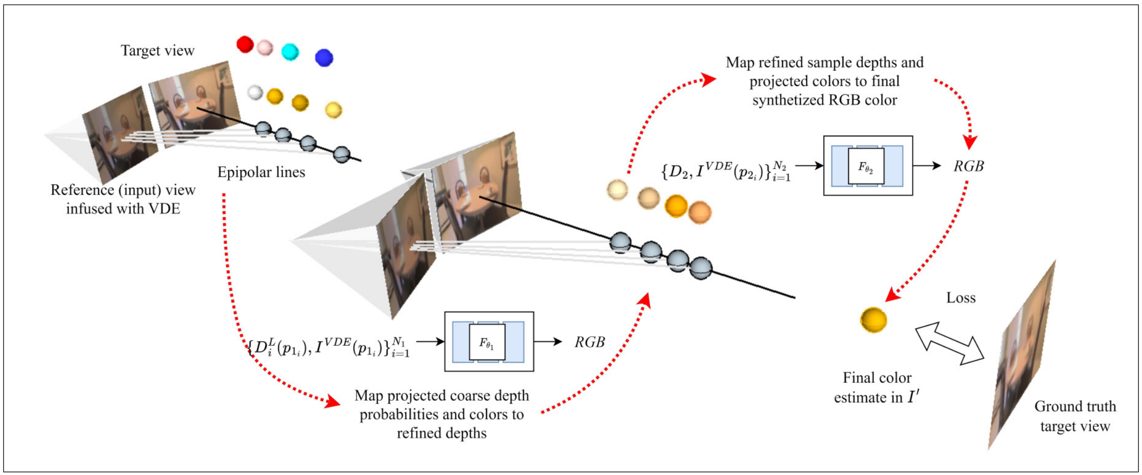
카메라 움직임 및 로컬 컨텍스트 사전 정보에 기초한 시점 의존적 효과를 고려한 자유 시점 이미지 합성에 관한 연구

A Study on Free-view Image Synthesis with View-Dependent Effects based on Camera Motion and Local Context Priors

Juan Luis Gonzalez Bello / KAIST 비디오 및 이미지 컴퓨팅 연구실

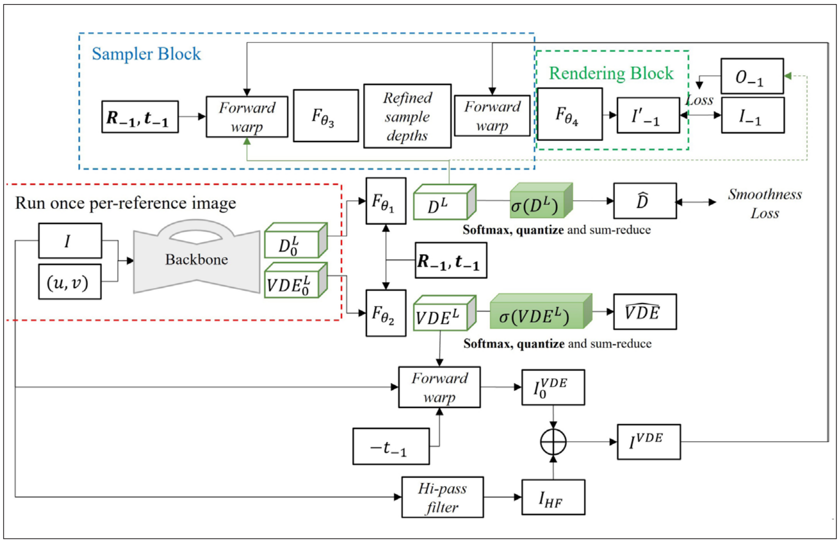
Ⅰ. Introduction

Recent advances in neural scene rendering have shed light on the usefulness of rendering view-dependent effects in the novel view synthesis task. In particular, NeRFs [1] have demonstrated that the radiance fields can be effectively learned in neural feature space via a multi-layer perceptron (MLP) that allows rendering geometry and view-dependent effects for scenes that are carefully captured from multiple views. However, NeRF purely relies on multi-view consistency and cannot exploit prior


Target view
Reference (input) view infused with VDE
Epipolar lines
$\{D_i^L(p_{1_i}), I^{VDE}(p_{1_i})\}_{i=1}^{N_1}$
F_{θ_1}
RGB
Map projected coarse depth probabilities and colors to refined depths
Map refined sample depths and projected colors to final synthetized RGB color
$\{D_2, I^{VDE}(p_{2_i})\}_{i=1}^{N_2}$
F_{θ_2}
RGB
Loss
Final color estimate in I'
Ground truth target view

<Figure 1> Conceptual illustration of the projection operations in our FVSVDE-Net pipeline. Our sampler block F_{θ_3} estimates refined sample depths from the coarsely projected colors and probabilities. New colors are projected into the refined sample depths and fed into the rendering head F_{θ_4}, which interpolates or generates the final color in the novel views.

knowledge, such as textures and depth cues that are common across natural scenes, limiting its use when a few or only one view is available. On the other hand, to leverage the 3D prior knowledge in multi-view datasets, PixelNeRF [2] proposed to train a multi-layer perception (MLP) which takes as input pixel locations and pixel-aligned features to generate the colors and opacities of the 3D points in the radiance fields. The pixel-aligned features obtained from a CNN backbone allow PixelNeRF to leverage the common priors among different scenes to render radiance fields but cause considerably limited quality. Other works, such as single-image MPIs [3] and MINE [4], have also proposed single-view based free-view synthesis but cannot model view-dependent effects (VDE). View-dependent effects depend on the material's reflectance, which is a function of the material properties and the angle of incidence of the light. Learning such material properties and light sources from a single image is a very ill-posed problem. Previous works, such as NeRFs of PixelNeRFs, learn to directly regress the colors of pixels given the viewing directions, while other methods, such as NeX [5], encode view-dependent effects into a given or learned basis. While these techniques are effective when learning from multiple input images, they are still limited to learning when a single image is given as input. Instead, to tackle the estimation of view-dependent effects in novel view synthesis, we propose to rely on the contents of the images and estimated (during training) or user-defined (during test time) camera motions to estimate photo-metrically realistic



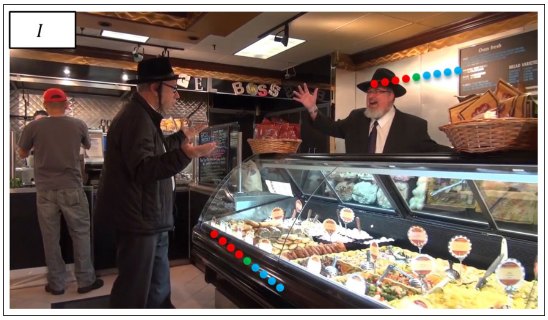
<Figure 2> FVSVDE-Net model for learning self-supervised single view-based free view synthesis with VDEs. The autoencoder network is run only once for any novel viewpoint. For each novel viewpoint, inexpensive adjustment (F_{θ_1}, F_{θ_2}), sampler (F_{θ_3}), and rendering (F_{θ_4}) 1×1 convolution blocks are run only one time.

view-dependent effects from a single image for the first time. In addition, in this study, we propose a new geometric rendering pipeline inspired by neural volumetric rendering (NVR) by approximating NVR with a single pass of a convolutional (or transformer-based) auto-encoder network, a sampler MLP block, and a rendering MLP block. In addition, we train our networks in a self-supervised manner, that is, under the conditions that no camera poses or depth GTs are given during training (as in previous works). We present extensive experiments and show that our proposed method can learn free view synthesis with view-dependent effects on the challenging RealEstate10k [6] and MannequinChallenge [7] datasets.

Ⅱ. Method

1. Rendering VDE from a single image

In natural image sequences, view-dependent effects (VDE), such as reflections, have almost 'no disparities.' That is, they seem to stay in similar image regions in both the current I and the next frame I_{+1}. In other words, VDEs 'follow' the camera motions relative to their reflective surfaces. To this extent, we propose to exploit this strong, yet simple, prior in VDEs, by generating a view-dependent appearance by re-sampling pixels in the opposite direction of the camera motion, as shown in 〈Figure 3〉. This operation is equivalent to an adaptive line-shaped adaptive kernel, whose orientation is given by the desired camera motion and the pixel location.



<Figure 3> The view-dependent effects of a novel view can be rendered at the target pixels (green dots) by re-sampling pixels in the image in the opposite direction to the camera motion (blue dots). On the other hand, the geometries of a novel view are rendered by re-sampling the resulting VDE image in the direction of the camera motion (red dots).

Given the relative camera motion in 〈Figure 3〉, which is for a camera pose with a positive Z-axis, the green dots in 〈Figure 3〉 depict the pixel to be estimated in the novel view, while the blue dots denote the sampling locations for generating view-dependent effects. The correct weights on the blue dots can effectively re-locate the specular regions in the scene to match those of the next frame I_{+1}. We also observe that relative VDE motion cannot be larger than the corresponding rigid flow generated by the scene depth and camera motion. That is, the disparity in view-dependent effects is inversely proportional to the scene depths.

Note that properly re-sampling image pixels in the same direction as the camera motion (from an epipolar line point of view), as denoted by the red dots in 〈Figure 3〉, can generate the novel view seen from the target camera pose by 'pulling' the structural details into the target view, in a similar spirit to the proposed methods in the prior works of [8,9,10]. In this work, we use re-sampling, but guided by our approximation of volumetric rendering, in both the original and the opposite camera motion directions to generate a novel view with VDEs seen from the target camera pose. In a similar manner as disparity logits D^L are utilized to approximate volumetric rendering in 〈Figure 2〉, VDE^L is used to perform forward warping, but in the opposite epipolar line direction, generating the view-dependent effects due to reflective surfaces. VDE^L re-samples the pixels in the input image I in the opposite direction of the camera translation to generate I^{VDE} in two stages. First I_0^{VDE} is obtained by progressively warping I and VDE^L with the 3D warping operation $g(\cdot)$, as given by

$$I_0^{VDE} = \sum_{n=0}^{N-1} g\left(I, vde_n, 0, -t_c, K\right) \odot g\left(VDE^{L_n}, vde_n, 0, -t_c, K\right) \quad (1)$$

where $-t_{\mathrm{c}}$ is the relative camera translation in the opposite direction, which causes the epipolar lines to sample pixels in the opposite directions. vde_n describes the hypothetical depth values that are used to progressively warp I and VDE^L, and is given by

$$vde_n = \frac{n}{N_{vde}}\left(\hat{D} - vde_{\min}\right) - vde_{\min}\,. \quad (2)$$

Note that vde_n is a function of the inverse depth (or disparity) $\hat{D}$ which is obtained from the estimated disparity logits D^L. The second stage to obtain I^{VDE} is to combine I_0^{VDE} with the high-frequency component of I, to enforce the same structural details in the view-dependent image I^{VDE} and I. I^{VDE} is the given by

$$I^{VDE} = I_0^{VDE} + \left(I - I * k_{5\times5}\right), \quad (3)$$

where $k_{5\times5}$ is a 5×5 box kernel that low-pass filters I. $I{*}k_{5\times5}$ is removed from I, roughly yielding the high frequency information of I. Finally, I^{VDE}, which contains the view-dependent effects induced by the target camera translation, is then forward warped (via our volumetric rendering approximation) by D^L into the target camera pose (rotation and translation). In other words, an approximated volumetric rendering Equation for free view synthesis with view-dependent effects can now be given by

$$I_c'\left(\mathbf{p}\right) = \sum_{i=1}^{N} D^{P_i}\left(\mathbf{p}\right) g\left(I^{VDE}, \mathbf{p}, 1/d_i, R_c, t_c, K\right) \quad (4)$$

where $I_c^{'}(\mathbf{p})$ is now infused with VDE. In order to learn free view synthesis with VDEs, $I_c^{'}(\mathbf{p})$ is compared against the corresponding reference views with an occlusion-aware synthesis loss function.

Similar to $\hat{D}$, a view-dependent effects activation map, which is useful for visualizing the most reflective regions in an image, can then be obtained by:

$$V\hat{D}E = \sum_{n=0}^{N-1} vde_n \sigma\left(VDE^L\right)_n \quad (5)$$

2. Network architecture / Rendering Pipeline

The proposed FVSVDE-Net in 〈Figure 2〉 only requires the backbone to be *run once per-reference image*. Once the 'primitive' disparity and VDE logits, denoted by D_0^L and VDE_0^L respectively, are estimated, the novel views are generated by running the relatively computationally inexpensive adjustment

(F_{θ_1}, F_{θ_2}), sampler (F_{θ_3}), and rendering (F_{θ_4}) 1×1 convolution blocks. Contrary to previous works that also incorporate MLP heads for rendering, we only need to run the MLP heads for a single pass, instead of running them once for each point in the target rays as in the previous works of [2,11].

3. The Adjustment Blocks

Due to the nature of our approximated volumetric rendering, the primitive disparity and VDE logits, D_0^L, and VDE_0^L, need to be adjusted for novel camera viewpoints by F_{θ_1} and F_{θ_2} as depicted in the center region of 〈Figure 2〉. This is because novel camera viewpoints represent different relative geometry distributions between the source (or input) view and the target view. For example, a relative camera motion of 5m (meters) in the Z-axis will mean that the first channel of D^L will no longer represent the minimum disparity (or maximum depth) relative to the input view, but instead, +5m into the Z-axis, as D^{L_0} will be projected by $d_o = d_{\min}$ relative to the target view, not the input reference view. These adjustments are carried out by the adjustment blocks F_{θ_1} and F_{θ_2}, which learn an encoding of camera poses and the re-arrangement of the probability logits volumes for D^L and VDE^L, respectively.

1) The Sampler Block for our FVSVDE-Net

The FVSVDE-Net's sampler block F_{θ_3} takes as input the projected probability logits and view-dependent image colors I^{VDE}, but maps them into N_2 refined_per-pixel sampling inverse depths $d_j(\mathbf{p})$ along with N_2 softmaxed weights $w_j(\mathbf{p})$ for approximated volumetric rendering in the subsequent rendering block.

2) The Rendering Block for our FVSVDE-Net

The FVSVDE-Net's rendering block consists of the approximated (but fine-grained) volumetric rendering of the estimated refined samples and weights. The final synthetic image color I' ($\mathbf{p}$) at pixel location $\mathbf{p}$ = (u,v) is then given by

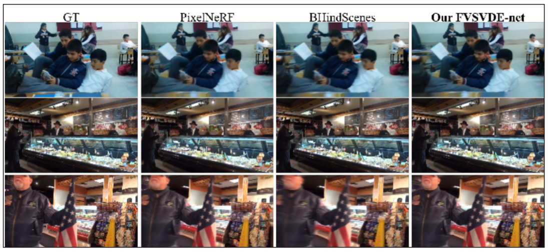
$$I'(\mathbf{p}) = \sum_{j=1}^{N_2} w_j(\mathbf{p}) g\left(I^{VDE}, \mathbf{p}, 1/d_i(\mathbf{p}), R_c, t_c, K\right) \quad (6)$$

4. Extending FVSVDE-Net for multiple inputs

Our FVSVDE-Net can be extended to multi-view inputs by accumulating the ray colors and probabilities from different source views in the sampler block.

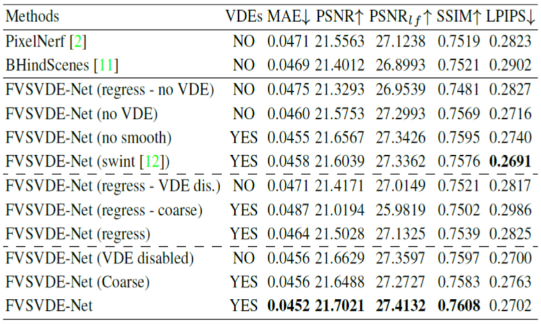
Ⅲ. Experiments and Results

After training our FVSVDE-Net for 50 epochs on the training splits of the RealEstate10k [6] and MannequinChallenge [7] datasets, our model is capable of rendering realistic novel views with VDEs with higher quality metrics and fewer artifacts than the previous single-view based methods of PixelNerF [2] and BHindScenes [11] as shown in 〈Figure 4〉 and 〈Tables 1,2〉.



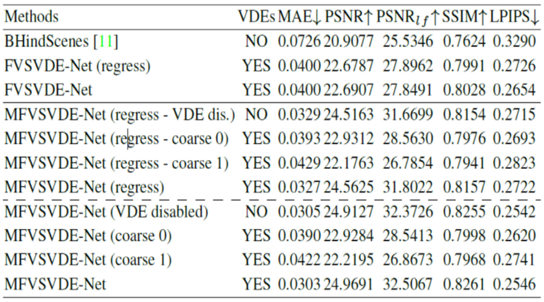
<Figure 4> Comparison among various single view-based free view synthesis methods and ours on the MannequinChallenge dataset [7].

<Table 1> Single view synthesis results on the Mannequin-Challenge dataset [7]. ↓ denotes the lower the better the metric. ↑ denotes the higher the better the metric.



Methods	VDEs	MAE↓	PSNR↑	PSNR$_{lf}$↑	SSIM↑	LPIPS↓
PixelNerf [2]	NO	0.0471	21.5563	27.1238	0.7519	0.2823
BHindScenes [11]	NO	0.0469	21.4012	26.8993	0.7521	0.2902
FVSVDE-Net (regress - no VDE)	NO	0.0475	21.3293	26.9539	0.7481	0.2827
FVSVDE-Net (no VDE)	NO	0.0460	21.5753	27.2993	0.7569	0.2716
FVSVDE-Net (no smooth)	YES	0.0455	21.6567	27.3426	0.7595	0.2740
FVSVDE-Net (swint [12])	YES	0.0458	21.6039	27.3362	0.7576	**0.2691**
FVSVDE-Net (regress - VDE dis.)	NO	0.0471	21.4171	27.0149	0.7521	0.2817
FVSVDE-Net (regress - coarse)	YES	0.0487	21.0194	25.9819	0.7502	0.2986
FVSVDE-Net (regress)	YES	0.0464	21.5028	27.1325	0.7539	0.2825
FVSVDE-Net (VDE disabled)	NO	0.0456	21.6629	27.3597	0.7597	0.2700
FVSVDE-Net (Coarse)	YES	0.0456	21.6488	27.2727	0.7583	0.2763
FVSVDE-Net	YES	**0.0452**	**21.7021**	**27.4132**	**0.7608**	0.2702

<Table 2> Multi view-based VS Single view-based free synthesis results on the RealEstate10k dataset [6].



Methods	VDEs	MAE↓	PSNR↑	PSNR$_{lf}$↑	SSIM↑	LPIPS↓
BHindScenes [11]	NO	0.0726	20.9077	25.5346	0.7624	0.3290
FVSVDE-Net (regress)	YES	0.0400	22.6787	27.8962	0.7991	0.2726
FVSVDE-Net	YES	0.0400	22.6907	27.8491	0.8028	0.2654
MFVSVDE-Net (regress - VDE dis.)	NO	0.0329	24.5163	31.6699	0.8154	0.2715
MFVSVDE-Net (regress - coarse 0)	YES	0.0393	22.9312	28.5630	0.7976	0.2693
MFVSVDE-Net (regress - coarse 1)	YES	0.0429	22.1763	26.7854	0.7941	0.2823
MFVSVDE-Net (regress)	YES	0.0327	24.5625	31.8022	0.8157	0.2722
MFVSVDE-Net (VDE disabled)	NO	0.0305	24.9127	32.3726	0.8255	0.2542
MFVSVDE-Net (coarse 0)	YES	0.0390	22.9284	28.5413	0.7998	0.2620
MFVSVDE-Net (coarse 1)	YES	0.0422	22.2195	26.8673	0.7968	0.2741
MFVSVDE-Net	YES	0.0303	24.9691	32.5067	0.8261	0.2546

Ⅳ. Conclusions

We presented a novel method that can learn to perform free view synthesis (FVS) and view-dependent effects (VDE) estimation on single images in a self-supervised manner from videos. We showed that our method generalizes well on unseen validation images and that it can generate plausible View Dependent Effects and Depth Maps from a single image. Our model, which we refer to as FVSVDE-Net, exploits the local scene contents and camera motion priors and is the first single-view-based method to explicitly model view-dependent effects. In addition, our FVSVDE-Net incorporates an approximation to volumetric rendering which we further improve by incorporating sampler and rendering modules for fine-grained ray sampling and rendering. Our FVSVDE-Net yields the most realistic synthetic images with view-dependent effects in comparison to the recent previous methods of PixelNeRF and BHindScenes.

We also showed that our rendering pipeline is flexible to accommodate multiple observations of the scene, AKA multiple views. With multiple input views, our fine-grained volumetric rendering pipeline is able to handle disocclusions and better estimate VDEs. However, we observed that the performance of multiple-view-based free view synthesis greatly depends on the pose estimation accuracy, as small errors in pose estimates can generate heavy 'double edge' artifacts.

참고문헌

[1] B. Mildenhall, et al., "Nerf: Representing scenes as neural radiance fields for view synthesis," European Conf. Computer Vision (ECCV), Aug. 2020.

[2] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," Conference on Computer Vision and Pattern Recognition (CVPR), June 2021.

[3] R. Tucker and N. Snavely, "Single-view view synthesis with multiplane images," CVPR, June 2020.

[4] J. Li, et al., "MINE: towards continuous depth MPI with nerf for novel view synthesis," Int. Conf. Computer Vision (ICCV), Oct 2021.

[5] S. Wizadwongsa, et al., "Nex: Real-time view synthesis with neural basis expansion," CVPR June 2021.

[6] T. Zhou, et al., "Stereo magnification: Learning view synthesis using multiplane images," SIGGRAPH, 2018.

[7] Z. Li, et al., "Learning the depths of moving people by watching frozen people," CVPR, June 2019.

[8] J. L. Gonzalez Bello and M. Kim, "Deep 3d pan via local adaptive" t-shaped" convolutions with global and local adaptive dilations," Int. Conf. Learning Representations (ICLR), April 2020.

[9] J. L. Gonzalez Bello and M. Kim, "Forget about the lidar: Self-supervised depth estimators with med probability volumes," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 12626-12637, 2020.

[10] J. L. Gonzalez and M. Kim, "Plade-net: Towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss," CVPR, June 2021.

[11] F. Wimbauer, N. Yang, C. Rupprecht, and D. Crsemers, "Behind the scenes: Density fields for single view reconstruction," arXiv preprint arXiv:2301.07668, 2023.

[12] Z. Liu, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," CVPR, June 2021.

Juan Luis Gonzalez Bello

Juan Luis Gonzalez Bello received a B.S. degree in Electro-Mechanical Engineering, specializing in digital systems, from the Mexican Autonomous National University (UNAM) in 2013. He worked in Procter and Gamble for three years in the engineering and regional engineering departments, managing projects and assisting in designing new advanced manufacturing machinery. Since 2017, he has joined the School of Electrical Engineering at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, where he is currently pursuing a Ph.D. degree. His research interests include novel-view synthesis, monocular/stereo depth estimation, and deep learning, with related papers published in top conference venues such as TPAMI, CVPR2021, NeurIPS2020, ICLR2020, and ICIP2019-2020 as the first author.