Camera Focal Length Prediction for Neural Novel View Synthesis from Monocular Video

Dipanita Chakraborty*, Werapon Chiracharit*, Kosin Chamnongthai*, and Minoru Okada†
* King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand
† Nara Institute of Science and Technology, Nara 630-0192, Japan
* E-mail: kosin.cha@kmutt.ac.th

***Abstract*—Novel view synthesis is a challenging task that generates multi-view images from a single-view object by reconstructing the depth and spatial information between the camera and the object. This task specifically facilitates the rendering of 2D objects into 3D representations from monocular video scenes. Existing methods suffer from depth information loss between the camera and object when provided with limited single-view input images, resulting in poor reconstruction accuracy in 3D space. Moreover, they lack a high-level depth feature map representation of scene information. Therefore, we propose a multilayer encoder-decoder architecture-based network that efficiently predicts the focal length between the object and camera from a mono-view image. Additionally, we utilize a combined feature extraction strategy from the estimated depth feature map and RGB input image to synthesize novel views. While the encoder network extracts semantic high-level features at multiple scales, the decoder network refines these combined features for synthesis. Our method effectively improves depth information retention while achieving good reconstruction performance. Experimental results evaluated on a benchmark dataset demonstrate the efficacy of our proposed method.**

I. Introduction

The need for an accurate camera focal length estimation directly facilitates facial alignment analysis for biometric applications [1], subject-to-camera distance estimation [2], automatic camera calibration, 3D-object reconstruction, novel view synthesis [3], and several other crucial computer vision applications. Among all these applications, novel view synthesis is a significant computer vision task that can generate new viewpoints of an object from single-viewed images. Traditional novel view synthesis methods use fixed camera intrinsic parameters, such as camera focal lengths. However, these methods are unable to synthesize novel views of an object from in-the-wild images due to the unavailability of camera intrinsic parameters. Hence, this computer vision problem of camera focal length estimation from in-the-wild images has gained the attention of many researchers in the past few years [4].

Some depth-based novel view synthesis methods have been proposed from single-viewed images using disocclussion filling approach [1], [5]. Kim and Kim [6] proposed a machine learning-based novel view synthesis module using pixel generation and flow prediction. They investigated the performance variations based on the skip connections. Their method of using 3 skip connections for flow predictions returned SSIM loss scores of 0.9181 and 0.8903 for object categories "Car" and "Chair", respectively. The results indicate that there is still room for improvement since the higher the SSIM score, the better the performance. A self-supervised learning-based method was proposed in [7] that used a semi-parametric approach to integrate rich visual information. This method synthesized novel views of "car" and "chair" objects from 12 evenly-spaced azimuthal angles. The Neural Radiance Field (NeRF)-based approach using sparse RGB-D images was proposed in [8]. The method solved the main limitation of the large dataset requirement in the conventional view synthesis systems. The highest mean SSIM score achieved by this proposed model was 0.6100. However, the main limitation is that these methods can not generate novel views of in-the-wild images due to the unavailability of camera intrinsic metadata.

An effective way to resolve this problem is by predicting the camera focal length from in-the-wild images. A fast camera focal length extraction method was proposed in [9] that used parallel particle swarm optimization. This method returned good performance while reducing the time consumption. A novel camera focal length determination method was proposed in [10] using a calibration pattern of the checkerboard. However, these methods lack semantic feature information which is needed for estimating the focal length between the camera and the captured image of a complex scene. To address this issue, a deep learning-based approach was proposed in [11] that used the DLA-34 network to perform high-level visual feature extraction first, followed by a 3D box parameters regression task to predict the camera focal length. The achieved accuracy of 56.31% on the benchmark KITTI dataset indicates further need for improvement in their proposed method.

In our proposed method, we consider integrating the camera focal length estimation module into the encoder-decoder network, depth estimation module, and view transformation module to enhance the performance of novel view synthesis. The main contributions of our proposed method are as follows:

1) We propose a camera focal length prediction module using an encoder network and a linear activation function that extracts the semantic high-level features at multi-scale.
2) We utilize the combined feature extraction strategy from the input RGB image and corresponding depth map to improve the view transformation module.
3) We modify the decoder network with convolutional operations to synthesize the novel views.

This article is presented as follows: Section II reviews the existing state-of-the-art research, Section III describes the proposed methodology in detail, Section IV explains implementation and experimental evaluation, and lastly Section V concludes the achievements and future scope of this research study.

II. Related Work

We introduce an encoder-decoder novel view synthesis model architecture with an integrated camera focal length prediction model. Therefore, we summarize a brief of existing methodologies into two sections, camera focal length estimation model and novel view synthesis model.

A. Camera Focal Length Estimation Model

A robust camera pose and camera focal length estimation method were proposed in [12] to solve the perspective-n-point (PnP) problem for uncalibrated cameras, also referred to as unknown camera intrinsic parameters. This method utilized exhaustive linearization to solve the UPnP (Uncalibrated PnP) equation and achieved good accuracy. However, the time consumption was comparatively high. To address this issue, Zhou et al. [13] proposed an efficient PnP solution for cameras with unknown camera focal lengths using a polynomial system. Later, Yin et al. [14] proposed a robust PnP solution for cameras with unknown focal lengths using Gröbner basis minimal solver combined with convex optimization. Although, this proposed method achieved higher accuracy with lower time consumption, in some cases the matching algorithm failed to provide sufficient key points to localize the PnP problem. He et al. [15] proposed a deep learning-based approach for depth estimation using embedded camera focal lengths from single-viewed images. Their method could efficiently reconstruct depth feature maps, however, their system could not achieve the fastest of all state-of-the-art methods in terms of lowest run time. Ponimarkin et al. [16] proposed a focal length estimation method using a render and compare strategy from in-the-wild RGB images. This method significantly reduced the loss error than the existing methods. However, this method suffered from false retrieval of some 3D models.

B. Deep View Synthesis Model

The novel view of an object can be efficiently reconstructed by generative neural networks using depth feature extraction. Liu et al. [17] proposed a deep generative model, called SCGN based on encoder-decoder network architecture. The advantage of this network it can encode appropriate high-level features without the requirement of geometrical property rectifications, resulting in reduced geometrical distortion issues. However, this method did not investigate single-viewed source images. Lei et al. [18] proposed a powerful deep view synthesis network, called DGCC-Net by combining gradual and cycle synthesis. Whereas the gradual conversion technique learns the progressive rotation trend using intermediate transformation between the source and target images to synthesize a clearer target view, the cycle network maps the synthesized target view back to the source view to promote better feature learning of the target view. The proposed methodology performed well in terms of reducing the computation complexity. However, the performance can be improved by incorporating various features along with the RGB source image. Addressing this issue, Jiang et al. [19] proposed an encoder-decoder architecture-based network that uses the source image and its corresponding warp image as the input. Additionally, this method introduced to channel attention block technique to reduce the errors from depth estimation and target view generation process. However, due to the unsupervised training for depth estimation, some missing pixels were observed in the generated target view.

In this article, we aim to improve the novel view synthesis model by embedding predicted camera focal length information and semantic high-level visual features at multi-scale to generate a depth feature map. Additionally, we apply a combined feature extraction strategy from estimated depth features and the corresponding RGB input image to generate novel views.

III. Methodology

This section presents the proposed camera focal length prediction method to generate a novel view of a single-viewed object, as shown in Fig. 1. We describe the method in four steps. First, we use an encoder network to extract semantic high-level features from the RGB input image, as shown in Fig. 2. These features are then utilized to predict the focal length between the object and the camera. In the second step, we concatenate the extracted semantic high-level features with the predicted focal length features and feed them to a decoder network. The decoder network then estimates the depth feature map from the input image. In the third step, we use another encoder network to extract combined features from the input image and its corresponding depth feature map. Finally, we use a perspective transformation module on the combined features to transform into a new viewpoint. Then another decoder network is applied to synthesize a novel view image from the transformed features, upsampling it to the original input image resolution. The following subsections A, B, C, and D provide detailed material about the proposed methodology.

A. Camera Focal Length Prediction with Encoder

The main objective of this step is to predict the camera's focal length from an in-the-wild image, where the camera's intrinsic properties, including the focal length, are unknown. This step is crucial because it enhances our understanding of the geometrical properties, spatial location, and depth information of objects in the image. As shown in Fig. 3, the distance between the camera lens and the sensor is referred to as the focal length (F), while the angle of view (θ) represents the scene that the camera sensor can capture. The angle of view (θ) can be determined using the focal length and sensor width (W), as in (1).

$$\theta = 2 \times tan^{-1} \frac{W}{2 \times F}. \tag{1}$$

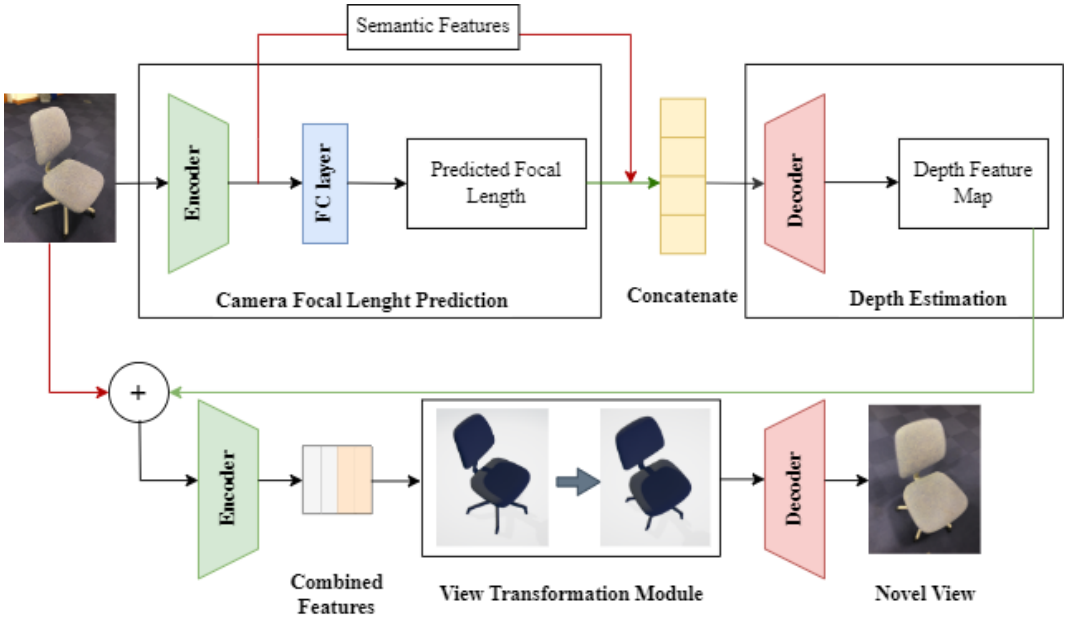

Fig. 1. Shows framework of the proposed methodology.

.

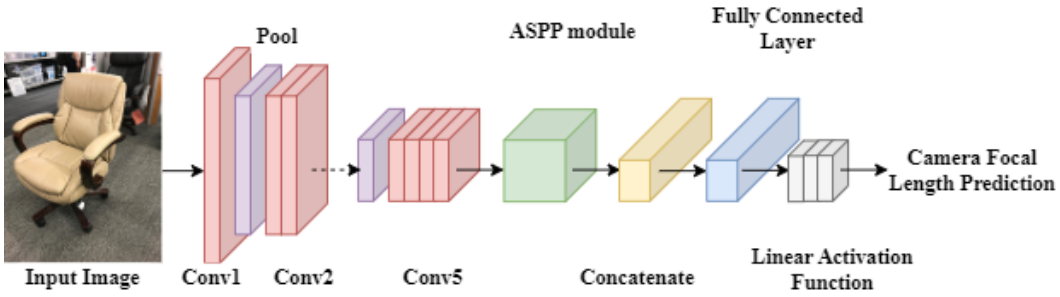

Fig. 2. Shows the proposed encoder network architecture based on DeepLabv3+ network, where Resnet-18 CNN is used as the backbone network.

.

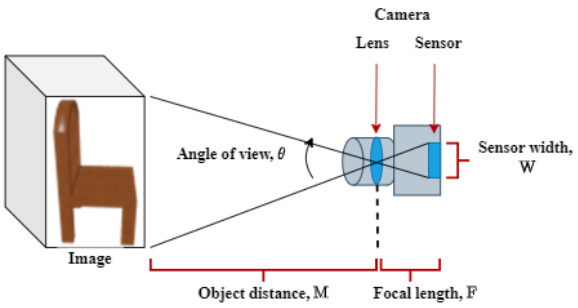

Fig. 3. Shows two of the camera intrinsic parameters, focal length and angle of view, which are used for novel view synthesis.

.

The camera focal length prediction approach directly facilitates depth estimation and 2D to 3D rendering of objects. We use an encoder network to extract the contextual features at a multi-scale. The encoder network is built upon a modified DeepLabv3+ network that encodes both low-level and high-level features from the source input image. We utilize the ResNet-18 convolutional neural network as the backbone of the encoder, as in [20]. As shown in Fig. 2, we extract the low-level feature from the Conv2 layer (64-channel feature map) and continue the backbone network flow until Conv5 layer. Then we feed the last feature map (512-channel feature map) to an **Atrous Spatial Pyramid Pooling (ASPP)** module. The ASPP module consists of one layer of point-wise convolution with the filter size 1×1 and 5 layers of depthwise convolution with the filter size 3×3. We employ different dilation rates of 2, 6, 12, 18, and 24 in the filters of depthwise convolutional layers which enables semantic high-level feature extraction at the multi-scale. The semantic high-level feature ($S_{i,j}$) with K numbers of Conv filters and $n-th$ dilated Conv layer at pixel position (i,j) can be defined as in (2).

$$S_{i,j} = \sum_{n=1}^{N} I_{i,j,nK} \times h_{i,j,nk}. \quad (2)$$

At the last layer, we use a fully connected (dense) layer, followed by a linear activation function to predict the focal length of the source image. In the next step, we combine this predicted focal length information with the semantic features for the depth estimation.

B. Depth Estimation with Decoder

The proposed decoder network generates the depth feature map using the extracted features from the encoder network. The depth map indicates the relative distances of objects in the image from the camera. After the prediction of the camera focal length, we concatenate the captured semantic high-level features with the corresponding focal length and feed this to the decoder network. The decoder refines the features, taking into account the focal length, to produce a depth map. We add 3×3 Conv. layer to refine the features. Finally, we add bilinear upsampling by factor 4 to the refined features, to make it compatible with the original image resolution. After that, we add a sigmoid activation function to estimate the depth feature map output of the corresponding input image.

C. Combined Feature Extraction

We combine the RGB input image and its corresponding depth feature map at this step. Similar to the previous encoder network architecture, we use another encoder network to extract combined features. This approach helps the network to learn semantic dense depth information of the input image. The combined feature extraction can be represented as in (3).

$$\beta_{combined} = Encoder_{combined}(I_{RGB} + S_{depth}). \quad (3)$$

Where, I_{RGB} and S_{depth} refer to RGB input image and estimated depth feature map generated using semantic high-level feature at multi-scale, respectively.

D. View Transformation Module

At this step, we transform the extracted combined features into a new viewpoint of the object based on the perspective transformation approach. The projection of the 3D point to the image plane (2D coordinates) can be defined as in (4).

$$p_{image} = T_l \rightarrow K \cdot [R|t] \cdot P_{world}. \quad (4)$$

where p_{image} and P_{world} are the homogeneous coordinates in the image plane and real-world coordinates, respectively. $[R|t]$ is combined rotation matrix R and a translation vector t. T_l represents the transformed features that belongs to p_{image}. The perspective projection matrix K includes the intrinsic parameters of the camera focal length F.

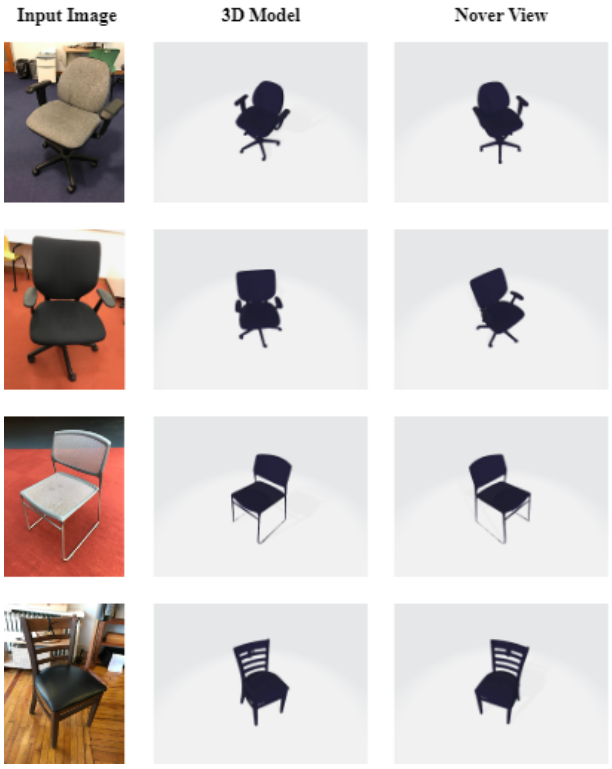

Fig. 4. Shows the RGB input image and novel views from Pix3D [22] dataset for object category "chair".

E. Generating Novel View

We use another decoder network that refines the transformed viewpoint features at this step. In this decoder network, we add two layers of Conv filters of size 3×3, followed by bilinear upsampling by factor 4 to generate the novel view of the same resolution as the input image. The synthesized image with a novel viewpoint using the transformed features can be represented as in (5).

$$\eta_{novel} = Decoder(T_l). \quad (5)$$

We discussed the implementation of the proposed methodology and evaluated results in the section IV.

IV. Experimental Results

Our proposed method is implemented in Python software on a personal computer with PyTorch library, NVIDIA Titan 1650i GPU of 8 GB memory, an Intel(R) Core(TM) i7-10750H CPU running at 2.60 GHz, a 64-bit operating system, and a 500 GB of solid-state drive (SSD). We use the Adam optimizer with momentum 0.9, LearningRate 0.001, and L2Regularization 0.005. We train and evaluate our proposed method on the benchmark ShapeNet [21] and Pix3D [22] datasets on the object category "chair". We divide the datasets for training, validation, and testing sets in a 70:20:10 ratio.

We use three loss functions to prevent overfitting: MSE for focal length loss, L1 loss for depth estimation, and SSIM loss

for novel view synthesis. Some examples of the RGB input image and generated novel views are represented in Fig. 4. The results show that our embedded camera focal length prediction model, and the combined feature extraction of RGB input and depth feature map technique performed well for novel view synthesis. However, our system encountered comparatively higher run time than the existing methods. This problem can be reduced by utilizing faster convolutional operations in the proposed network.

V. Conclusions

We propose a deep neural network model based on an encoder-decoder architecture, which includes an embedded camera focal length prediction component. This model can efficiently reconstruct new viewpoints from a single input image with unknown focal lengths. Our experimental results demonstrate good performance in depth feature map estimation by utilizing high-level semantic features at multiple scales. In the future, we plan to evaluate the proposed methodology on different object categories from benchmark datasets.

Acknowledgment

The authors highly appreciate the anonymous reviewers from APSIPA ASC 2024 for their valuable reviews and helpful comments. This research was supported by King Mongkut's University of Technology Thonburi's postdoctoral fellowship in the 2024 (2567) fiscal year under KIRIM project ID 27928 and agreement number 06/2567.

References

[1] X. Li, J. Liu, J. Baron, K. Luu, and E. Patterson, "Evaluating effects of focal length and viewing angle in a comparison of recent face landmark and alignment methods," *J. Image Video Process.*, vol. 2021, no. 9, Mar. 2021, doi:10.1186/s13640-021-00549-3.

[2] E. Bermejo, E. F.-Blanco, A. Valsecchi, P. Mesejo, O. Ibáñez, and K. Imaizumi, "FacialSCDnet: A deep learning approach for the estimation of subject-to-camera distance in facial photographs," *Expert Syst. Appl.*, vol. 210, Dec. 2022, doi:10.1016/j.eswa.2022.118457.

[3] I. Ahn and C. Kim, "A Novel Depth-Based Virtual View Synthesis Method for Free Viewpoint Video," *IEEE Trans. Broadcast.*, vol. 59, no. 4, pp. 614-626, Dec. 2013, doi: 10.1109/TBC.2013.2281658.

[4] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry, and J. Sivic, "Focal Length and Object Pose Estimation via Render and Compare," *in Proces. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, 2022, pp. 3815-3824, doi: 10.1109/CVPR52688.2022.00380.

[5] J. Thatte and B. Girod, "A Statistical Model for Disocclusions in Depth-based Novel View Synthesis," *in Proc. IEEE Visual Communications and Image Processing*, Sydney, Australia, 2019, pp. 1-4, doi: 10.1109/VCIP47243.2019.8966071.

[6] J. Kim and Y. M. Kim, "Novel View Synthesis With Skip Connections," *in Proc. International Conference on Image Processing*, Abu Dhabi, United Arab Emirates, 2020, pp. 1616-1620, doi: 10.1109/ICIP40778.2020.9191076.

[7] A. Palazzi, L. Bergamini, S. Calderara, and R. Cucchiara, "Warp and Learn: Novel Views Generation for Vehicles and Other Objects," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2216-2227, Apr. 2022, doi: 10.1109/TPAMI.2020.3030701.

[8] Y. -J. Yuan, Y. -K. Lai, Y. -H. Huang, L. Kobbelt, and L. Gao, "Neural Radiance Fields From Sparse RGB-D Images for High-Quality View Synthesis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8713-8728, Jul. 2023, doi: 10.1109/TPAMI.2022.3232502.

[9] C. Zheng, H. Qiu, C. Liu, X. Zheng, C. Zhou, Z. Liu, and J. Yang, "A Fast Method to Extract Focal Length of Camera Based on Parallel Particle Swarm Optimization," *in Proc. Chinese Control Conference*, Wuhan, China, 2018, pp. 9550-9555, doi: 10.23919/ChiCC.2018.8483981.

[10] C. T. Nguyen, R. D. Khlynov, A. A. Gorbachev, V. A. Ryzhova, S. N. Yarishev, I. A. Konyakhin, T. S. Djamiykov, M. B. Marinov, "Determining the Focal Length of a Video Camera Using a Calibration Pattern," *International Scientific Conference Electronics*, Sozopol, Bulgaria, 2022, pp. 1-4, doi: 10.1109/ET55967.2022.9920285.

[11] H. Xiao, K. Li, Y. Zhu, and J. Zhang, "3D Object Detection Based on Long and Short Focal Length Cameras," *International Conference on Computer Vision, Image and Deep Learning*, Zhuhai, China, 2024, pp. 545-548, doi: 10.1109/CVIDL62147.2024.10604155.

[12] A. P.-Sanchez, J. A.-Cetto, and F. M.-Noguer, "Exhaustive Linearization for Robust Camera Pose and Focal Length Estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2387-2400, Oct. 2013, doi: 10.1109/TPAMI.

[13] B. Zhou, Z. Chen, and Q. Liu, "An Efficient Solution to the Perspective-n-Point Problem for Camera With Unknown Focal Length," *IEEE Access*, vol. 8, pp. 162838-162846, 2020, doi: 10.1109/ACCESS.2020.3021313.

[14] X. Yin, L. Ma, X. Tan, and D. Qin, "A Robust Visual Localization Method With Unknown Focal Length Camera," *IEEE Access*, vol. 9, pp. 42896-42906, 2021, doi: 10.1109/ACCESS.2021.3065953.

[15] L. He, G. Wang, and Z. Hu, "Learning Depth From Single Images With Deep Neural Network Embedding Focal Length," *IEEE Trans. on Image Process.*, vol. 27, no. 9, pp. 4676-4689, Sept. 2018, doi: 10.1109/TIP.2018.2832296.

[16] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry, and J. Sivic, "Focal Length and Object Pose Estimation via Render and Compare", 2022, arXiv:2204.05145.

[17] Z. Liu, W. Jia, M. Yang, P. Luo, Y. Guo, and M. Tan, "Deep View Synthesis via Self-Consistent Generative Network," *IEEE Transactions on Multimedia*, vol. 24, pp. 451-465, 2022, doi: 10.1109/TMM.2021.3053401.

[18] J. Lei, B. Liu, B. Peng, X. Cao, Q. Huang, and N. Ling, "Deep Gradual-Conversion and Cycle Network for Single-View Synthesis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 6, pp. 1665-1675, Dec. 2023, doi: 10.1109/TETCI.2023.3272003.

[19] L. Jiang, G. Schaefer, and Q. Meng, "An Improved Novel View Synthesis Approach Based on Feature Fusion and Channel Attention," *in Process. IEEE International Conference on Systems, Man, and Cybernetics*, Prague, Czech Republic, 2022, pp. 2459-2464, doi: 10.1109/SMC53654.2022.9945244.

[20] D. Chakraborty, W. Chiracharit, and K. Chamnongthai, "Semantic Scene Object-Camera Motion Recognition for Scene Transition Detection Using Dense Spatial Frame Segments and Temporal Trajectory Analysis," *IEEE Access*, vol. 12, pp. 21673-21697, 2024, doi: 10.1109/ACCESS.2024.3363233.

[21] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su et al., "ShapeNet: An information-rich 3D model repository," arXiv preprint arXiv:1512.03012, 2015.

[22] X.Sun, J.Wu, X.Zhang, Z.Zhang, C.Zhang, T.Xue,J.B.Tenenbaum, and W. T. Freeman, "Pix3D: Dataset and methods for single-image 3D shape modeling," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 29742983.