Semantic 3D Reconstruction

Institute for Visual Computing
Department of Computer Science
ETH Zürich

Martin Oswald

Motivation

Scene Understanding & Accurate Mapping

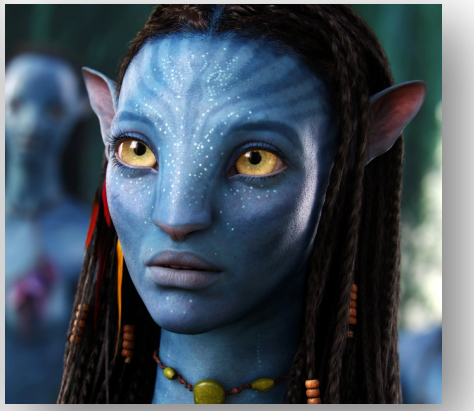
Robotics

3D Modeling

Realistic Content Creation

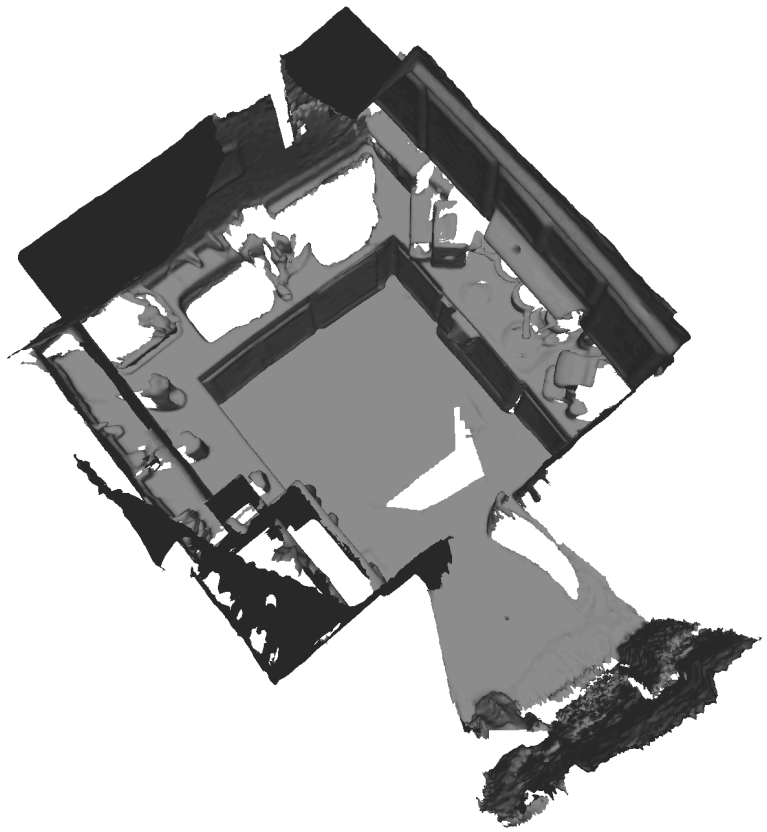
Computer Games

Virtual/Augmented Reality



Movies

Goals of Semantic 3D Reconstruction



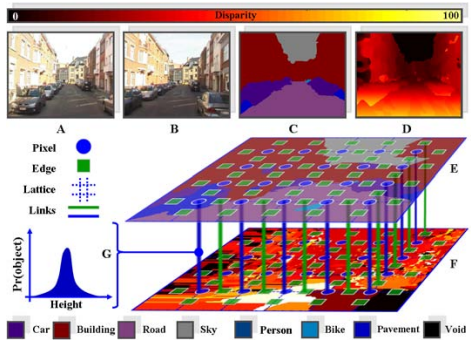
Geometry

Appearance

Semantics

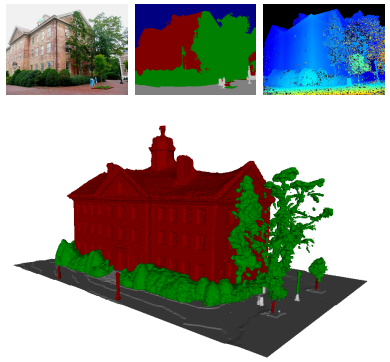
History of Semantic 3D Reconstruction

First combining geometry & depth



[Ladicky et al. BMVC]

First full 3D + joint optimization

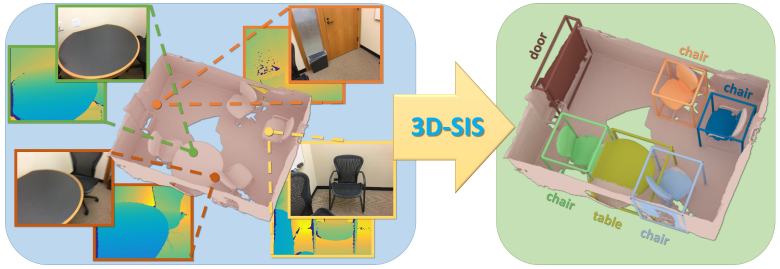


[Häne et al. CVPR]

ScanNet dataset

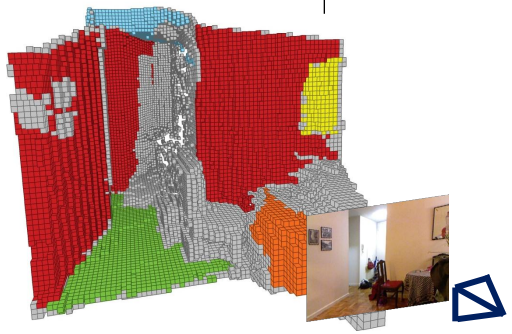
[Dai et al. CVPR]

3D Instance Segmentation

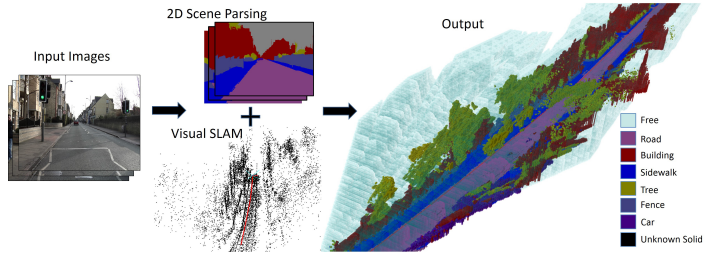


[Hou et al. CVPR]

2010 2013 2013 2014 2017 2017 2018 2019 **time**

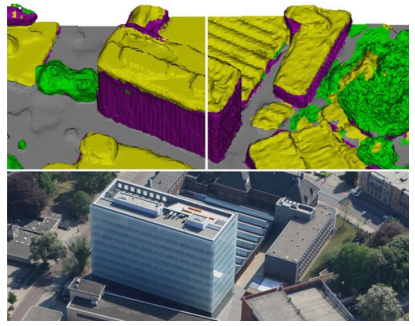


[Kim et al. ICCV]



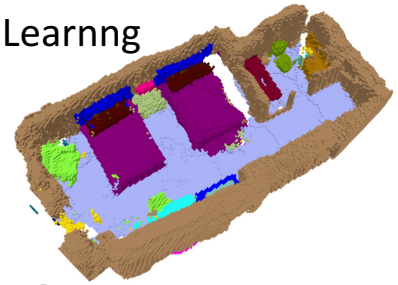
[Kundu et al. ECCV]

Semantic Refinement



[Blaha et al. ICCV]

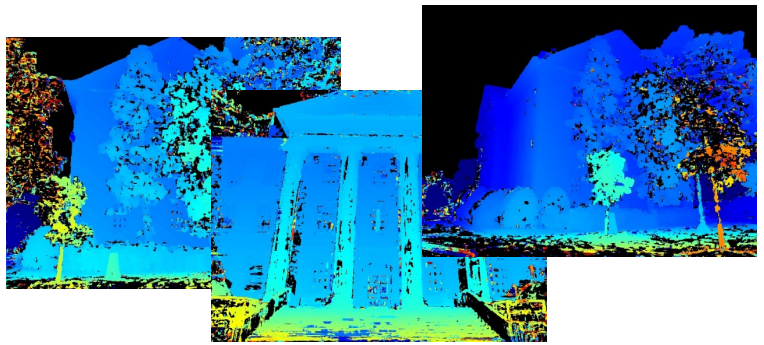
Surface Optimization + Learnng



[Cherabier et al. ECCV]

Semantic 3D Reconstruction

[Häne et al., CVPR 2013]

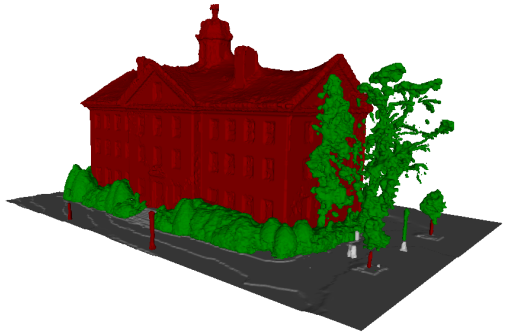


Class Likelihoods

Depth Maps

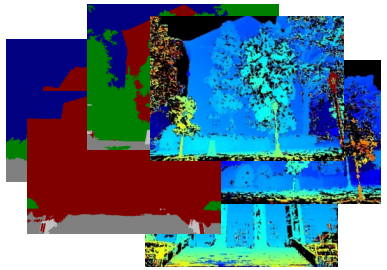
Joint Fusion, Convex Optimization

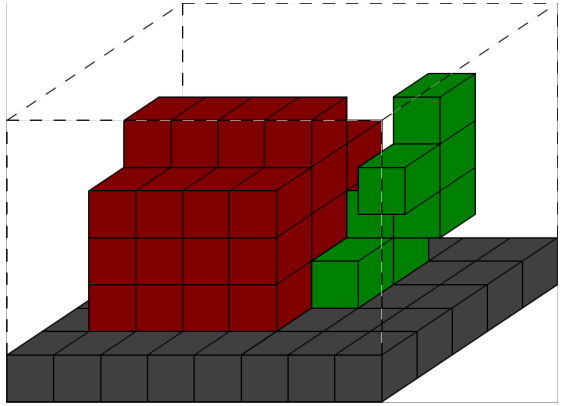
Dense Semantic 3D Model

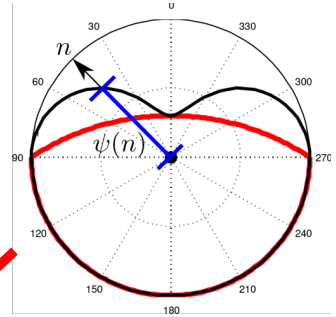


Semantic 3D Reconstruction

[Häne et al., CVPR 2013]







Data Term: Described as per-voxel unary potentials

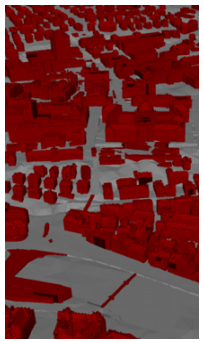
Regularization Term: Class-specific, direction dependent, surface area penalization

$$E(x) = \sum_{s \in \Omega} \left(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi^{ij}(x_s^{ij} - x_s^{ji}) \right)$$

$$\text{subject to } x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k$$

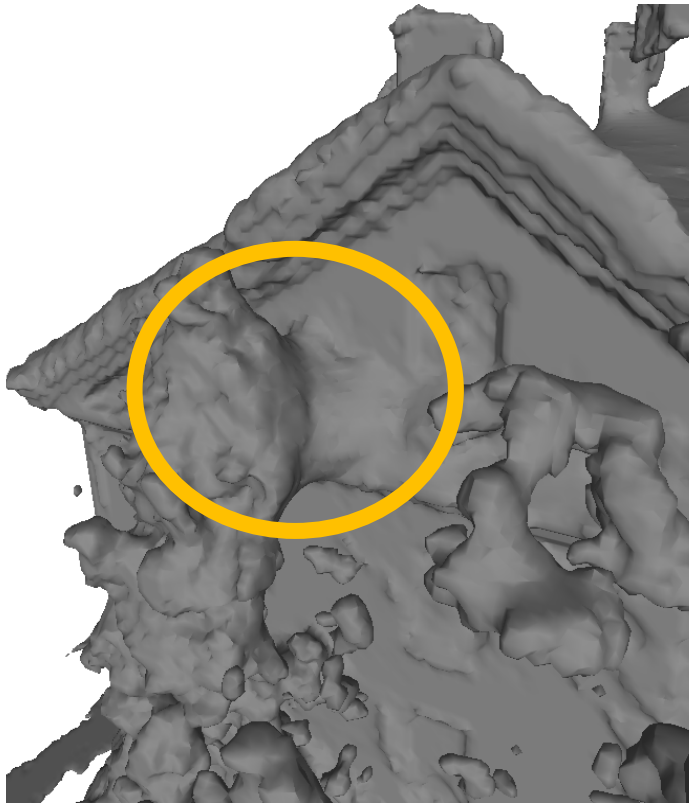
$$x_s^i \geq 0, \quad \sum_i x_s^i = 1, \quad x_s^{ij} \geq 0$$

↑ Learned from **training data**

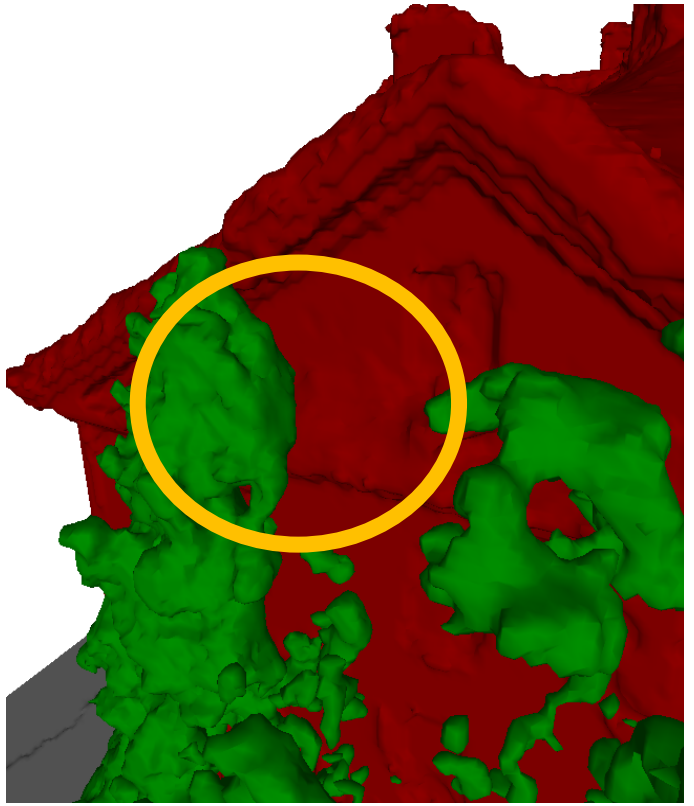


Semantic 3D Reconstruction

[Häne et al., CVPR 2013]



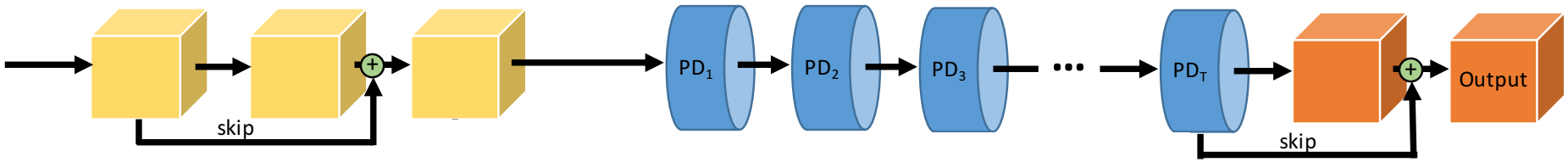
Geometry Only



Geometry + Semantics

Semantic 3D Reconstruction

[Cherabier, Schönberger et al., ECCV 2018]



Multi-label segmentation/
3D reconstruction

$$\underset{u}{\text{minimize}} \quad \int_{\Omega} \left(\|Wu\|_2 + fu \right) d\mathbf{x} \quad \text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum_{\ell} u_{\ell}(\mathbf{x}) = 1$$

Sattel-point problem

$$\underset{u}{\text{minimize}} \; \max_{\|\xi\|_{\infty} \leq 1} \langle Wu, \xi \rangle + \langle f, u \rangle + \nu \left(1 - \sum_{\ell} u_{\ell} \right)$$

Iterate primal-dual
update steps

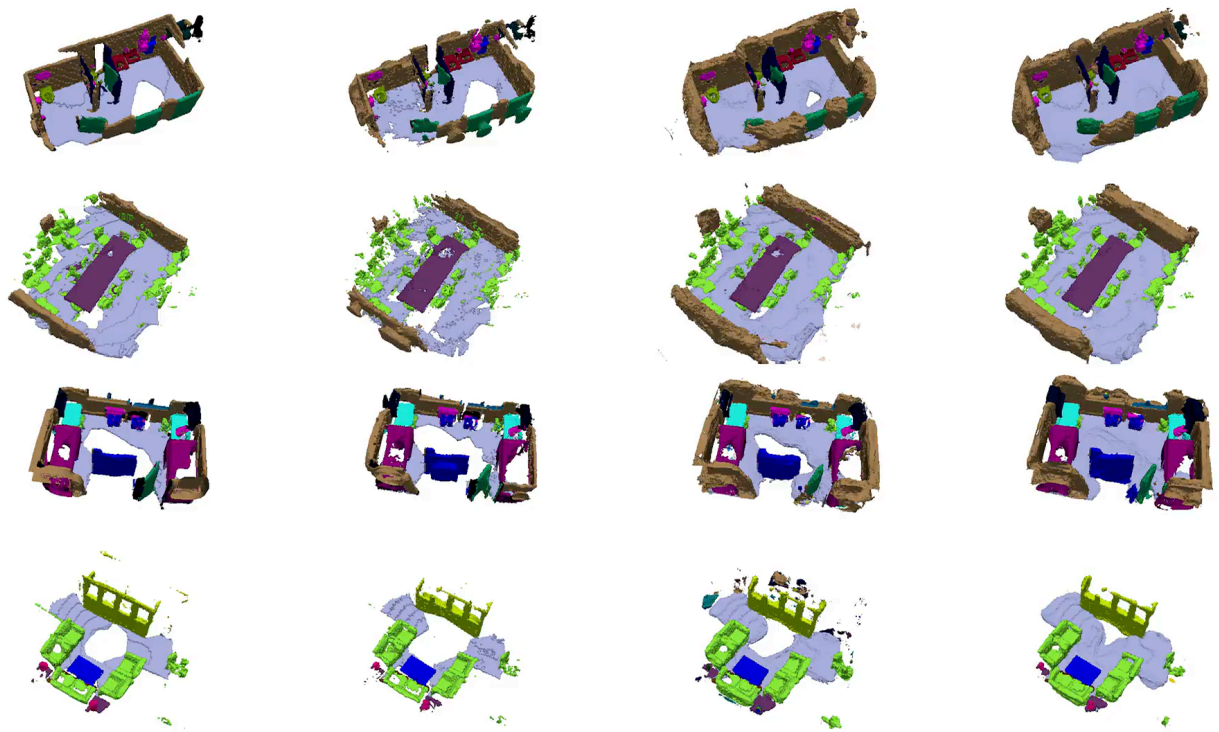
$$\nu^{t+1} = \nu^t + \sigma \left(\sum_{\ell} \bar{u}_{\ell}^t - 1 \right) \qquad u^{t+1} = \Pi_{[0,1]} \left[u^t + \tau (W^* \xi^{t+1} - f) \right]$$

$$\xi^{t+1} = \Pi_{\|\cdot\| \leq 1} \left[\xi^t + \sigma W \bar{u}^t \right] \qquad \bar{u}^{t+1} = 2u^{t+1} - u^t$$

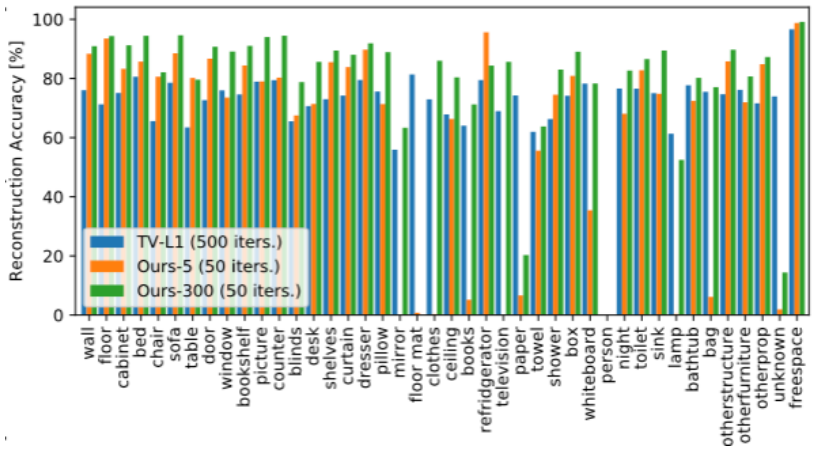
Semantic 3D Reconstruction

[Cherabier, Schönberger et al., ECCV 2018]

Results on ScanNet dataset (40 labels)



Ground Truth | TV-L1 (50 iter.) | Ours-5 (50 iter.) | Ours-300 (50 iter.)

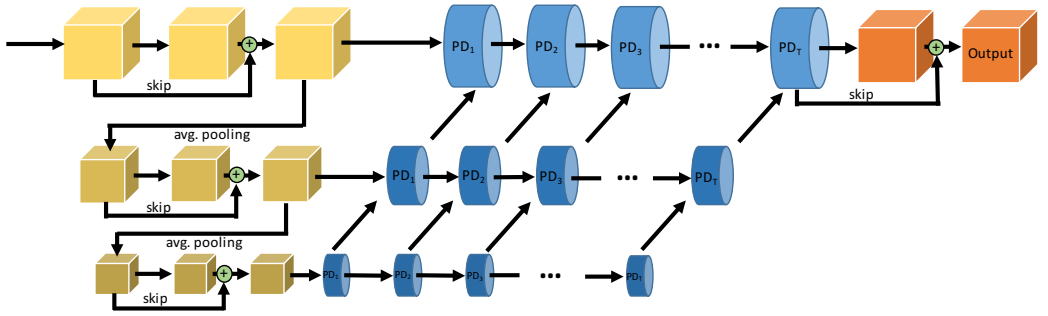


Methods	Overall	Freespace	Occupied	Semantic
Input data	59.8	39.1	99.7	68.4
TV-L1 (50 it.)	92.8	71.0	91.4	87.8
TV-L1 (500 it.)	95.8	86.4	92.3	88.5
C2F (50 it.)	21.0	26.7	99.9	31.4
Ours-5 (50 it.)	96.7	95.8	93.9	86.4
Ours-300 (0 it.)	97.3	97.6	92.3	90.2
Ours-300 (50 it.)	**98.7**	**98.6**	**94.4**	**91.5**

$$\underset{u}{\text{minimize}} \quad \int_{\Omega} \left(\|Wu\|_2 + fu \right) d\mathbf{x} \quad \text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum\nolimits_{\ell} u_{\ell}(\mathbf{x}) = 1$$

Multi-Sensor Depth Map Fusion

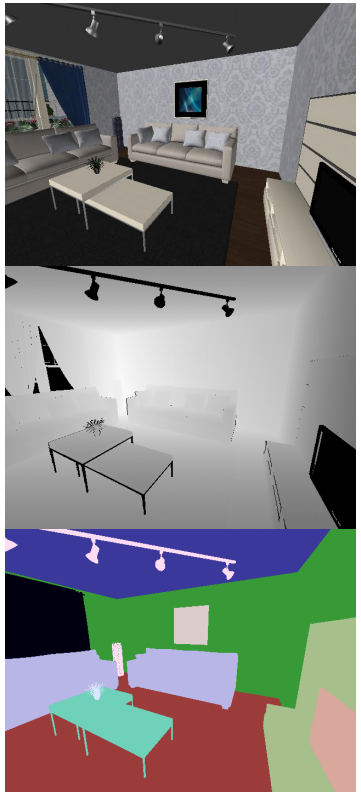
[Roszumnyi et al., ICCVW 2019]



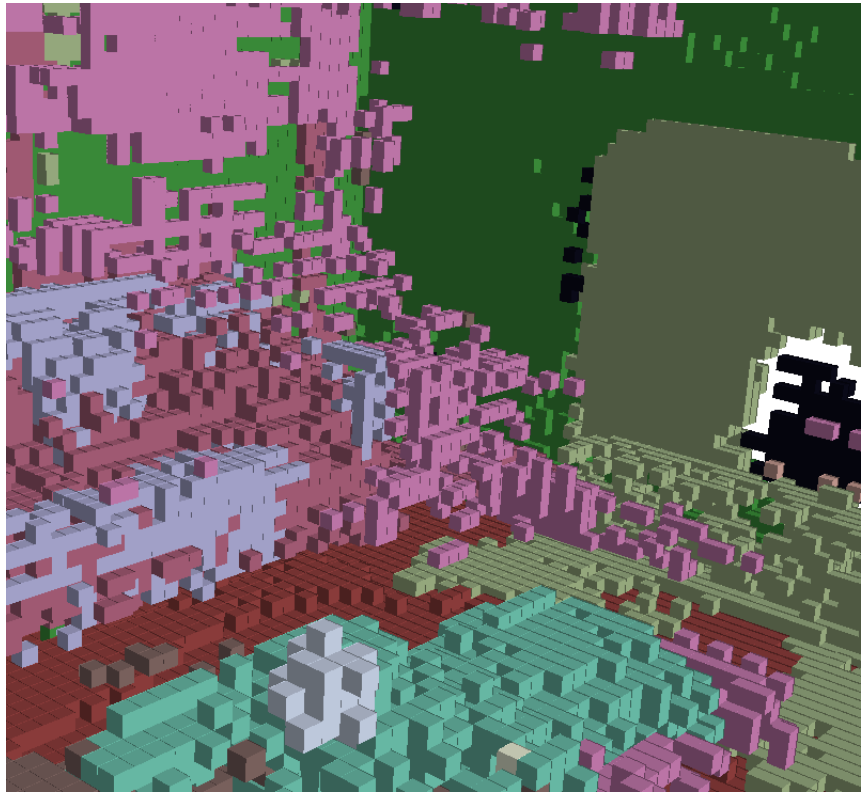
Semantic 3D Reconstruction

Multi-Sensor Depth Map Fusion

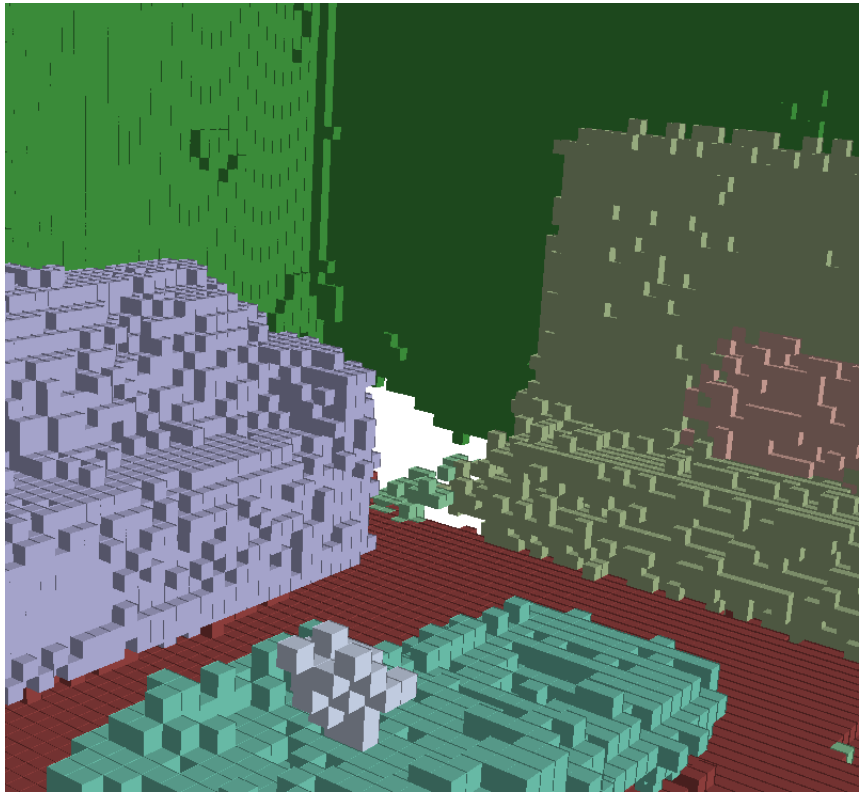
[Roszumnyi et al., ICCVW 2019]



Inputs



Standard TSDF Fusion + Semantics

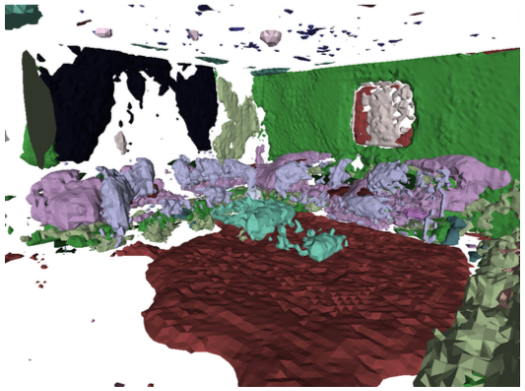


Learned Semantic Fusion

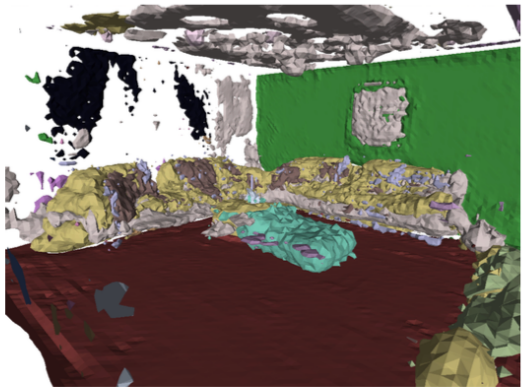
Multi-Sensor Depth Map Fusion

[Roszumnyi et al., ICCVW 2019]

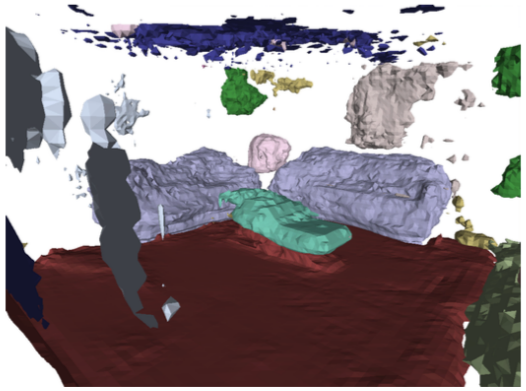
Expert System for Stero - Semantic Accuracy



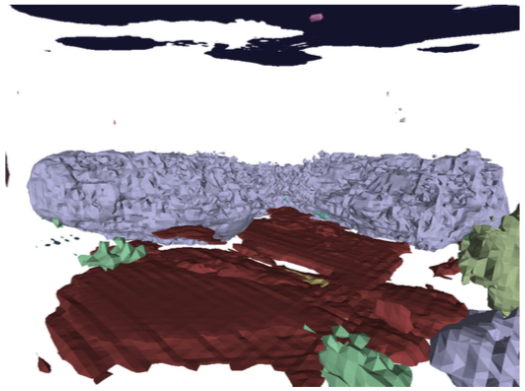
A) SGBM [4]: 0.71



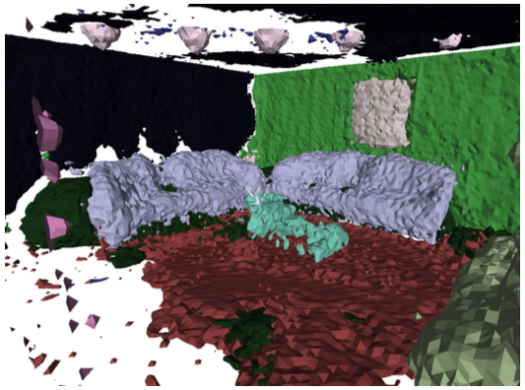
B) BM [1]: 0.71



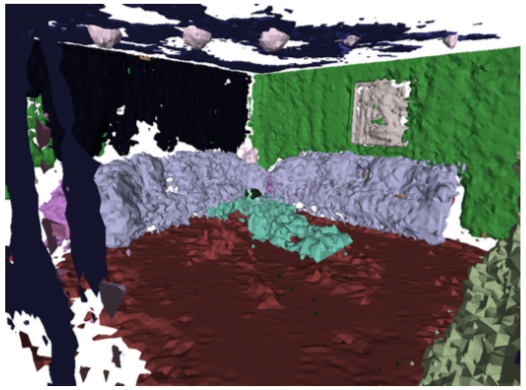
C) PSMNet [2]: 0.69



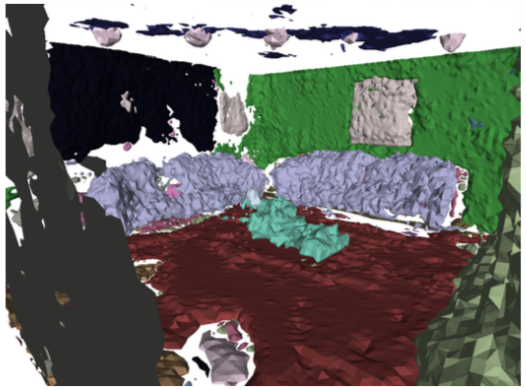
D) FCRN monocular [5]: 0.44



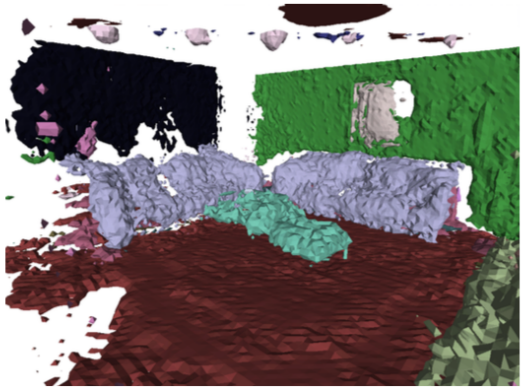
A+B+C (*d*): 0.72



A+B+C (*d*, *g*): 0.725



A+B+C (*d*, *g*, *n*): 0.735



A+B+C+D (*d*, *g*): 0.73

3D Instance Segmentation

[Lahoud et al., ICCV 2019]

Input Scene

3D Instance Segmentation

Type 1

Adopt 2D instance label from the 2D images

Type 2

Bounding box prediction then mask

Type 3

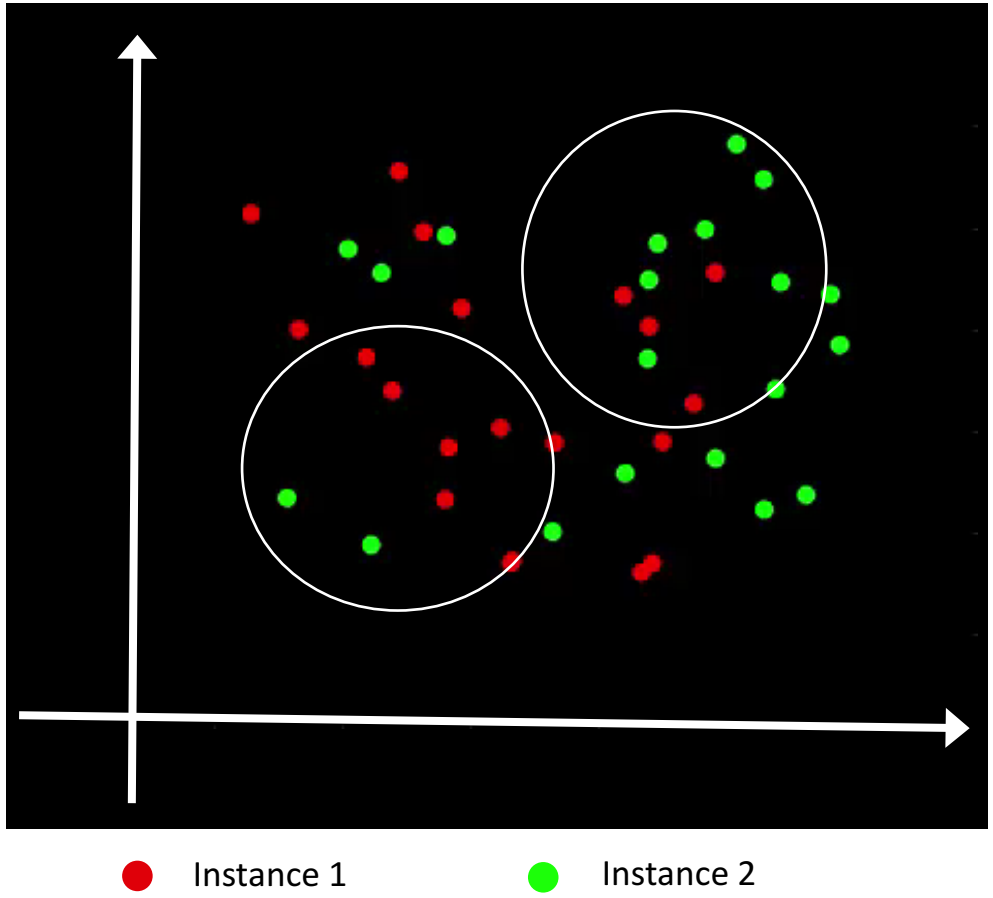
Learn a feature embedding (3D point, voxel,…)

Metric Feature Learning

Learn a feature embedding which groups parts of the same object

[SSC-Net: Song S et al. *Semantic scene completion from a single depth image, CVPR 2017*]

Learn Feature Embedding



Introduced for 2D instance segmentation by [De Brabandere et al., ArXiv 2017]

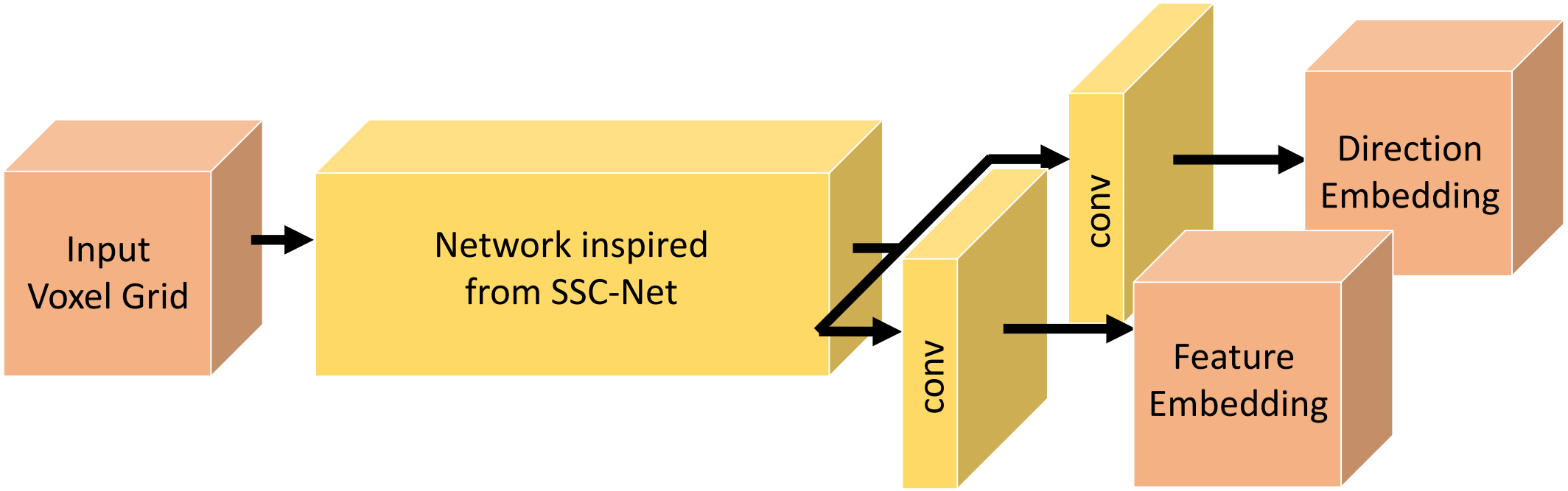
Applied to 3D point clouds in concurrent work of [Pham et al., CVPR 2019]

Difficult to learn with:

(1) Larger areas with many instances

(2) Similar instances outside the receptive field

Multi-task Metric Feature Learning

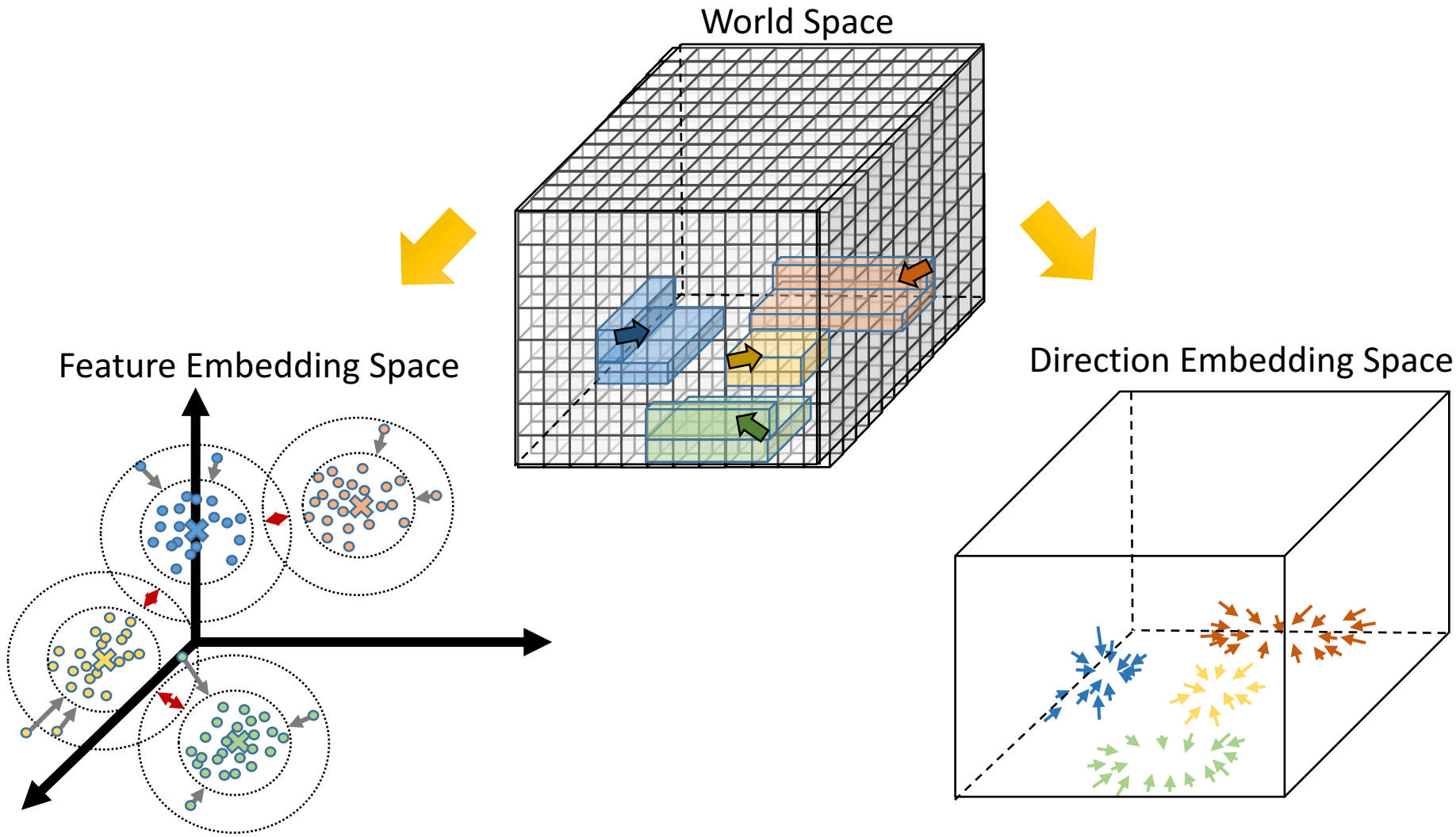


3D instance-labeling problem as a multi-task learning strategy:

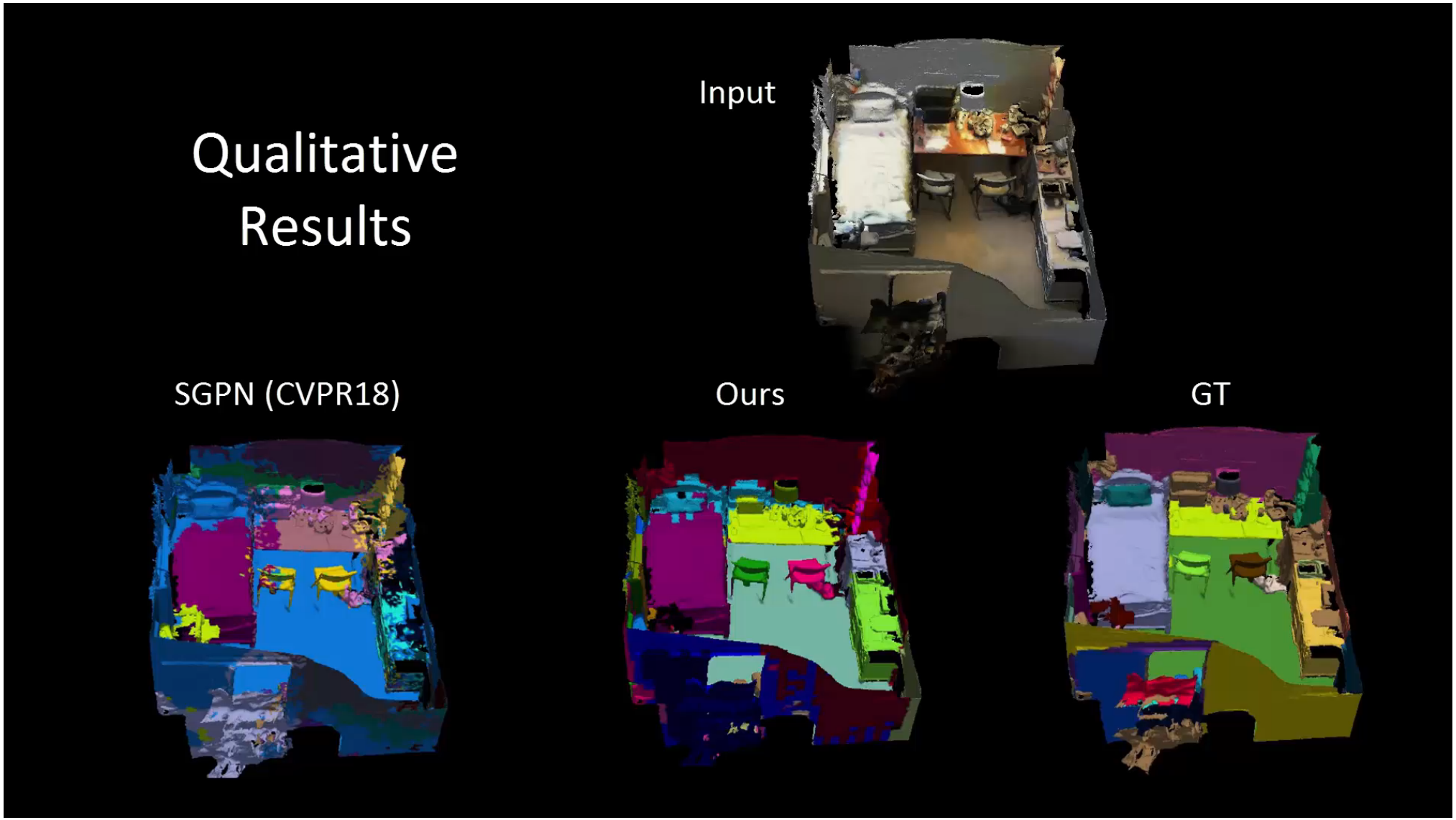
(1) Learn a feature embedding which groups parts of the same object

(2) Estimate directional information of the instances' center of mass

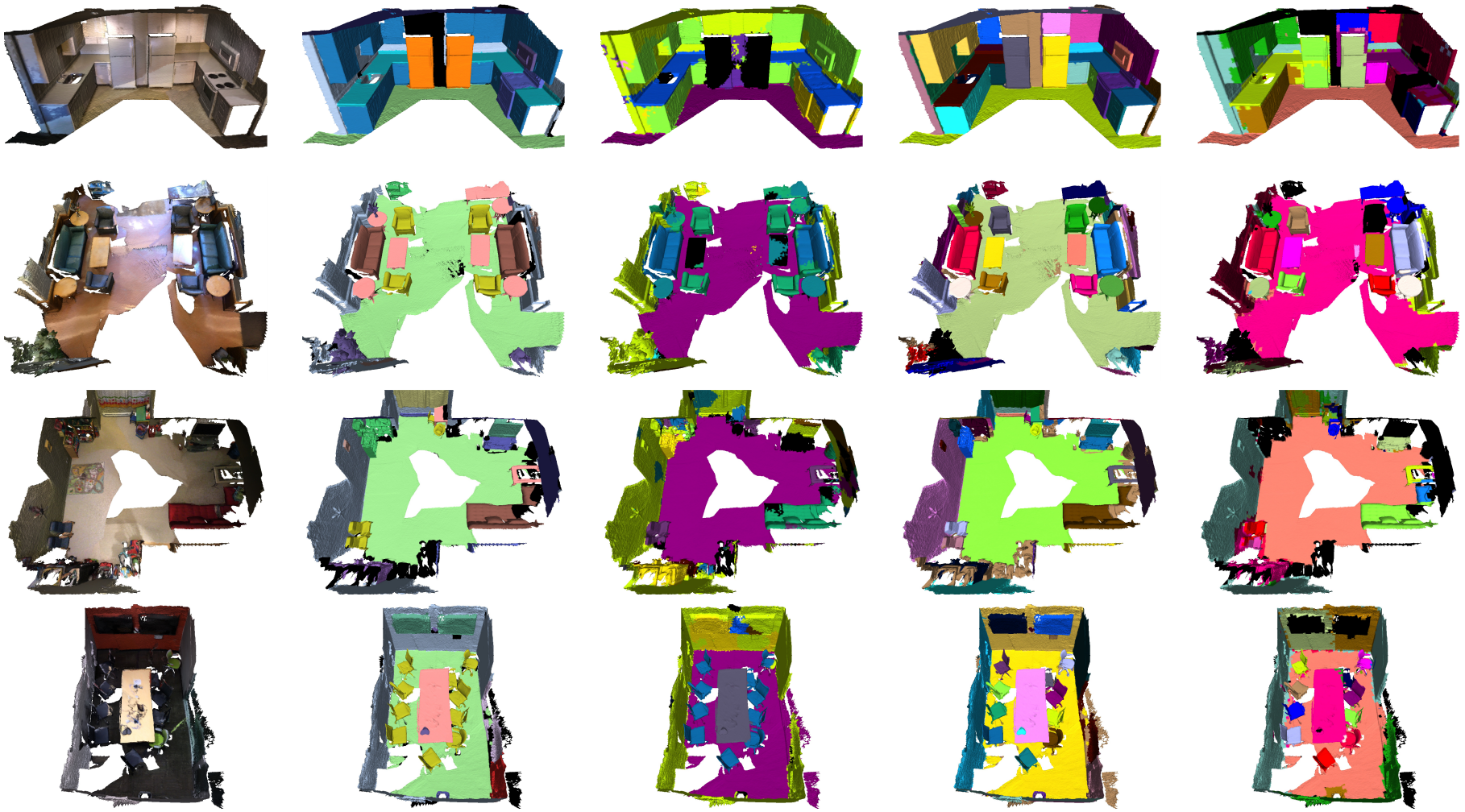
Multi-task Metric Feature Learning



3D Instance Labeling Results

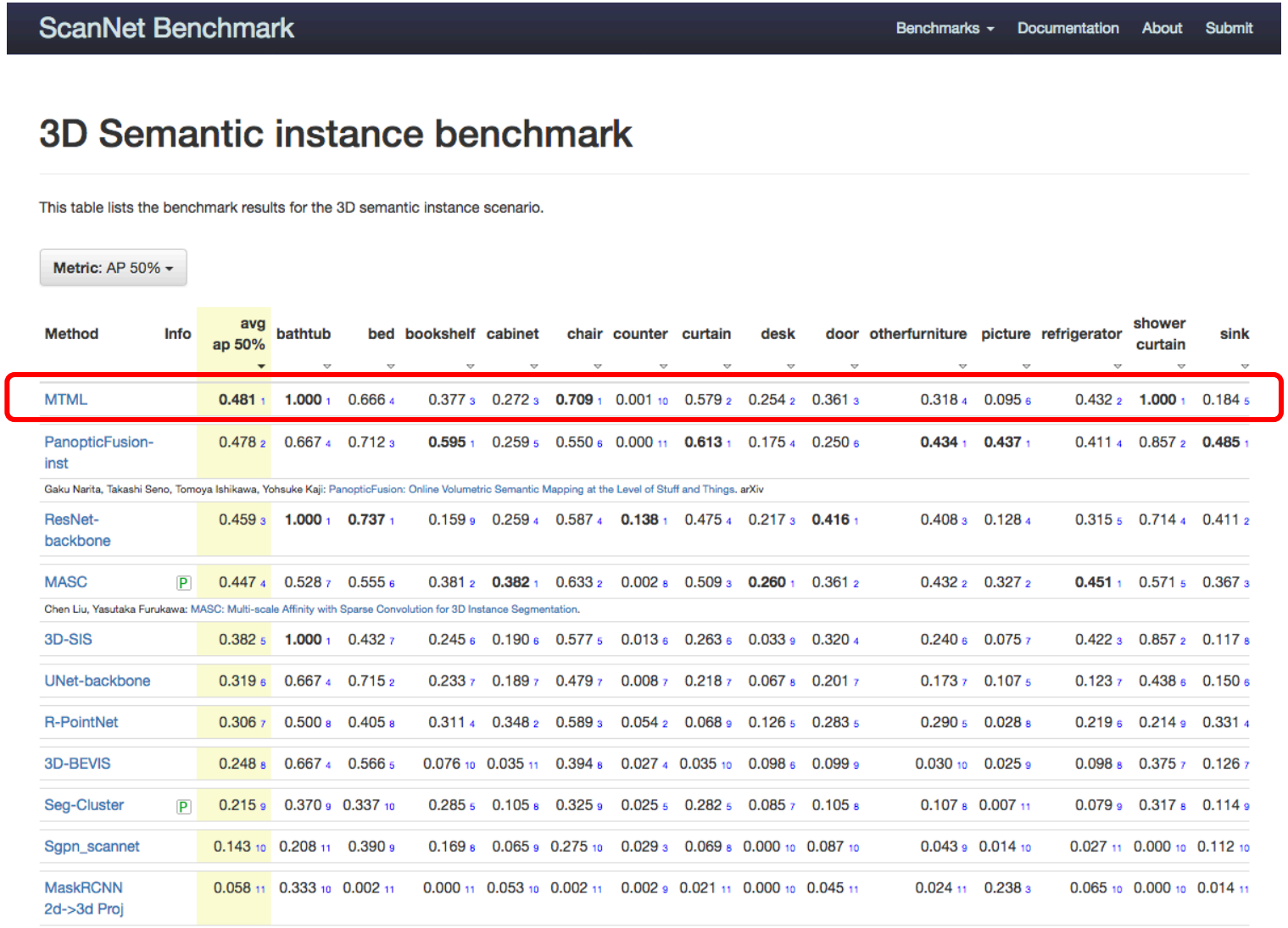


3D Instance Labeling Results



Input (RGB) | Semantic GT | SparseConvNet [14] | Instance GT | Ours

3D Instance Labeling Results



ScanNet Benchmark — Benchmarks ▾ Documentation About Submit

3D Semantic instance benchmark

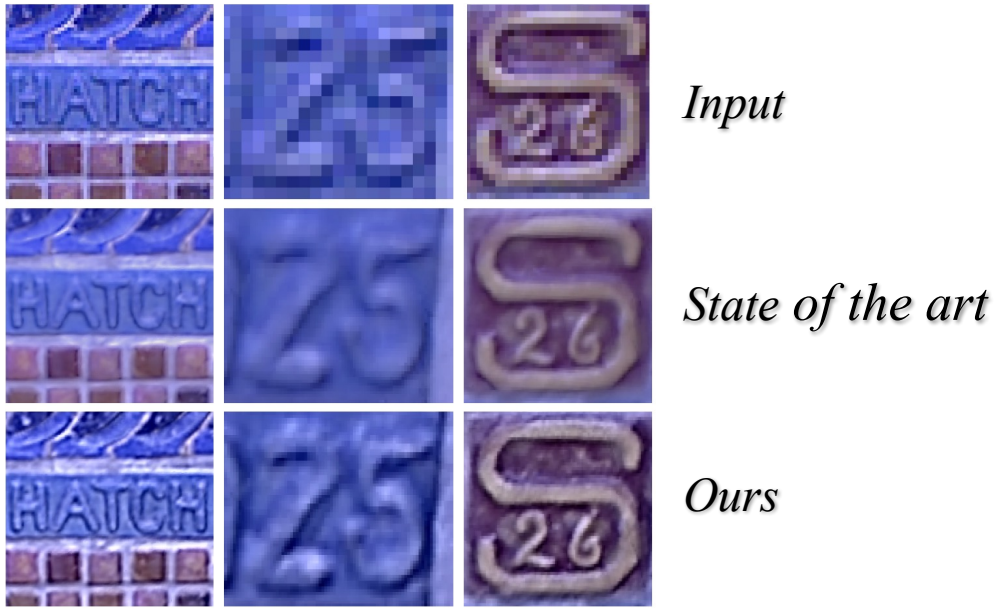
This table lists the benchmark results for the 3D semantic instance scenario.

Metric: AP 50% ▾

Method	Info	avg ap 50%	bathtub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	otherfurniture	picture	refrigerator	shower curtain	sink
MTML		**0.481** 1	**1.000** 1	0.666 4	0.377 3	0.272 3	**0.709** 1	0.001 10	0.579 2	0.254 2	0.361 3	0.318 4	0.095 6	0.432 2	**1.000** 1	0.184 5
PanopticFusion-inst		0.478 2	0.667 4	0.712 3	**0.595** 1	0.259 5	0.550 6	0.000 11	**0.613** 1	0.175 4	0.250 6	**0.434** 1	**0.437** 1	0.411 4	0.857 2	**0.485** 1
Gaku Narita, Takashi Seno, Tomoya Ishikawa, Yohsuke Kaji: PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things. arXiv																
ResNet-backbone		0.459 3	**1.000** 1	**0.737** 1	0.159 9	0.259 4	0.587 4	**0.138** 1	0.475 4	0.217 3	**0.416** 1	0.408 3	0.128 4	0.315 5	0.714 4	0.411 2
MASC	P	0.447 4	0.528 7	0.555 6	0.381 2	**0.382** 1	0.633 2	0.002 8	0.509 3	**0.260** 1	0.361 2	0.432 2	0.327 2	**0.451** 1	0.571 5	0.367 3
Chen Liu, Yasutaka Furukawa: MASC: Multi-scale Affinity with Sparse Convolution for 3D Instance Segmentation.																
3D-SIS		0.382 5	**1.000** 1	0.432 7	0.245 6	0.190 6	0.577 5	0.013 6	0.263 6	0.033 9	0.320 4	0.240 6	0.075 7	0.422 3	0.857 2	0.117 8
UNet-backbone		0.319 6	0.667 4	0.715 2	0.233 7	0.189 7	0.479 7	0.008 7	0.218 7	0.067 8	0.201 7	0.173 7	0.107 5	0.123 7	0.438 6	0.150 6
R-PointNet		0.306 7	0.500 8	0.405 8	0.311 4	0.348 2	0.589 3	0.054 2	0.068 9	0.126 5	0.283 5	0.290 5	0.028 8	0.219 6	0.214 9	0.331 4
3D-BEVIS		0.248 8	0.667 4	0.566 5	0.076 10	0.035 11	0.394 8	0.027 4	0.035 10	0.098 6	0.099 9	0.030 10	0.025 9	0.098 8	0.375 7	0.126 7
Seg-Cluster	P	0.215 9	0.370 9	0.337 10	0.285 5	0.105 8	0.325 9	0.025 5	0.282 5	0.085 7	0.105 8	0.107 8	0.007 11	0.079 9	0.317 8	0.114 9
Sgpn_scannet		0.143 10	0.208 11	0.390 9	0.169 8	0.065 9	0.275 10	0.029 3	0.069 8	0.000 10	0.087 10	0.043 9	0.014 10	0.027 11	0.000 10	0.112 10
MaskRCNN 2d->3d Proj		0.058 11	0.333 10	0.002 11	0.000 11	0.053 10	0.002 11	0.002 9	0.021 11	0.000 10	0.045 11	0.024 11	0.238 3	0.065 10	0.000 10	0.014 11

Learned Multi-View Texture Super-Resolution

Audrey Richard, Ian Cherabier, Martin R. Oswald, Vagia Tsiminaki,
Marc Pollefeys, Konrad Schindler



Input

State of the art

Ours

ETH zürich
IBM Research | Zurich
Microsoft

Appearance Modeling

- Reconstruct accurate 3D geometry → *3D Modeling*
- Generate high-fidelity surface texture → ***Appearance Modeling***

Limited to the resolution of the input images !

Appearance Modeling

- Reconstruct accurate 3D geometry → *3D Modeling*
- Generate high-fidelity surface texture → ***Appearance Modeling***

→ **We need super-resolution (SR) techniques that exploit multi-view information**

Super-resolution (SR)

View redundancy

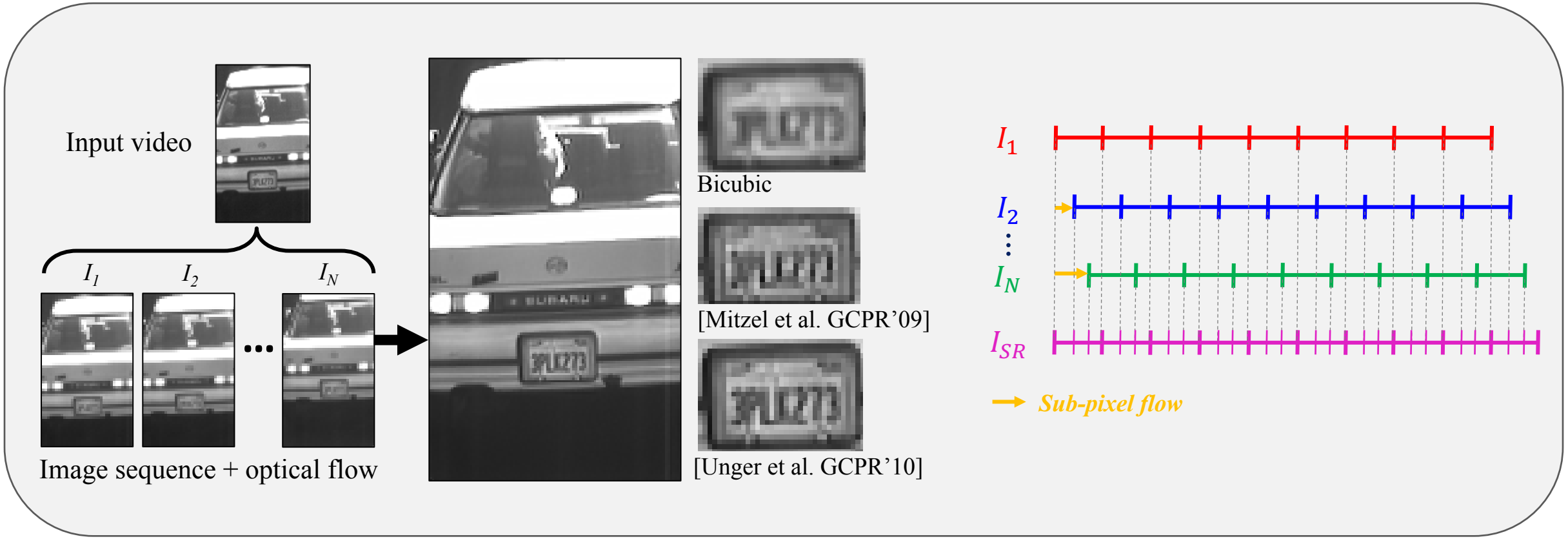
2D Images

Properties

- Leveraging input data (oversampling)

Super-resolution (SR)

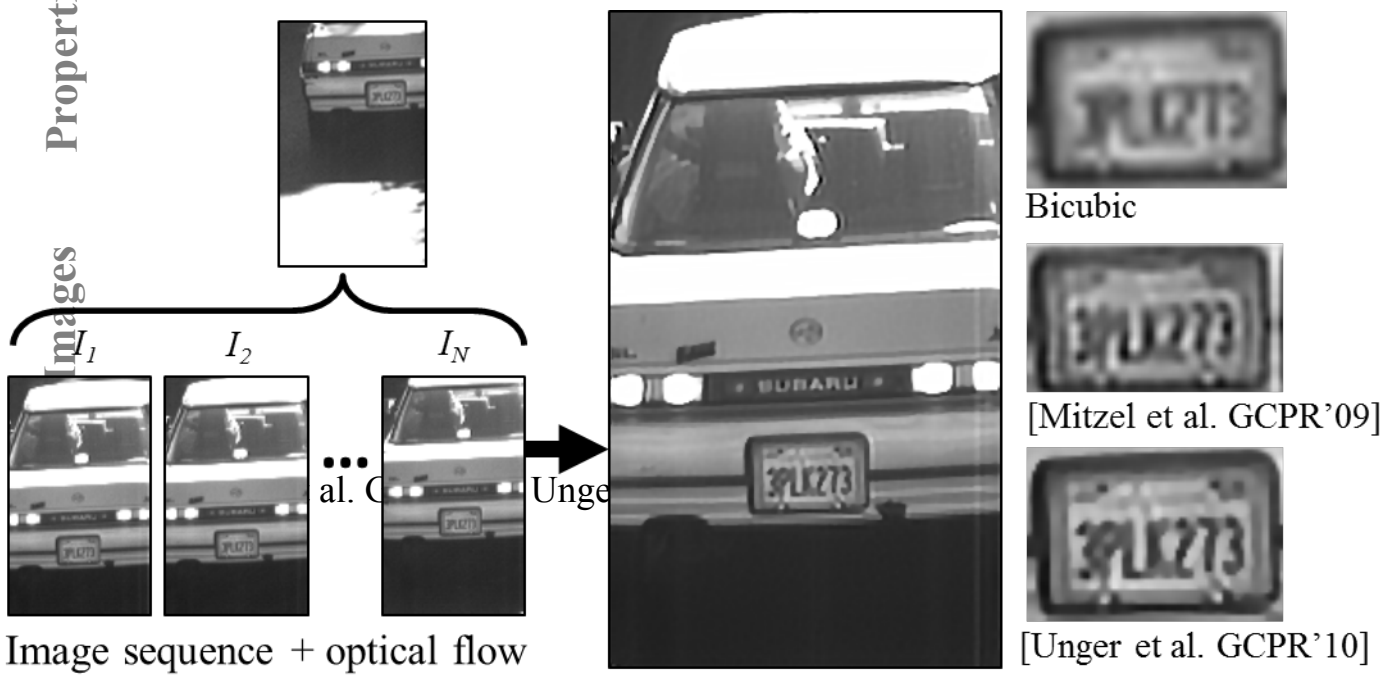
View redundancy

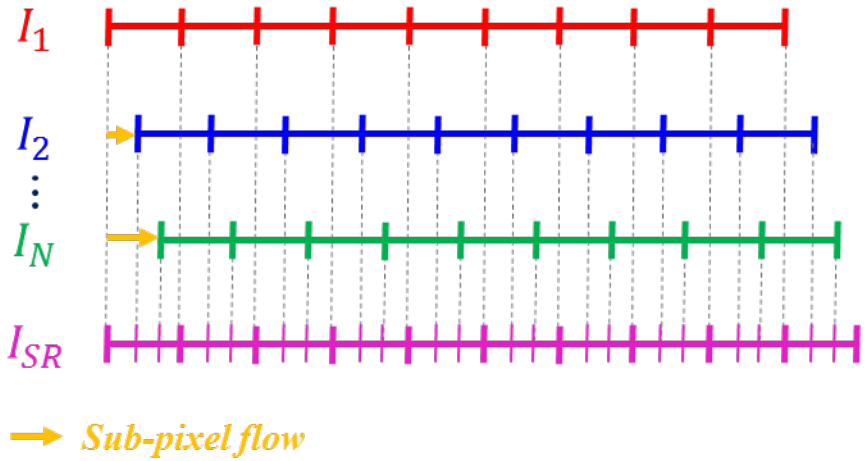


Super-resolution (SR)

View redundancy

- Leveraging input data (oversampling)





Super-resolution (SR)

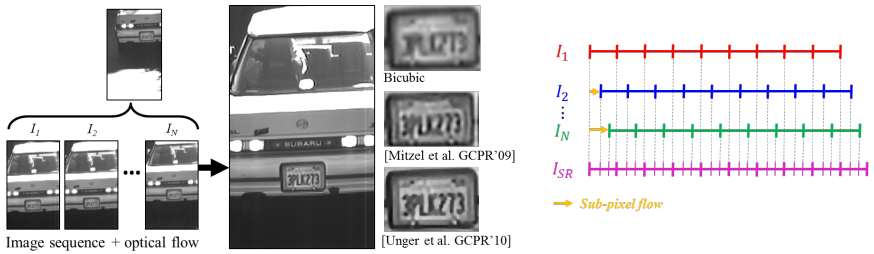
Properties

2D Images

Texture

View redundancy

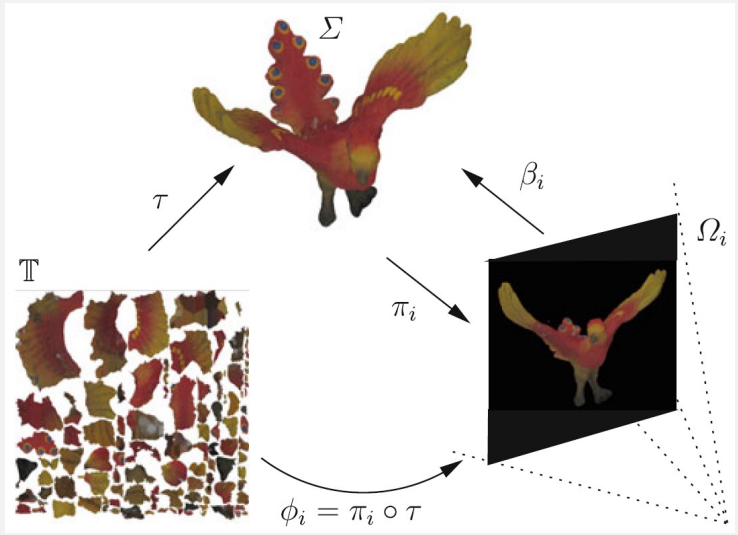
- Leveraging input data (oversampling)

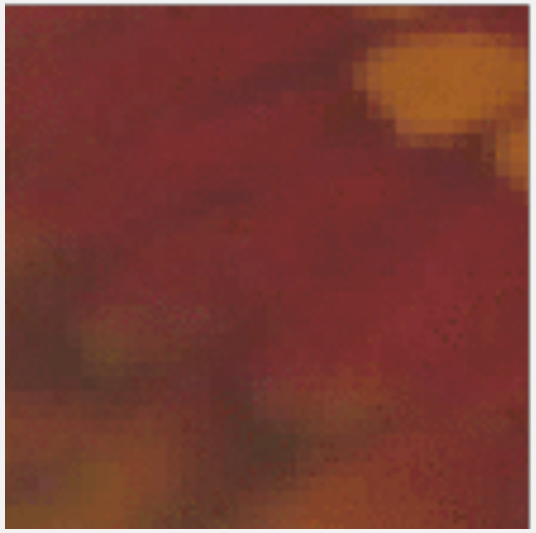


[Mitzel et al. GCPR'09, Unger et al. GCPR'10]

Super-resolution (SR)

View redundancy





Input image close-up

Weighted average

Super-resolution result

Super-resolution (SR)

View redundancy

- Leveraging input data (oversampling)

[Goldluecke et al. ICCV'09, Tsiminaki et al. CVPR'14]

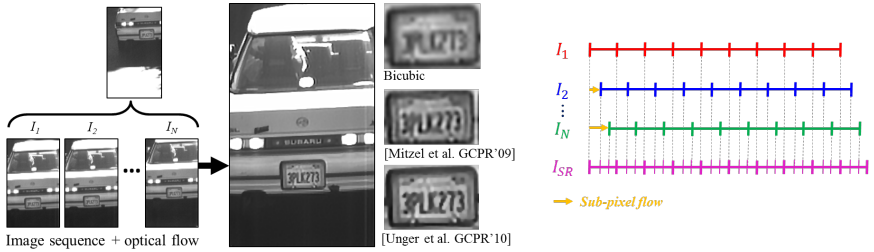
Super-resolution (SR)

View redundancy

Properties

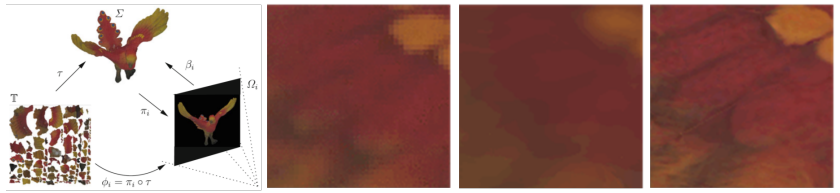
- Leveraging input data (oversampling)
- Typically x2 or x4 upsampling factor
- **Does not work for missing data**

2D Images



[Mitzel et al. GCPR'09, Unger et al. GCPR'10]

Texture



[Goldluecke et al. ICCV'09, Tsiminaki et al. CVPR'14]

Prior knowledge

- Learn structure and shape properties from datasets

Super-resolution (SR)

Prior knowledge

bicubic
(21.59dB/0.6423)

SRResNet
(23.53dB/0.7832)

SRGAN
(21.15dB/0.6868)

original

Super-resolution (SR)

View redundancy

- Leveraging input data (oversampling)
- Typically x2 or x4 upsampling factor

Properties

2D Images

Texture

[Goldluecke et al. ICCV'09, Tsiminaki et al. CVPR'14]

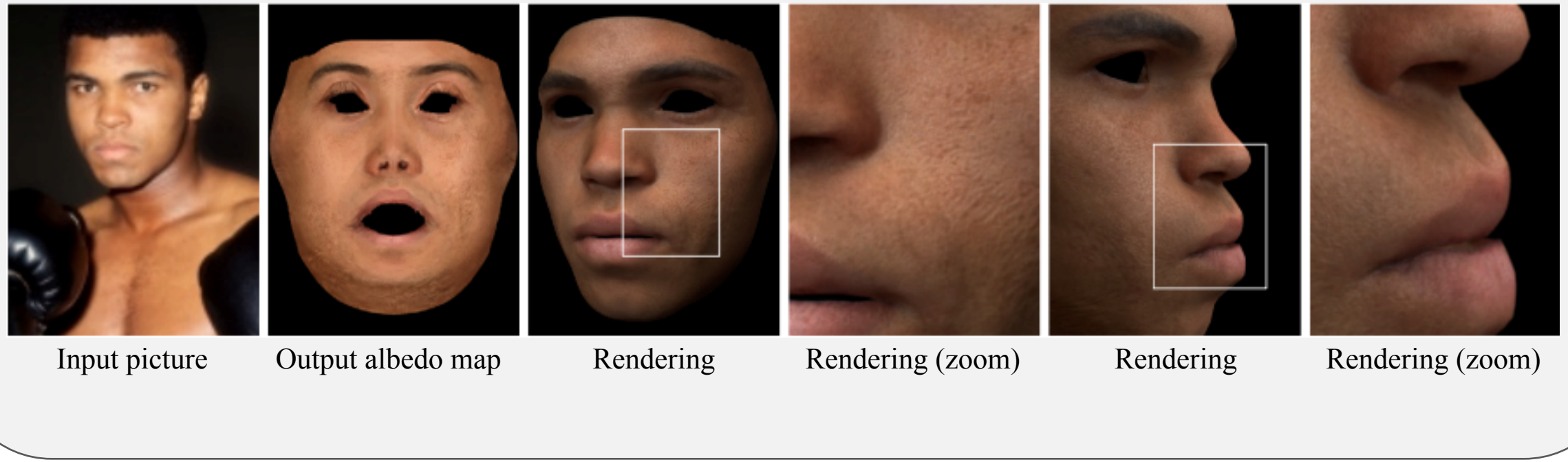
Prior knowledge

- Learn structure and shape properties from datasets

[..., Hu... ...al. ACCV'14, etc.]

Super-resolution (SR)

Prior knowledge



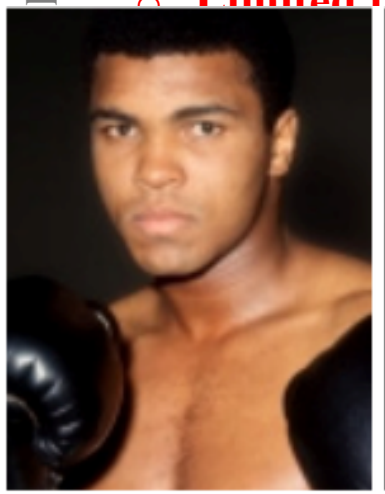
Input picture | Output albedo map | Rendering | Rendering (zoom) | Rendering | Rendering (zoom)

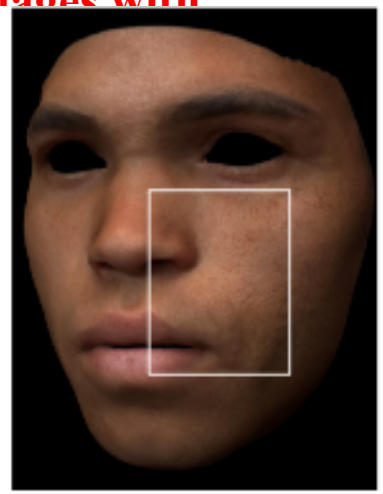
Super-resolution (SR)

View redundancy

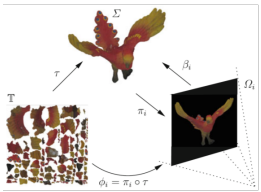
- Leveraging input data (oversampling)
- Typically x2 or x4 upsampling factor
- Limited to fixed number images with

perties





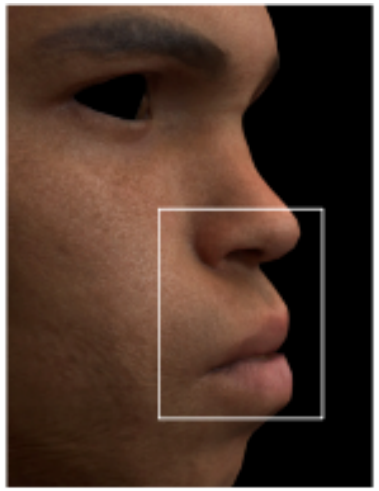
Texture

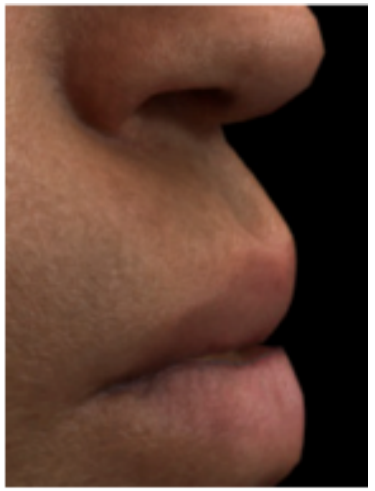


[Goldluecke et al. ICCV'09, Tsiminaki et al. CVPR'14]

Prior knowledge

- Learn structure and shape properties from datasets





[Saito et al. CVPR'17, Huynh et al. CVPR'18]

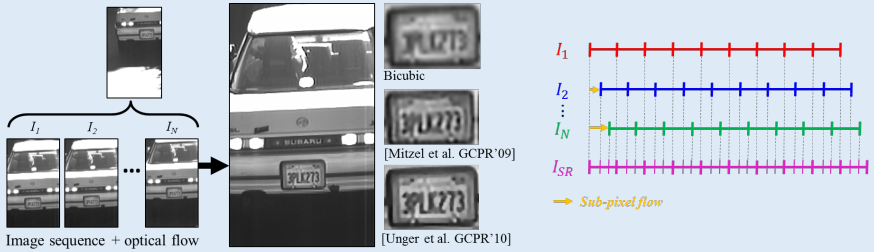
Super-resolution (SR)

Properties

2D Images

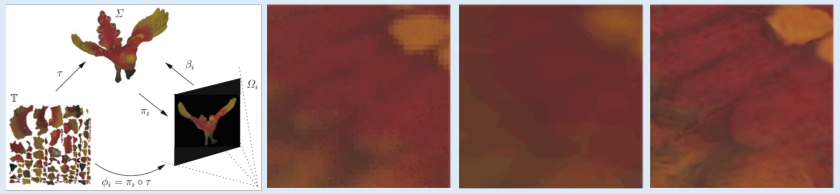
View redundancy

- Leveraging input data (oversampling)
- Typically x2 or x4 upsampling factor
- **Does not work for missing data**



[Mitzel et al. GCPR'09, Unger et al. GCPR'10]

Texture

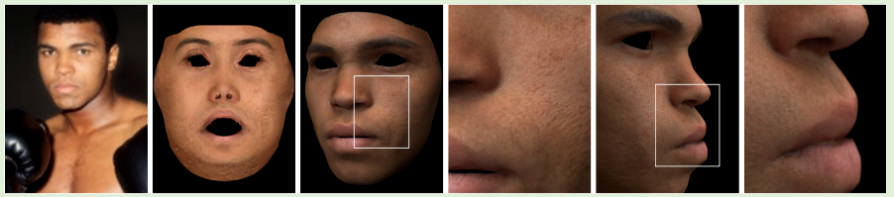


[Goldluecke et al. ICCV'09, Tsiminaki et al. CVPR'14]

Prior knowledge

- Learn structure and shape properties from datasets
- Typically x2 – x4 upsampling factor
- **Educated guess, not based on target scene**

[Ledig et al. CVPR'17, Hui et al. CVPR'18, Timofte et al., etc.]



[Saito et al. CVPR'17, Huynh et al. CVPR'18]

Our Approach

LR input images

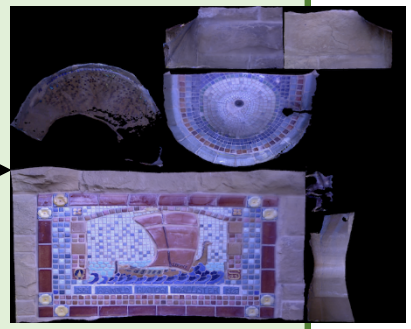
Multi-view Aggregation Subnet (MVA)

Based on [Tsiminaki et al. CVPR'14]

Intermediate HR atlas

Single-image Prior Subnet (SIP)

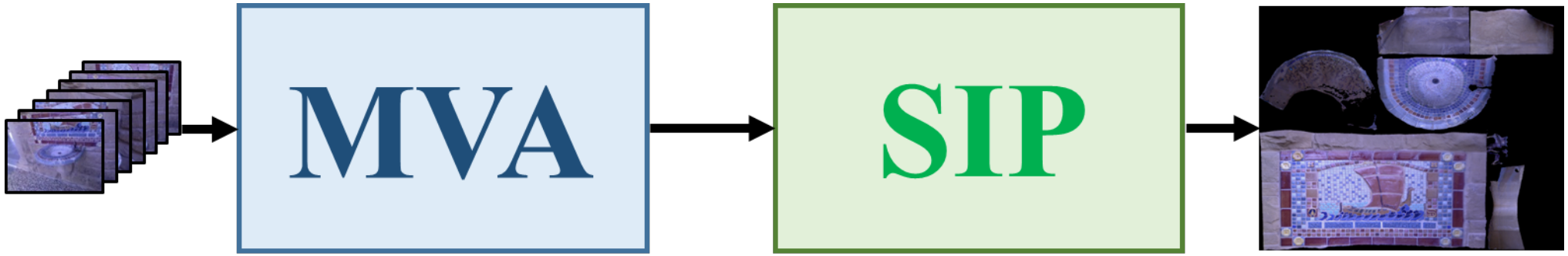
Based on [Hui et al. CVPR'18]



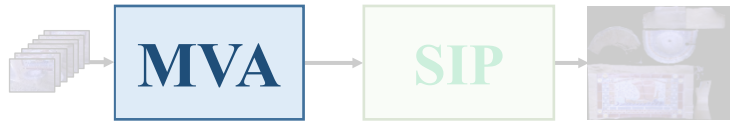
HR atlas (super-resolved)

Unify the two SR concepts

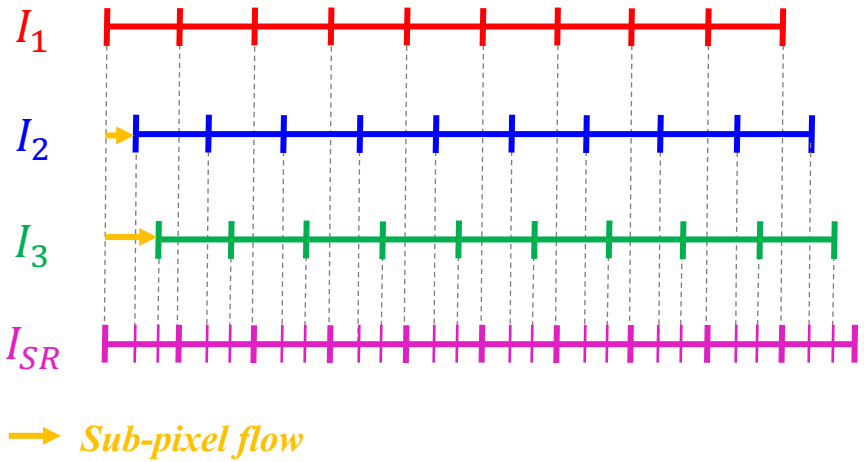
Our Approach

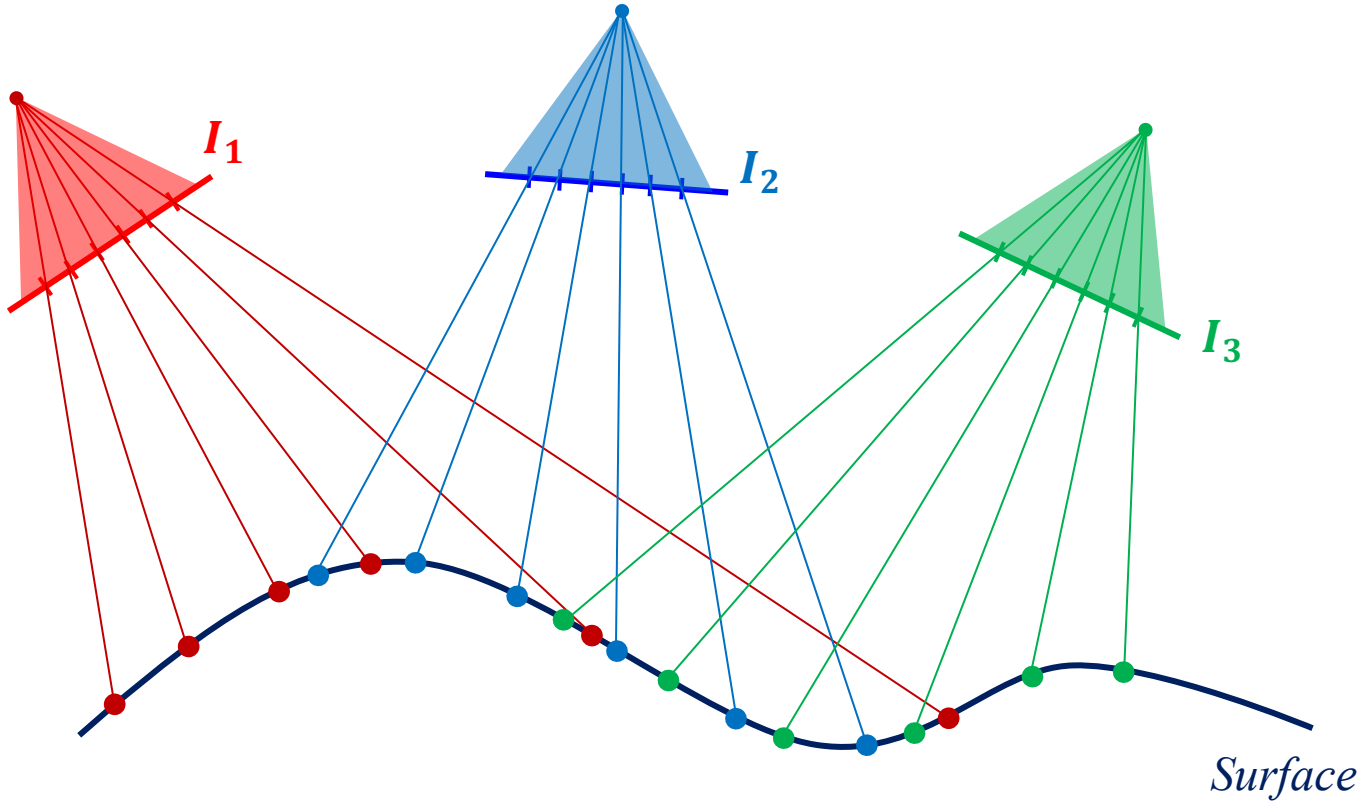


Redundancy Principle

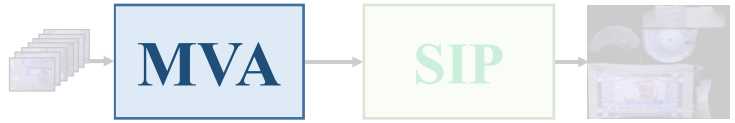


Sampling the surface at different positions



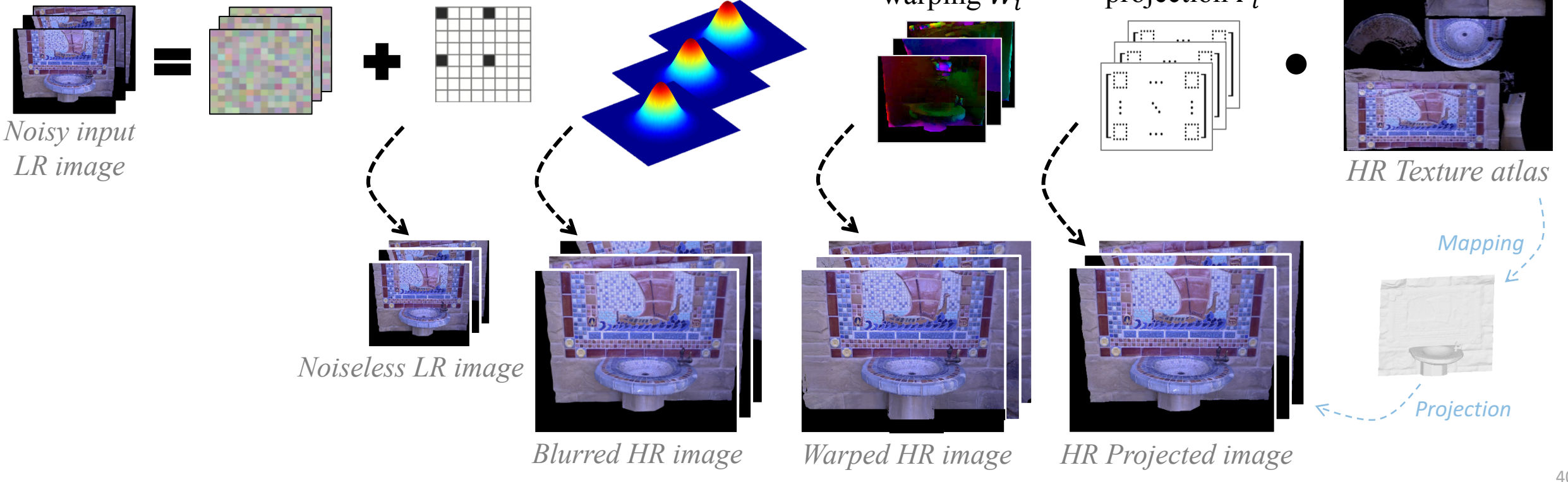


Generative Model

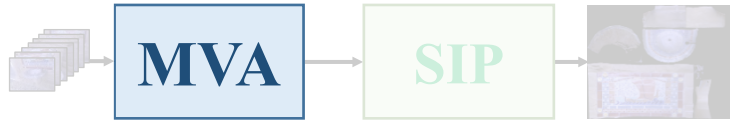


$$\mathrm{I}_i = \qquad\qquad \mathrm{T}$$

I_i Noise e_i Downsampling D Blurring K_i Optical flow warping W_i Appearance projection P_i $\mathbf{T}$

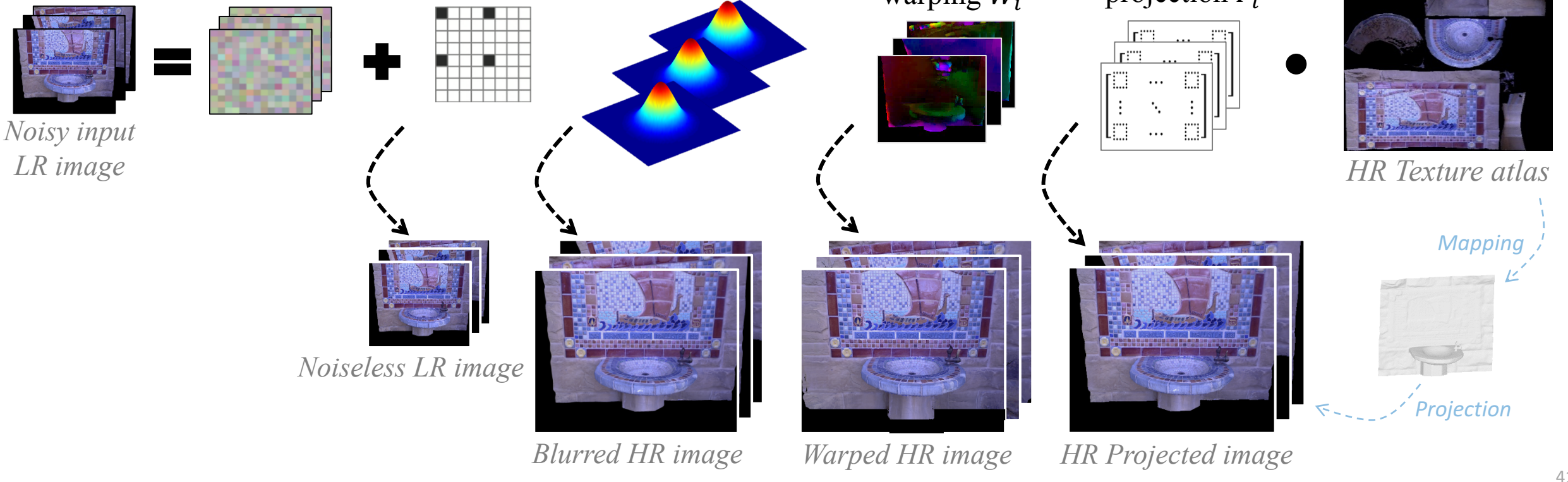


Generative Model

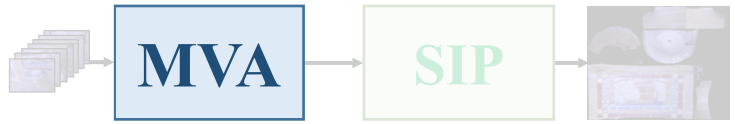


$$I_i - DK_iW_iP_i \cdot T$$

I_i Noise e_i Downsampling D Blurring K_i Optical flow warping W_i Appearance projection P_i T



SR Multi-view Energy



$$\underset{\mathrm{T}}{\text{minimize}} \sum_{i=1} \left\| \mathrm{I}_i - \mathrm{DK}_i\mathrm{W}_i\mathrm{P}_i \cdot \mathrm{T} \right\|_1$$

Re-projection error

Weighted TV

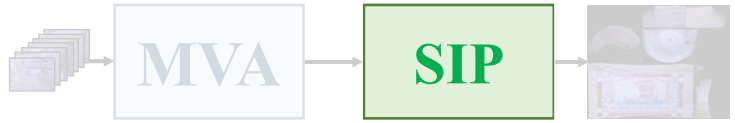
- $\boldsymbol{K_i}$: blurring kernel per view with standard deviation σ_i
- A small CNN ***adjusts*** $\boldsymbol{\sigma_i}$ for each view i

- **g** is manually set in classical SR methods
- We ***locally estimate*** it with a small CNN

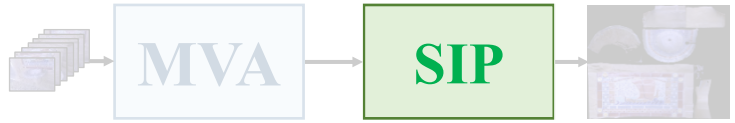
➡ **Variational optimization** numerically solved with 1st order primal-dual

➡ **Unrolling fixed number of optim steps**: each update cycle represents one network layer in MVA

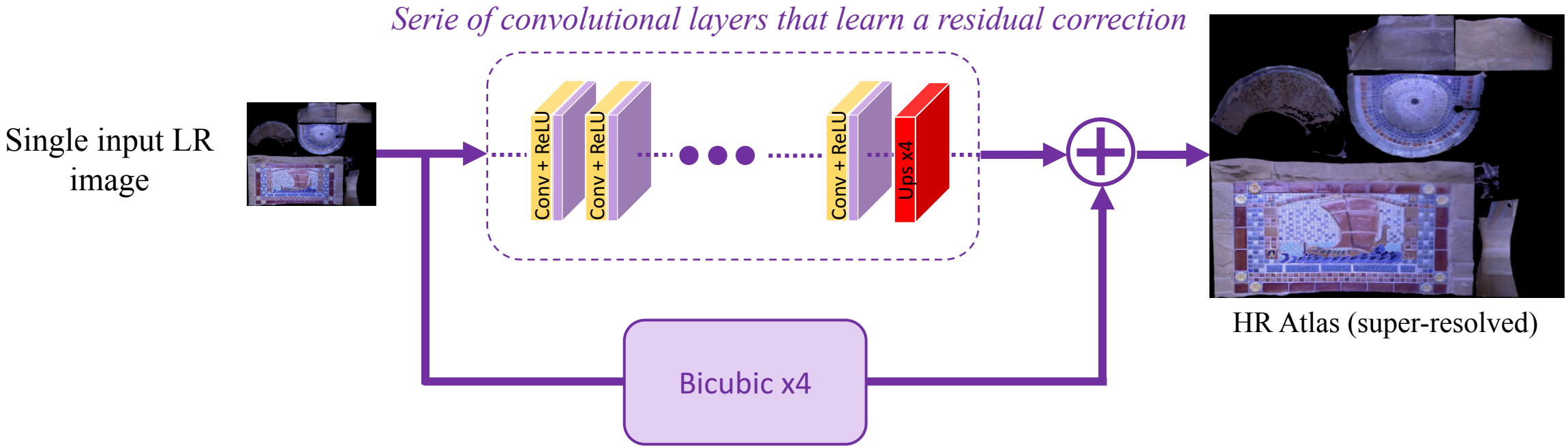
SIP Subnet



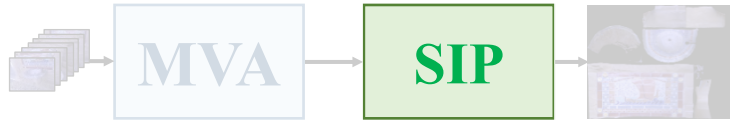
SIP Subnet



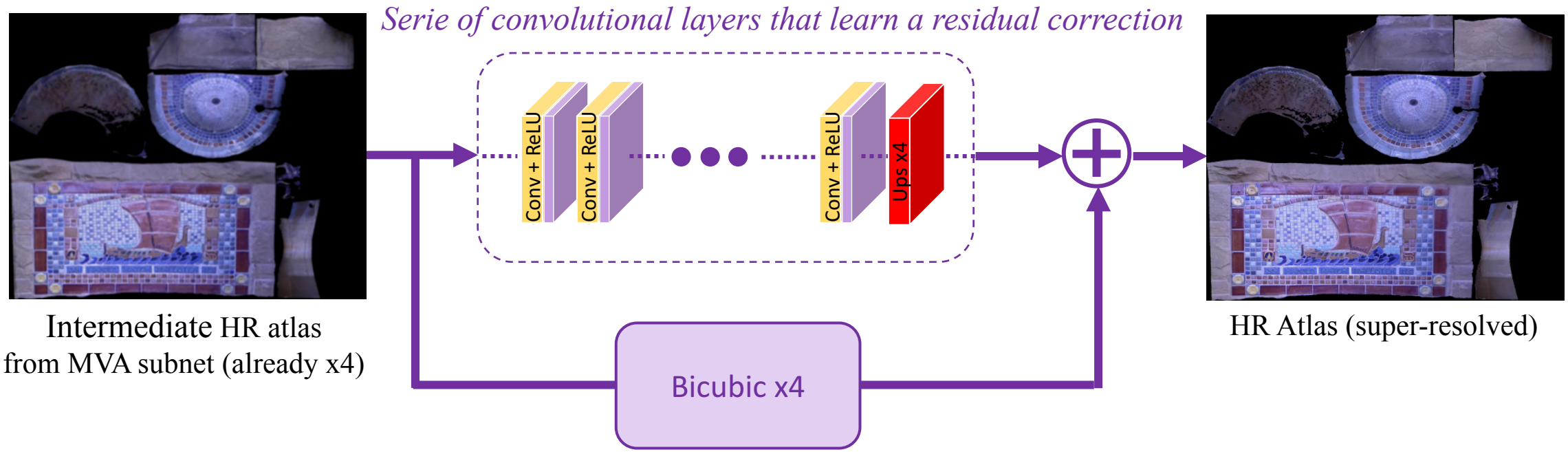
Classical ResNet architecture (e.g. x4) :



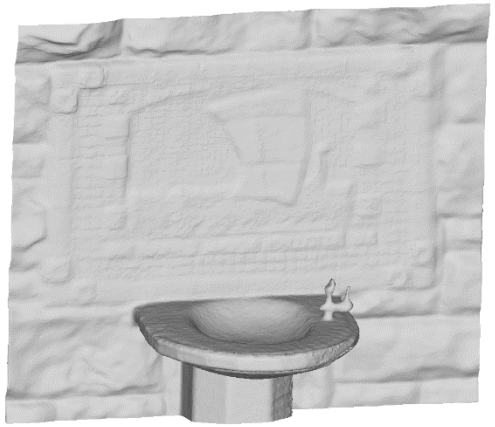
SIP Subnet



Classical ResNet architecture (e.g. x4) :

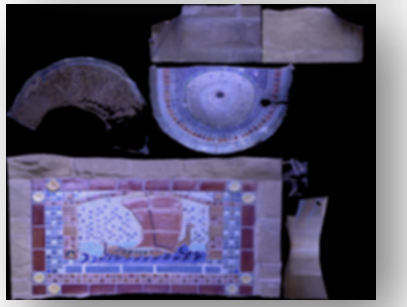


Data

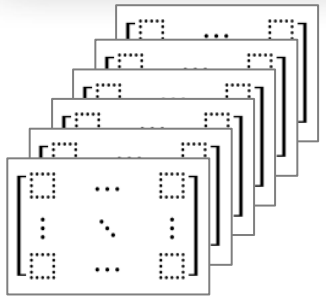


Set of calibrated low-resolution images and 3D mesh

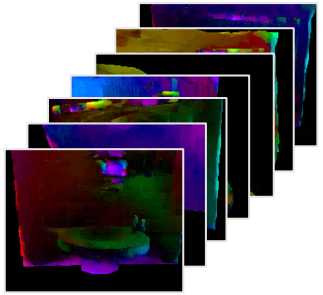
Ground truth atlas



Initial blurry atlas

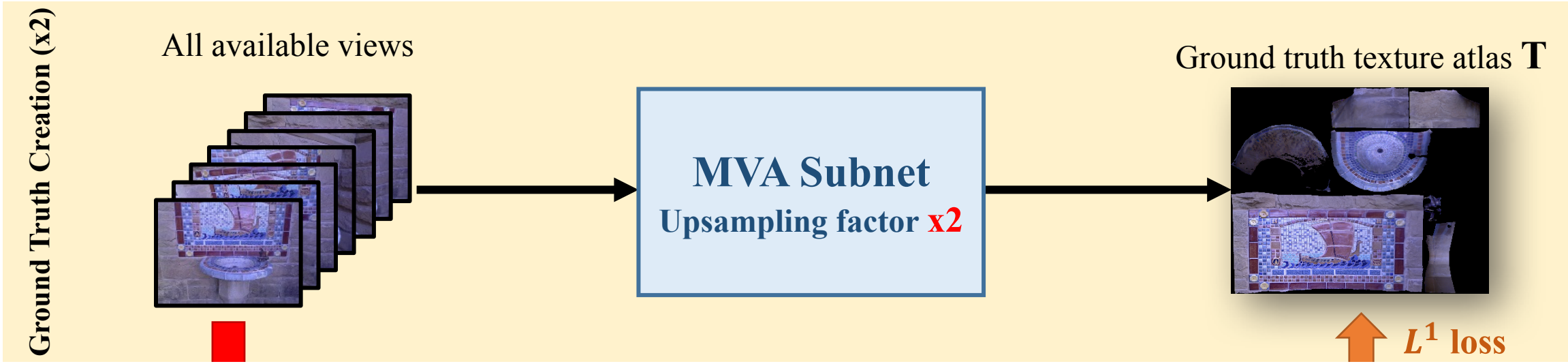


Appearance projection operator



Optical flow

Training Setup

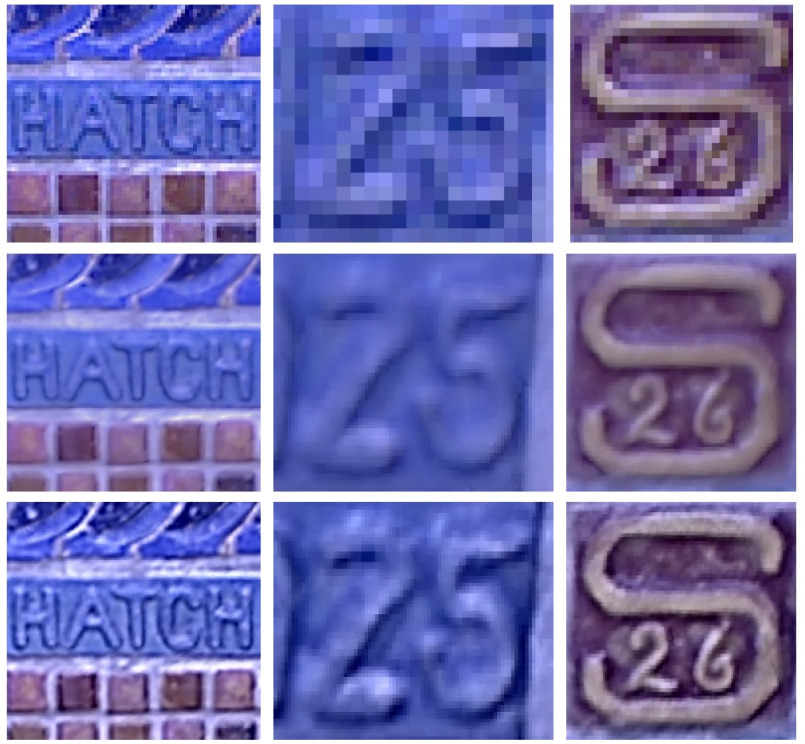


Comparison with State of the Art

- Upsampling factor x2 *(SOTA results only available for* x2*)*

Comparison with State of the Art

- Upsampling factor x2 *(SOTA results only available for* x2*)*



Input

State of the art

Ours

Quantitative Comparison

- Upsampling factor x4, same initial texture atlas
- Evaluated on: *Beethoven* 33 views, *Bird* 20 views, *Bunny* 36 views

	Beethoven			Bird			Bunny			Average		
	SSIM	*PSNR*	*SRE*	*SSIM*	*PSNR*	*SRE*	*SSIM*	*PSNR*	*SRE*	*SSIM*	*PSNR*	*SRE*
Tsiminaki *et al.* [55] with L^1-dataterm	0.934	40.488	14.482	0.942	41.379	21.914	0.923	38.596	15.292	0.933	40.155	17.229
Hui *et al.* [22] (pre-trained on [1])	0.933	39.505	13.410	0.941	41.999	22.499	0.916	37.888	14.525	0.930	39.797	16.811
MVA subnet	0.931	40.260	14.253	**0.944**	41.901	22.436	0.922	38.573	15.268	0.933	40.245	17.319
SIP subnet (trained only on [1])	0.914	37.041	2.009	0.921	38.738	16.766	0.869	35.252	7.006	0.901	37.010	8.594
Ours (no pre-training, σ_i fixed)	0.941	39.231	13.224	0.940	39.146	19.680	0.929	38.198	14.894	0.937	38.858	15.933
Ours (SIP-Net pre-trained on [1])	**0.948**	**43.309**	**17.304**	0.943	**44.634**	**25.171**	**0.932**	**39.690**	**16.386**	**0.941**	**42.544**	**19.620**

Baselines

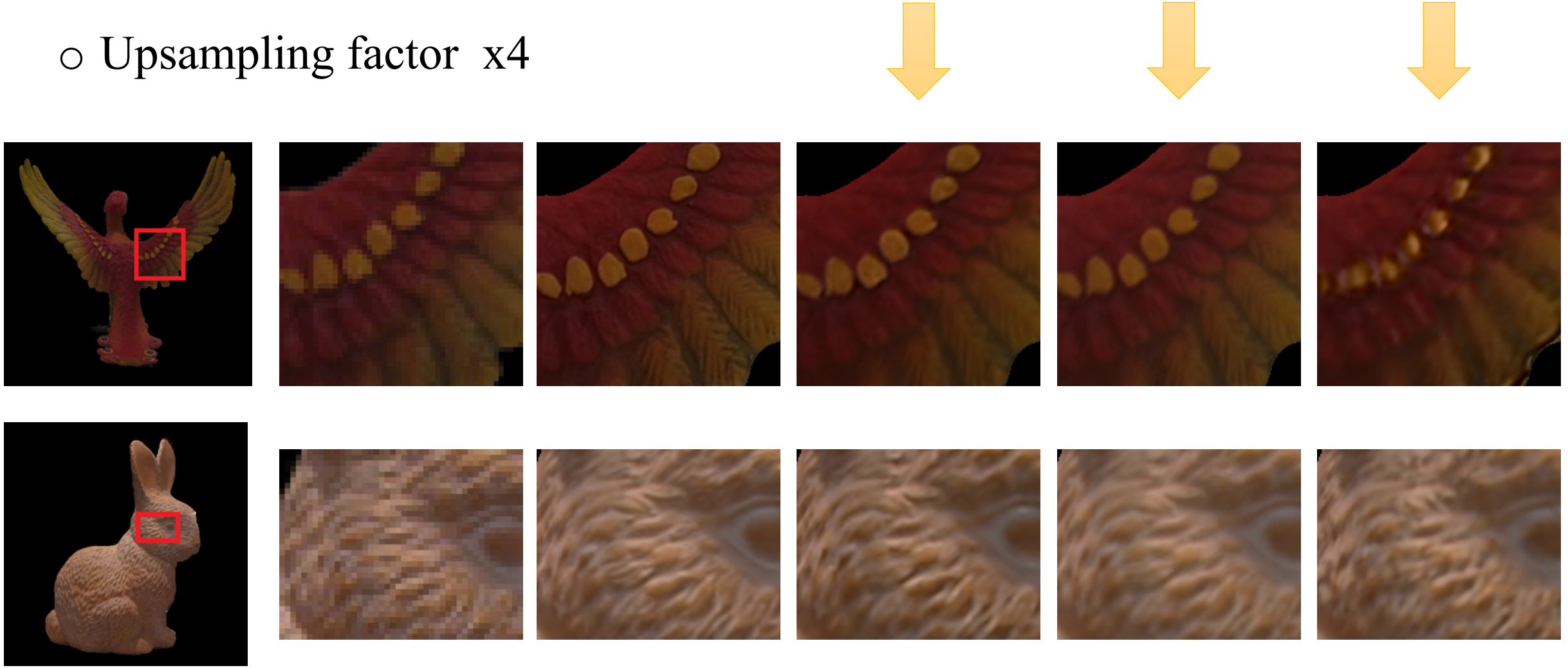
- [Tsiminaki et al. CVPR'14] with L1-dataterm

- [Hui et al. CVPR'18] pre-trained on DIV2K dataset

Our proposed network

Ablation Study

- Upsampling factor x4



Input | Groundtruth | **Ours (MVA + SIP)** | MVA | SIP

Take Home Messages on SR

- Combine the *two SR concepts*
- Handle *arbitrary* number of input images in neural network
- *End-to-end* trainable

SOTA Texturing [Waechter et al.]

SOTA SR Texturing [Tsiminaki et al.]

Ours

Mixed-Reality Lab

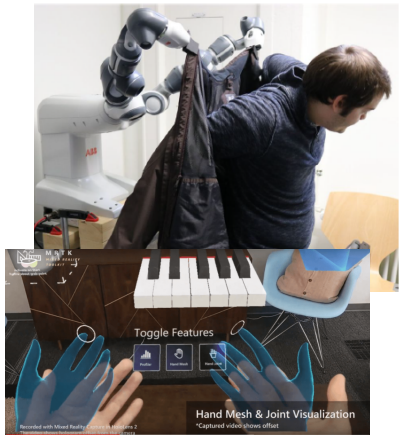
Practical Course

Lecturers:

Federica Bogo, Microsoft MR&AI Lab

Martin Oswald, CVG, ETH Zurich

- 10 credits
- Lectures & guest lectures
- Student groups work on projects



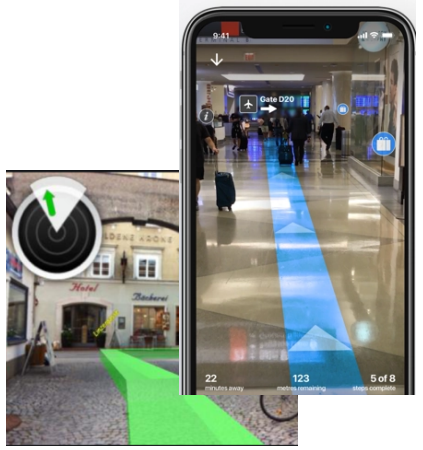
MR Immersive Telemanipulation

Interactive Airport Experience

Art Planning & Exploration

MR Surgery Assistence

Science & Education

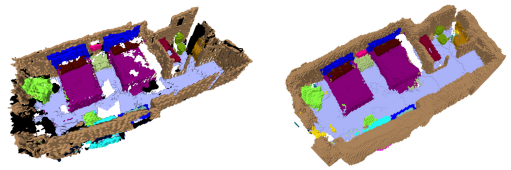


MR Navigation

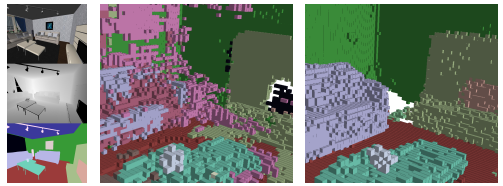
Thank you!

Semantic 3D Reconstruction with …

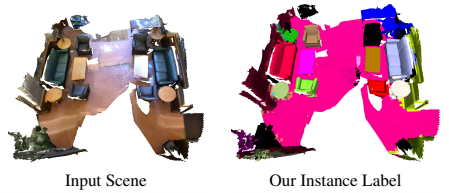
- **Learning Priors for Semantic 3D Reconstruction**
  (Ian Cherabier, Martin Oswald, Marc Pollefeys, Andreas Geiger –ECCV'18)



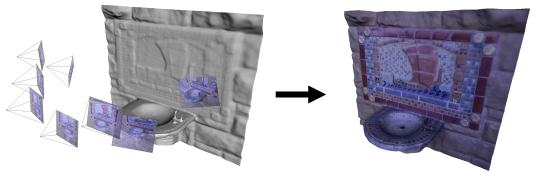
- **Multi-Sensor Data Fusion**
  (Denys Rozumny, Ian Cherabier, Marc Pollefeys, Martin Oswald – ICCVW'19)



- **3D Instance Segmentation**
  (Jean Lahoud, Bernard Ghanem, Marc Pollefeys, Martin Oswald – ICCV'19)

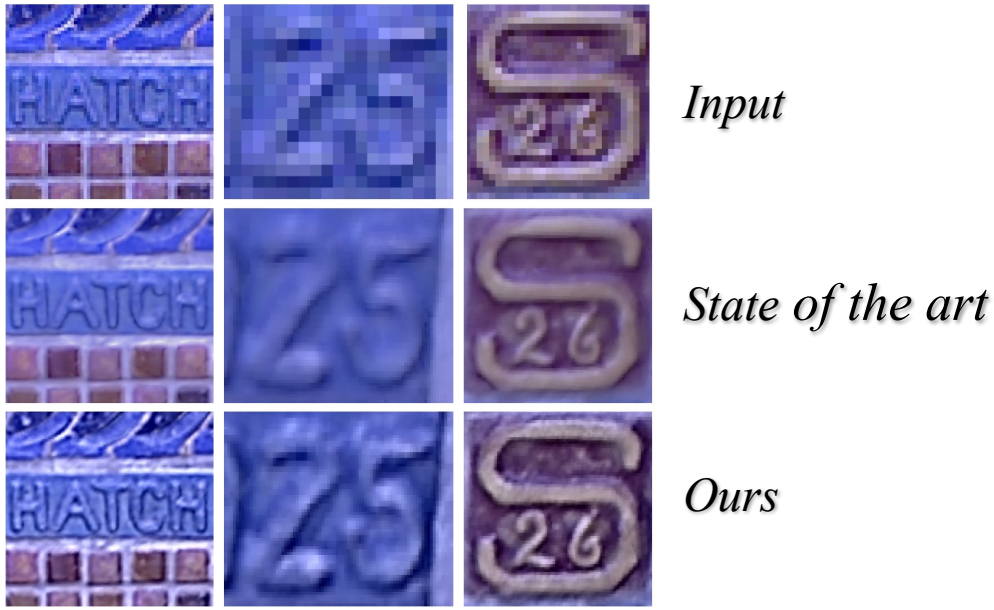


- **Learned Multi-view Texture Super-resolution**
  (Audrey Richard, Ian Cherabier, Martin Oswald, Vagia Tsiminaki,
  Marc Pollefeys, Konrad Schindler– 3DV'19)



Backup Slides

Audrey Richard, Ian Cherabier, Martin R. Oswald, Vagia Tsiminaki,
Marc Pollefeys, Konrad Schindler

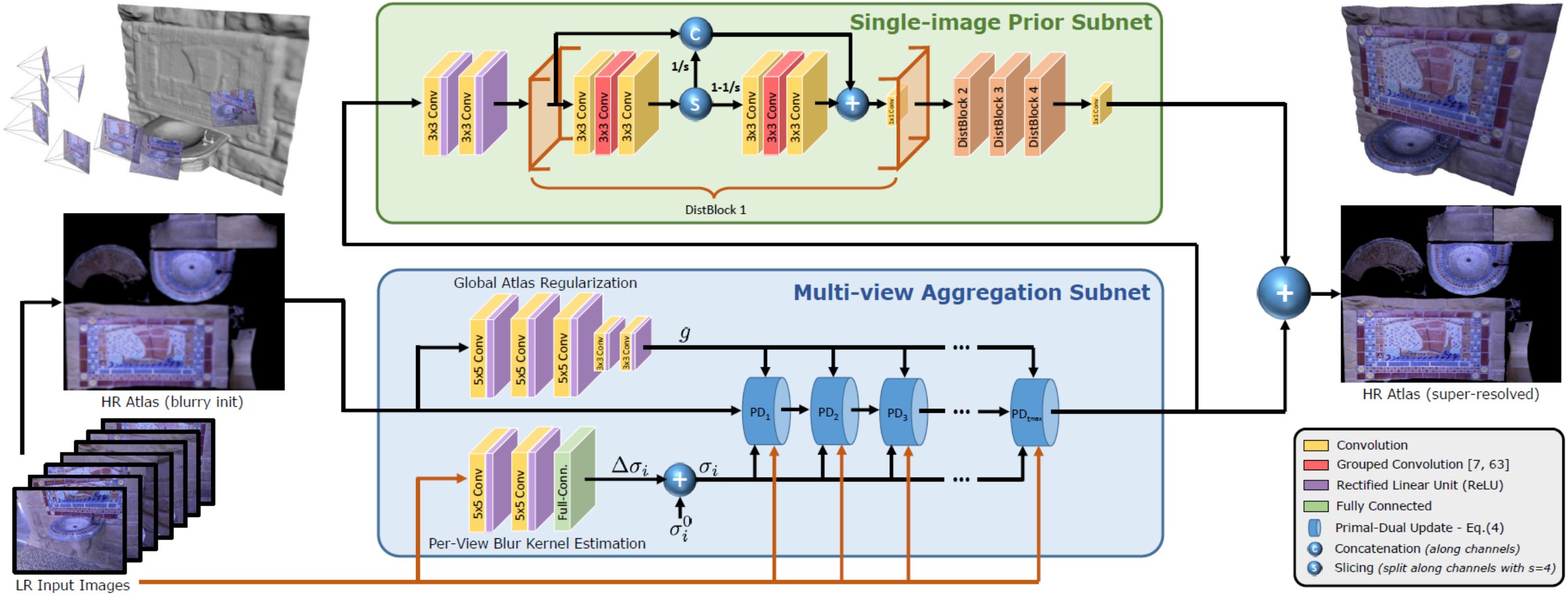


Input

State of the art

Ours

Network Architecture



57

Numerical Optimization

- Saddle-point problem :

$$\min_{\mathrm{T}} \max_{\substack{\|\phi_i\|_\infty \leq 1 \\ \|\xi\|_\infty \leq 1}} \sum_{i=1}^{N} \langle \mathrm{DK}_i \mathrm{W}_i \mathrm{P}_i \cdot \mathrm{T} - \mathrm{I}_i, \phi_i \rangle + \langle \mathrm{g} \cdot \nabla \mathrm{T}, \xi \rangle \qquad (1)$$

- Update steps at each iteration :

$$\phi_i^{t+1} = \Pi_{\|\cdot\| \leq 1} \left[\phi_i^t + \eta \left(\mathrm{DK}_i \mathrm{W}_i \mathrm{P}_i \bar{\mathrm{T}}^t - \mathrm{I}_i \right) \right] \qquad (2a)$$

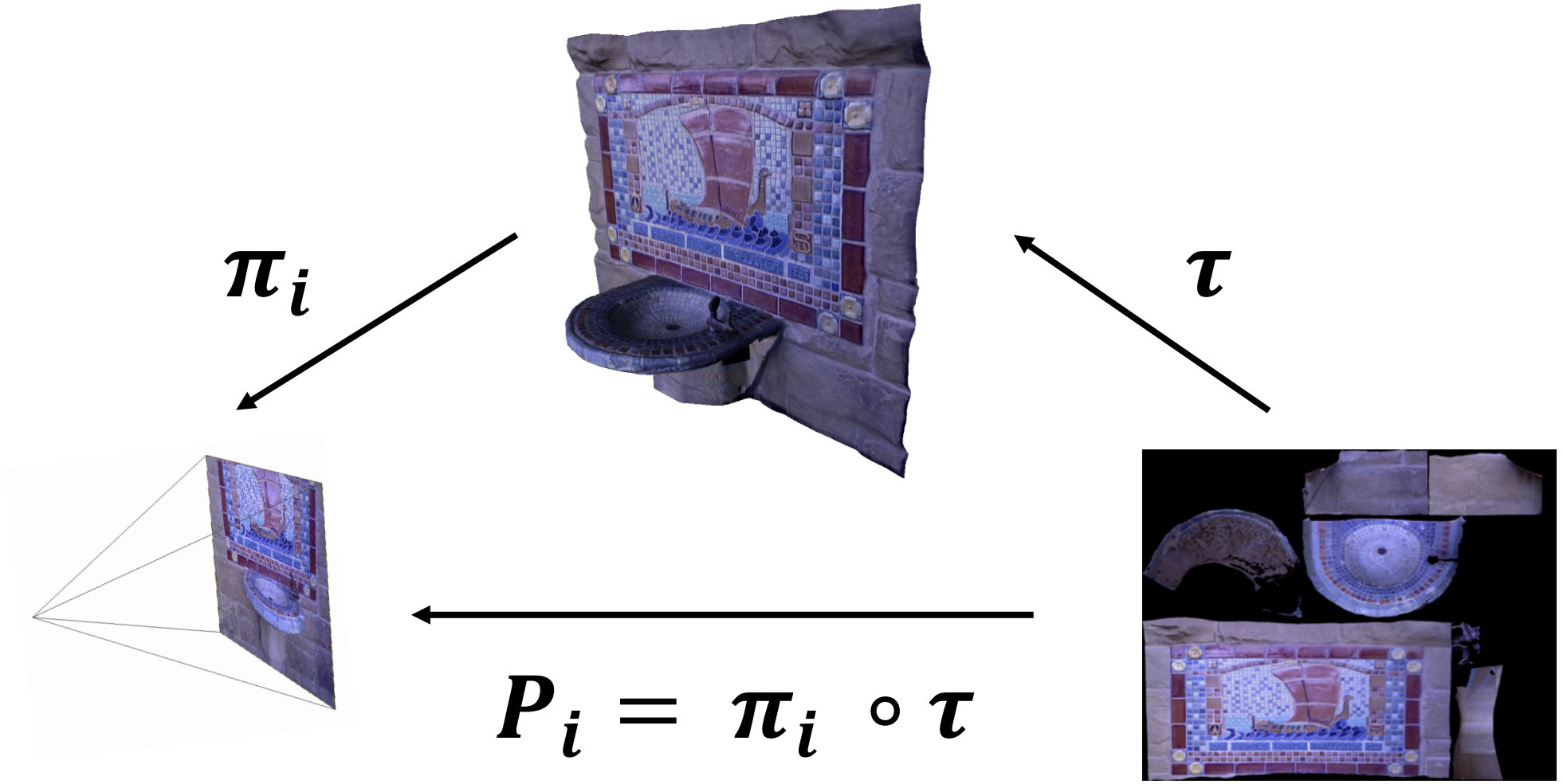
$$\xi^{t+1} = \Pi_{\|\cdot\| \leq 1} \left[\xi^t + \eta \cdot \mathrm{g} \cdot \nabla \bar{\mathrm{T}}^t \right] \qquad (2b)$$

$$\mathrm{T}^{t+1} = \mathrm{T}^t + \tau \left(\mathrm{g} \cdot \operatorname{div} \xi^{t+1} - \sum_{i=1}^{N} \mathrm{P}_i^\intercal \mathrm{W}_i^\intercal \mathrm{K}_i^\intercal \mathrm{D}^\intercal \phi_i^{t+1} \right) \qquad (2c)$$

$$\bar{\mathrm{T}}^{t+1} = 2\mathrm{T}^{t+1} - \mathrm{T}^t \qquad (2d)$$

58

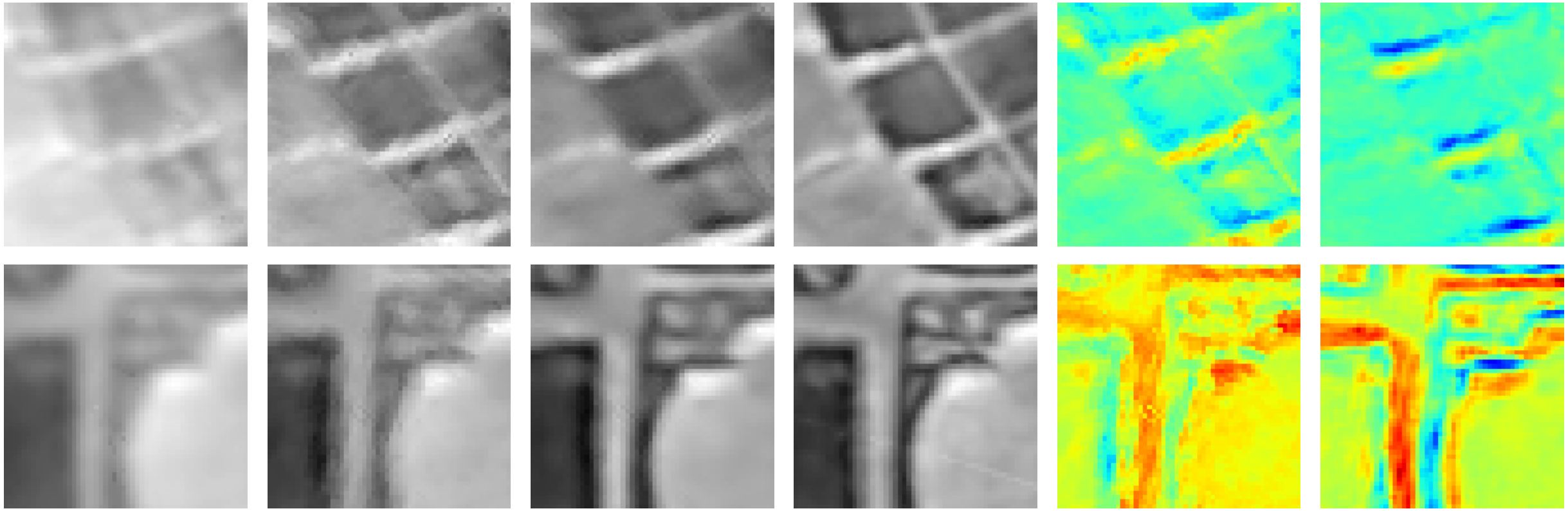
Projection Operator



59

Network Behavior

- Upsampling factor x4



Input atlas | MVA | MVA + SIP | GT | Residual MVA | Residual SIP