dyson Robotics Lab   Imperial College London

Semantic Neural Representation for Scene Understanding

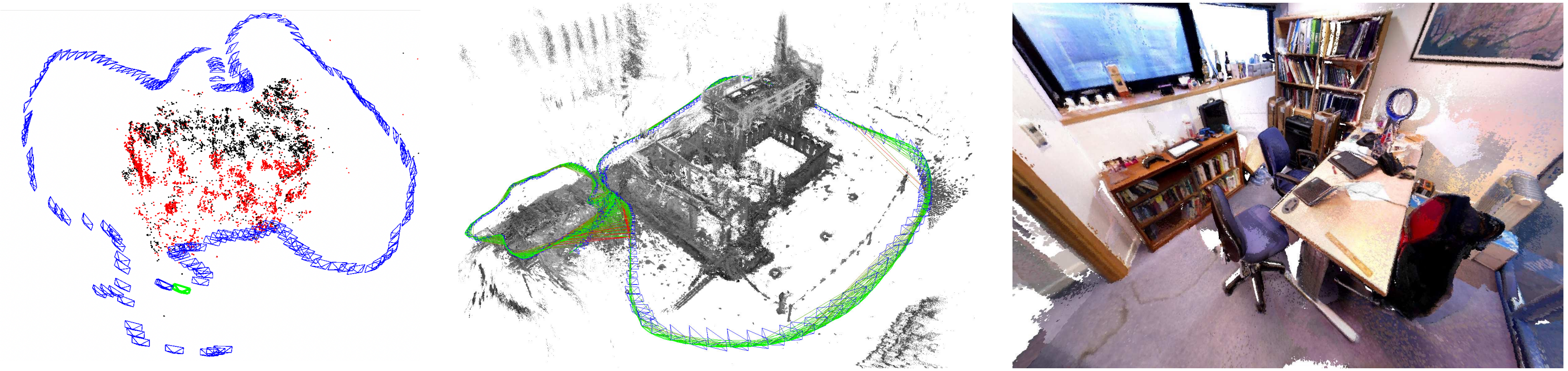
Shuaifeng Zhi

4th -year PhD Student

Dyson Robotics Lab at Imperial College London

GAMES Webinar 21st October 2021

Scene Representations in vSLAM



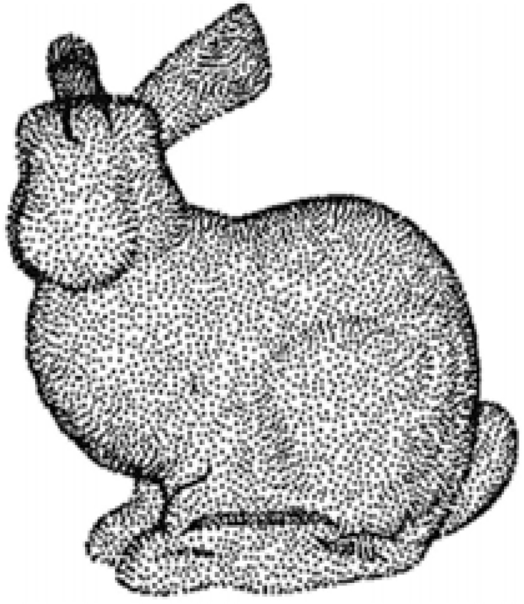
$$(x, y, z) \rightarrow \text{DNNs} \rightarrow (R, G, B, \sigma)$$

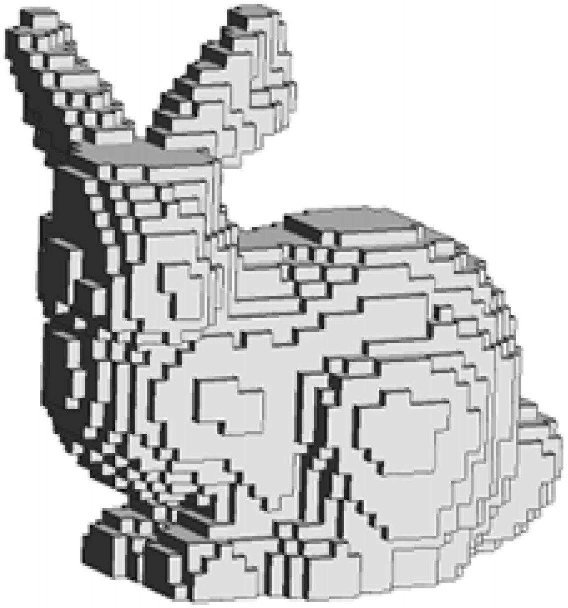
DNNs

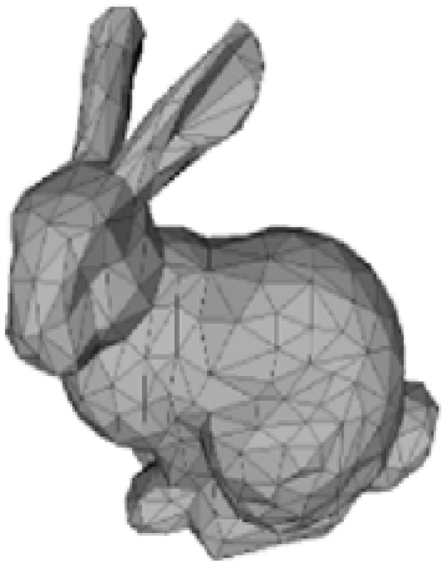
- Scene representation concerns the environmental attributes that can be captured in a SLAM system's world model.

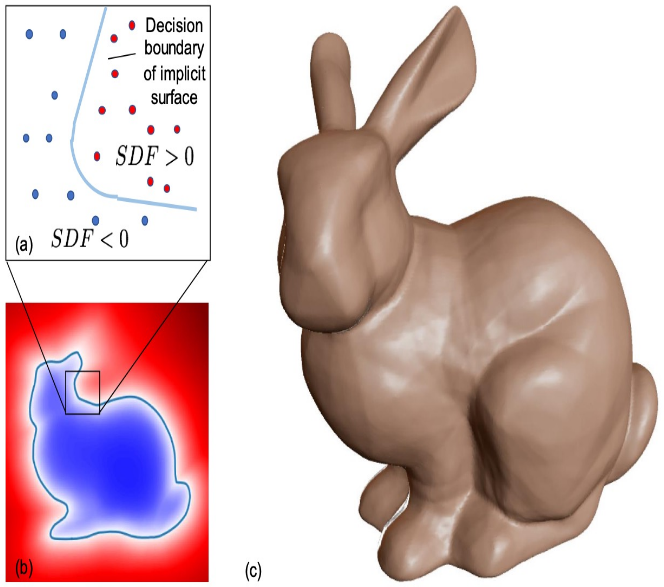
- Scene representations in vSLAM have gradually progressed from <u>sparse point sets</u> to <u>dense geometric 3D maps</u> and more recently to <u>neural representations</u>.

Classical Geometric Scene Representation









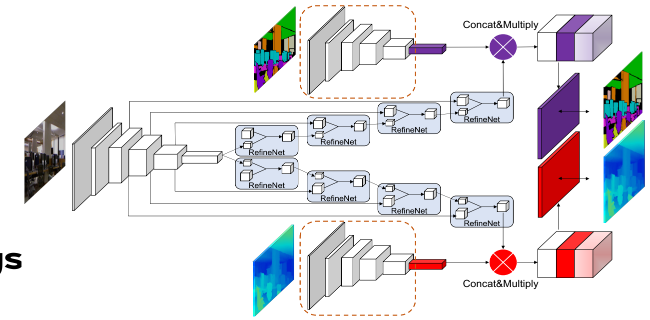
Point Cloud | Voxel | Mesh | Signed Distance Field

Explicit | Implicit

Neural Scene Representation

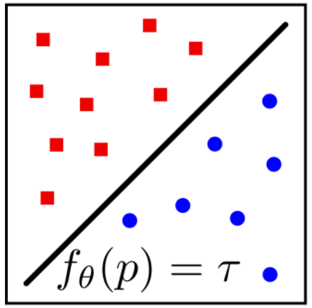
Explicit Representation

- GQN [Eslami et al. 2018]
- CodeSLAM [Bloesch et al. 2018]
- SceneCode [Zhi et al. 2019]
- DeepVoxels [Sitzmann et al. 2019]
- Neural Volumes [Lombardi et al 2019]
- Latent Fusion [Park, et al. 2020]



Implicit Representation

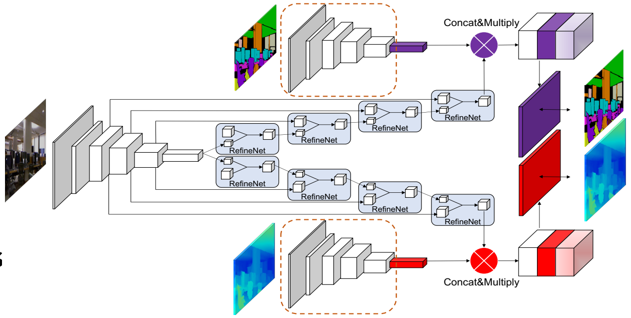
- SRN [Sitzmann etal. 2019]
- DeepSDF [Park et al. 2019]
- PIFu [Shunsuke et al. 2019]
- CON [Mescheder et al. 2020]
- NeRF [Mildenhall et al. 2020]
- NeRF Explosion...



Neural Scene Representation

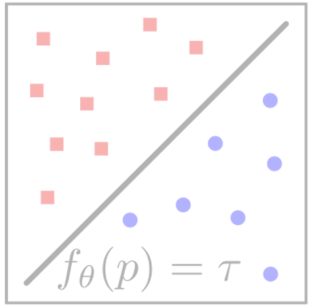
Explicit Representation

- GQN [Eslami et al. 2018]
- CodeSLAM [Bloesch et al. 2018]
- SceneCode [Zhi et al. 2019]
- DeepVoxels [Sitzmann et al. 2019]
- Neural Volumes [Lombardi et al 2019]
- Latent Fusion [Park, et al. 2020]



Implicit Representation

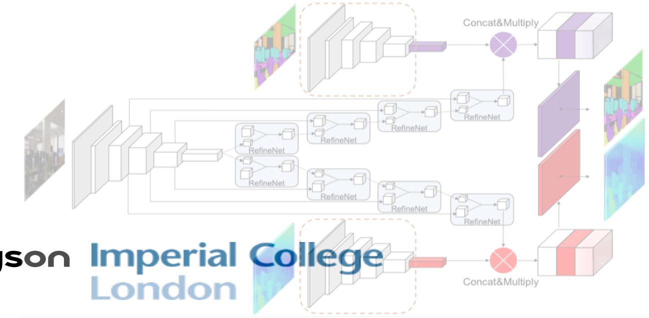
- SRN [Sitzmann etal. 2019]
- DeepSDF [Park et al. 2019]
- PIFu [Shunsuke et al. 2019]
- CON [Mescheder et al. 2020]
- NeRF [Mildenhall et al. 2020]
- NeRF Explosion...



Neural Scene Representation

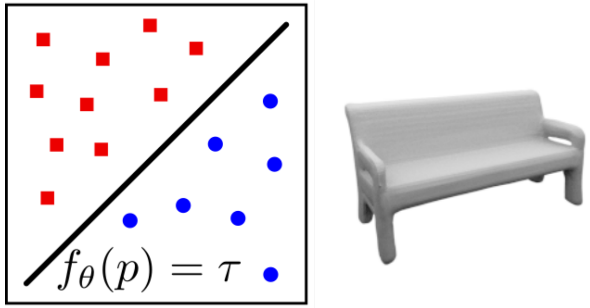
Explicit Representation

- GQN [Eslami et al. 2018]
- CodeSLAM [Bloesch et al. 2018]
- SceneCode [Zhi et al. 2019]
- DeepVoxels [Sitzmann et al. 2019]
- Neural Volumes [Lombardi et al 2019]
- Latent Fusion [Park, et al. 2020]



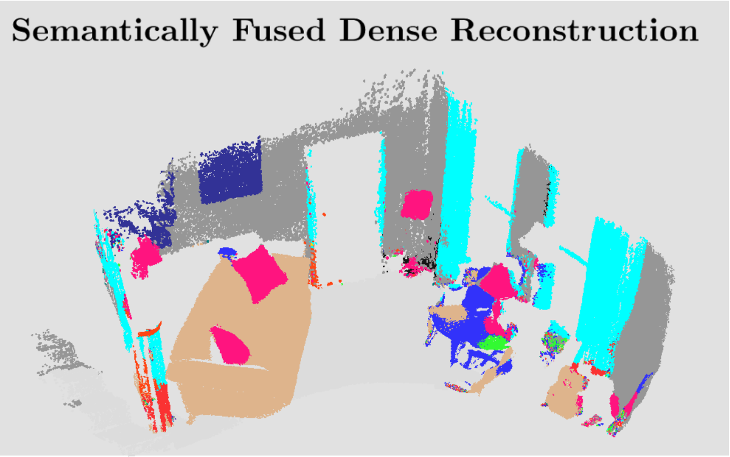
Implicit Representation

- SRN [Sitzmann etal. 2019]
- DeepSDF [Park et al. 2019]
- PIFu [Shunsuke et al. 2019]
- CON [Mescheder et al. 2020]
- NeRF [Mildenhall et al. 2020]
- NeRF Explosion...



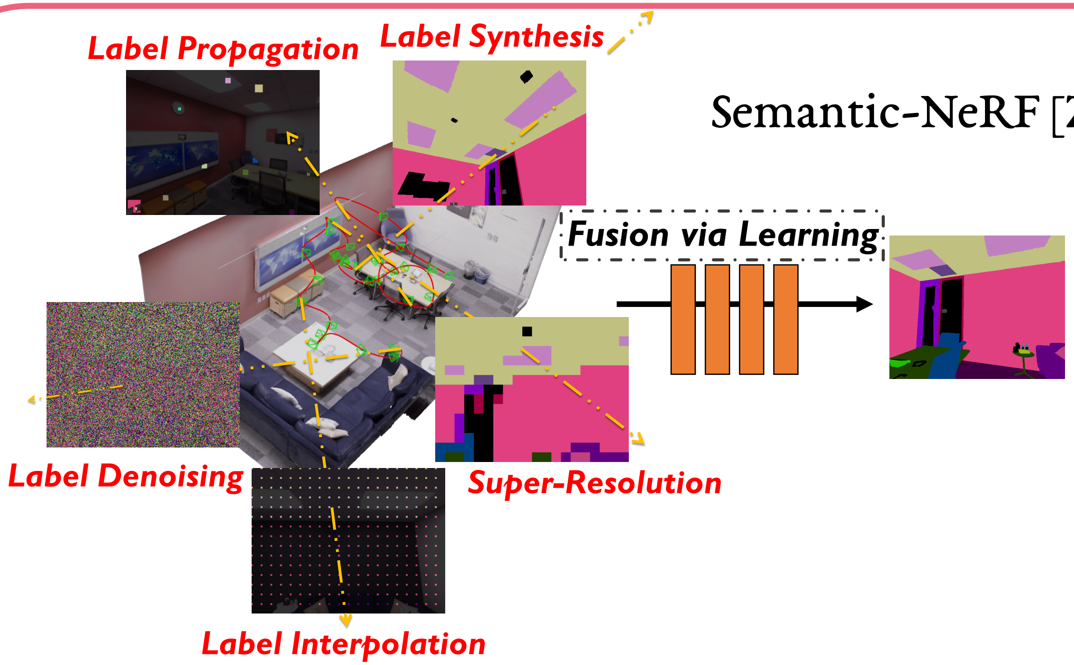
Existing Semantic Scene Representations

SemanticFusion [McCormac et al. 2017]



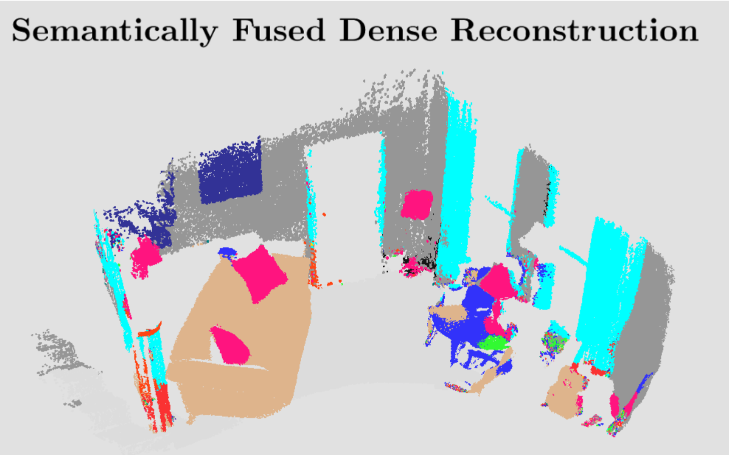
SceneCode [Zhi et al. 2019]

Semantic-NeRF [Zhi et al. 2021]



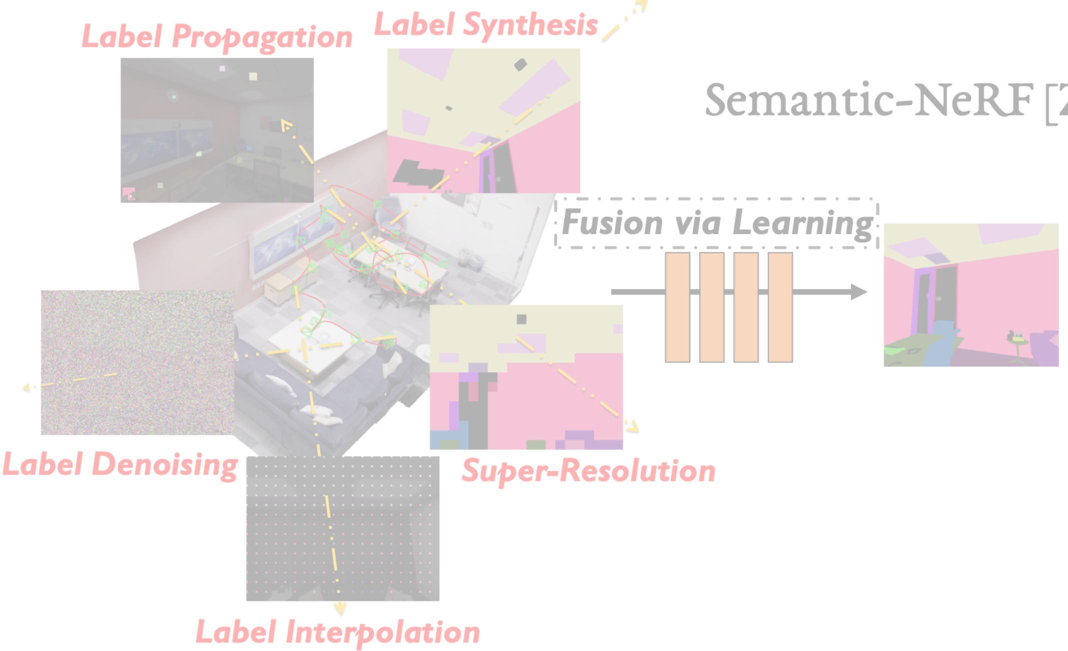
Existing Semantic Scene Representations

SemanticFusion [McCormac et al. 2017]

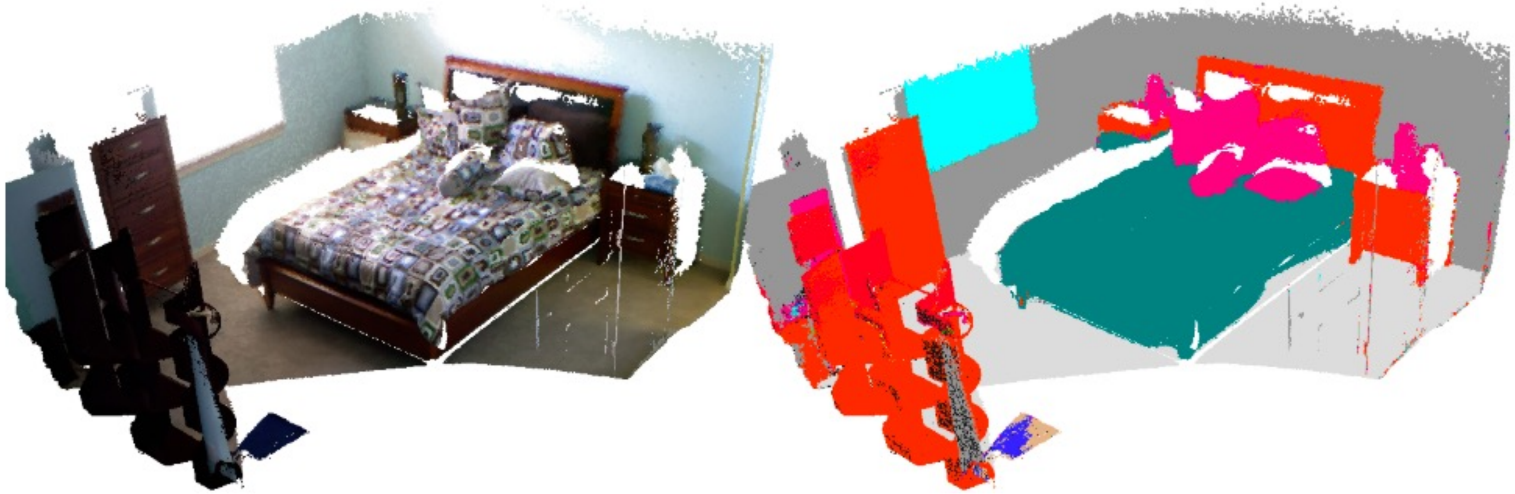


SceneCode [Zhi et al. 2019]

Semantic-NeRF [Zhi et al. 2021]



SemanticFusion



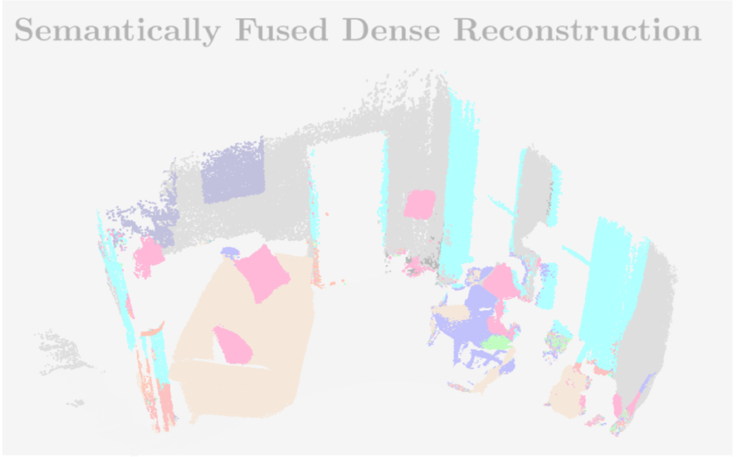
In many semantic mapping systems,

- mature geometric SLAM systems are used and their semantic representation relies on the geometric one.
- 3D dense map elements are associated with 2D/3D semantic predictions.
- semantics of each map element is individually processed.

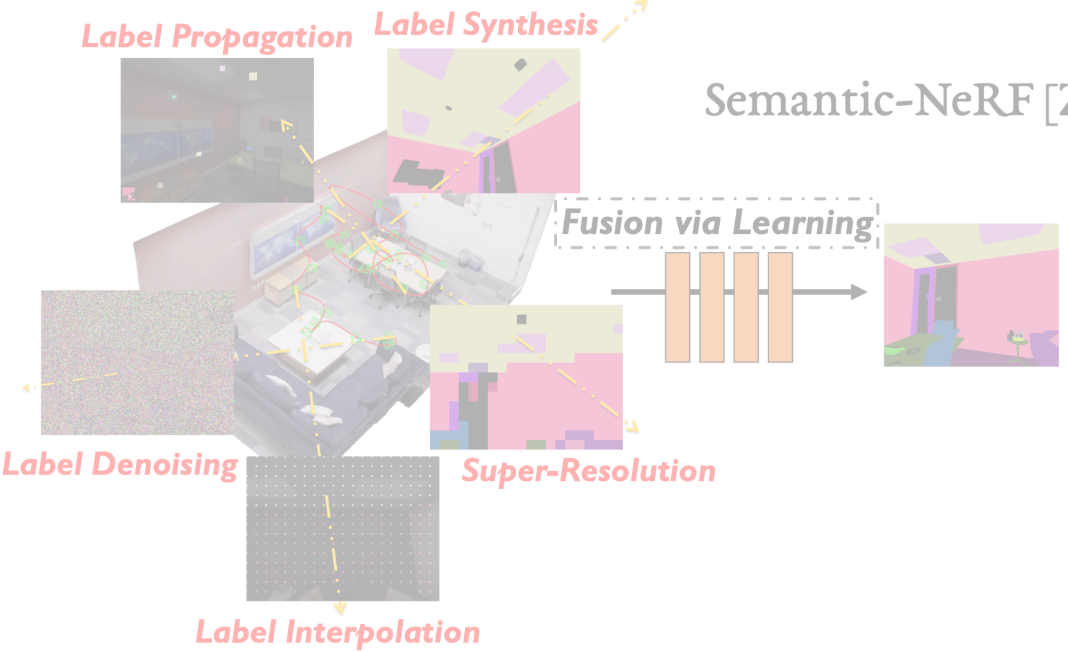
Existing Semantic Scene Representations

SceneCode [Zhi et al. 2019]

SemanticFusion [McCormac et al. 2017]



Semantic-NeRF [Zhi et al. 2021]

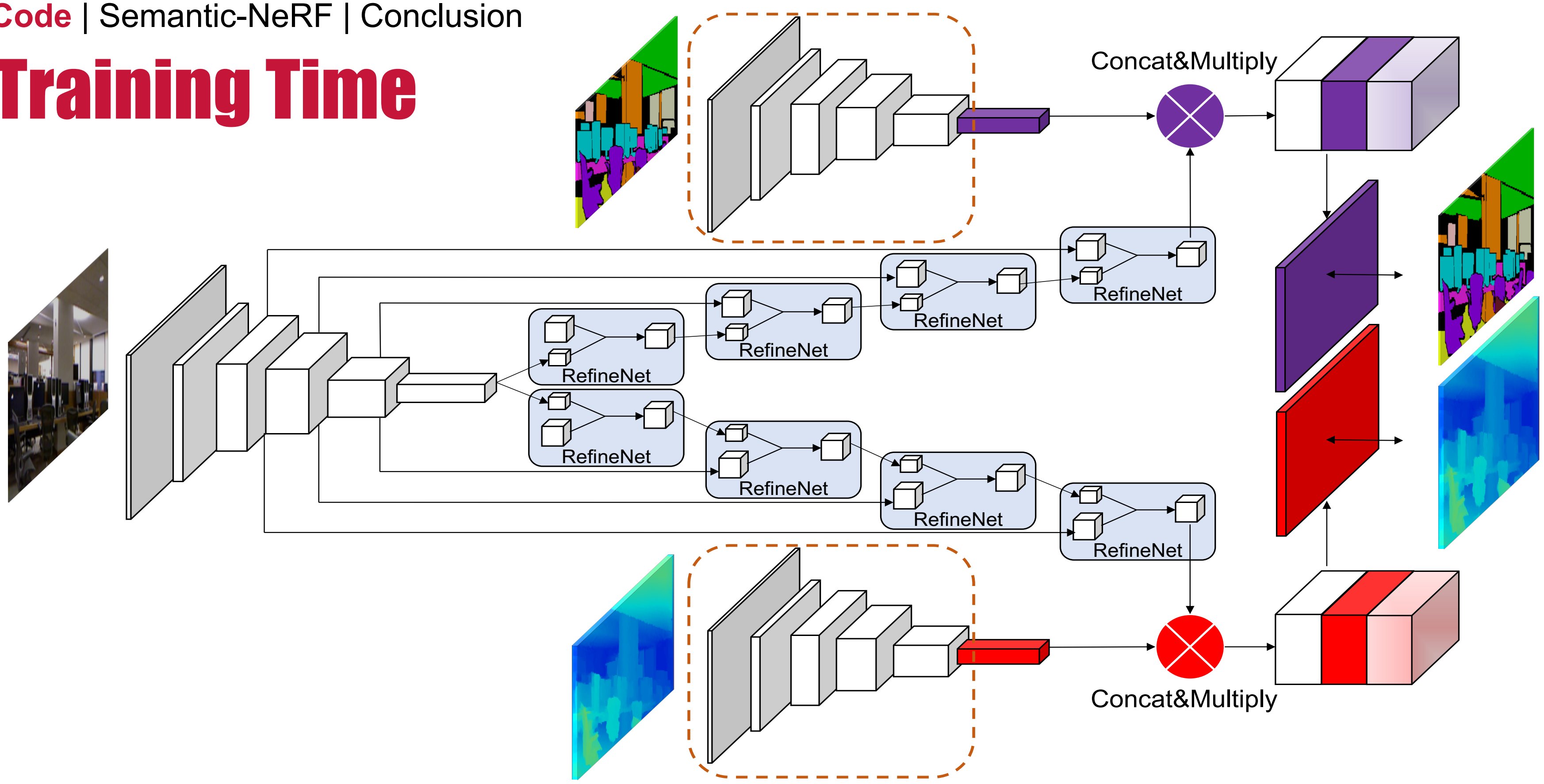


If dense geometry can be represented by a compact code,
how about dense semantic labelling?

SceneCode

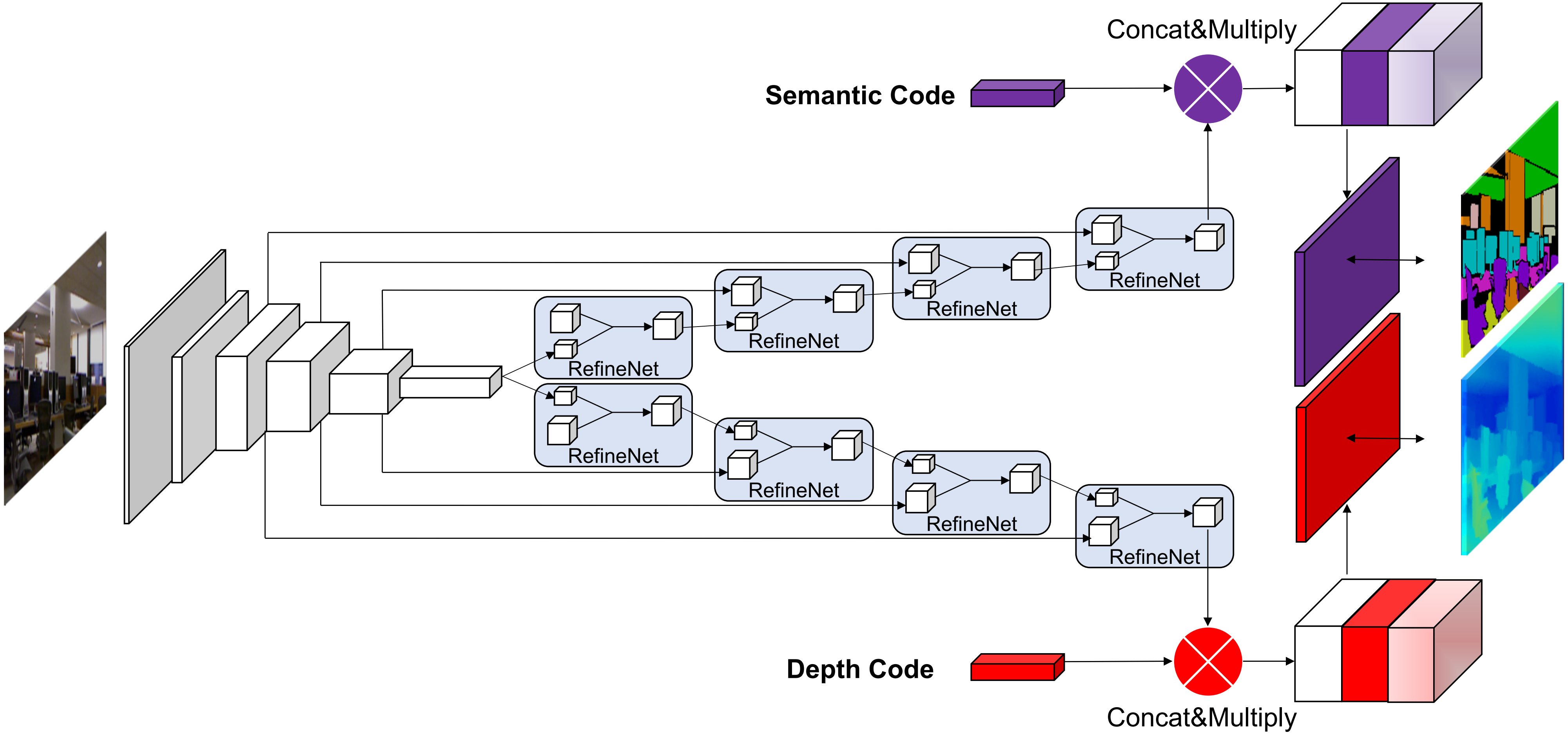
- Introduce **a compact and optimisable semantic representation** using an image-conditioned variational auto-encoder.
- Propose **a new multi-view semantic label fusion method** maximising semantic consistency.
- Build a monocular dense semantic 3D reconstruction system, where geometry and semantics are tightly coupled into a joint optimisation framework.

Network-Training Time



- Compact and optimisable code representations of depth and semantics via a CVAE.
- Allow inference-time refinement via photometric and semantic costs.

Network-Test Time

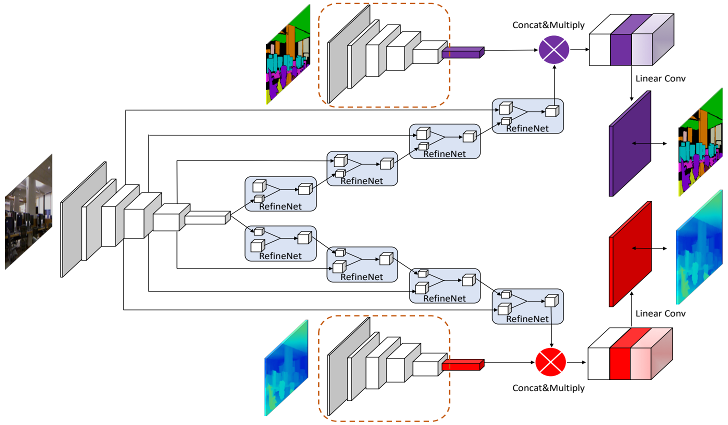


- Full-zero codes are used for both initialisation and monocular predictions.

Why Linear Decoder?

A linear VAE-decoder after the **'Concat &Multiply'** operation makes the output **non-linear** w.r.t. input colour images while **linear** w.r.t. latent codes.

$$D(c_d, I) = D_0(I) + J_d(I)c_d$$
$$S(c_s, I) = S_0(I) + J_s(I)c_s$$

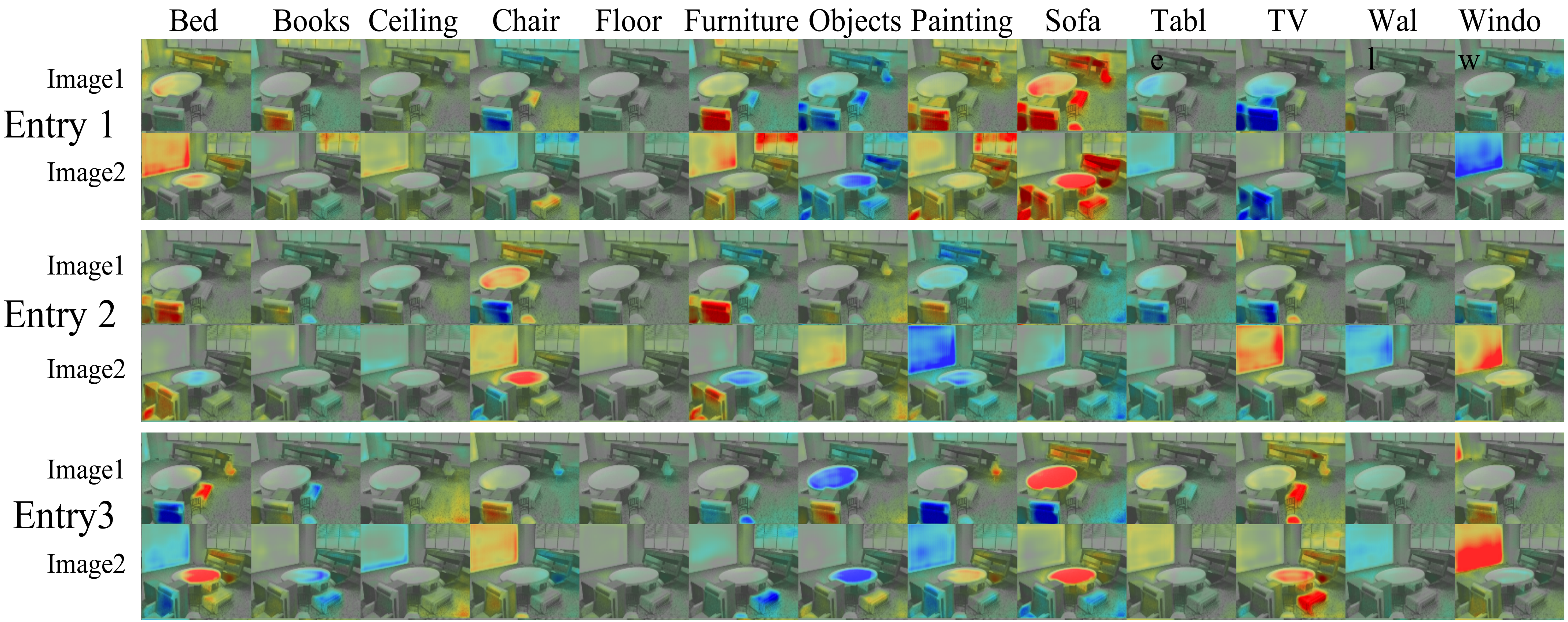


- $J_{s/d}$ is the learned linear Jacobians
- c_d and c_s are code-representations of depth and semantic
- D_0 and S_0 are monocular predictions with full-zero codes, i.e.,

$$D_0(I) = D(0, I),\ S_0(I) = S(0, I)$$

What Has Been Learned by Semantic Code Representations?

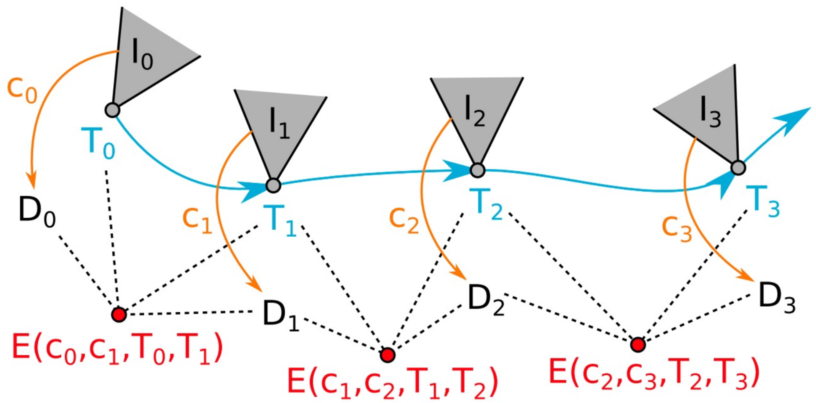
Visualisation of Semantic Code-Jacobians



Exploring the Latent Space

Multi-view Fusion via Code-optimisation

Dense Geometry Refinement



Depth code can be refined by **minimising** both **photometric error r_i** and **geometric error r_z** :

$$r_i = I_A\left[\boldsymbol{u}_A\right] - I_B\left[w\left(\boldsymbol{u}_A, \boldsymbol{c}_d^A, \boldsymbol{T}_{BA}\right)\right]$$

$$r_z = D_B[w(\boldsymbol{u}_A, \boldsymbol{c}_d^A, \boldsymbol{T}_{BA})] - [\boldsymbol{T}_{BA}\pi^{-1}(\boldsymbol{u}_A, D_A[\boldsymbol{u}_A])]_Z$$

Multi-view Fusion via Code-optimisation

Dense Semantics Refinement

$$r_s^{'} = DS\left(S_A\left[\mathbf{u}_A\right], S_B\left[w\left(\mathbf{u_A}, \boldsymbol{c}_d^A, \boldsymbol{T}_{BA}\right)\right]\right)$$

However, simply maximising semantic consistency has trivial solutions, e.g., wrong but consistent labels compared to ground truth annotations.

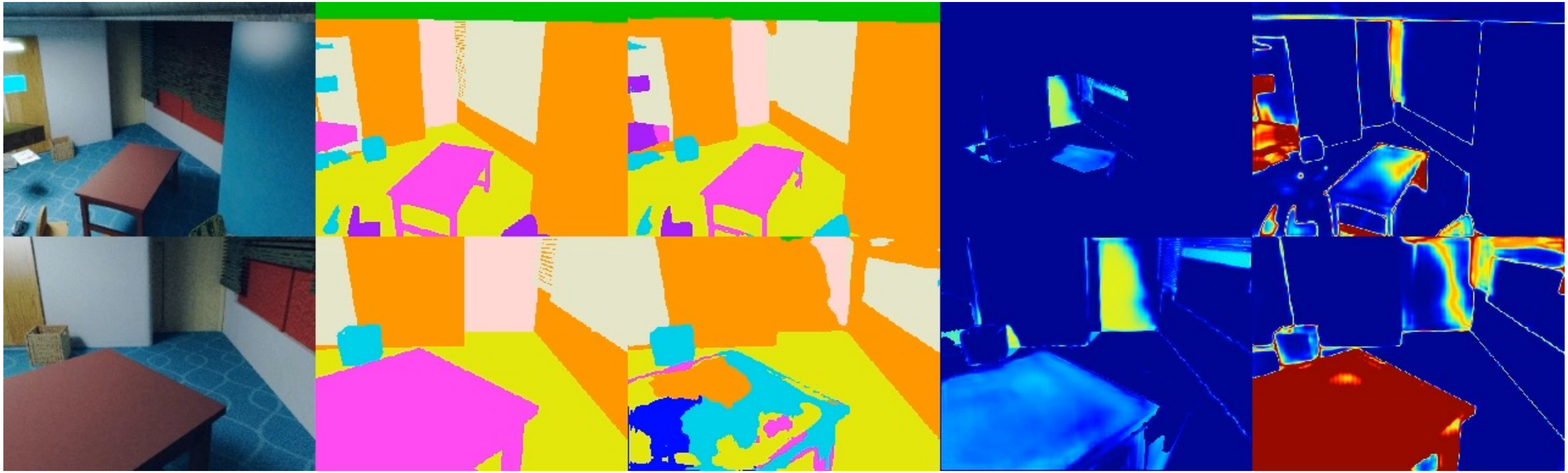
We explicitly introduce *zero-code regularisation* term to avoid this:

$$r_s = r_s^{'} + \lambda\left(\|\boldsymbol{c_s}_{\boldsymbol{A}}\|_2^2 + \|\boldsymbol{c_s}_{\boldsymbol{B}}\|_2^2\right)$$

SceneCode-Multiview Semantic Label Fusion

W/ zero-code prior

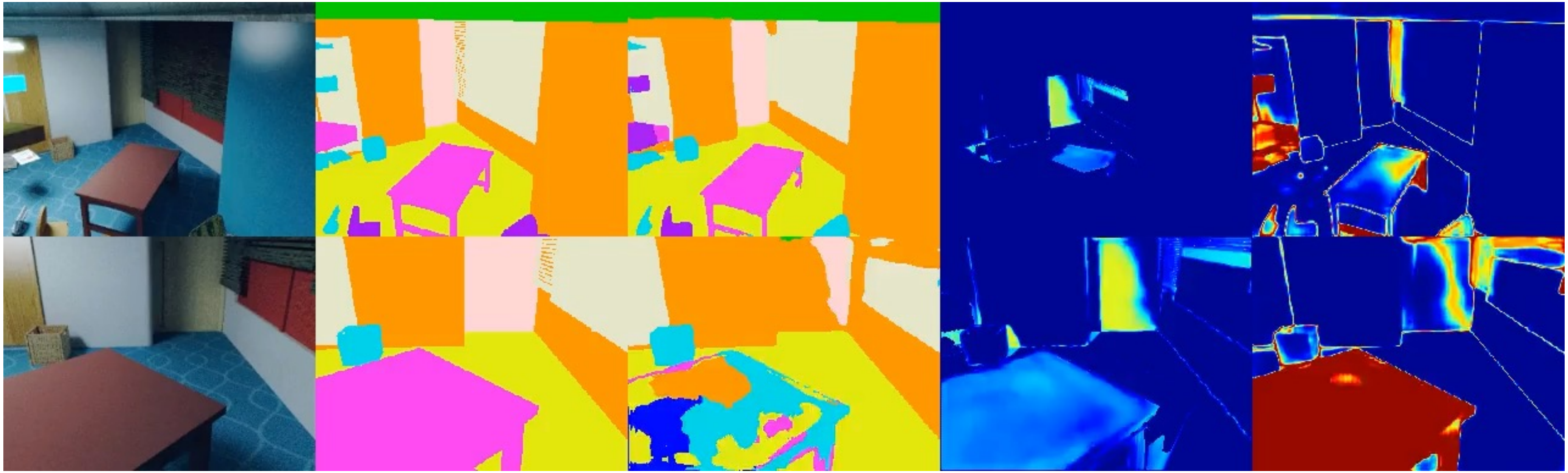
Input Image	GT Label	Opt. Label	Sem. Error	Entropy

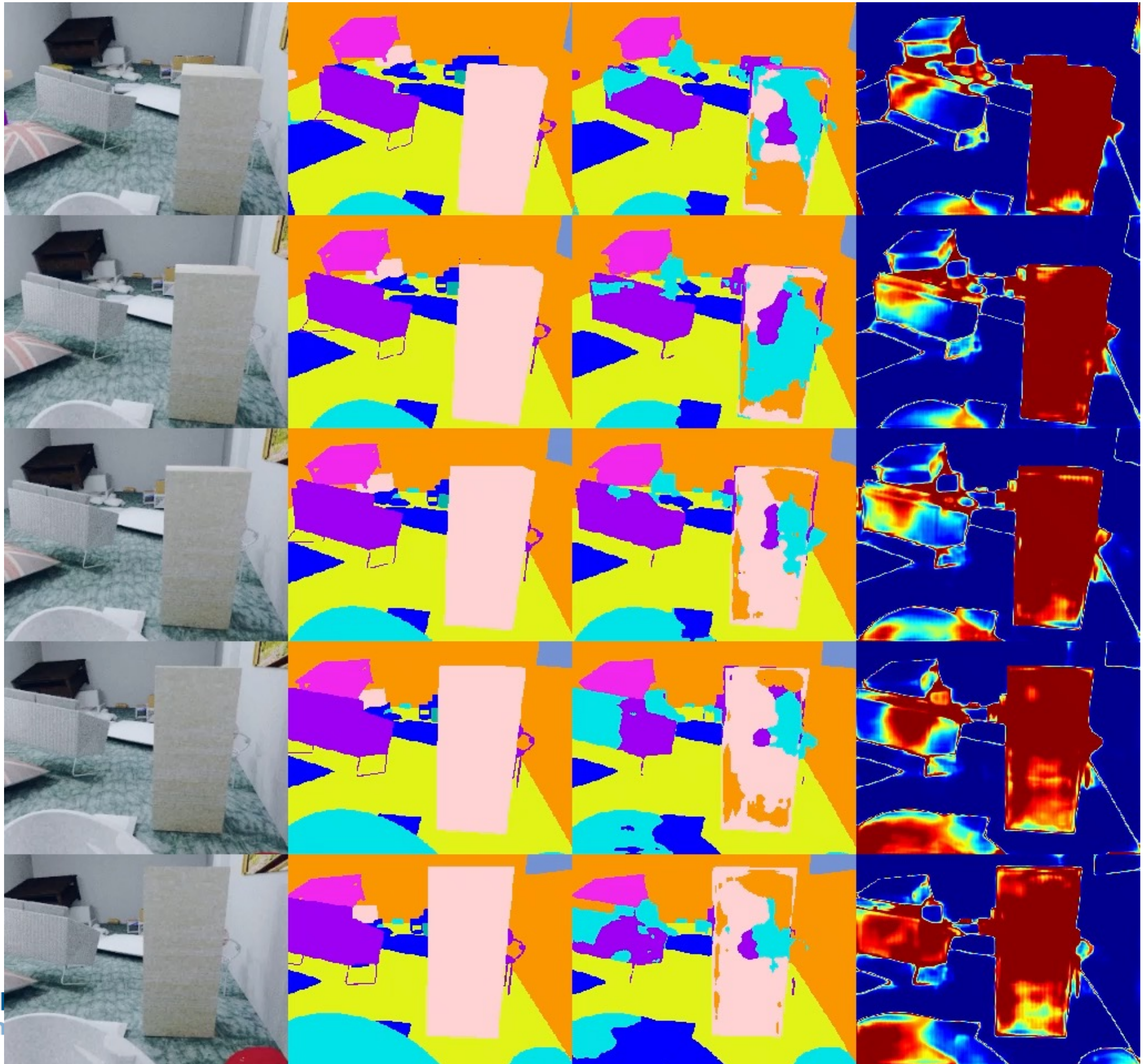


SceneCode-Multiview Semantic Label Fusion

W/O zero-code prior

Input Image	GT Label	Opt. Label	Sem. Error	Entropy

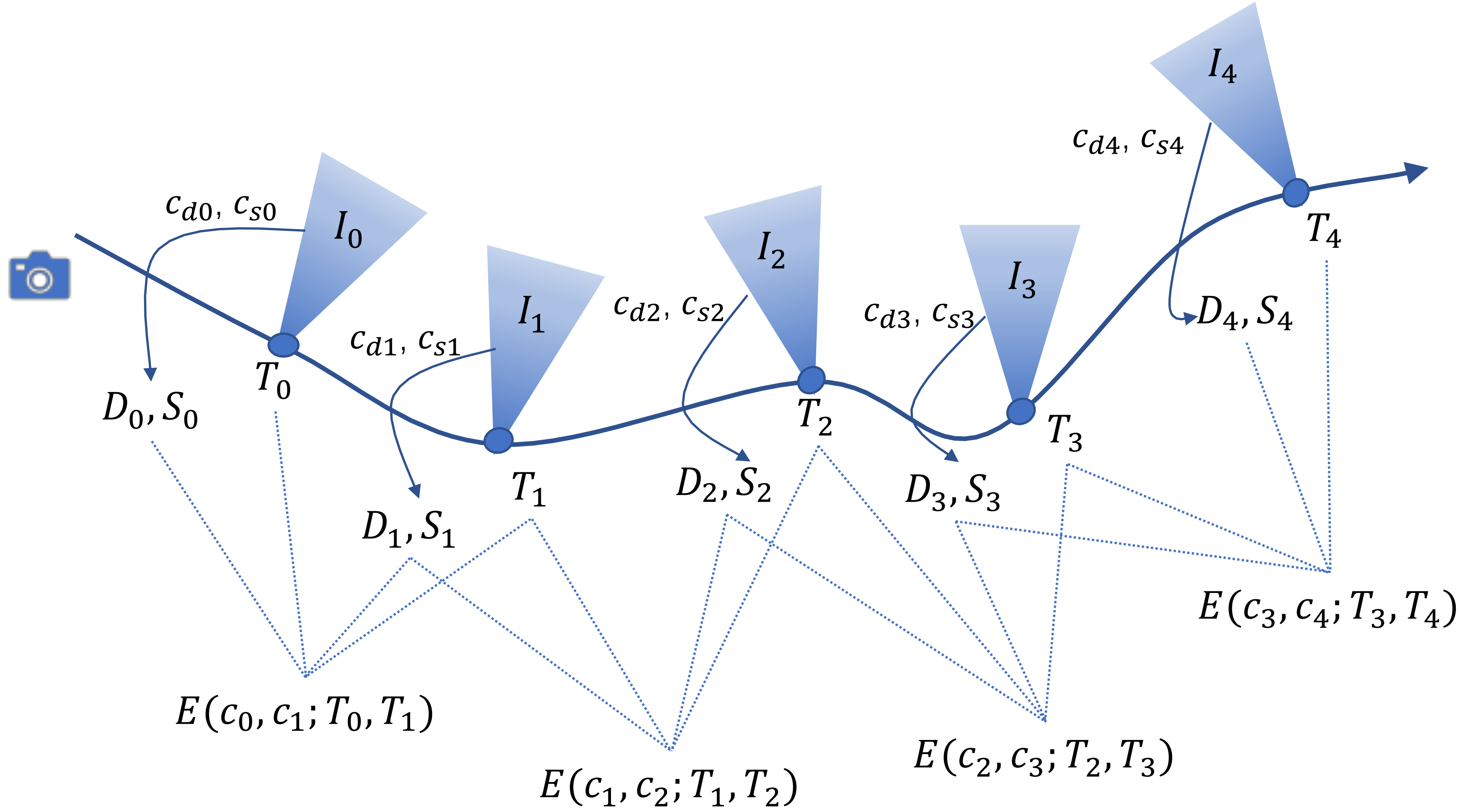




dyson Robotics Lab

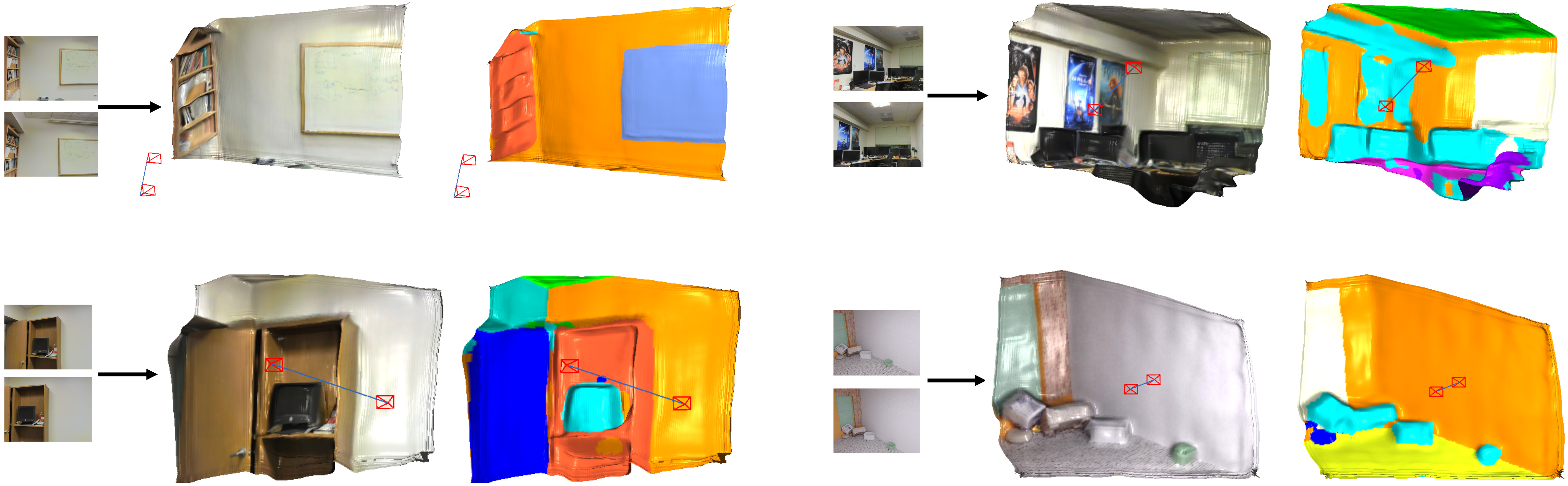
Imperial London

Monocular Dense Semantic SLAM



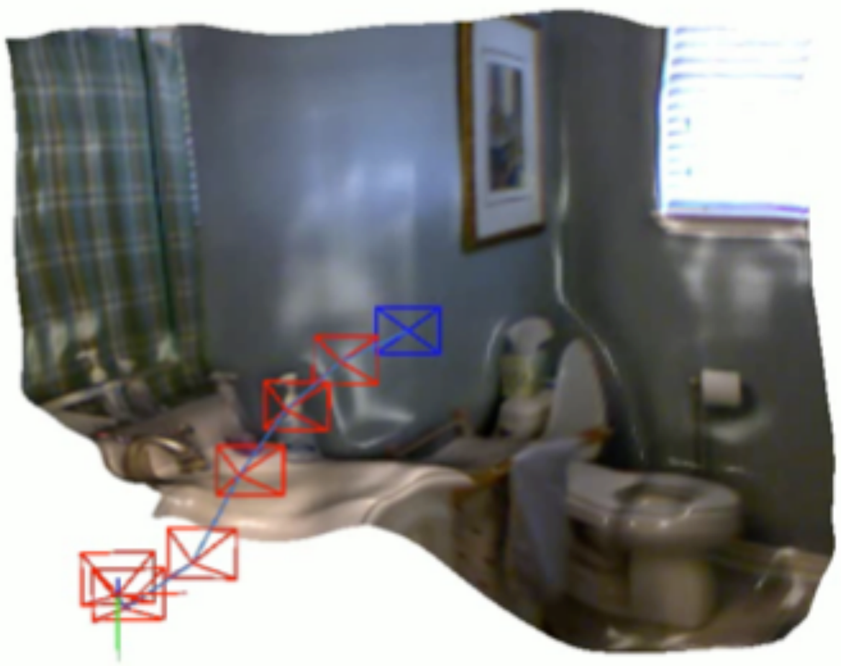
Monocular Dense Semantic SLAM

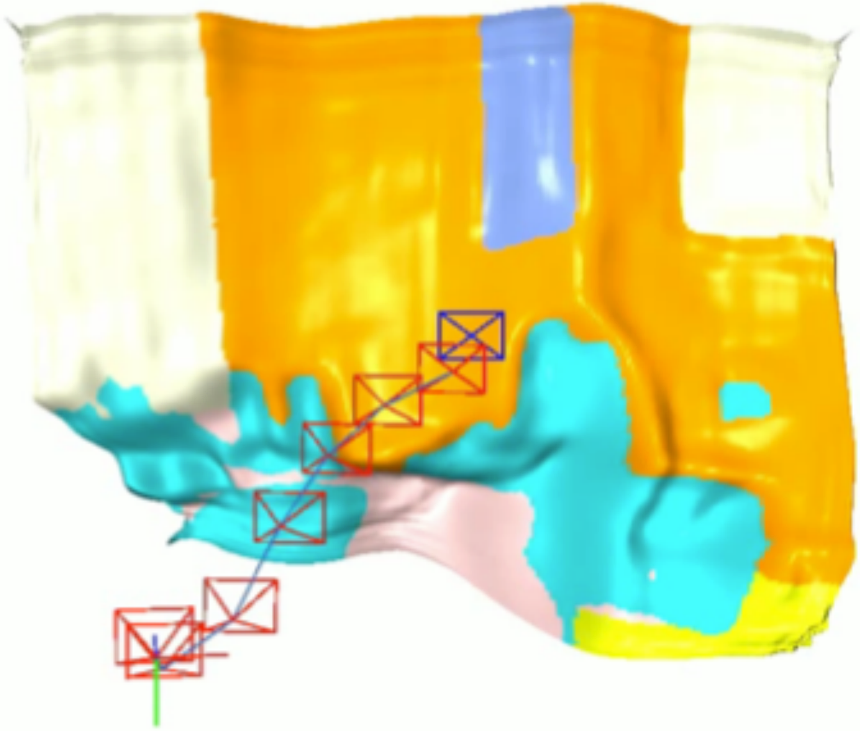
Two-frame SfM



Monocular Dense Semantic SLAM

Key-frame based Monocular SLAM



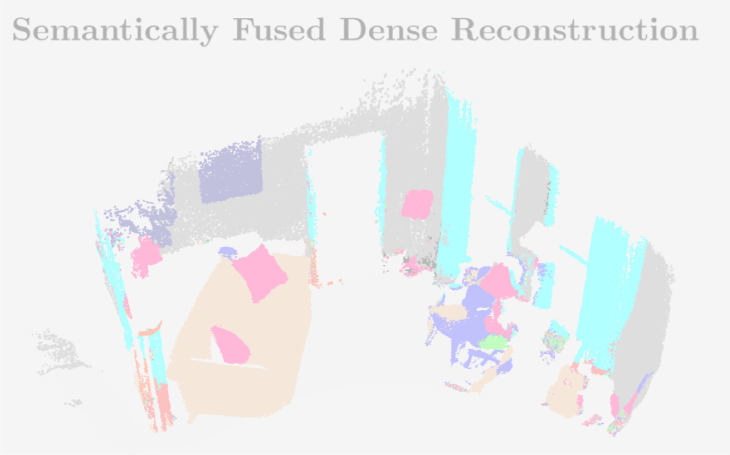


Semantic Structure from Motion

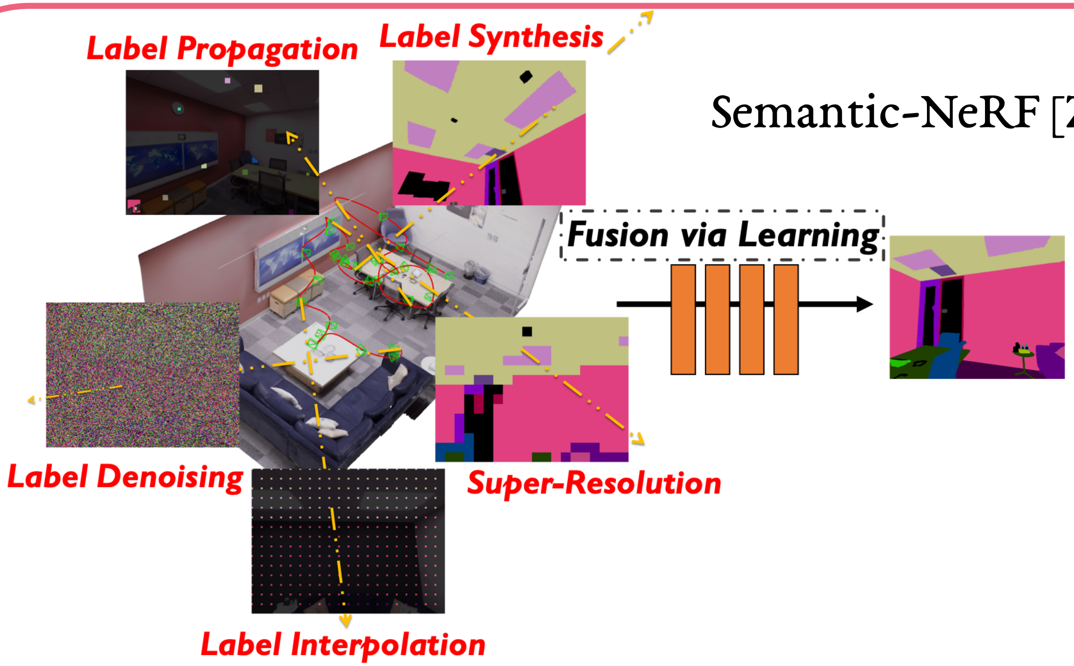
Existing Semantic Scene Representations

SceneCode [Zhi et al. 2019]

SemanticFusion [McCormac et al. 2017]

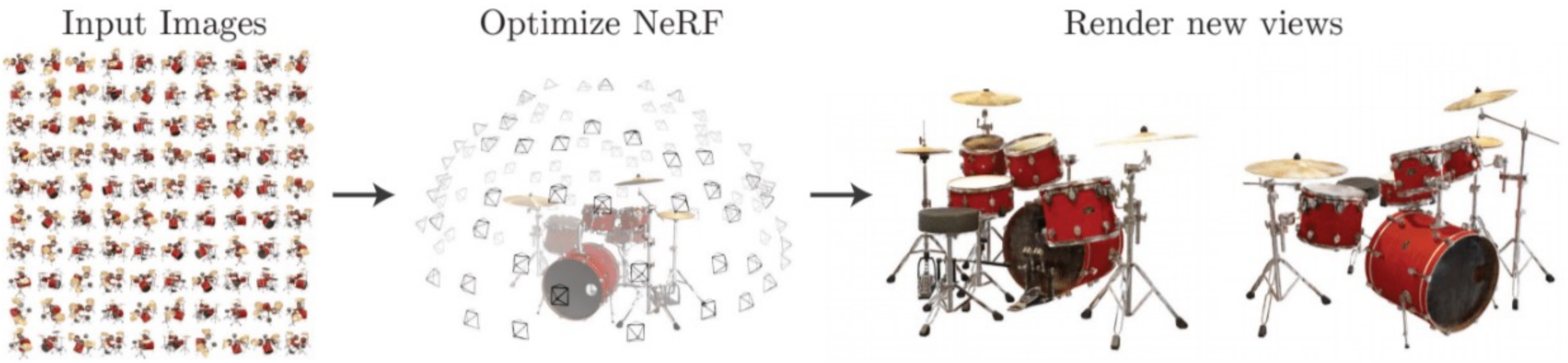


Semantic-NeRF [Zhi et al. 2021]



Neural Radiance Fields (NeRF)

NeRF use MLPs to represent the 3D scene, which can be treated as a continuous volumetric representation.



Neural Radiance Fields (NeRF):

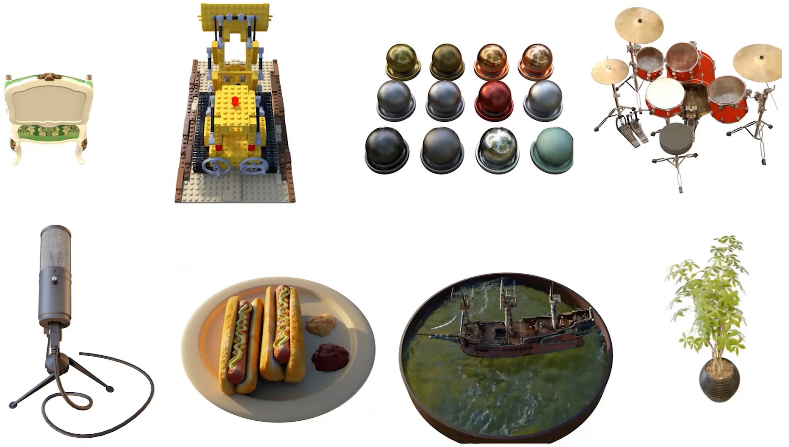
Encode scenes as a mapping on the 5D manifold of 3D positions and viewing directions.

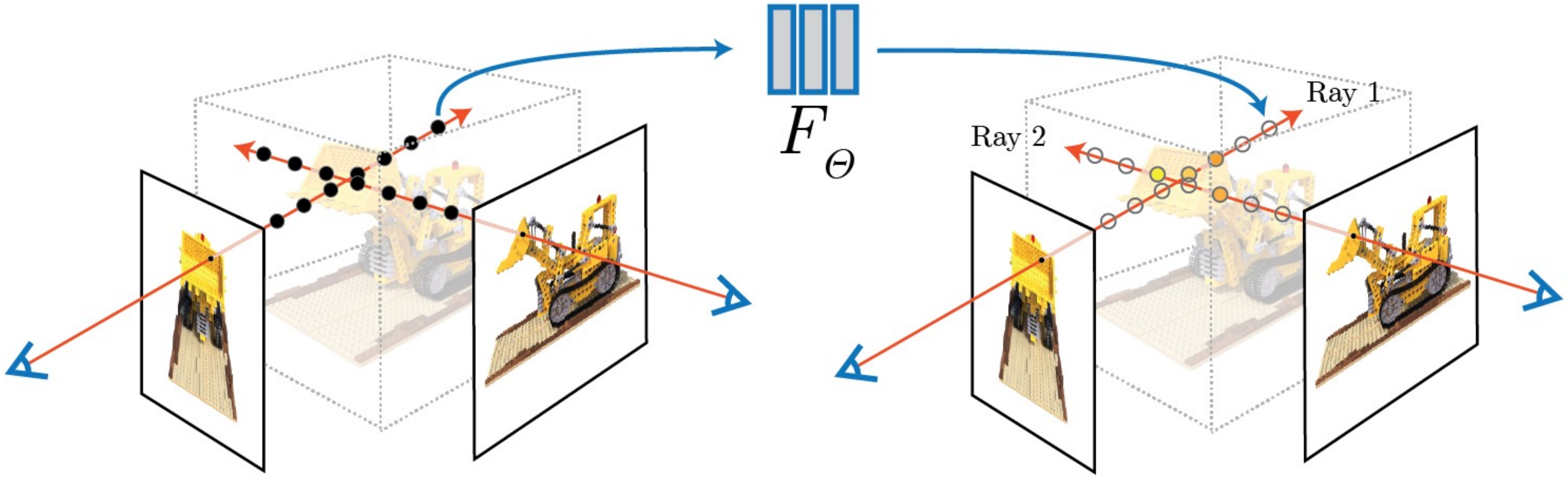
Implicit Scene Representation

(x, y, z, θ, ϕ)

(R, G, B, σ)

MLPs
(Multi-layer Perceptrons)



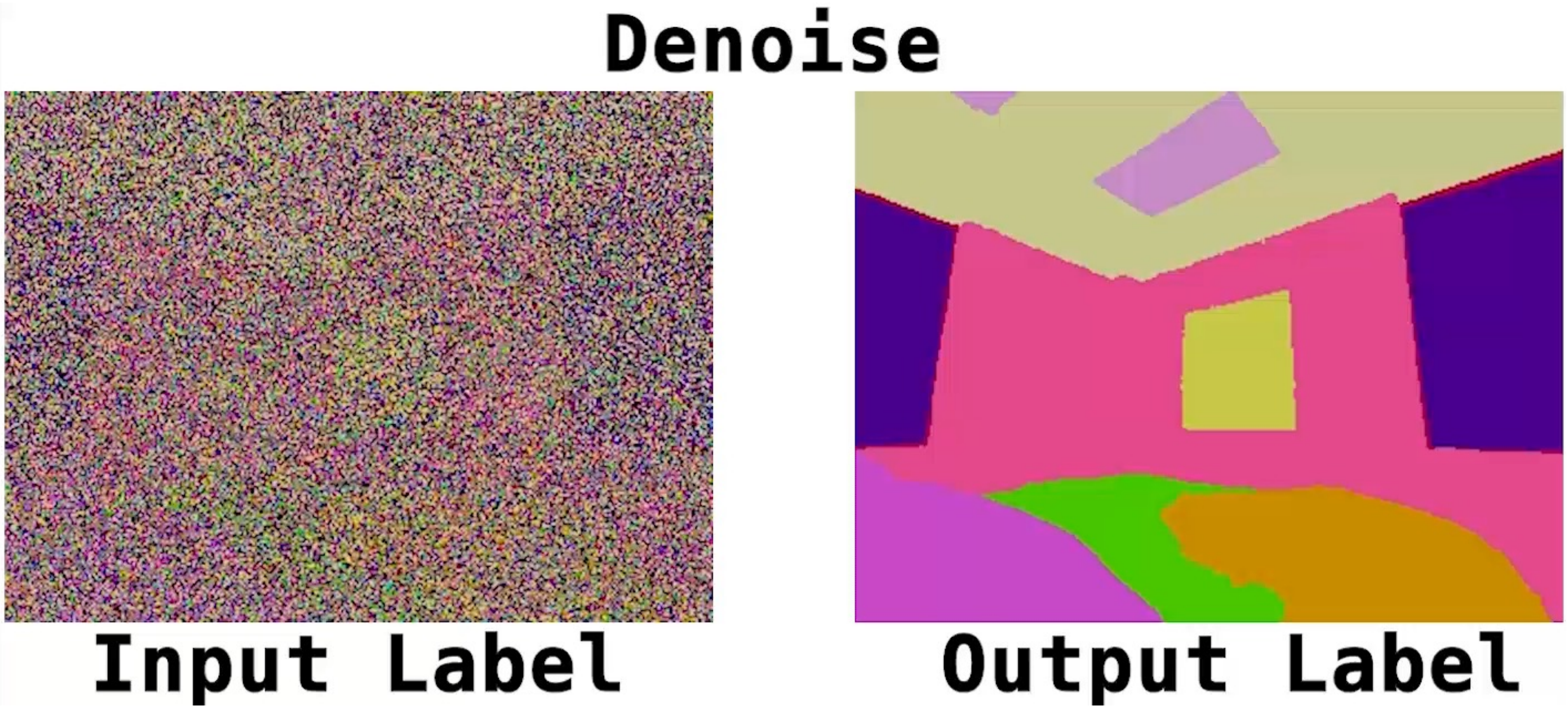


NeRF computes the colour of a single pixel $\underline{\boldsymbol{C}(\boldsymbol{r})}$ using volume rendering :

$$\hat{\mathrm{C}}(\mathbf{r}) = \sum_{k=1}^{K} \hat{T}(t_k)\,\alpha(\sigma(t_k)\delta_k)\,\mathbf{c}(t_k), \text{ where } \hat{T}(t_k) = \exp\left(-\sum_{k'=1}^{k-1} \sigma(t_k)\delta_k\right)$$

Semantic-NeRF

In-Place Scene Labelling and Understanding with Implicit Scene Representation



Semantic-NeRF

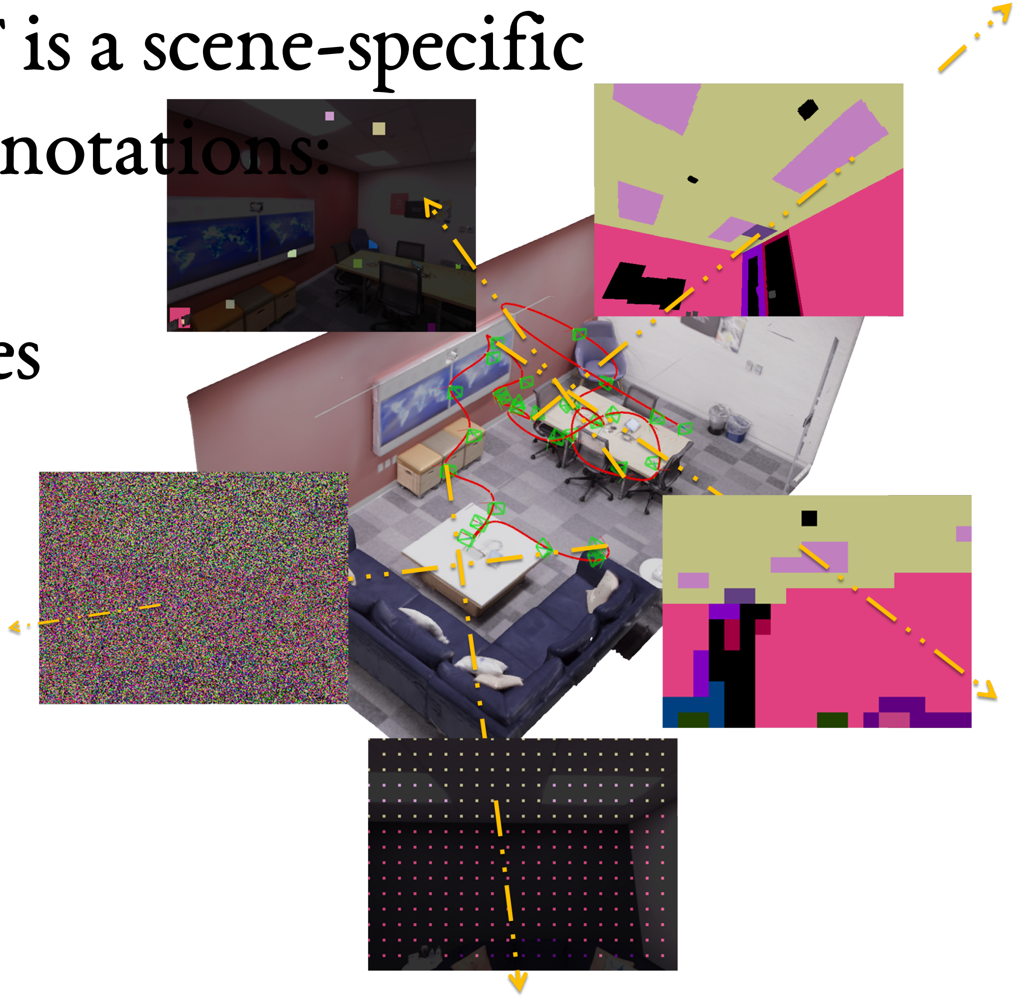
Why Semantic-NeRF:

- Most existing semantic representations relied on geometric ones.
- Semantic labelling is highly correlated with radiance and geometry
- Supervised semantic representation requires expensive annotation and shows unsatisfying generalisation in unseen or open-set environments.

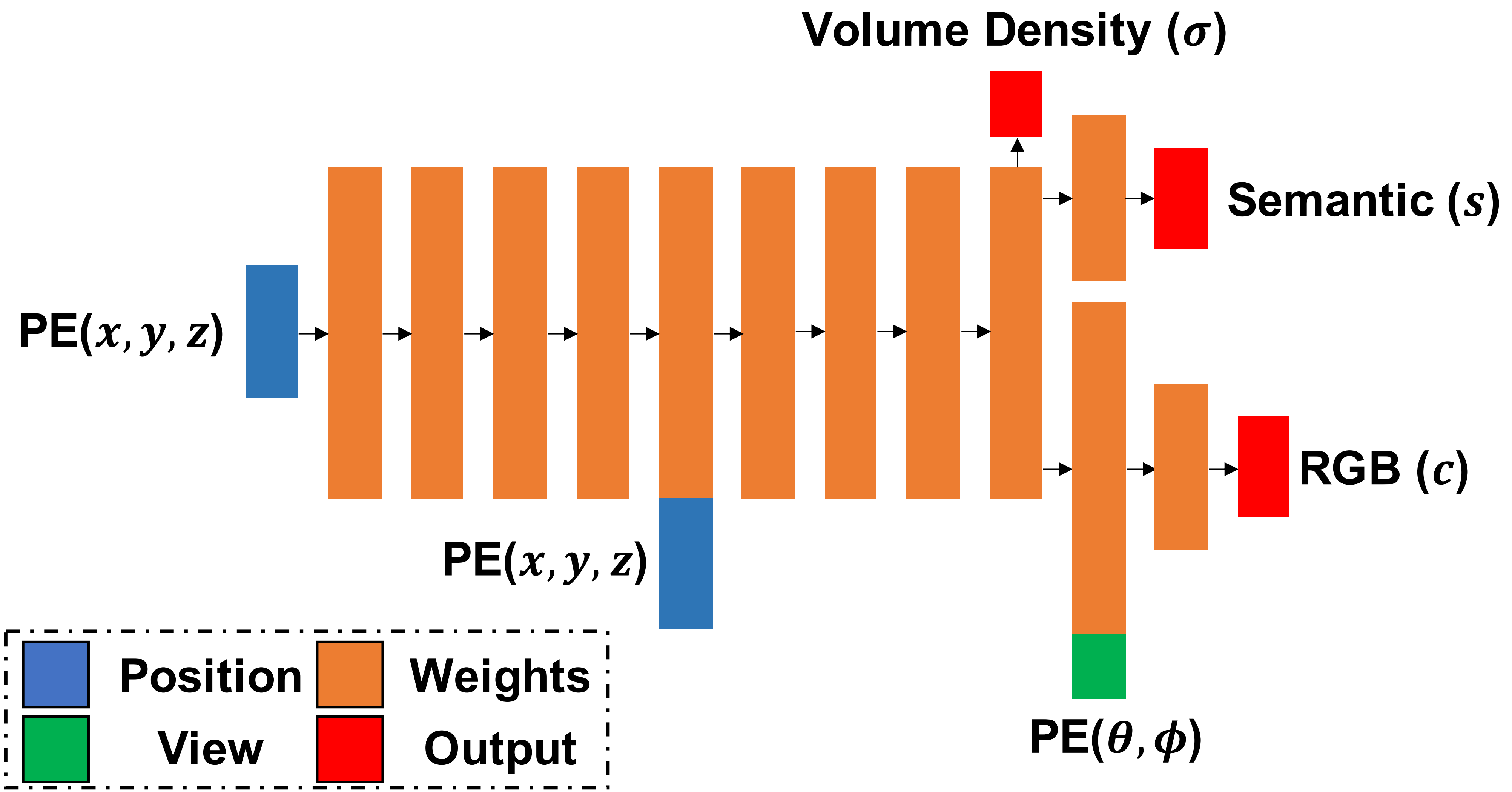
Semantic-NeRF Set-up

Without any prior training, Semantic-NeRF is a scene-specific representation learned with only in-place annotations:

- Multi-view RGB Images with camera poses
- Semantic Annotations
  - Dense Labels
  - Sparse Labels
  - Noisy Labels
  - Coarse Labels
  - Imperfect Labels



Semantic-NeRF Network Architecture



Volume Rendering of Colour and Semantics:

$$\mathbf{c} = F_{\Theta}(\mathbf{x}, \mathbf{d}), \quad \mathbf{s} = F_{\Theta}(\mathbf{x})$$

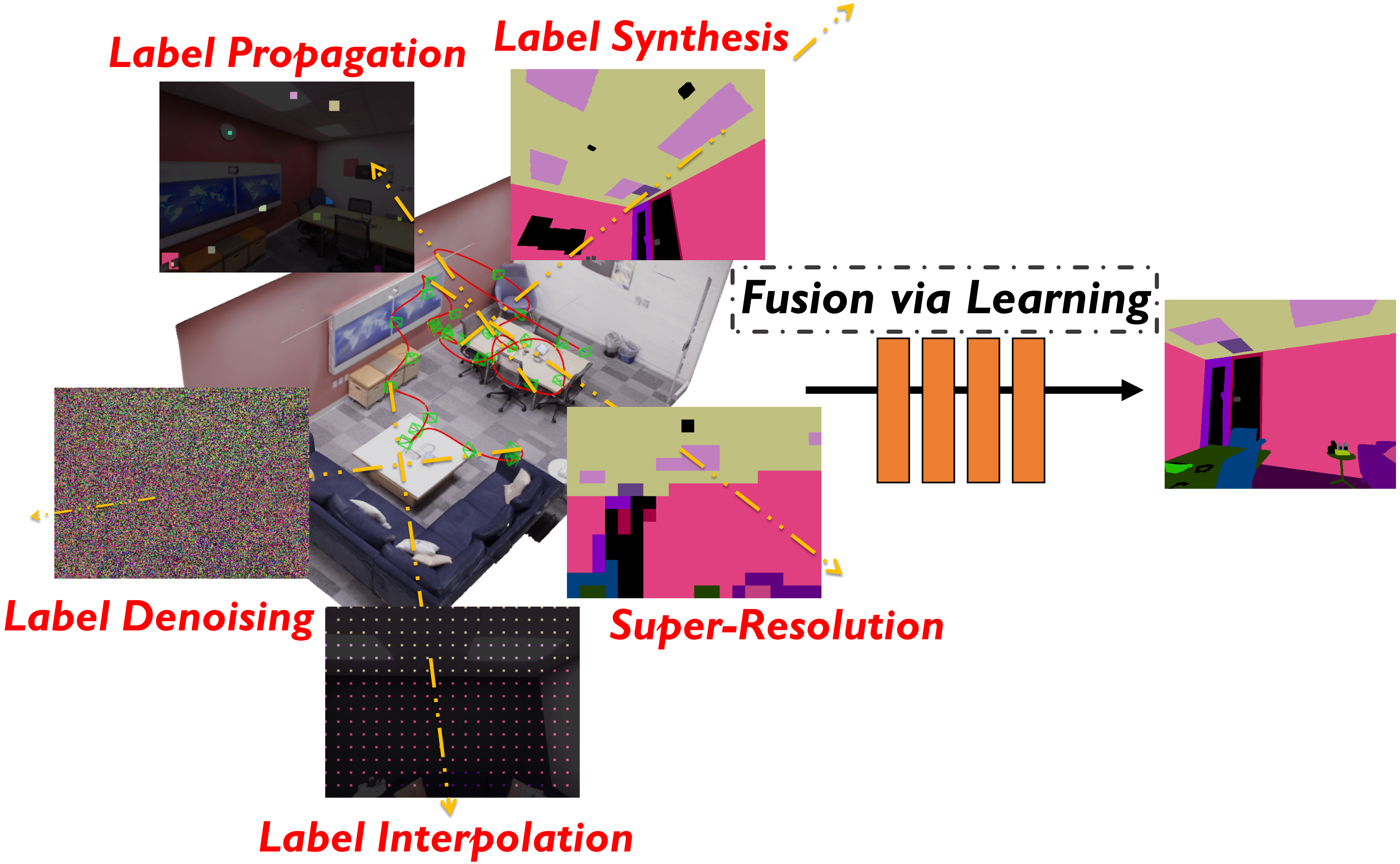
Colour:

$$\hat{\mathrm{C}}(\mathbf{r}) = \sum_{k=1}^{K} w_i c_i, \quad w_i = T_i(1 - \exp(-\sigma_i \delta_i))$$

Semantic:

$$\hat{\mathrm{S}}(\mathbf{r}) = \sum_{k=1}^{K} w_i s_i, \quad w_i = T_i(1 - \exp(-\sigma_i \delta_i))$$

Semantic-NeRF can fuse various types of annotations via training, leading to accurate dense labels.



Applications of Semantic-NeRF

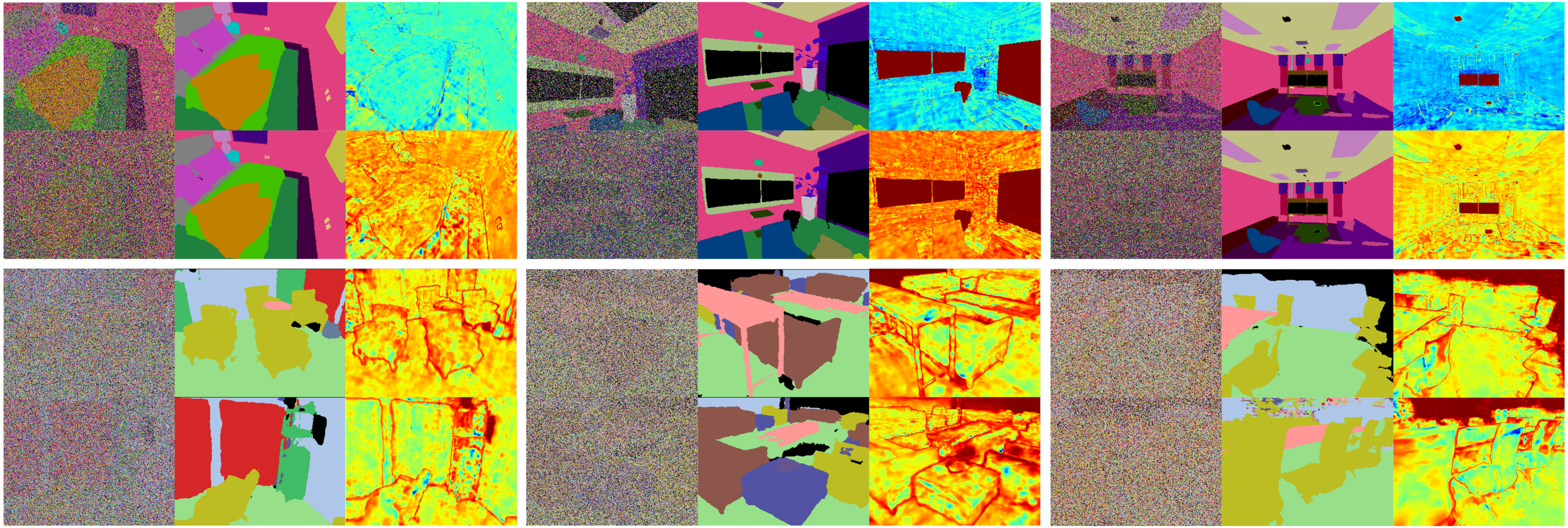
- Semantic View Synthesis with Sparse Labels
- Semantic Label Denoising
- Semantic Label Super-Resolution
- Semantic Label Propagation
- Multi-view Semantic Fusion
- Semantic 3D Reconstruction using Posed Images

Semantic View Synthesis with Sparse Labels

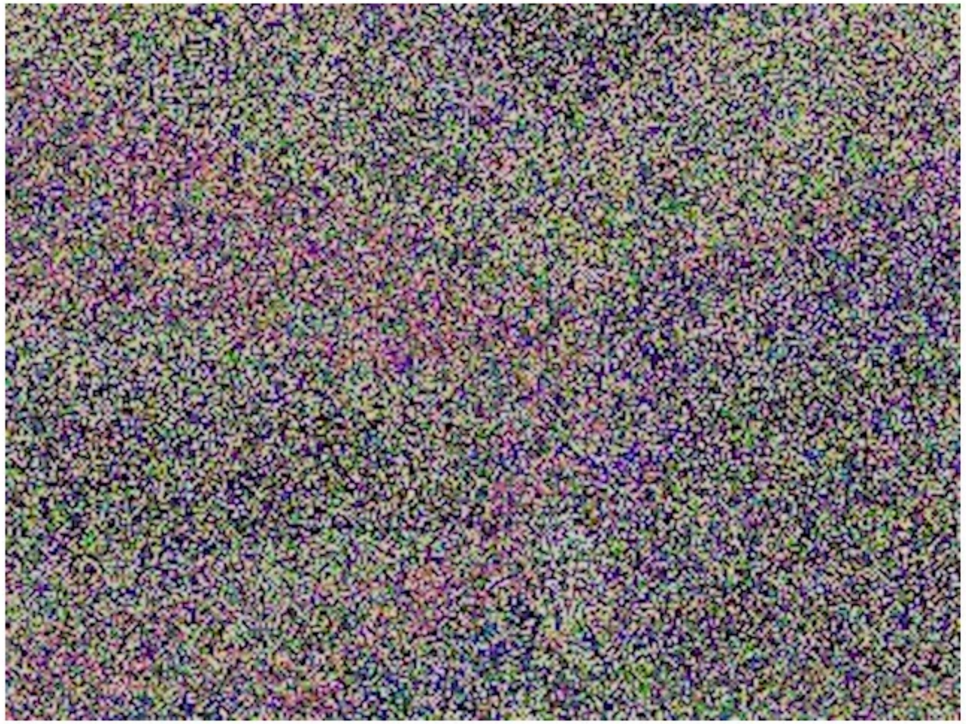
Ground Truth Rendering Entropy

Pixel-Wise Label Denoising

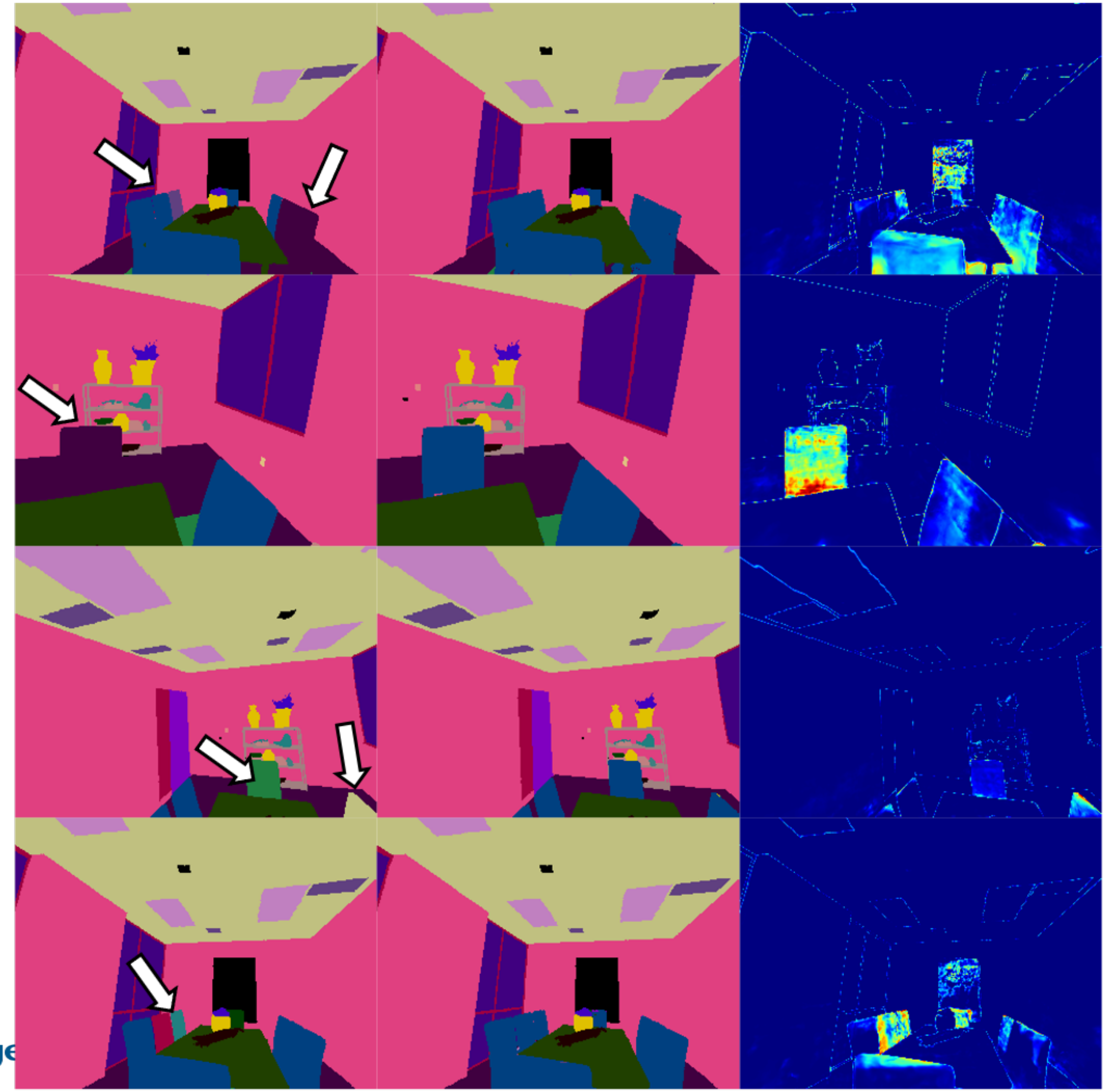
Within each block, from left to right are:
noisy training labels, denoised labels and entropy.



Pixel-Wise Label Denoising



Region-Wise Label Denoising



Semantic Label Super-Resolution

Semantic Label Propagation

Single-click per class/frame

Partial Label

Propagated Label

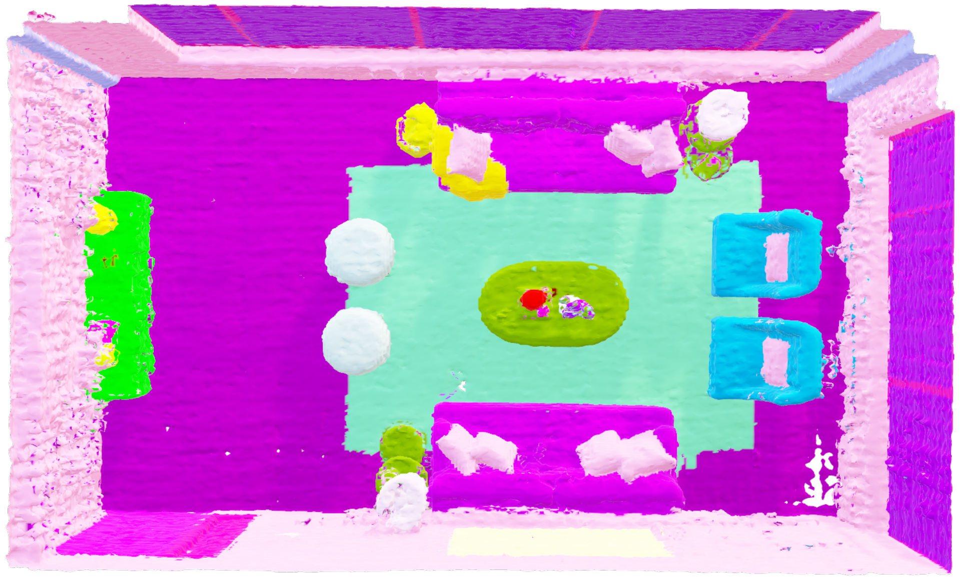
Multi-view Semantic Fusion

Can Semantic-NeRF improve monocular CNN predictions?

Semantic Fusion	mIoU	Avg Acc	Total Acc
Monocular	0.659	0.763	0.855
Bayesian Fusion *	0.668	0.764	0.865
Average Fusion *	0.586	0.703	0.814
Bayesian Fusion †	0.666	0.761	0.862
Average Fusion †	0.586	0.708	0.808
NeRF-Training (Ours)	**0.680**	**0.772**	**0.870**

* Using ground truth depth for data association.
† Using learned depth of Semantic-NeRF for data association.

Semantic 3D Reconstruction using Posed Images



Conclusion

- We have presented several methods to learn semantic scene representations using either external or in-place supervision.
- A monocular semantic mapping system and an online interactive scene understanding system are built on top of proposed representations.
- Better methods to describe intrinsic semantic error and higher efficiency of NeRF-like models are required to improve their practicability in real-world applications.
- Enabling mutual benefits of geometry and semantics is a promising direction.

Thanks!