# EpipolarNVS: leveraging on Epipolar geometry for single-image Novel View Synthesis

*Gaétan Landreau[1,2] and Mohamed Tamaazousti[2]*

**Meero[1], Paris, France - Université Paris-Saclay, CEA-LIST[2], F-91190 Palaiseau, France**

gaetan.landreau@meero.com - mohamed.tamaazousti@cea.fr

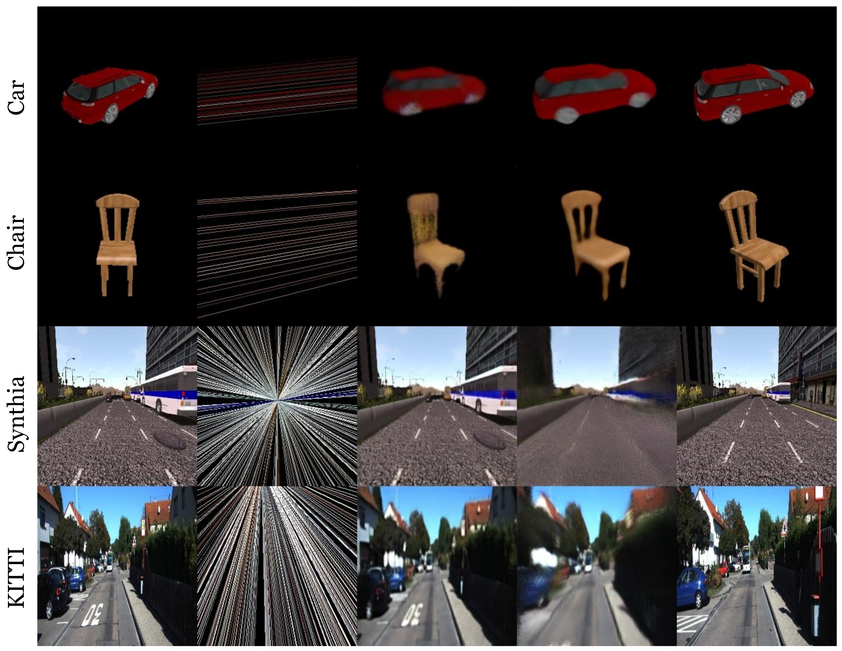
BMVC 2022

## Motivation

Single-image Novel View Synthesis (NVS) inherently requires feeding at least **extrinsic camera matrices** [R|t] to the neural network. We claim that providing such information as is to express the geometrical relationship between the source and target image is suboptimal.

We leverage on **epipolar geometry** to interpret [R|t] as an **RGB image**. Such a design gives the network some prior information regarding where source image pixels should reproject on the target image to predict and also implicitly gets considerations for the intrinsic matrix K.

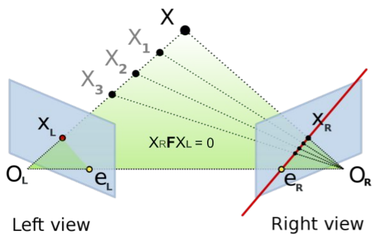
## Qualitative Results



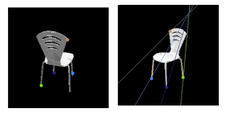

On **synthetic images**, our method allows to better synthesize complex structures (e.g on the chair's back) and object shape (e.g on the car). On **real-world scenes**, the overall motion is better retrieved (e.g the location of the bus from Synthia) while the global perspective (e.g the sign "30" on the ground of KITTI) is correctly keep as-is.

## Overview: General architecture

**Epipolar geometry** indirectly encodes the viewpoint transformation that exists between two cameras.

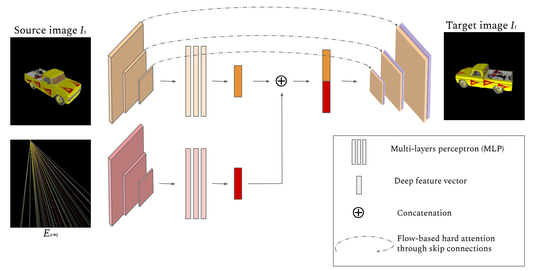



***Epipolar geometry concept:*** *A pixel-to-line correspondence is drawn between two camera through a fundamental matrix* ***F****.*



***A synthetic example:*** *Few pixels sampled on a source image with their corresponding epipolar lines.*

***A practical example:*** *Pixels are sampled according to a regular grid on the source image to produce the RGB image Es→t .*




***Overall architecture:*** *In our architecture, the camera viewpoint transformation is encoded as an RGB image, that is passed through an encoder that shares the same architecture as the one responsible for encoding the source image.*

**Extrinsic relative camera transformation** is no longer provided as-is to the network. A finite set of coloured epipolar lines are sufficient to condition the network.

## Quantitative Results

| Modality | Method | Car | | | Chair | | |
|---|---|---|---|---|---|---|---|
| | | *MAE* (↓) | SSIM (↑) | PSNR (↑) | *MAE* (↓) | SSIM (↑)) | PSNR (↑) |
| *Multi-views* | [3] | 0.078 | 0.935 | - | 0.141 | 0.911 | - |
| | [4] | 0.139 | 0.875 | - | 0.223 | 0.882 | - |
| | [6] | 0.148 | 0.877 | - | 0.229 | 0.871 | - |
| *Single-view* | [1] | 0.119 | 0.913 | - | 0.202 | 0.889 | - |
| | [5] | - | 0.900 | 23.17 | - | **0.911** | **23.72** |
| | [2] | 0.026 | 0.892 | 21.18 | 0.045 | 0.865 | 17.89 |
| | Ours | **0.016** | **0.928** | **24.23** | **0.032** | 0.901 | 19.55 |

***Performances on ShapeNet synthetic dataset:*** *We achieve state of the art results on ShapeNet-Car class and competitives results on ShapeNet-Chair.*

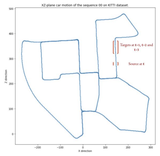
| Modality | Method | Synthia | | | KITTI | | |
|---|---|---|---|---|---|---|---|
| | | *MAE* (↓) | SSIM (↑) | PNSR (↑) | *MAE* (↓) | SSIM (↑) | PNSR (↑) |
| *Multi-views* | [3] | 0.118 | 0.737 | - | 0.163 | 0.691 | - |
| | [4] | 0.175 | 0.612 | - | 0.295 | 0.505 | - |
| *Single-view* | [6] | 0.221 | **0.636** | - | 0.418 | 0.504 | - |
| | [2] | 0.065 | 0.632 | **19.81** | 0.087 | 0.602 | 16.84 |
| | Ours | **0.065** | 0.631 | 19.44 | **0.082** | **0.609** | **17.11** |

***Performances on real-world Synthia & KITTI datasets:*** *We achieve state of the art results on KITTI and manage to get extremely competitive results on Synthia.*

Our method reaches extremely competitive results with a camera pose encoding design that is unified between all the four datasets we get consideration for.

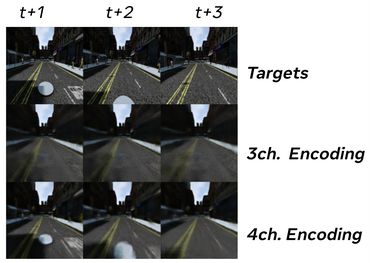
## Focus n°1 : The extended encoding

Our epipolar-based encoding can be improved on real-world datasets by **adding a fourth channel.** It contains a pseudo-depth that leverages the main displacement direction.



***Source image:*** *Source view sampled at t from a Synthia scene.*

***3 consecutives views:*** *This ablation study compare the influence of the 4th channel on the novel view that was generated given the source image.*



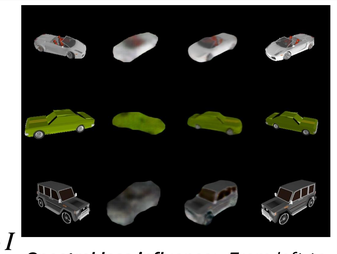

## Focus n°2 : The spectral loss

Such spectral loss function aims to **preserve high frequencies** in the generated images. Such term is added during training to the original L1 loss used in [2].

***Equations:*** *General expression of the Spectral loss function.*

$$\begin{cases} I^{lf} = I \circledast w_{gauss} \\ I_{hf} = I - I^{lf} = (\delta - w_{gauss}) \circledast I \end{cases}$$

$$\mathcal{L}_{spectral} = ||I_t^{hf} - \hat{I}_t^{hf}||_2^2$$



***Spectral loss influence:*** *From left to right: Source, w/o Spectral loss, w/ Spectral loss, Target. Details are much better retrieved when the Spectral loss is used during training.*

## References

- [1] Park Eunbyung & al. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017.
- [2] Kim Juhyeon and Young Min Kim. Novel view synthesis with skip connections. In *ICIP*, 2020.
- [3]Shao-HuaSun & al. Multi- view to novel view: Synthesizing novel views with self-learned confidence. In *ECCV*, 2018.
- [4] Maxim Tatarchenk & al. Single-view to multi-view: Reconstructing unseen views with a Convolutional Network *ArXiv*, 2015.
- [5] Alex Yu & al. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021.
- [6] TinghuiZhou & al. View synthesis by appearance flow. In *ECCV*, 2016.