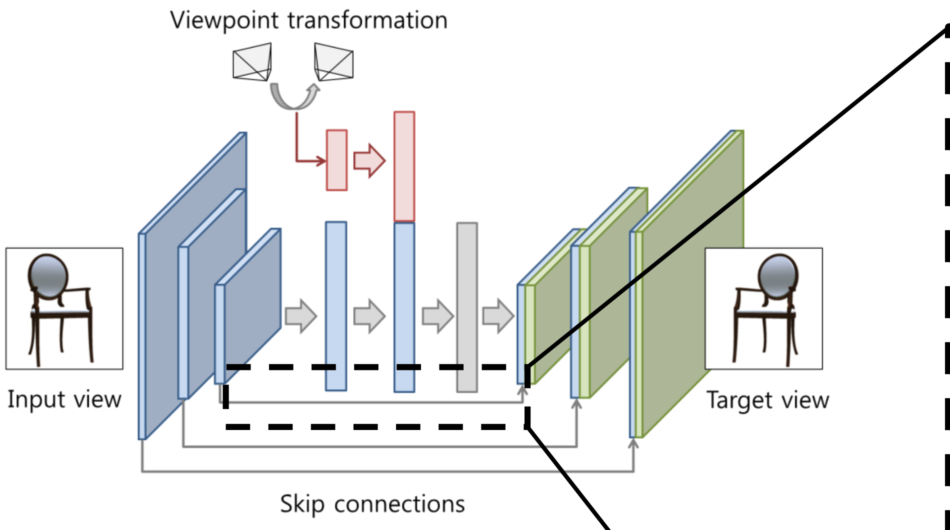
$$x_d^l = o^l \oplus x_{d,0}^l$$

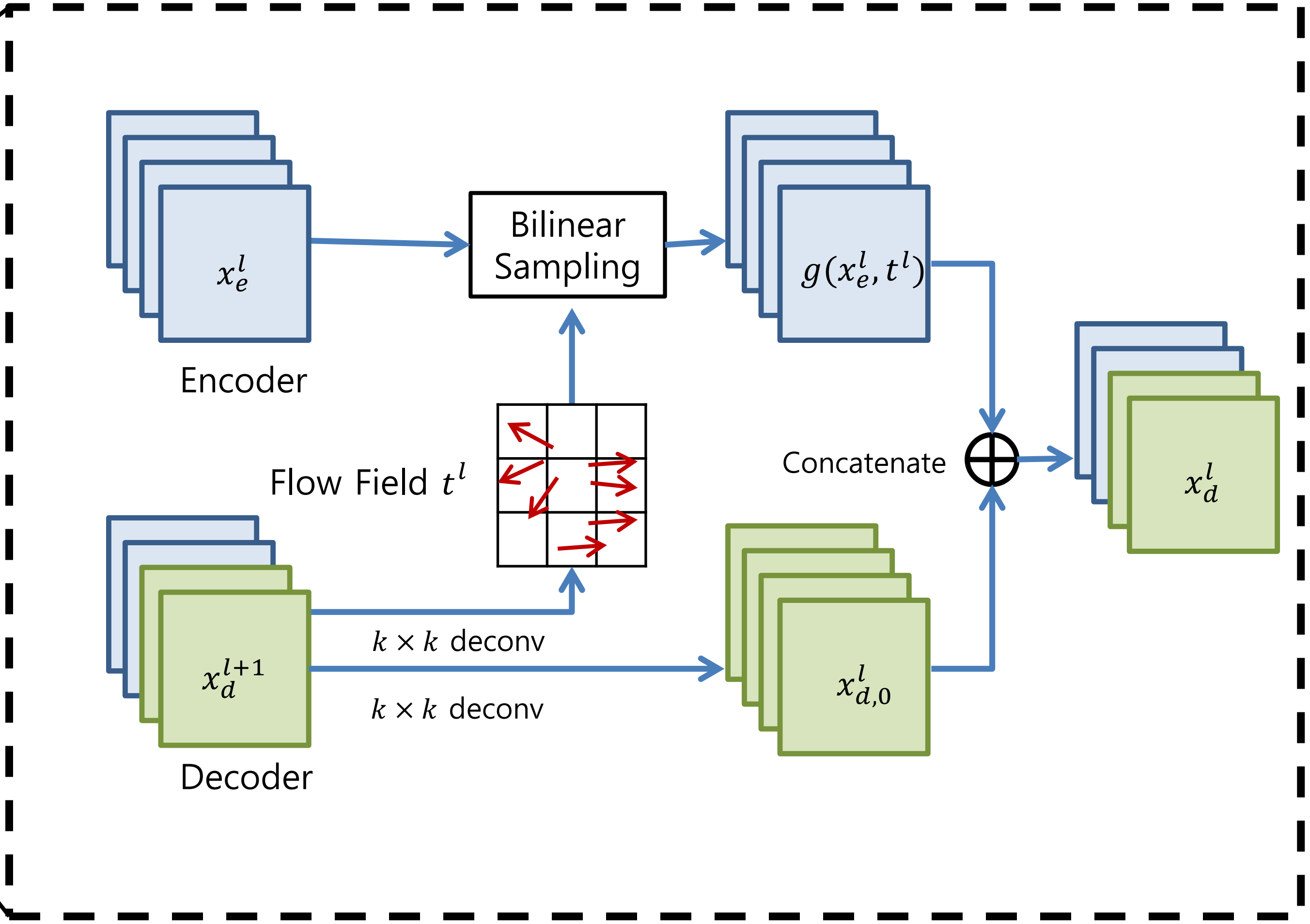
$\otimes$: element-wise multiplication

ψ^l, o^l are function of x_e^l and $x_{d,0}^l$



Flow-Based Hard Attention



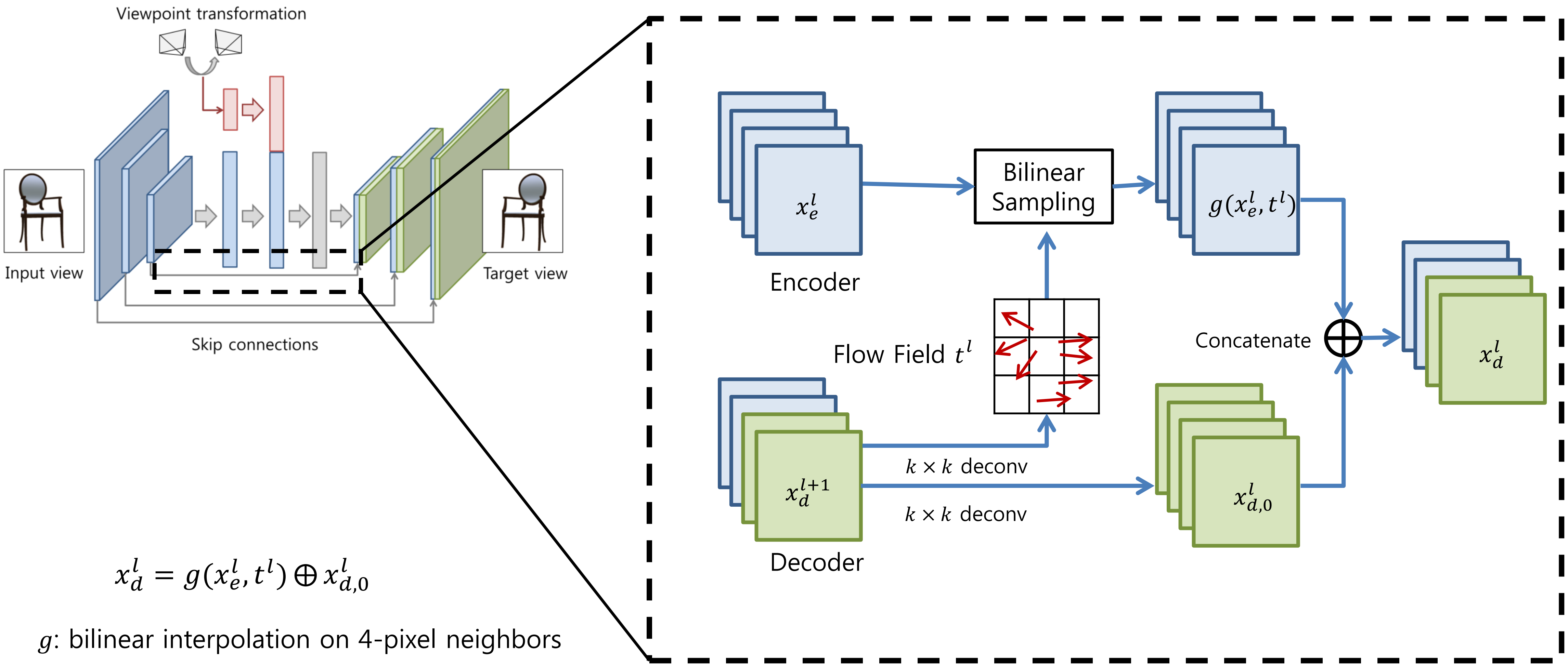


$$x_d^l = g(x_e^l, t^l) \oplus x_{d,0}^l$$

g: bilinear interpolation on 4-pixel neighbors

Flow can be thought as a hard attention

Flow-Based Hard Attention



$$x_d^l = g(x_e^l, t^l) \oplus x_{d,0}^l$$

g: bilinear interpolation on 4-pixel neighbors

Flow can be thought as a hard attention

EXPERIMENTS & RESULTS

Data set

- Two ShapeNet object classes (car, chair)
  - 500 models for training and 198 models for testing.
  - Each model rendered at 18 different azimuths with fixed distance, elevation
- Real scene (KITTI) and synthetic scene (Synthia)
  - 6 DoF as a pose representation.
  - 80% for training and 20% for testing.
  - Maximum 10 frame difference is allowed.

Chair

Car

KITTI

Synthia

Dataset modified from previous work. (Sun, 2018)

256 by 256 sized

Baselines

- Test on pixel generation and flow prediction module.
- Various skip connection strategies were tested.

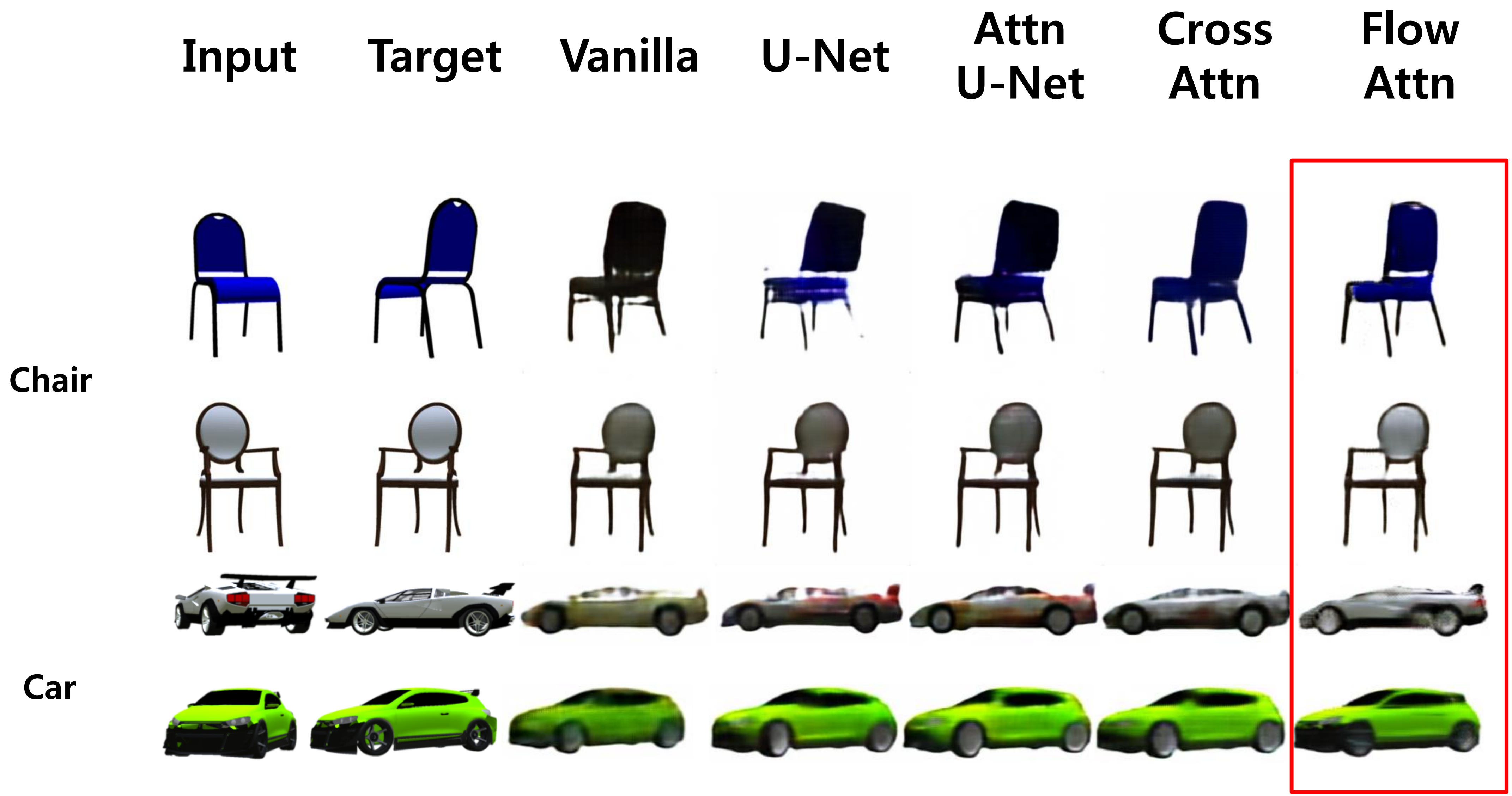
	Attention Type	Memory
Vanilla	-	$O(1)$
U-Net	-	$O(1)$
Attn U-Net (Oktay, 2018)	ith pixel to ith pixel	$O(HW)$
Cross Attn	ith pixel to image	$O(H^2W^2)$
Flow Attn	ith pixel to jth pixel	$O(HW)$

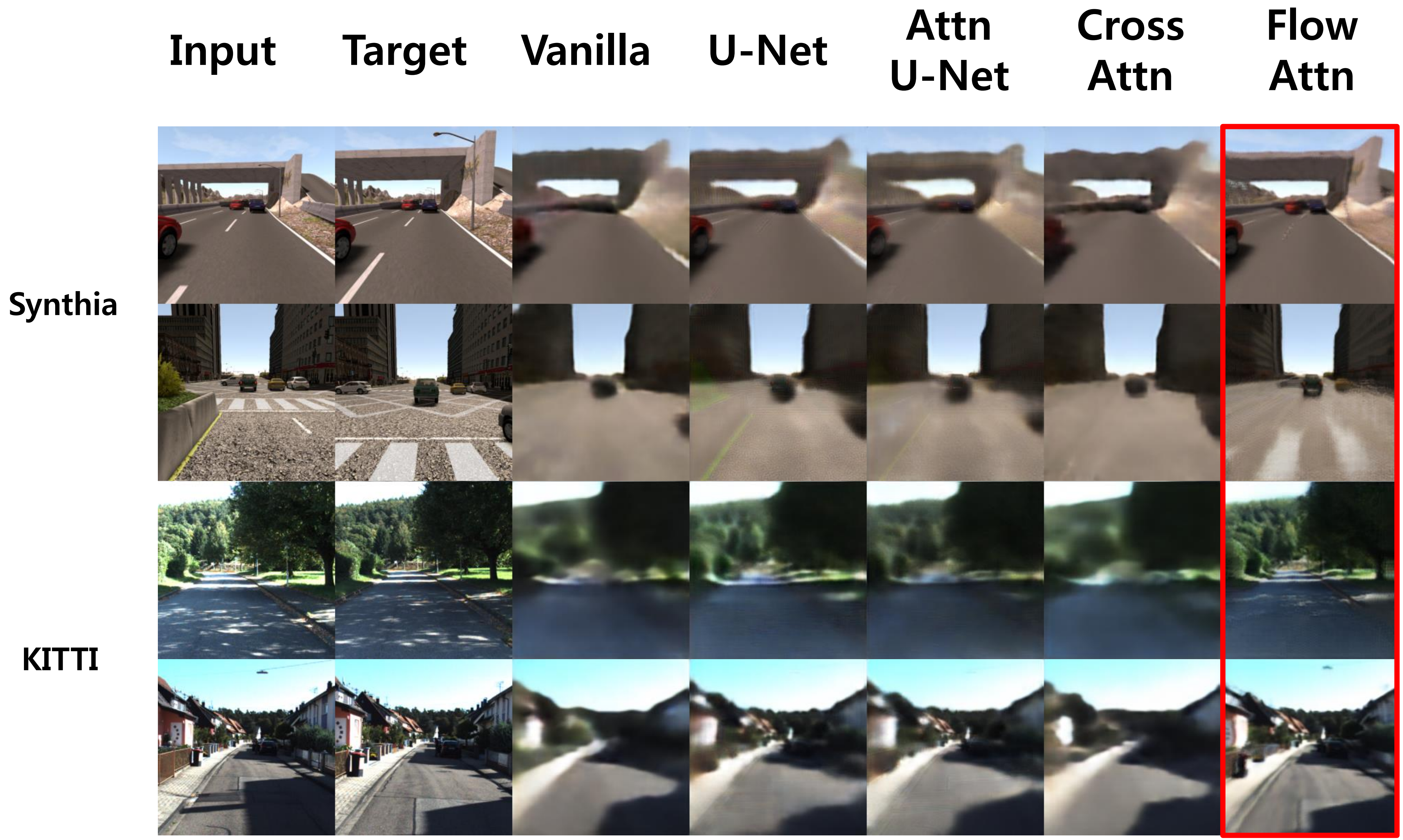
Results – Pixel Generation

Method	Car		Chair		Synthia		KITTI	
	$L1$	SSIM	$L1$	SSIM	$L1$	SSIM	$L1$	SSIM
Vanilla	0.0332	0.8910	0.0622	0.8535	0.0599	0.7324	0.0947	0.6681
U-Net	0.0327	0.8935	0.0623	0.8559	0.0544	0.7521	0.0838	0.6842
Attn U-Net	0.0330	0.8926	0.0629	0.8550	0.0548	0.7575	0.0835	0.6870
Cross Attn	0.0322	0.8961	0.0614	0.8573	0.0600	0.7331	0.0969	0.6659
Flow Attn	**0.0259**	**0.9091**	**0.0499**	**0.8725**	**0.0512**	**0.7597**	**0.0776**	**0.6939**

L1 : lower is better

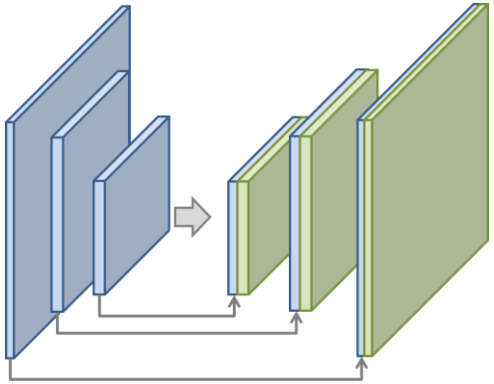
SSIM : higher is better


Input
Target
Vanilla
U-Net
Attn U-Net
Cross Attn
Flow Attn
Chair
Car

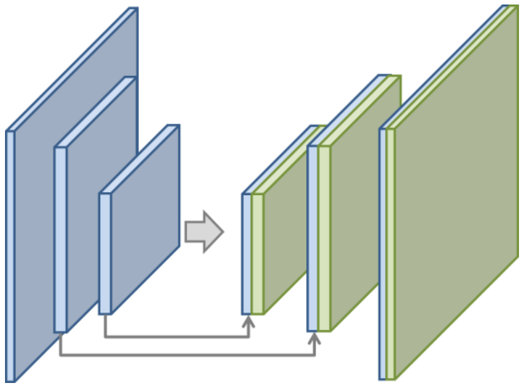

Input
Target
Vanilla
U-Net
Attn U-Net
Cross Attn
Flow Attn
Synthia
KITTI

Results – Flow Prediction

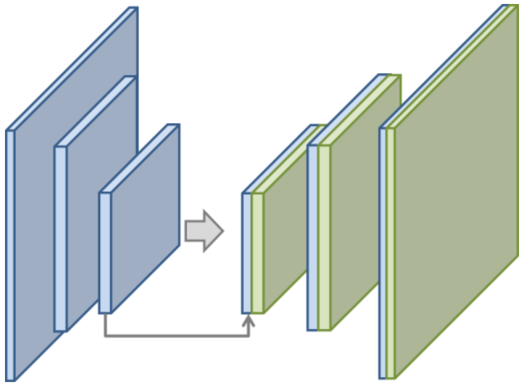
- All of the above methods seems not that helpful.
- Instead, we find that reduced numbers of the skip connections N_s by removing the outermost layers brings marginal improvement.



$N_s = 3$

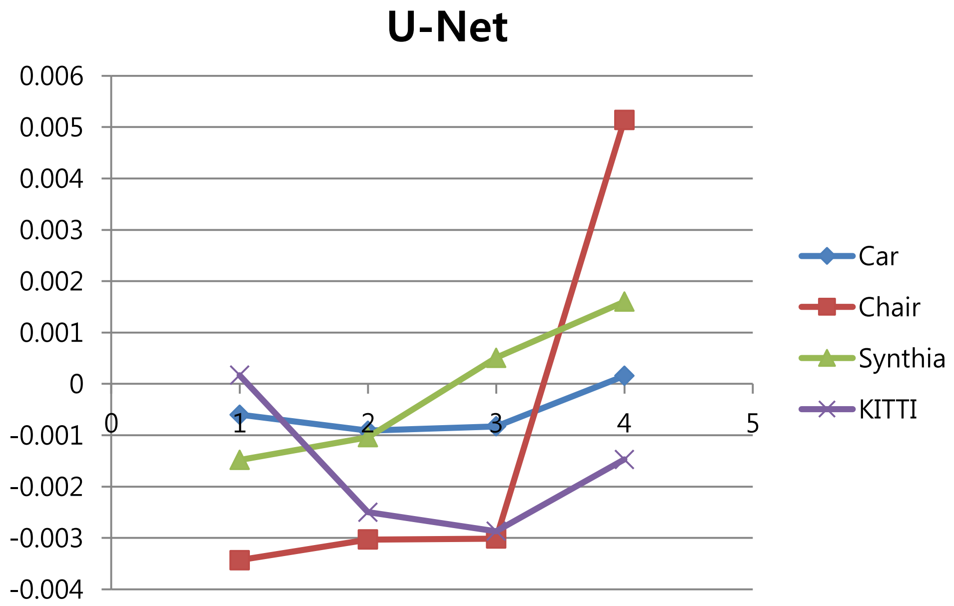


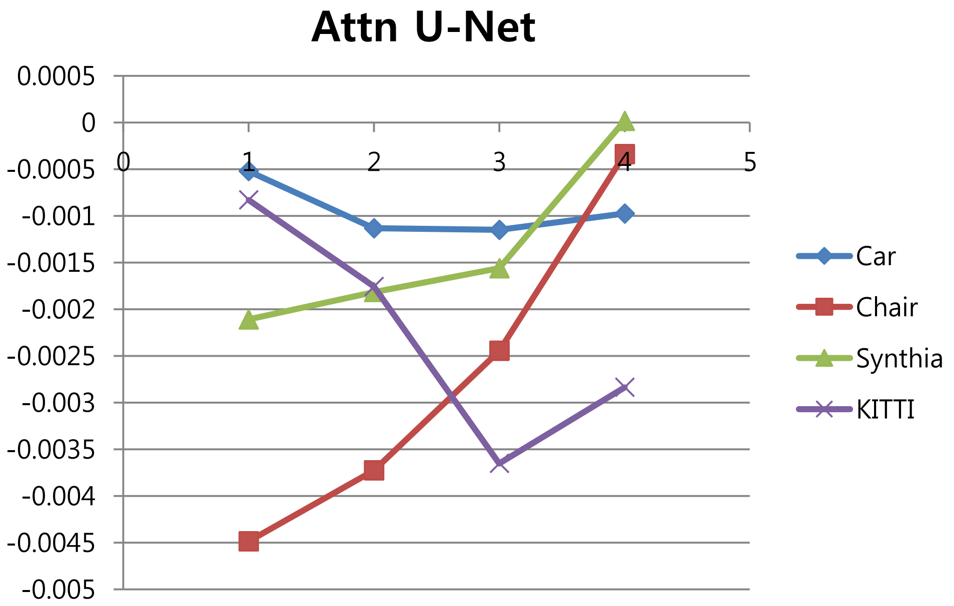
$N_s = 2$

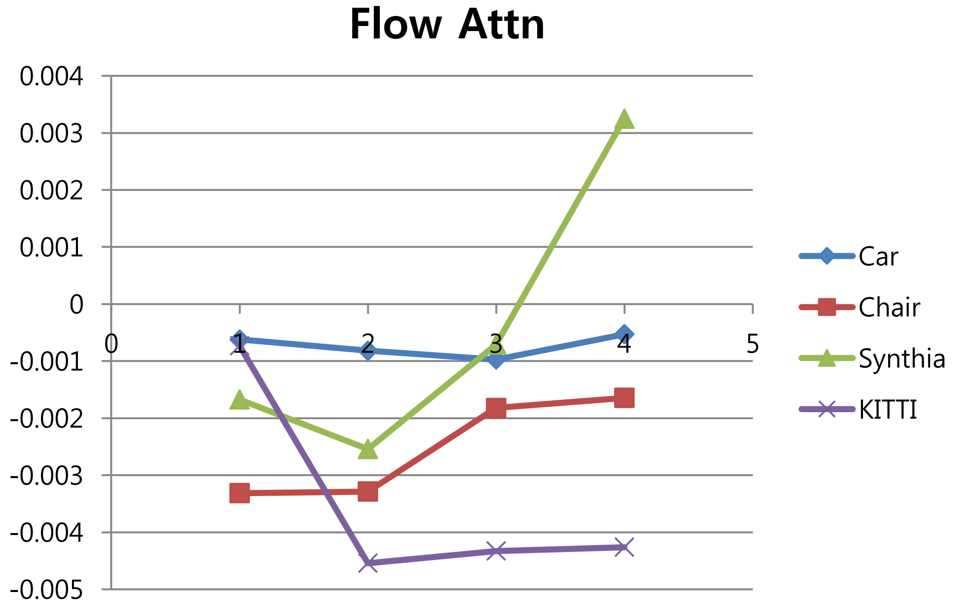


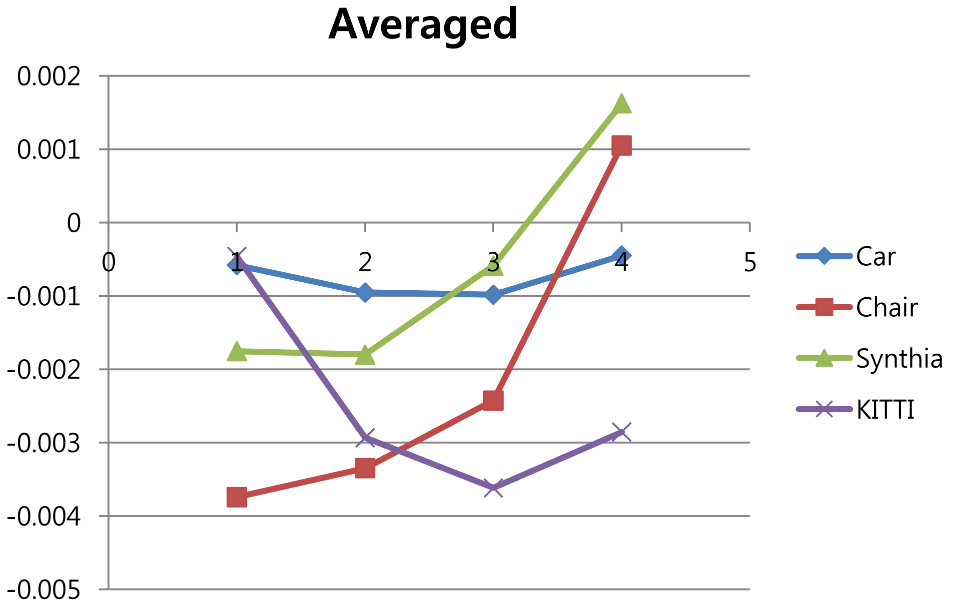
$N_s = 1$

※ y-axis means relative value to vanilla's L1 loss (lower is better), x-axis means N_s









Channel-wise Averaged Hidden Layers

Pixel Generation　　　　Flow Prediction

Layer Size	x_e^l	$g(x_e^l, t^l)$	$x_{d,0}^l$	x_e^l	$g(x_e^l, t^l)$	$x_{d,0}^l$

128

64

Similar　　　　**Different**

- x_e^l : encoder's feature
- $x_{d,0}^l$: decoder's feature
- $g(x_e^l, t^l)$: rearranged x_e^l

Conclusion

- We investigate how **skip connections** affect two widely used novel view synthesis module, pixel generation and flow prediction.
- We propose how the **skip connections** can be effectively applied on image-to-image translation under significant geometric change.

Code is available on github.

https://github.com/juhyeonkim95/NovelViewSynthesis

If you have any question, please email us.

E-mail : cjdeka3123@snu.ac.kr

Thank You!