Extreme View Synthesis

Presenter: Alex Matasa

Inchang Choi
Orazio Gallo
Alejandro Troccoli
Min H. Kim
Jan Kautz



Outline

- Introduction
- Related Work
- Overview of the Solution
- Synthesis of a Novel View
- Image Refinement
- Evaluation and Results
- Conclusion

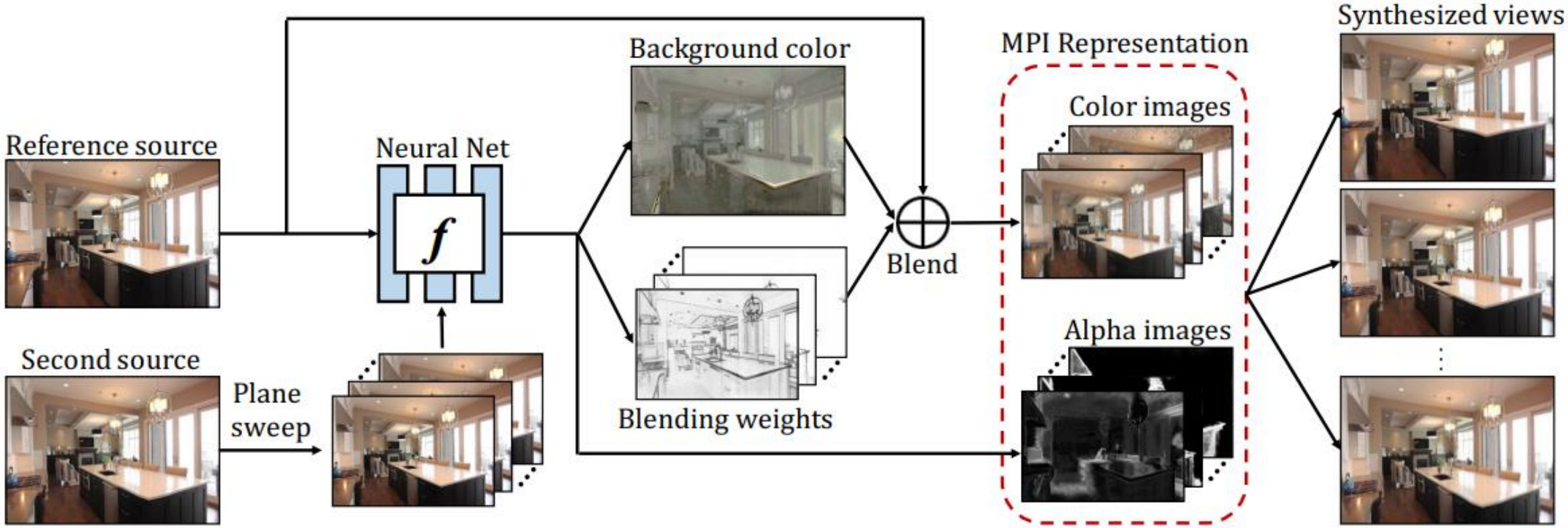


What is Novel View Synthesis?

- Capture visual content to render a new perspective
- Enables telepresence applications
- Examples
  - Virtual and mixed reality
  - Navigation of 3D environments in 2D (Google Maps)

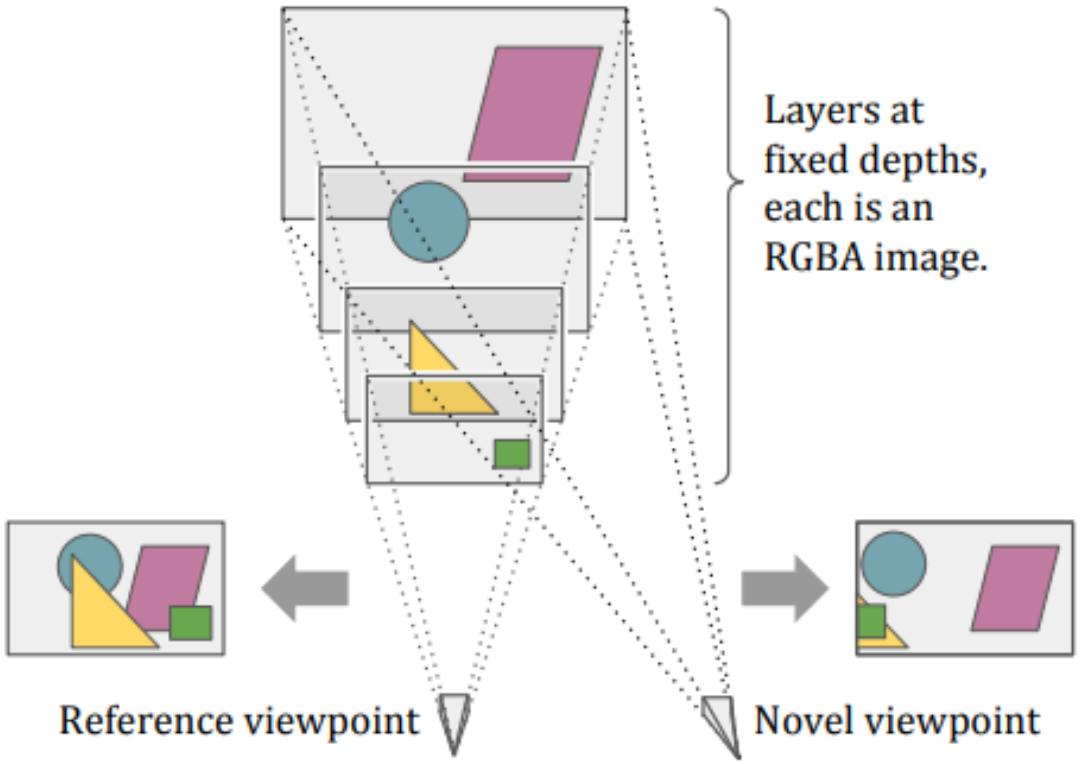


Stereo Magnification: Learning view synthesis using multiplane images, Zhou et al.



Stereo Magnification: Multiplane Image

- Each plane encodes an RGB image and alpha map
- Describes the fronto-parallel scenes at various depths
- Novel view synthesized using homography and alpha compositing



DeepMVS: Learning multi-view stereopsis, Huang et al.

- 3-part process
- Repeated for each neighboring image
- Patch matching
- Intra-volume aggregation
- Inter-volume aggregation
- Outputs a refined disparity map

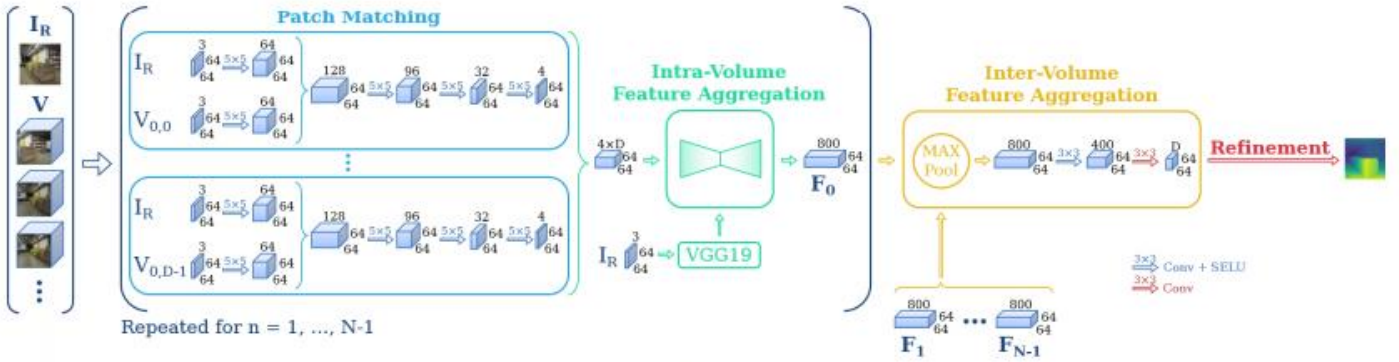
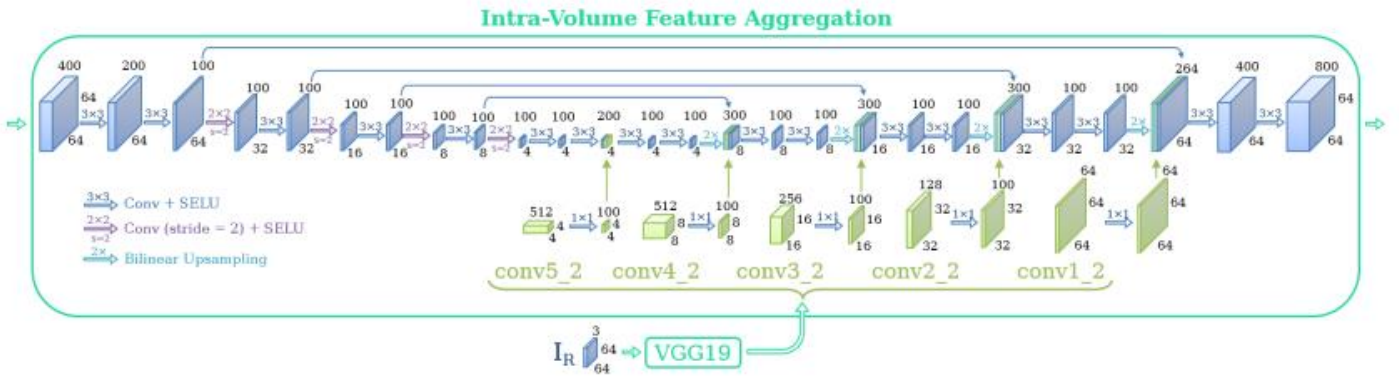
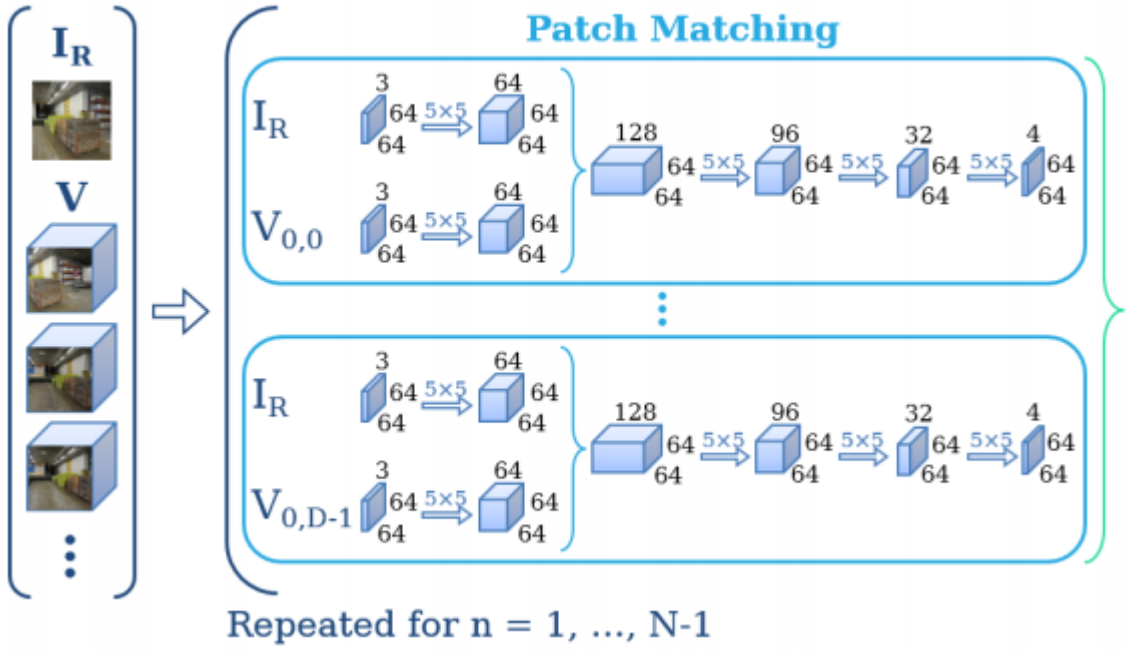


Figure 2. DeepMVS network architecture.



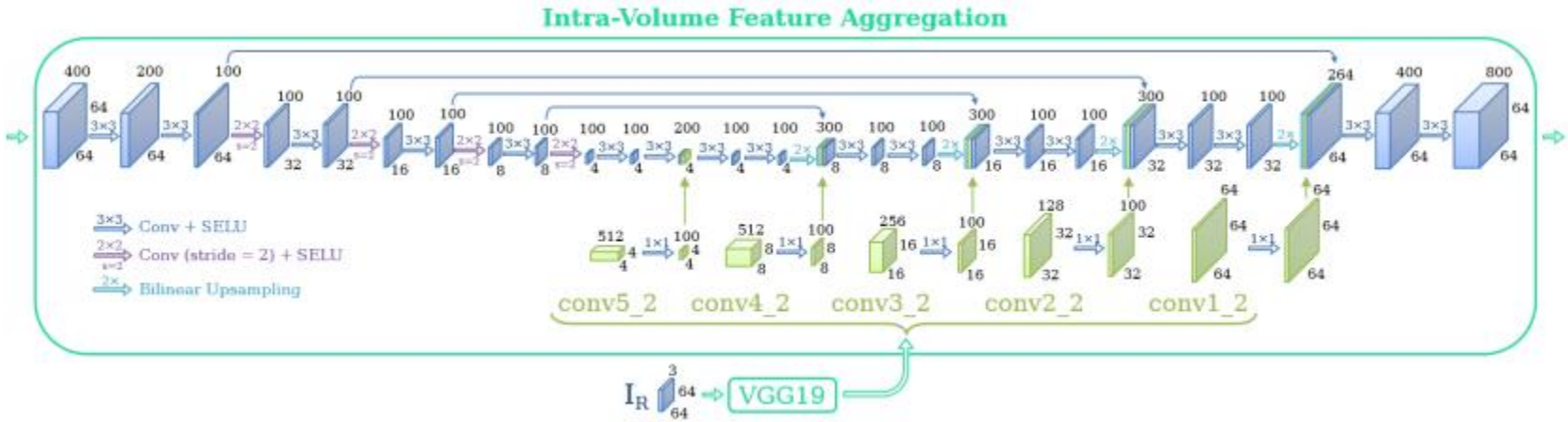
DeepMVS: Patch Matching

- Reference Image + Single patch $V_{n,d}$ from plane sweep volumes
- Extract 64-channel features from both patches
- Concatenate features then down sample creating a 4-channel patch
- Repeat for NxD plane swept images



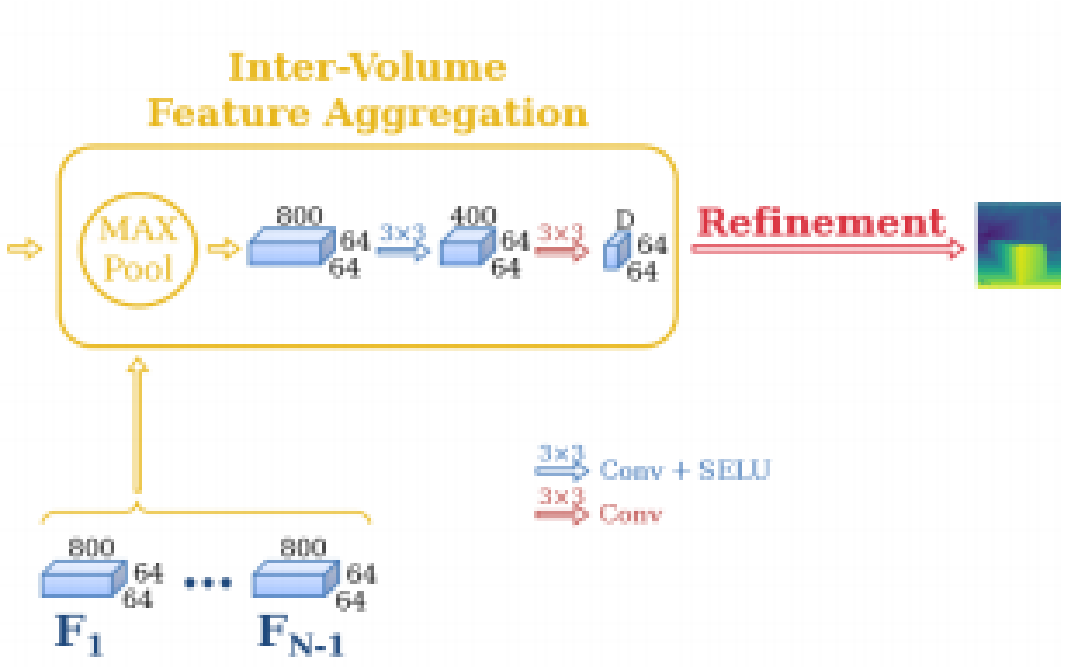
DeepMVS: Intra-Volume Feature Aggregation

- UNet architecture
- For each neighboring image, concatenate 4-channel patches for D disparity levels
- Semantic features from VGG19 are concatenated at each encoding level
- Outs an 800-channel volume containing disparity prediction info

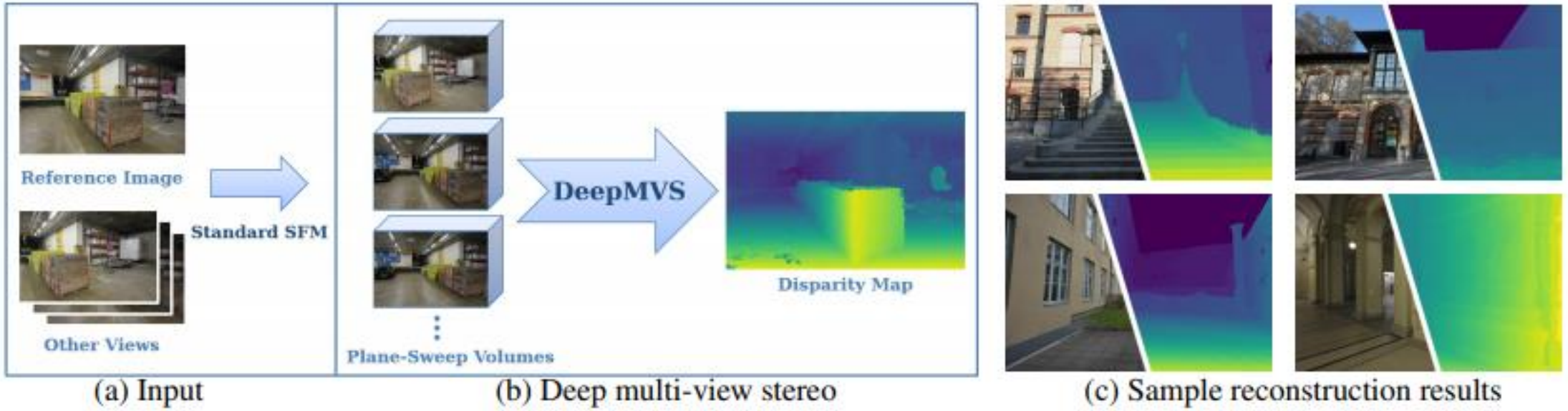


DeepMVS: Inter-Volume Feature Aggregation

- Max pool the generated volumes allowing for arbitrary amount of images
- Invariant to image order
- Down sample to D-channel pixel map



DeepMVS: Disparity Maps



(a) Input (b) Deep multi-view stereo (c) Sample reconstruction results

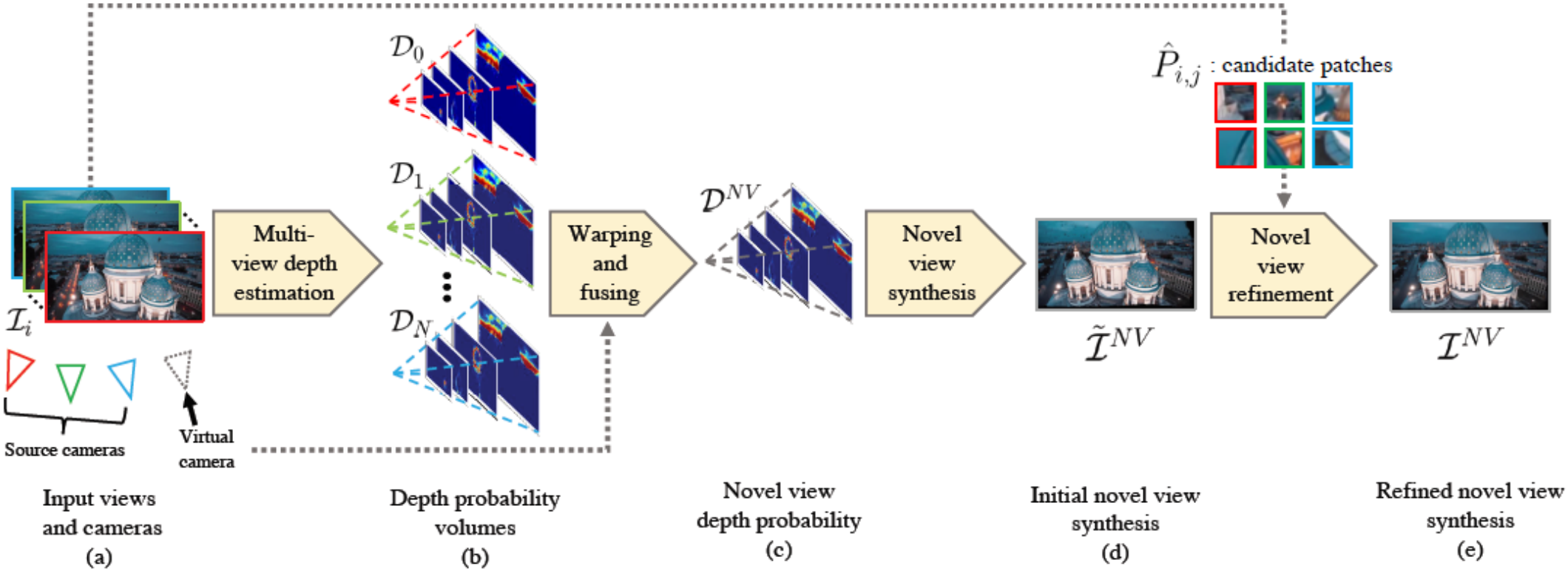
What is trying to be solved?

- Reduce data needed for synthesis
- Image synthesis of ‘Extreme Views’
  - Small number of cameras – as little as two
  - Significantly different viewpoints
- Generating views that create disocclusions
- Increasing baseline magnification

Proposed Solution

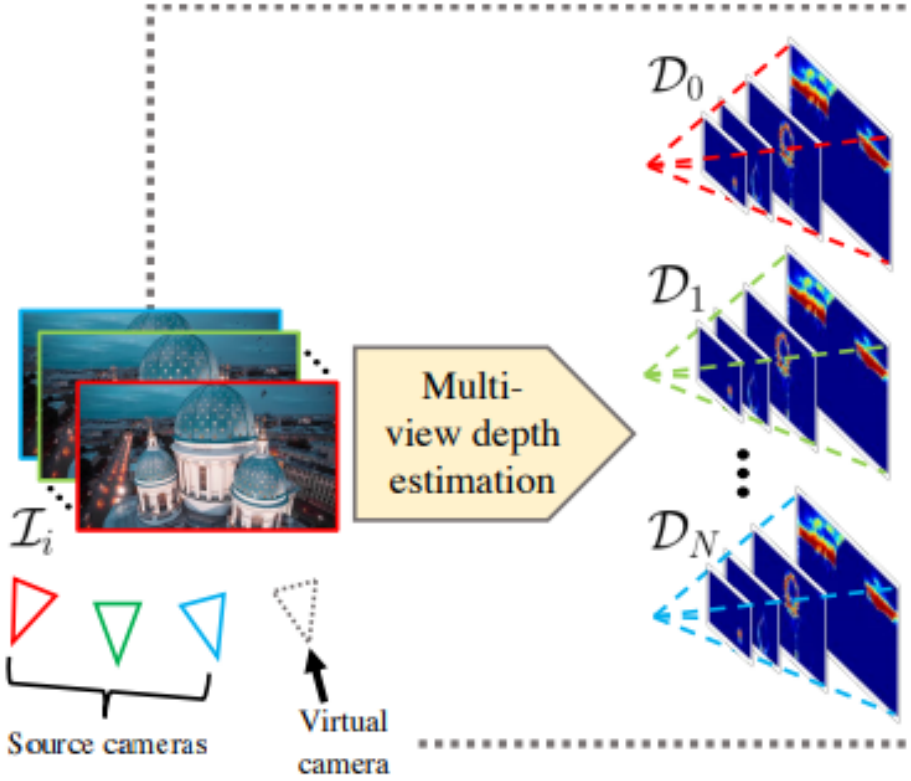
1. Estimate ‘Depth Probability Volumes’ using DeepMVS
2. Fuse then warp these volumes to form a ‘Novel View Depth Probability’
3. Create a novel view synthesis from the depth probability
4. Refine the novel view

Proposed Architecture

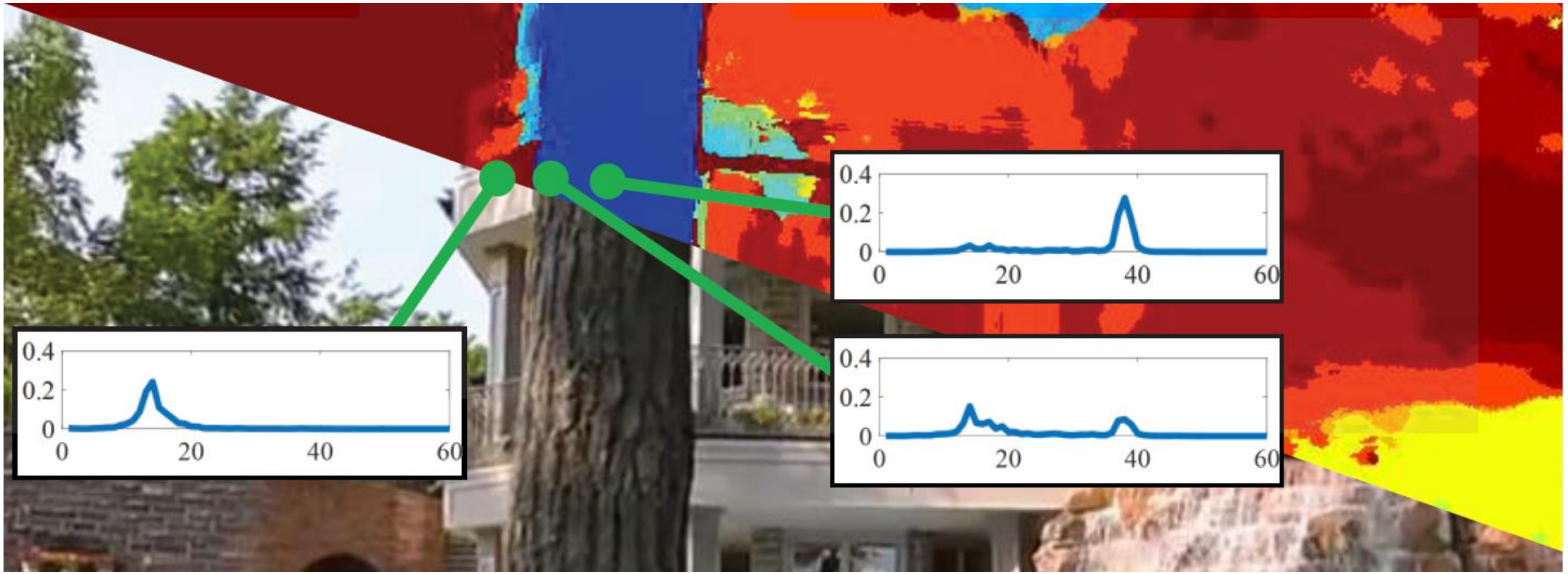


Estimating the Depth Probability Volume

- DeepMVS
- Treat depth estimation as a classification problem where:
  - n_d depth values (classes)
  - Define a probability distribution over n_d along $R_i(x_i, y_i)$
  - The collection of these ray probabilities is the 'depth probability volume' denoted as $D_i \in R^{h \times w \times n_d}$
  - Cross-entropy loss using ground truth one-hot vectors (0 or 1)

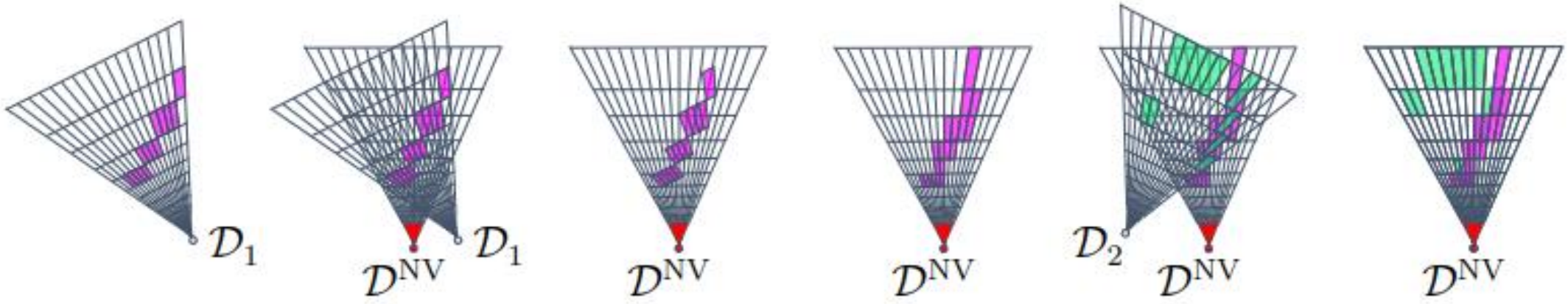


Depth Probability Distributions and Discontinuities



Estimating Depth Probability of the Novel View

- Simple solution: average intersecting rays
- More efficient to resample D_i with respect to D^{NV}
  - Nearest neighbor sampling
  - Assume 1 if an object at a disparity (zero if not), accumulate these values, then normalize

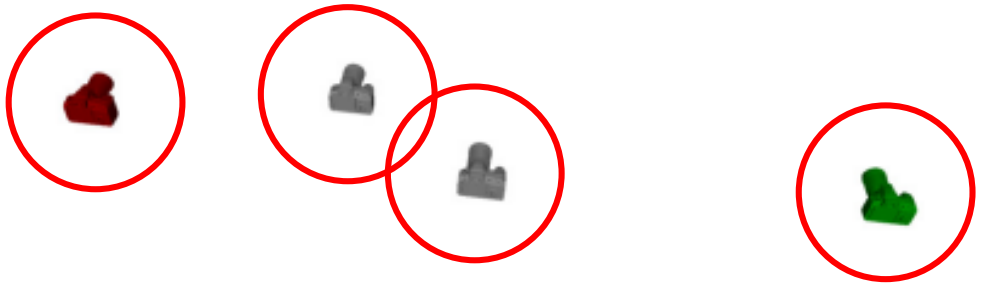


Synthesis of a Novel View

$$\widetilde{\mathcal{I}}^{\mathrm{NV}}(x,y)\Big|_{d=0} = \mathrm{R}\left(\left\{\mathcal{I}_i(x_i,y_i)\cdot \mathbb{1}_{\{\mathcal{D}^{\mathrm{NV}}(x,y,0)>t\}}\right\}_{i=1:N}\right)$$

- Pixels rendered independently back to front
- $\mathbb{1}$ is the indicator function
- I_i are the coordinates that coordinates that correspond with (x,y) in NV
- *R* is a weighting value based on camera parameters and principal axes
- As depth is closer to the camera, overwrite pixels where $D^{NV}(x,y,d)$ is above *t*



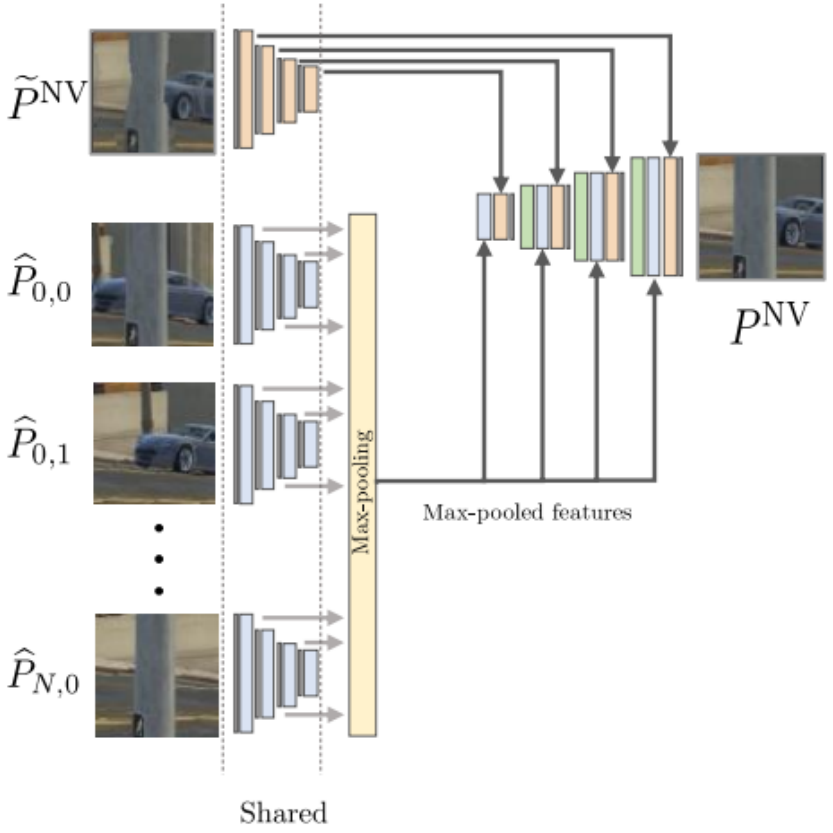


Artifacts Present After Synthesis



Refinement Network

- Akin to denoising
- For a pixel *p* in I^{NV}, extract 64x64 patch around *p*
- Varying number of patches needed to create P^{NV} - these patches are warped
  - Patches gathered using educated selection of a plane sweep
- Encode max pooled features at various layers using skip connections



Patch Warping

- Warp generated patch based on homography induced by the plane at disparity *d*
  - Images of same planar space from different views
  - Affine transformation

$$\widehat{P}_{i,j} = \mathbf{W}(P_{i,j}, H_{i\rightarrow \mathrm{NV}}^{d=\bar{d}})$$

Refinement Network Training

- *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, Johnson et al.
- Deep feature matching loss
- Feature reconstruction not style reconstruction
- Any 2d conv base network can be used
- Φ_j = activations at jth layer
- $C_j H_j W_j$ = feature map shape
- Squared and normalized

$$\ell^{\phi,j}_{feat}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2$$

Refinement Example



Dataset

- MVS-Synth dataset
  - 120 sequences of 1920x1080 images from the game GTA V
  - 100 frames per sequence
  - Ground truth disparity maps
  - Extrinsic and intrinsic camera parameters
  - Urban setting similar to real world

Evaluation

- Quantitative results use MVS-Synth
  - Structural Similarity Index (SSIM)
  - Peak Signal-to-Noise Ratio (PSNR)
  - SSIM and PSNR measure quality loss due to encoding
  - Comparison against Stereo Magnification (Similar constraints but not the same)
- Qualitative results use real images

Quantitative Results

Metric	Ours Warped	Ours Refined	SM
Mean SSIM	0.851	**0.877**	0.842
Mean PSNR	24.6dB	**27.38dB**	25.49dB

- Execution time (average) given two views as input
  - Depth probability volumes – 40s
  - View synthesis – 30s
  - Refinement Network – 28s

Comparison to Stereo Magnification



Comparison to Soft3D



Soft3d, Penner et al.



Extreme View Synthesis

Limitations

- Refinement network cannot hallucinate pixels
  - GAN-based approach could solve this
- Certain novel views may be affected by depth quantization artifacts



Conclusion

- Heavily influenced from DeepMVS
- Synthesize novel views from extreme cases
  - As little as 2 images
  - Varying viewpoints
- Limited quantitative results
- Public code
  - Penner and Zhang propose Soft3D which has similar results, but the code is not available