NOVEL VIEW SYNTHESIS WITH SKIP CONNECTIONS

Juhyeon Kim and Young Min Kim

Department of Electrical and Computer Engineering
Seoul National University, Seoul, South Korea

ABSTRACT

Novel view synthesis is the task of synthesizing an image of an object at an arbitrary viewpoint given one or a few views of the object [1]. The output image of novel view synthesis exhibits a significant structural change from the input. Because of the large change, the skip connections or U-Net architecture, which can sustain the multi-level characteristics of the input images, cannot be directly utilized for the novel view synthesis [2]. In this paper, we investigate several variations of skip connection on two widely used novel view synthesis modules, pixel generation [1] and flow prediction [3]. For pixel generation, we find that the combination of the skip connections with flow-based hard attention is helpful. On the other hand, flow prediction enjoys marginal benefit from skip connections in deeper layers. Our pipeline suggests how to make use of skip connections on tasks that involve large geometric changes.

Index Terms— novel view synthesis, attention mechanism, skip connection, image-to-image translation.

1. INTRODUCTION

Novel view synthesis is the task that synthesizes an image of an object at an arbitrary camera pose given source images. While humans can easily infer invisible parts of an object from a view of an object, it is a challenging task for a computer especially when the view-point change is drastic.

Considering the camera pose as a domain, novel view synthesis can be thought as an image-to-image translation task [4]. One of the most successful techniques in image-to-image translation is the skip connections, also known as the U-Net architecture [5, 6, 7]. In contrast to the typical image-to-image translation which changes the style or texture of an image while maintaining the global structure, the novel view synthesis endures dramatic geometrical transformation. Therefore, as Park *et al.* [2] pointed out, such residual connections cannot be simply applied to novel view synthesis.

Relatively early works on the novel view synthesis use the encoder-decoder structure without any skip connections [1, 8, 3]. The works can be divided into two groups; pixel generation and flow prediction. *Pixel generation* methods [1, 8] directly generate output pixels from the input image. On the other hand, *flow prediction* generates a dense flow field that maps each pixel in input and output images [3].

There are a few variations; the input image can first be deformed by the flow network and further refined [2]; or the pixel generation and the flow prediction networks can be used in parallel and aggregated into a single final image using a confidence map [4]. Park *et al.* [2] use skip connections only at the refinement network, where the deformation between the input and output is minimal. Sun *et al.* [4] use skip connections only for the pixel generation module. There are also other methods using 3D kernel [9, 10] or depth-guided warping [11], that skip connections are not used.

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. In this paper, we study the effect of various types of skip connections and find the appropriate way to apply residual connections to novel view synthesis.

For pixel generation, we find that using flow-based hard attention mechanism to skip connection is effective. Inspired by the recent success of attention guided skip connections [12, 13], we are the first to suggest a light-weighted attention mechanism to be applied in the skip connection for the task that handles large geometric changes. As expected, the combination of the skip connections and flow-based hard attention mechanism can deliver multi-level features from the encoder layer to the adequate places in the decoder layer even under severe view point changes.

For flow generation module, delivering low-level features even using several attention mechanisms is not as effective as for pixel generation, because of the intrinsic difference between input (pixel) and output (flow). Instead, we try an alternative way, varying the number of skip connections and demonstrate that only connecting deeper feature layer can be helpful.

We present thorough evaluations on various methods to apply skip connections on the image generation modules involving geometric changes. The correct combination of attention and layers can improve the performance of the existing architecture.

This work was supported by the New Faculty Startup Fund from Seoul National University, and the BK21 Plus program of the Creative Research Engineer Development for IT, Seoul National University in 2020.

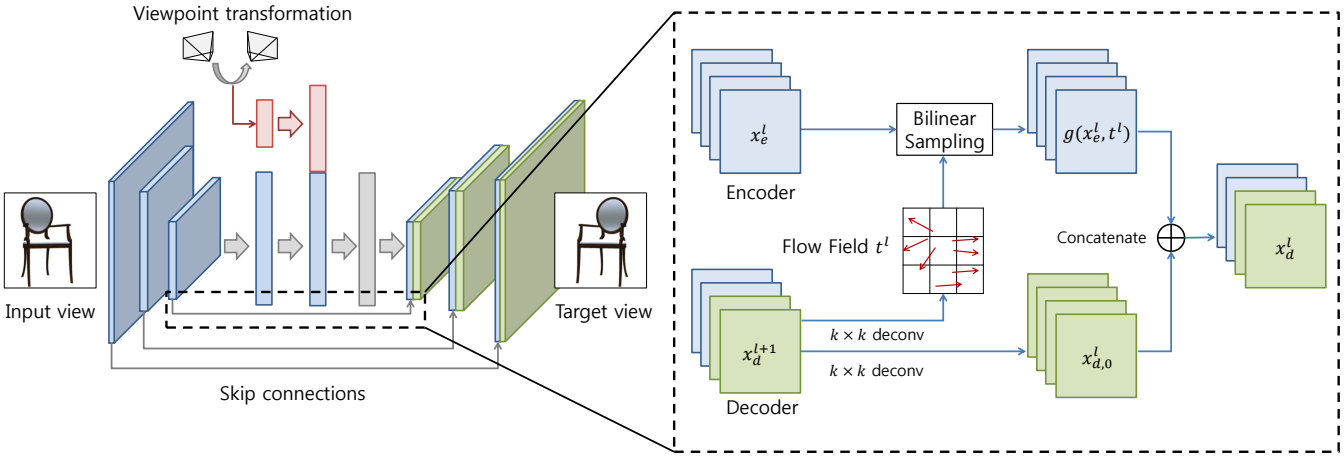

Fig. 1. Flow-based hard attention. It applies flow-guided skip connection to the vanilla novel view synthesis module.

2. PROPOSED METHOD

2.1. U-Net Architecture

Given an image, the conventional encoder-decoder architecture is composed of multiple layers of encoder x_e^l followed by layers of decoder x_d^l. The encoder and decoder are symmetric; to advocate the symmetric structure, we notate the encoder layers as $\{x_e^1, x_e^2, \cdots x_e^L\}$ and index the decoder layers the in reverse order, $\{x_d^L, x_d^{L-1} \cdots x_d^1\}$. The U-Net architecture (Fig. 1, left) connects the corresponding symmetric layers from the encoder (blue) to the decoder (green) to propagate the information of original feature when the decoder tries to decipher the abstract information. The residual connection between the same location at the paired layers was sufficient when the global structure is preserved, such as texture transfer or image segmentation [5, 6, 7].

In novel view synthesis case, we have to additionally incorporate the view transformation between input and output. We flatten the final encoder layer, concatenate the view information and pass it through several fully connected layers [1, 3], as shown in red in Fig. 1. The original U-Net [6] does not contain any intermediate fully connected layer, but for convenience, we just call this architecture also U-Net.

2.2. Flow-Based Hard Attention

An attention mechanism used in [12] controls the magnitude of the features with the element-wise multiplication. However this method seems vulnerable to dramatic shape change of novel view synthesis. We want to utilize attention mechanism when there is a significant deformation between the input and the output. Our first attempt is to find the mapping from every location in the encoder layer to the corresponding decoder layer. We call it *cross attention* and the details can be found in the appendix. However, it is pixel-to-image attention, so cannot be applied to large images because of memory limitation.

Table 1. Variation of architectures with skip connections

Method	Attention Type	Memory
Vanilla	-	$O(1)$
U-Net (Sec. 2.1)	-	$O(1)$
Attn U-Net [12]	ith pixel to ith pixel	$O(HW)$
Cross Attn (Sec. 6.1)	ith pixel to image	$O(H^2W^2)$
Flow Attn (Sec. 2.2)	ith pixel to jth pixel	$O(HW)$

To overcome this problem, we propose a *flow-based hard attention* or just *flow attention* that relates a pixel from the encoder layer to a pixel in the paired decoder layer, inspired by [3]. The overall structure is depicted on the right side of Fig. 1. Given the previous layer x_d^{l+1} from the decoder side, we train the network to generate two outputs; the flow field, t^l, and the decoder layer with the l-th feature map as the vanilla implementation, $x_{d,0}^l$. The dimensions of flow field t^l is the size of the l-th decoder layer, where individual components are two-dimensional vector. According to the vector fields, t^l samples features from the paired encoder layer x_e^l, acting as a hard attention between x_e^l and x_d^l. The final decoder feature is the concatenation of the sampled encoder feature $g(x_e^l, t^l)$ and $x_{d,0}^l$,

$$x_d^l = g(x_e^l, t^l) \oplus x_{d,0}^l. \tag{1}$$

Here, g is a bilinear interpolation on 4-pixel neighbors.

While Zhou *et al.* [3] applied dense flow field only at the last layer, we utilized it on every hidden layer, transferring the encoder features effectively to the decoder side. Compared to the cross attention which maps individual pixels to the image, the flow-based hard attention is pixel-to-pixel mapping, which requires much less memory (Table 1).

3. EXPERIMENTS

We demonstrate the performance of various skip connections for novel view synthesis by augmenting the baseline modules of both pixel generation [1] and flow prediction [3].[1,2] The tested architectures are listed in Table 1. The vanilla model is the original encoder-decoder model without skip connections. Starting from 256×256 image, individual layers of encoder extract features from the previous layer by reducing the size by the factor of two. U-Net connects the inner layers of the same size from 8 to 128 as described in Sec. 2.1. Attn U-Net [12] applies the attention to the pixels at the same position of the corresponding layer. Cross attention can calculate pixel to image attention, but because of memory issue, we excluded skip connections for the size 128 layer. For flow prediction module [3], the innermost layer starts from size 16, while others are same.

Data set. We tested our algorithm with objects and scenes as provided by [4]. The testing was conducted in an exhaustive way for both cases, considering all possible pairs and averaged. We used two ShapeNet [14] object classes (car, chair) with total 698 models, where 500 models were used for training and 198 models were used for testing. Each model is rendered at 18 different azimuths (20° shift) with a fixed distance and elevation. Pose information is incorporated to the encoder as sin cosine value [1] or one-hot encoded vector [3]. We also tested with real-world (KITTI [15]) and synthetic scenes (Synthia [16]) where the difference of the input and target frame is limited by 10. We used the 6 DoF information (translation and Euler rotation) as a pose representation. We used 80% for training and 20% for testing [4]. Note that we only used a part of ShapeNet data because of the file I/O speed issue in random accessing. The whole data set is used for scene data.

Training and testing details. We trained each model $100,000$ times with learning rate 5e-5. At each iteration, we randomly selected 16 source/target image pairs. We used $L1$ loss for simplicity, but perceptual or adversarial loss could be added. For evaluation, we used $L1$ loss (the lower the better) and SSIM (the higher the better) as previous works [2, 4].

4. RESULT

4.1. Pixel Generation

The quantitative results for pixel generation are presented in Table 2. Compared to the vanilla architecture, U-Net, Attn U-Net, Cross-Attn do not demonstrate noticeable improvement in the ShapeNet case. Figure 2 shows that they manage to approximate the color, but fail to accurately transfer the texture. In the case of the scene, where the shape difference between the source and the target is relatively small, there is a slight performance improvement for U-Net and Attn U-Net. While cross-attention is expected for better performance incorporating deformation, the improvement is limited as we excluded skip connection on the outermost layer due to the memory problem. [3] The flow attention not only produces the best quantitative results but also reliable synthesized images in both object and scene cases. Figure 3 shows that the encoder features are successfully modified to fit to the decoder features $x^l_{d,0}$.

Table 2. Performance of variants of skip connections for pixel generation.

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

4.2. Flow Prediction

The improvement of flow prediction model is marginal with the full attention mechanism (N_s=4 in Table 3). However, interestingly, removing the skip connections of the outer layers turns out to be helpful. We tested reduced numbers of skip connections N_s by removing the outermost layers. (*i.e.*, $N_s = 1$ means skip connection on size 16 layer, $N_s = 2$ means on size $16, 32$ layers and so on.) As indicated in Table 3, reduced skip connections slightly improve the results. Note too many removals may converge to the vanilla encoder-decoder model. We present only the results for flow attention, but similar tendencies were observed for the other architectures except cross attention.

We guess these results are due to the domain discrepancy between the input and output. The skip connections of outer layers deliver low-level features deduced from *pixels* to regress *flows*, therefore direct connections are not effective even if it has been rearranged. However, at inner layers, the domain differences are diluted as more abstract features are in

[1] [2, 4] were excluded because the two modules are used in combination.

[2] Our code and full result are available at github.

[3] Cross attention showed more noticeable improvement in our previous work [17] with smaller image size (64×64) that applying it to the outermost layer is manageable.

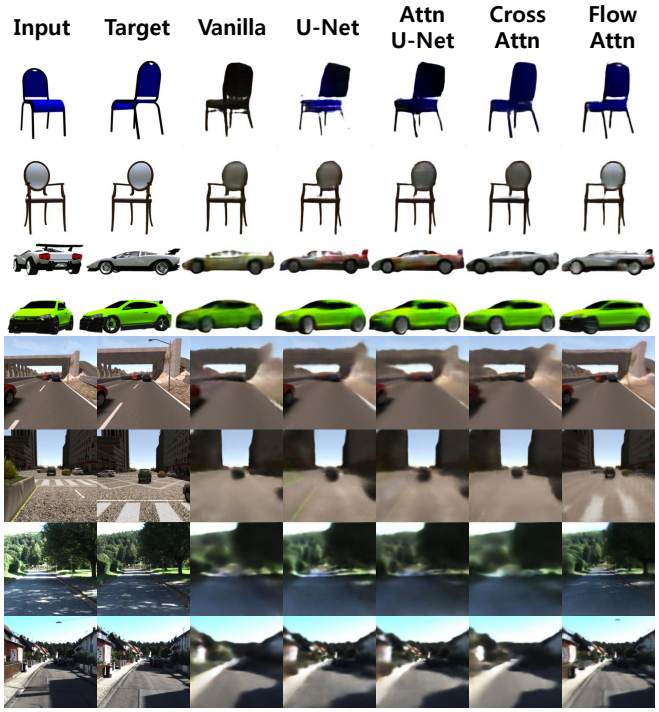

Fig. 2. Qualitative results of pixel generation module with different skip connections and attention mechanisms. For scenes, the top two lines are Synthia, and the bottom two lines are KITTI.

action, therefore connecting them would be beneficial. Also inner layers have next layers to be processed, have chance to be properly converted into flows. Figure 3 shows that the transformed encoder feature $g(x_e^l, t^l)$ is geometrically similar to the decoder feature $x_{d,0}^l$ for pixel generation, where as flow prediction seems to be unsuccessful. This indicates that rearranging the encoder feature to adapt for the decoder feature is difficult for flow prediction, which indirectly supports our conjecture on the domain discrepancy.

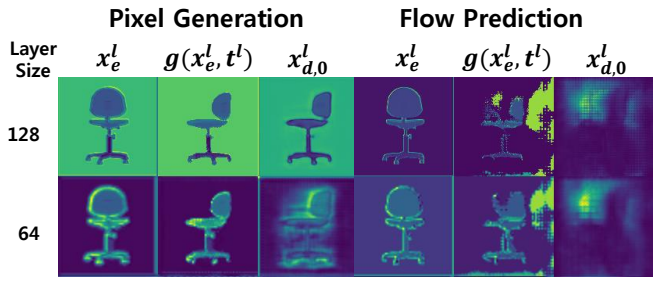

Fig. 3. Channel-wise averaged hidden layers for modules that uses skip connections with flow attention ($N_s = 4$).

Table 3. Performance of skip connections for flow prediction. N_s means number of skip connections from the innermost hidden layer.

Method	N_s	Car		Chair	
		L_1	SSIM	L_1	SSIM
Vanilla	-	0.0256	0.9168	0.0448	0.8898
Flow-Attn	4	0.0251	0.9172	0.0449	0.8885
	3	**0.0246**	**0.9181**	0.0426	0.8903
	2	0.0248	0.9176	0.0415	0.8933
	1	0.0250	0.9175	**0.0415**	**0.8936**
Method	N_s	Synthia		KITTI	
		L_1	SSIM	L_1	SSIM
Vanilla	-	0.0580	0.7372	0.0931	0.6470
Flow Attn	4	0.0612	0.7060	0.0888	0.6458
	3	0.0573	0.7221	0.0887	0.6454
	2	**0.0554**	**0.7372**	**0.0885**	0.6471
	1	0.0563	0.7356	0.0923	**0.6482**

5. CONCLUSION

In this paper, we investigated how skip connections affect novel view synthesis for both pixel generation [1] and flow prediction [3]. The conventional U-Net architecture [6], while successful in image-to-image translation, cannot accommodate significant global shape change in the original form. Applying our suggested attention mechanism improved the performance for pixel generation, transferring low-level information at the correct locations given the large deformation between input and output. For flow prediction, on the other hand, the partial skip connections acting on inner layers are helpful alleviating possible domain discrepancy. Our work propose how the skip connections can be applied under significant geometric change, and can be applied to other tasks, such as video frame interpolation or optical flow estimation.

6. APPENDIX

6.1. Cross Attention

We define the cross attention β_{ij}^l between i-th pixel of the l th encoder layer x_e^l to the j-th pixel of the corresponding decoder layer x_d^l as following:

$$\beta_{ij}^l = \frac{\exp s_{ij}^l}{\sum_{i=1}^N \exp s_{ij}^l}, \quad \text{where} \quad s_{ij}^l = f(x_e^l)_i^T g(x_d^l)_j. \quad (2)$$

f and g in Eq. (2) are 1×1 convolution to each layer, $f(x_e^l) = W_f^l x_e^l$ and $g(x_d^l) = W_d^l x_d^l$. The cross attention β_{ij}^l are re-normalized from s_{ij}^l over all input pixels in Eq. (2). Then, we get jth pixel information o_j^l using attention $o_j^l = \sum_{i=1}^N \beta_{ij}^l x_{e,i}^l$. We get final output by concatenating o^l

to original decoder feature layer: $y_d^l = o^l \oplus x_d^l$. This method is pixel-to-image attention. It is similar to the method of self-attention GAN [18], but the attention is calculated between the encoder and decoder instead of itself.

7. REFERENCES

[1] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox, "Multi-view 3d models from single images with a convolutional network," in *European Conference on Computer Vision*. Springer, 2016, pp. 322–337.

[2] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg, "Transformation-grounded image generation network for novel 3d view synthesis," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 3500–3509.

[3] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros, "View synthesis by appearance flow," in *European conference on computer vision*. Springer, 2016, pp. 286–301.

[4] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim, "Multi-view to novel view: Synthesizing novel views with self-learned confidence," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 155–171.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[7] Vladimir Iglovikov and Alexey Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[8] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee, "Weakly-supervised disentangling with recurrent transformations for 3d view synthesis," in *Advances in Neural Information Processing Systems*, 2015, pp. 1099–1107.

[9] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang, "Hologan: Unsupervised learning of 3d representations from natural images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7588–7597.

[10] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo, "Transformable bottleneck networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7648–7657.

[11] Xu Chen, Jie Song, and Otmar Hilliges, "Monocular neural image based rendering with continuous view control," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4090–4100.

[12] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[13] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim, "Unsupervised attention-guided image-to-image translation," in *Advances in Neural Information Processing Systems*, 2018, pp. 3693–3703.

[14] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al., "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[16] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

[17] Juhyeon Kim and Young Min Kim, "Novel view synthesis using attention mechanism," in *32nd Workshop on Image Processing and Image Understanding*. Korean Institute of Information Scientists and Engineers, 2020.

[18] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.