Eurographics 2012
Cagliari, Italy
May 13-18
33rd ANNUAL CONFERENCE OF THE EUROPEAN ASSOCIATION FOR COMPUTER GRAPHICS

Coherent Spatiotemporal Filtering, Upsampling and Rendering of RGBZ Videos

Christian Richardt[1,2] Carsten Stoll[1] Neil A. Dodgson[2]
Hans-Peter Seidel[1] Christian Theobalt[1]

[1] max planck institut informatik
[2] UNIVERSITY OF CAMBRIDGE

Eurographics
EUROPEAN ASSOCIATION FOR COMPUTER GRAPHICS

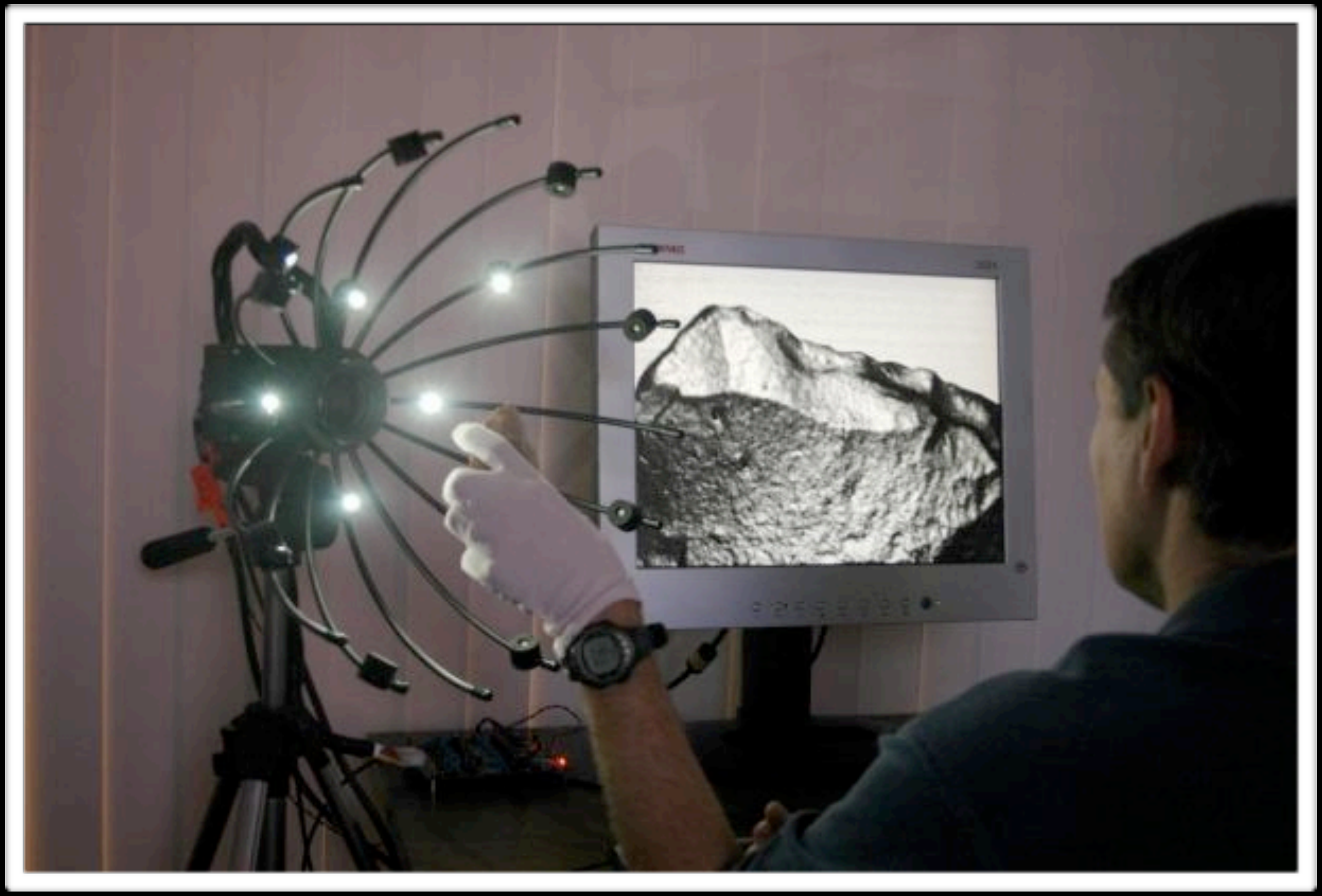
Unfiltered depth map

Introduction & motivation

- our work tackles the noisy, low-resolution depth data: filter colour + depth to upsample and denoise depth
- capturing colour + depth enables a variety of compelling, previously impossible video effects
- prototype video camera + video processing algorithms = effective and robust capture of RGBZ video
- result: dynamic, temporally coherent scene geometry, calculated at interactive frame rates

Related work – geometry capture

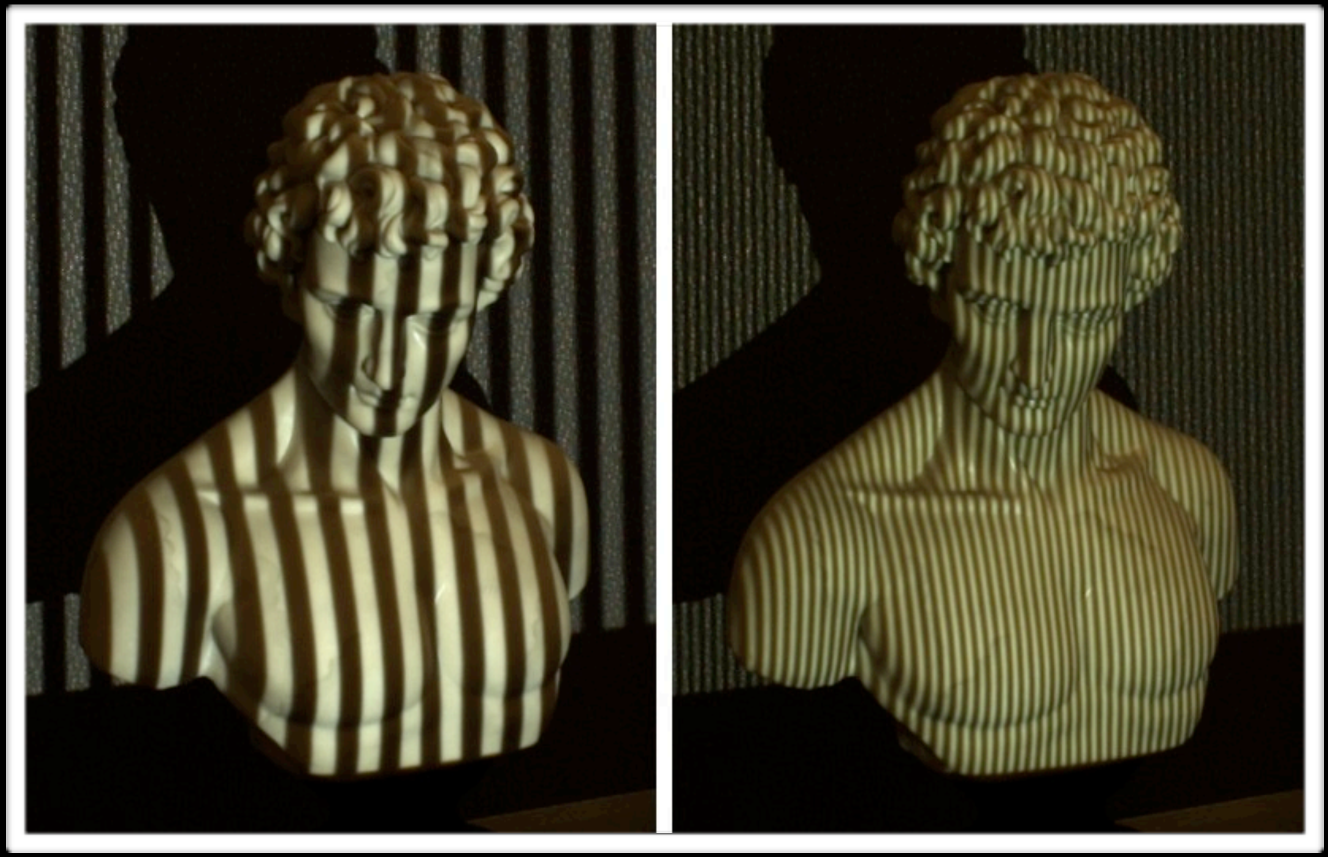
- four main approaches to capture dynamic geometry:
  1. photometric stereo / shape-from-shading



[Malzbender et al. 2006]

Related work – geometry capture

- four main approaches to capture dynamic geometry:
  1. photometric stereo / shape-from-shading
  2. active stereo / structured light



[Lanman and Taubin 2009]

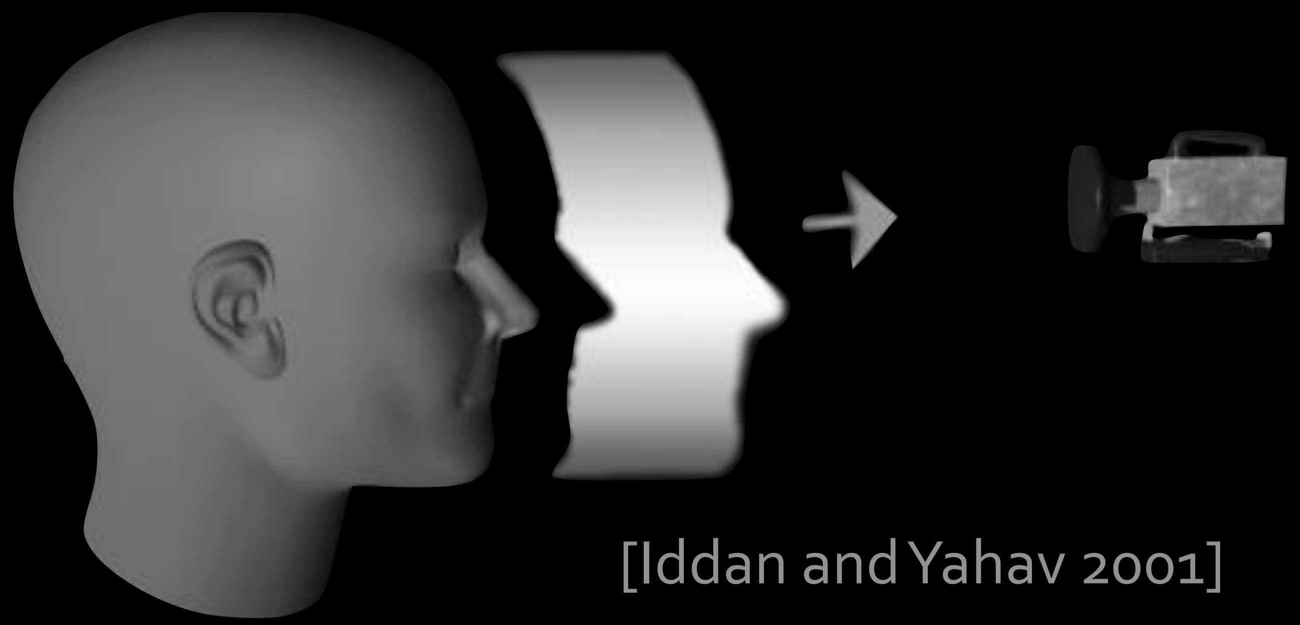
Related work – geometry capture

- four main approaches to capture dynamic geometry:
  1. photometric stereo / shape-from-shading
  2. active stereo / structured light
  3. structure-from-motion & stereo vision

[Scharstein and Szeliski 2003]

Related work – geometry capture

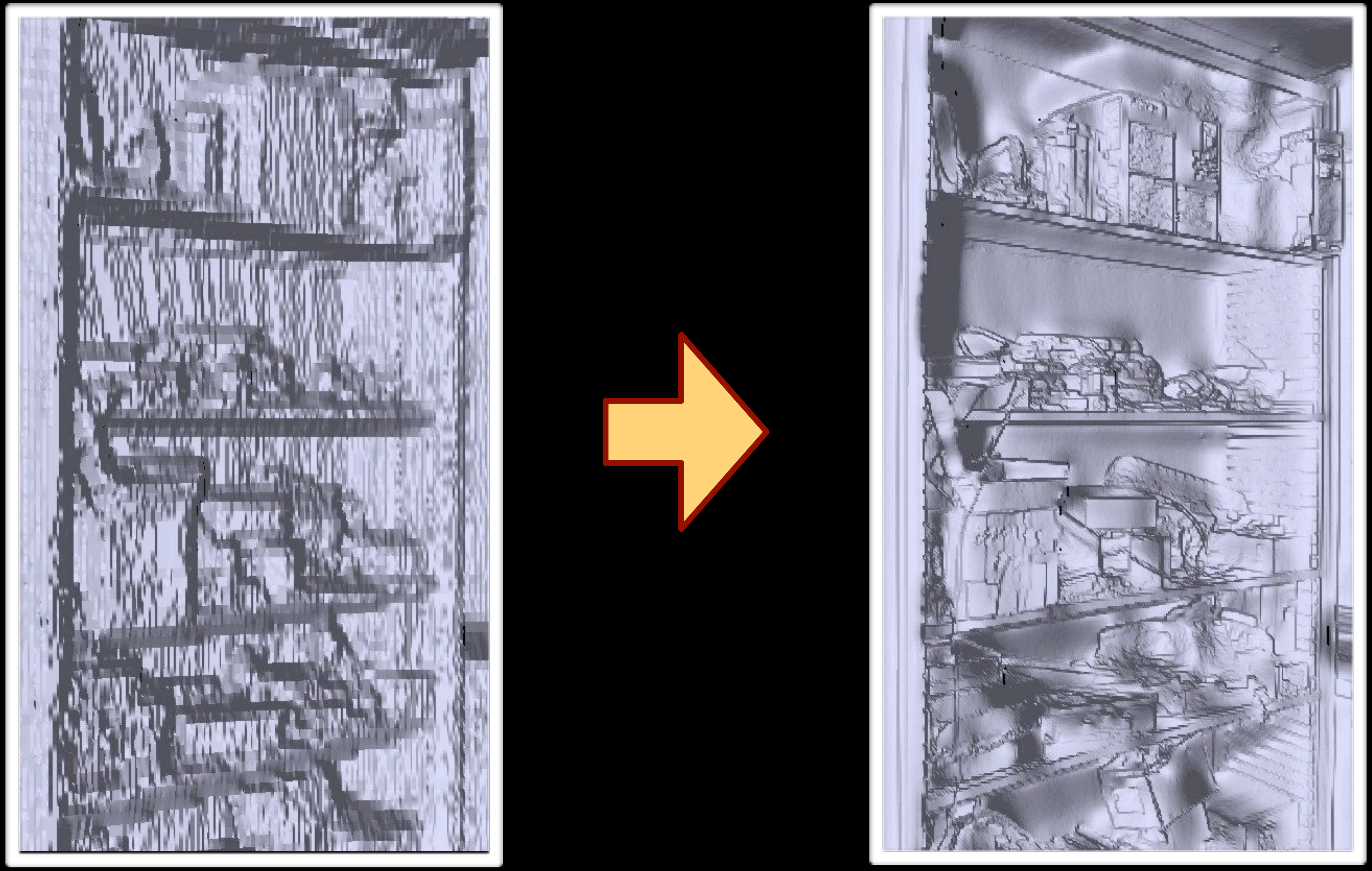
- four main approaches to capture dynamic geometry:
  1. photometric stereo / shape-from-shading
  2. active stereo / structured light
  3. structure-from-motion & stereo vision
  4. time-of-flight cameras



[Iddan and Yahav 2001]

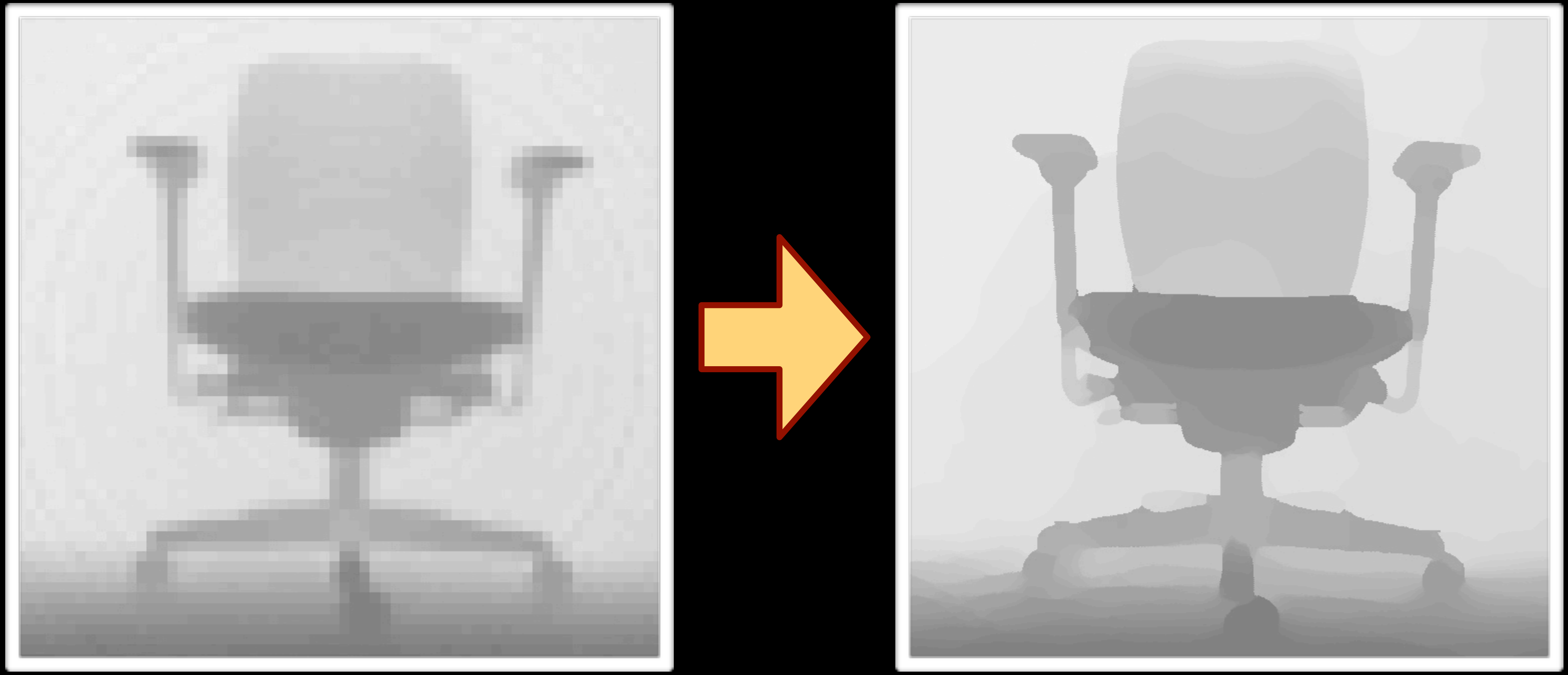
Related work – depth upsampling

- Markov random fields
  [Diebel and Thrun 2006]



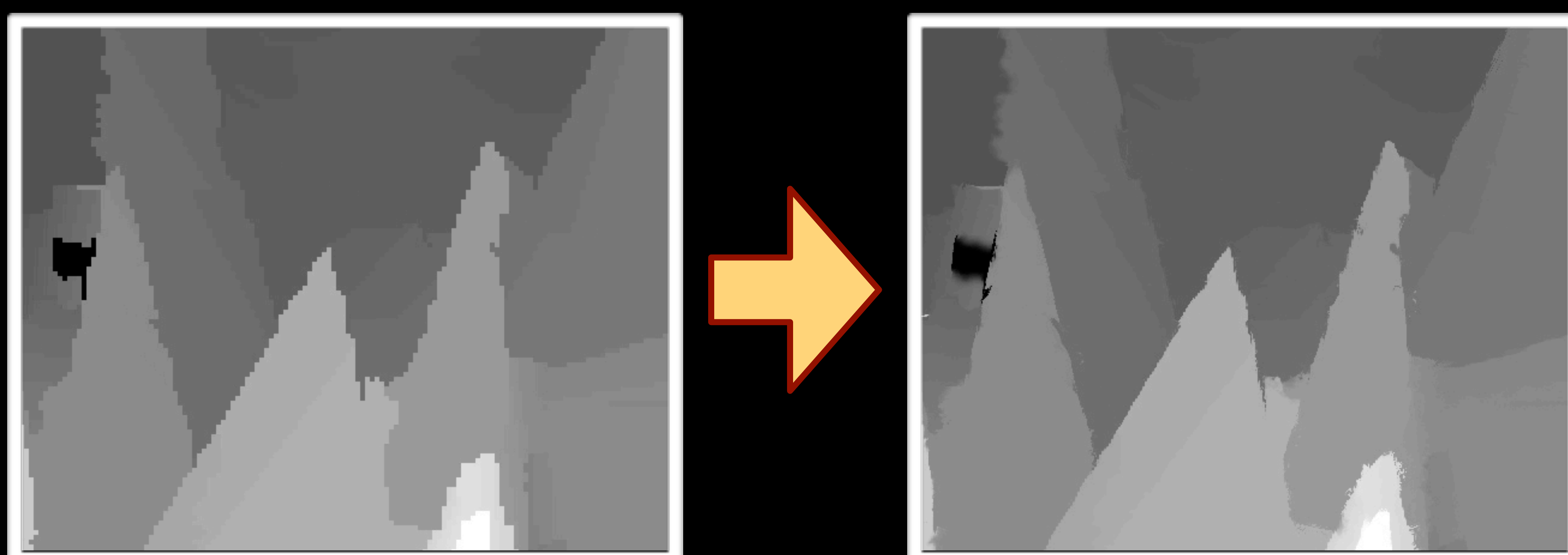
Related work – depth upsampling

- Markov random fields
  [Diebel and Thrun 2006]
- spatial-depth super-resolution
  [Yang et al. 2007]



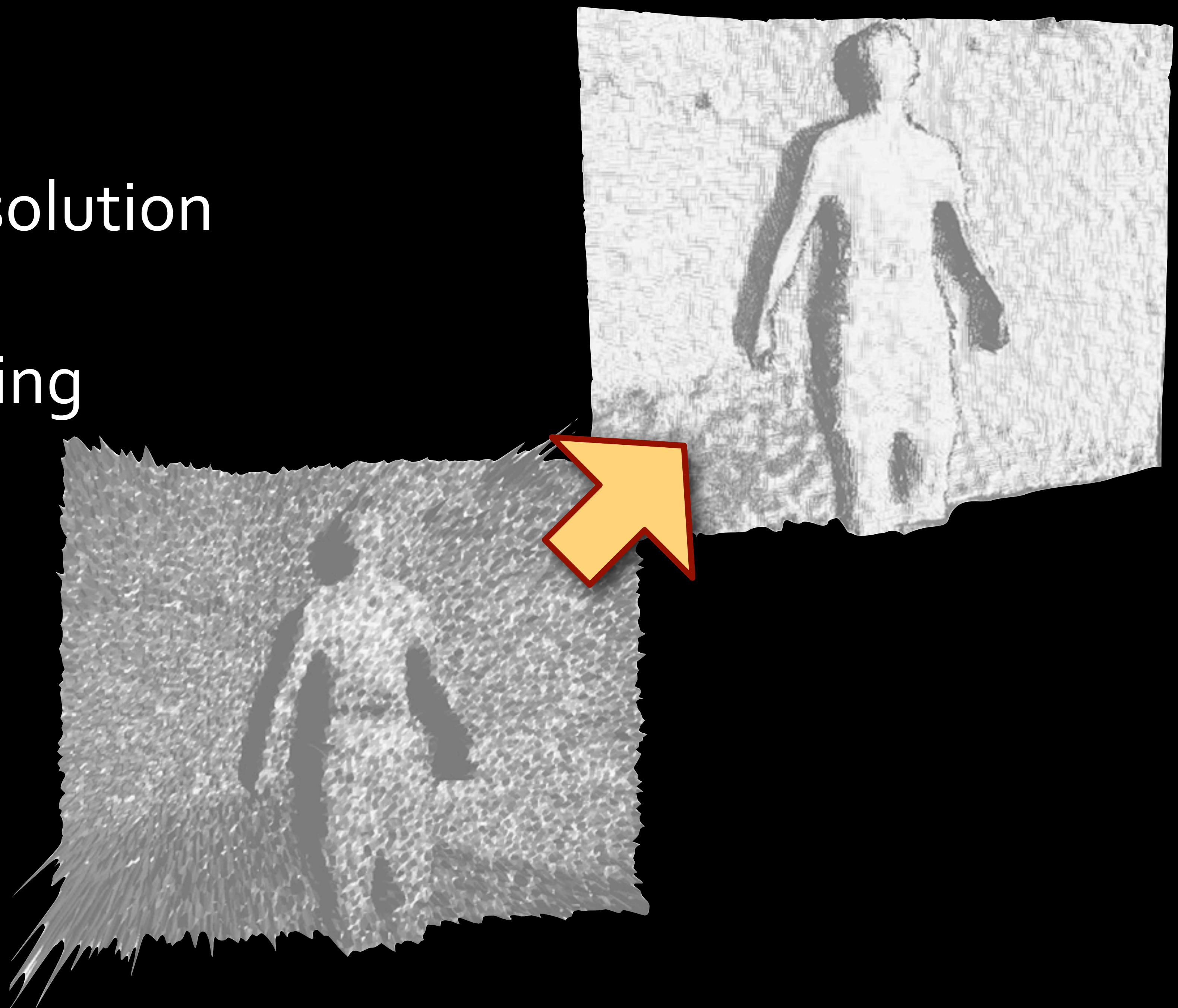
Related work – depth upsampling

- Markov random fields
  [Diebel and Thrun 2006]
- spatial-depth super-resolution
  [Yang et al. 2007]
- joint-bilateral upsampling
  [Kopf et al. 2007]



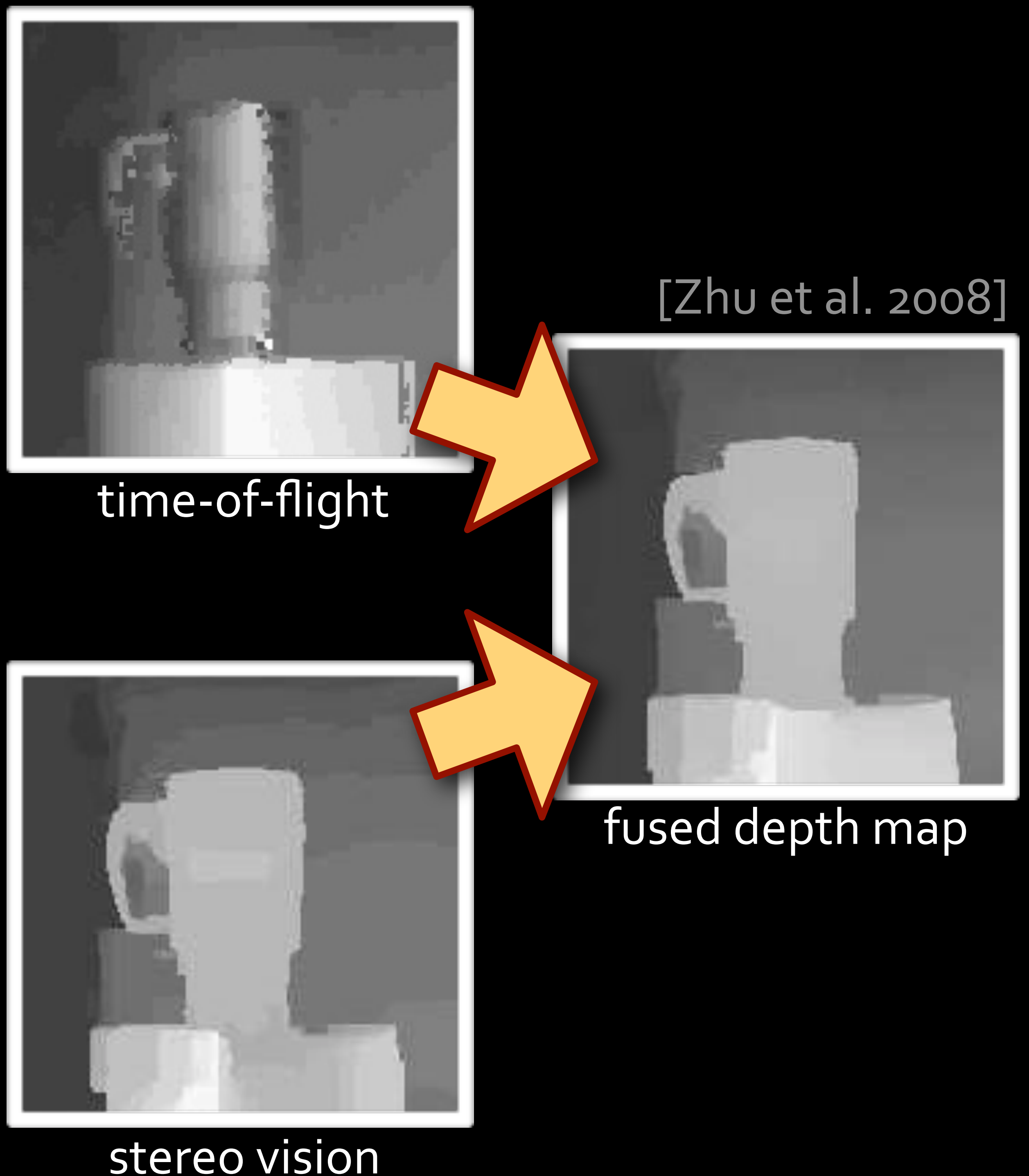
Related work – depth upsampling

- Markov random fields
  [Diebel and Thrun 2006]
- spatial-depth super-resolution
  [Yang et al. 2007]
- joint-bilateral upsampling
  [Kopf et al. 2007]
- noise-aware filtering
  [Chan et al. 2008]



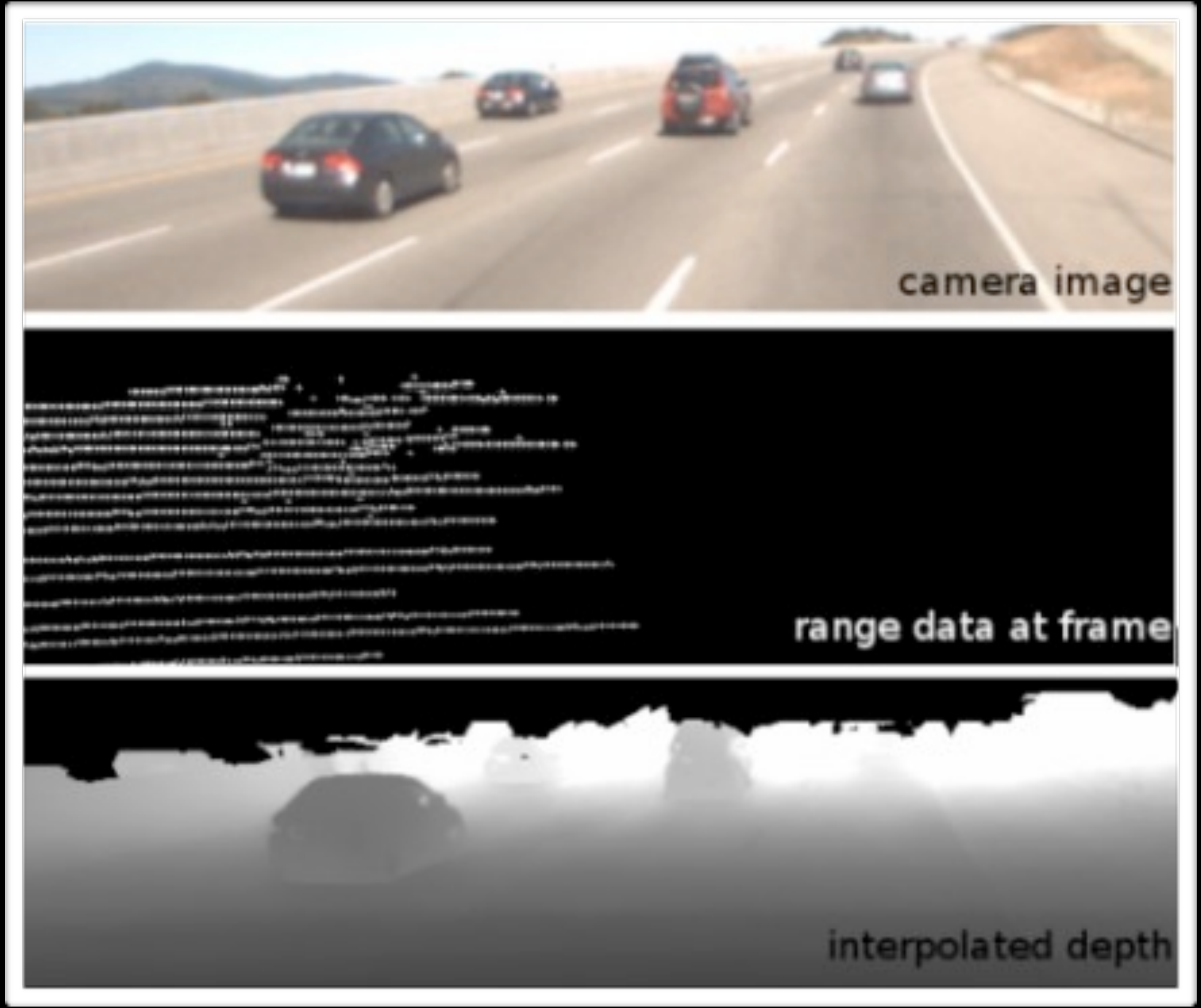
Related work – depth upsampling

- Markov random fields
  [Diebel and Thrun 2006]
- spatial-depth super-resolution
  [Yang et al. 2007]
- joint-bilateral upsampling
  [Kopf et al. 2007]
- noise-aware filtering
  [Chan et al. 2008]
- time-of-flight + stereo
  [Beder et al. 2007, Zhu et al. 2008]



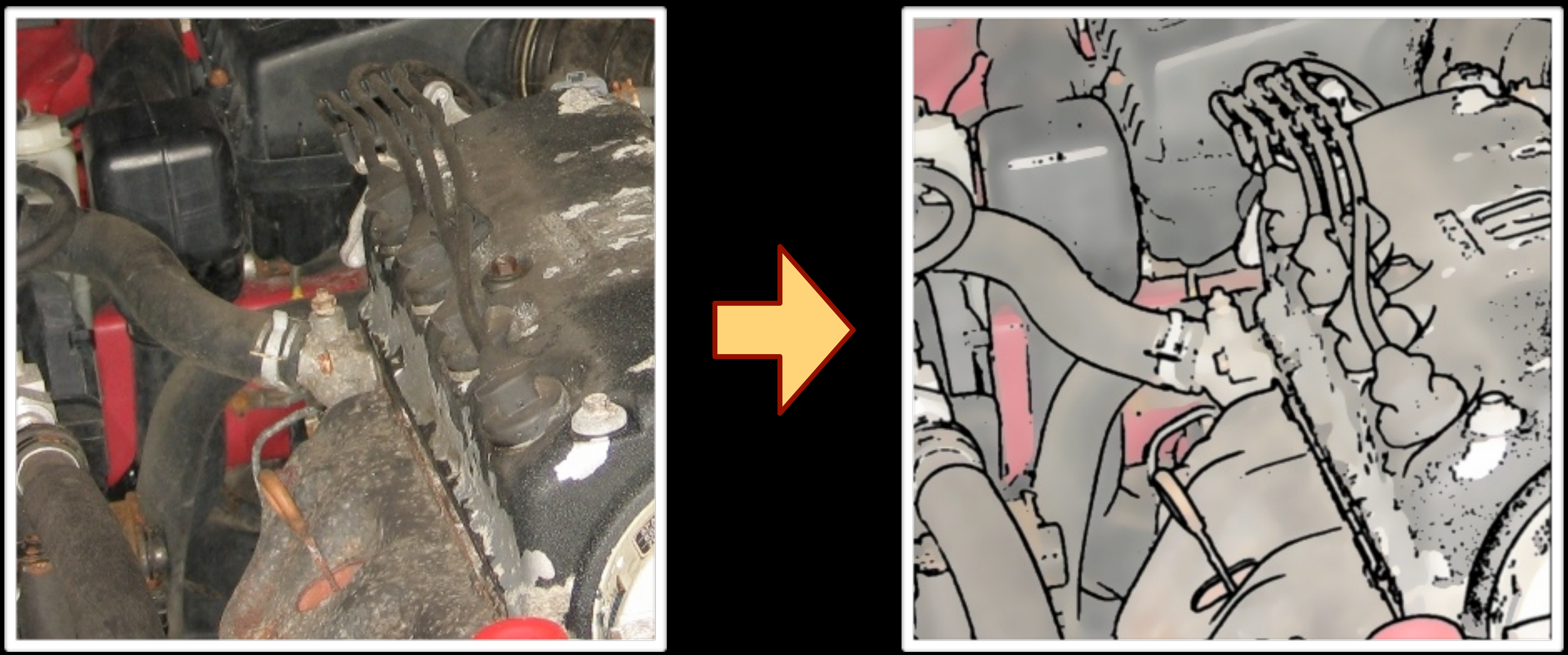
Related work – depth upsampling

- Markov random fields
  [Diebel and Thrun 2006]
- spatial-depth super-resolution
  [Yang et al. 2007]
- joint-bilateral upsampling
  [Kopf et al. 2007]
- noise-aware filtering
  [Chan et al. 2008]
- time-of-flight + stereo
  [Beder et al. 2007, Zhu et al. 2008]
- upsampling dynamic range data
  [Dolson et al. 2010]



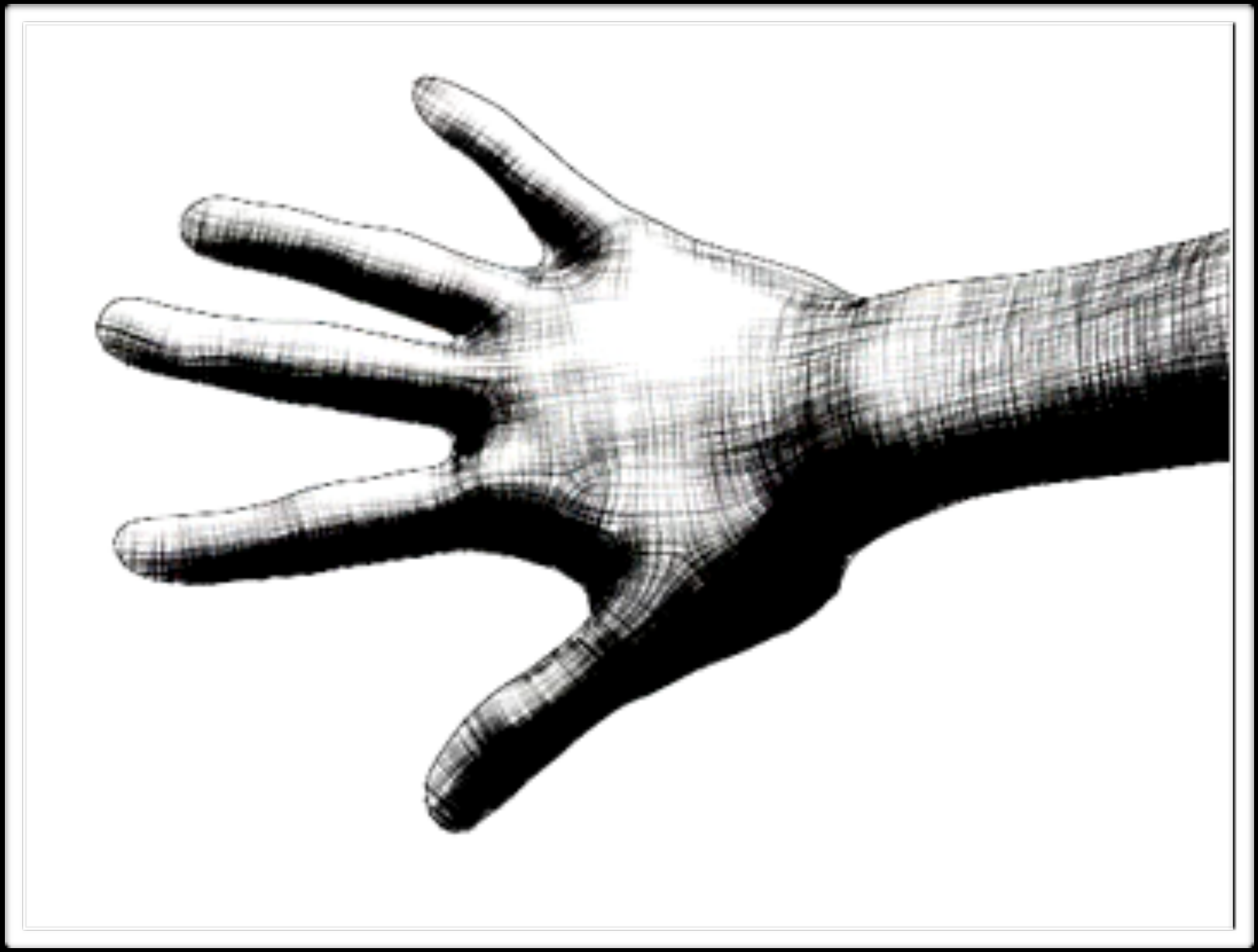
Related work – depth-based stylisation

- NPR camera
  [Raskar et al. 2004]



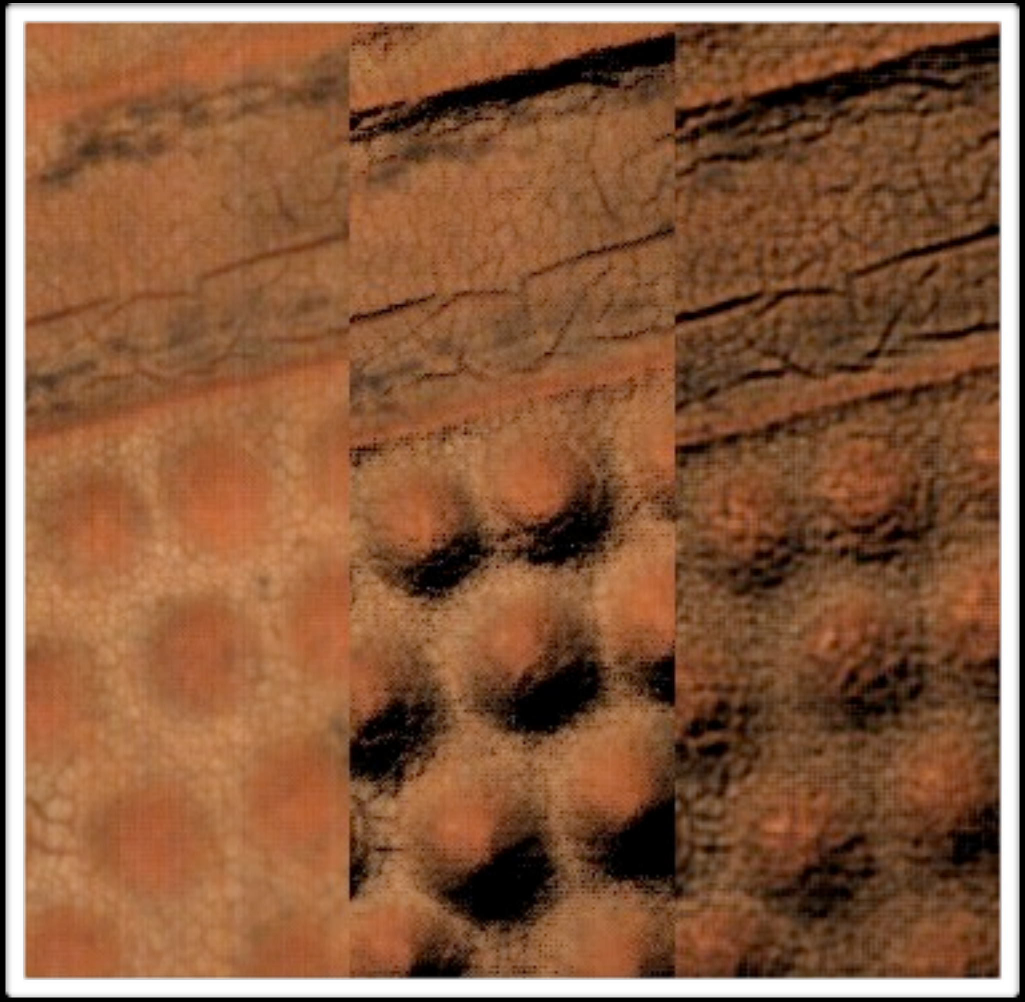
Related work – depth-based stylisation

- NPR camera
  [Raskar et al. 2004]
- 2.5-D video stylisation
  [Snavely et al. 2006]



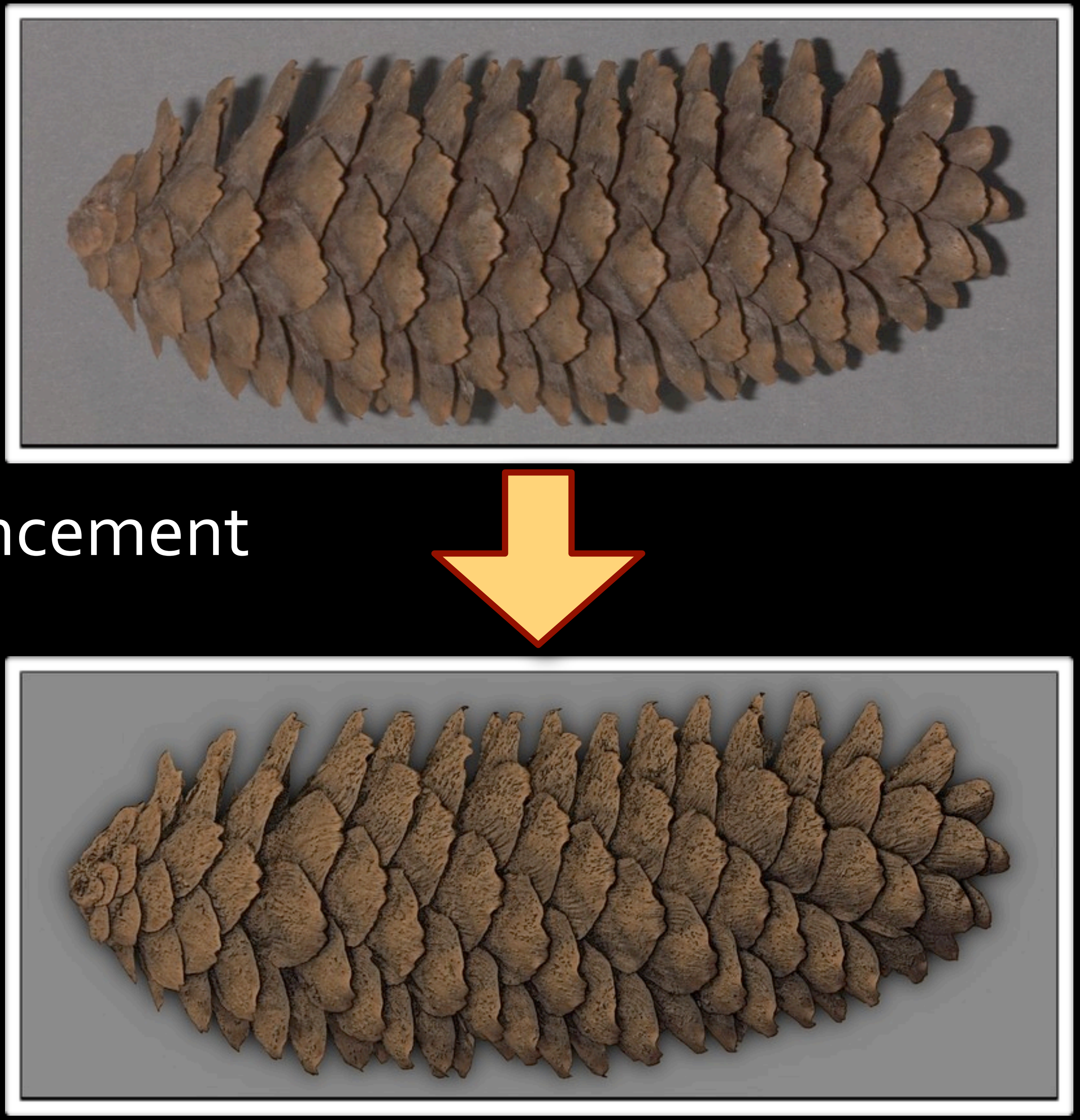
Related work – depth-based stylisation

- NPR camera
  [Raskar et al. 2004]
- 2.5-D video stylisation
  [Snavely et al. 2006]
- photometric surface enhancement
  [Malzbender et al. 2006]



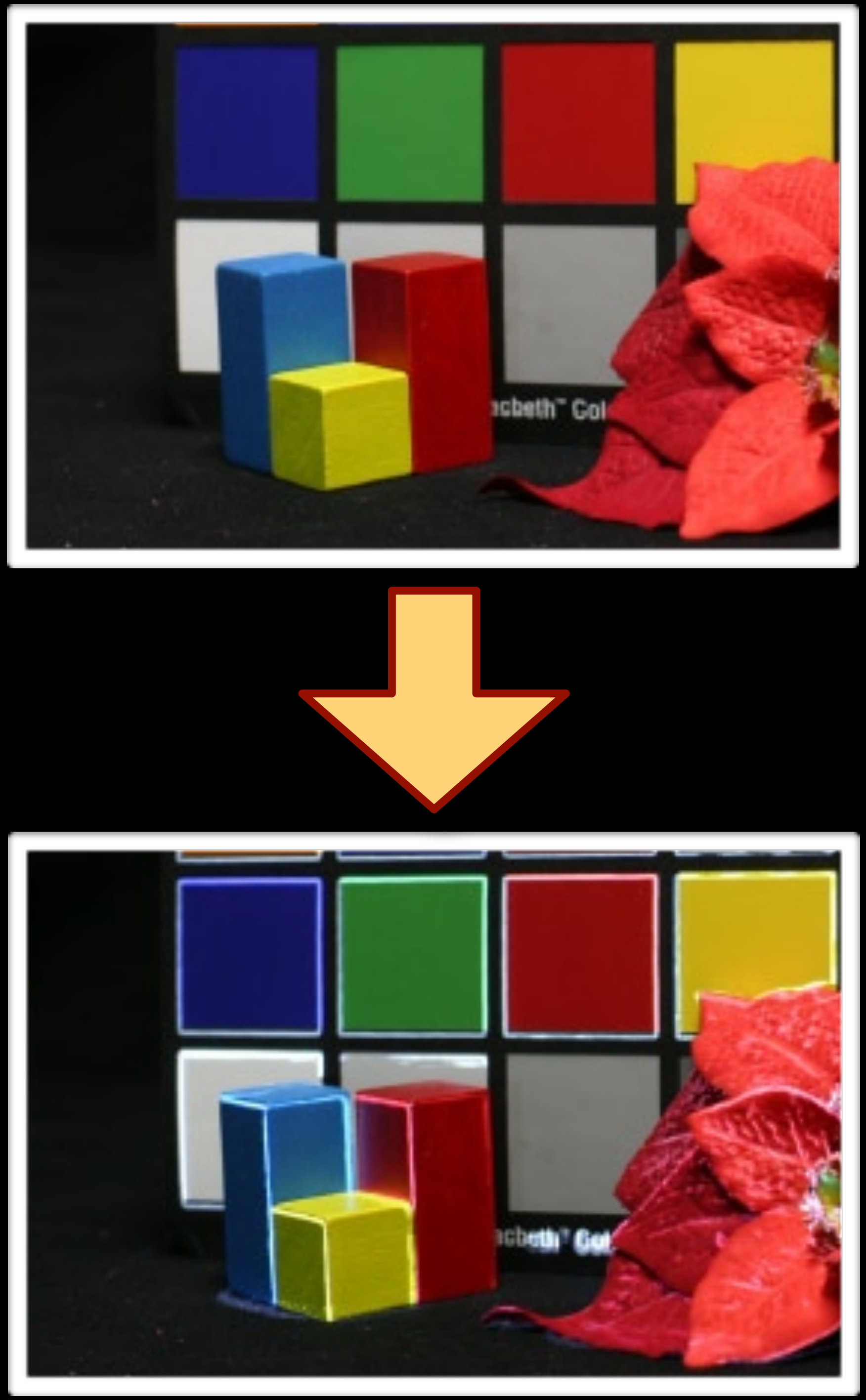
Related work – depth-based stylisation

- NPR camera
  [Raskar et al. 2004]
- 2.5-D video stylisation
  [Snavely et al. 2006]
- photometric surface enhancement
  [Malzbender et al. 2006]
- Images with normals
  [Toler-Franklin et al. 2007]



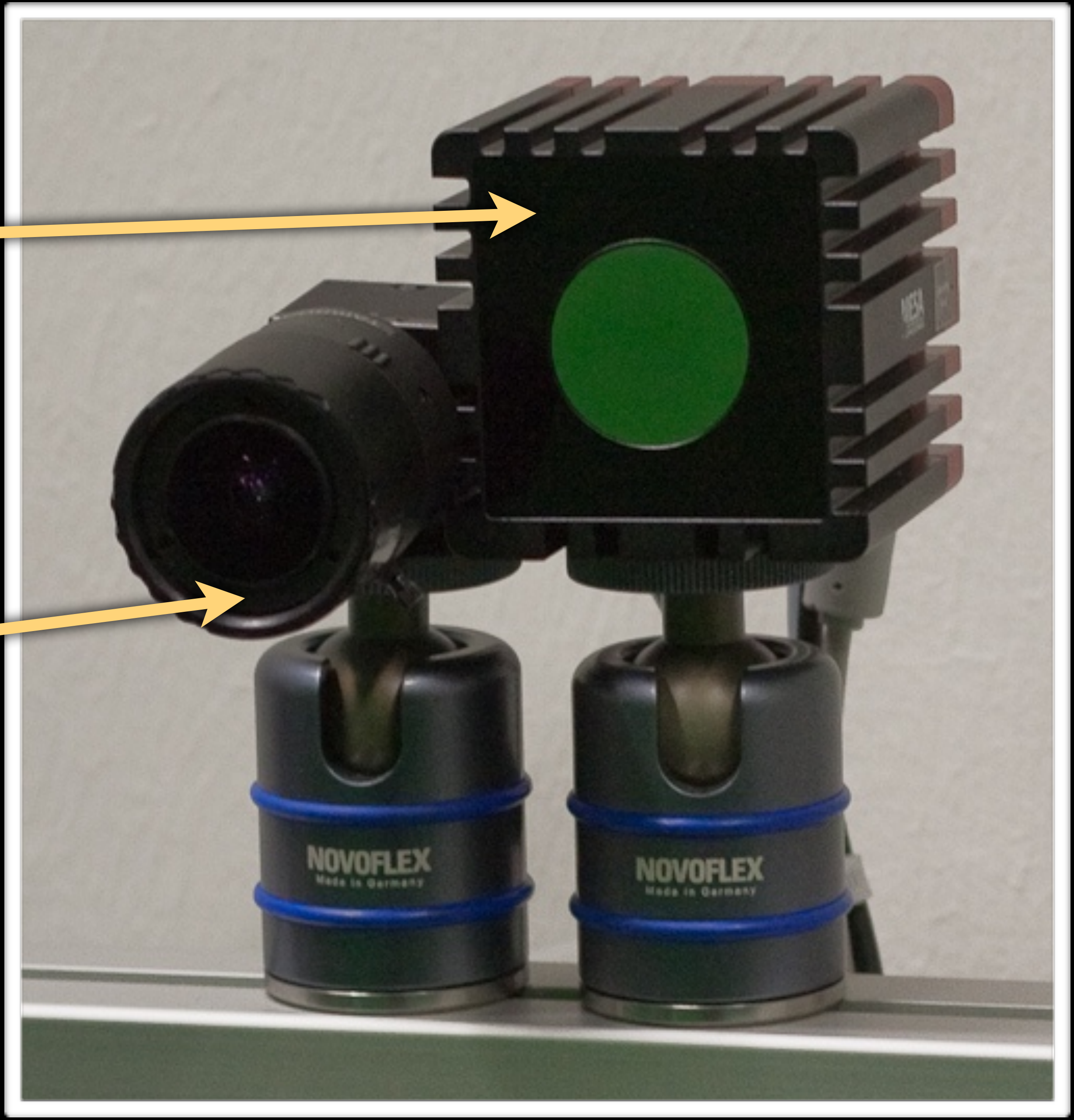
Related work – depth-based stylisation

- NPR camera
  [Raskar et al. 2004]
- 2.5-D video stylisation
  [Snavely et al. 2006]
- photometric surface enhancement
  [Malzbender et al. 2006]
- Images with normals
  [Toler-Franklin et al. 2007]
- context-aware light source
  [Wang et al. 2010]



Prototype camera hardware

- depth sensor:
  - MESA Imaging SR4000
  - 176 × 144 resolution
- video camera:
  - PointGrey Flea2
  - 1024 × 768 resolution
- hardware synchronised

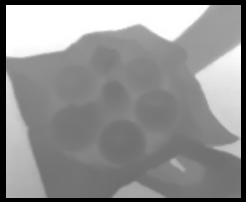


Microsoft Kinect

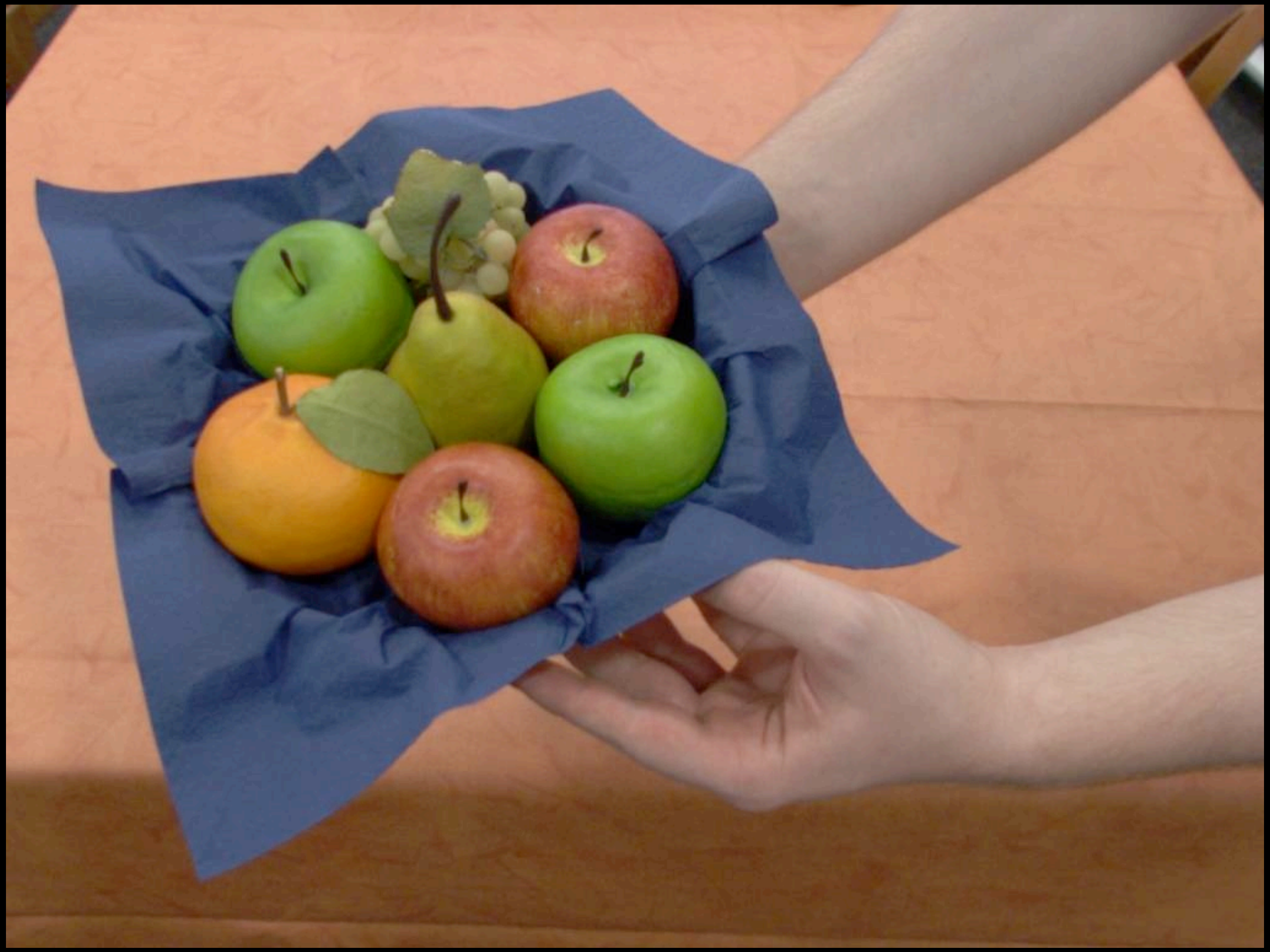
- low-cost IR-based active stereo + colour camera in one case
- our approach is also applicable to the Kinect
- but our prototype gives us full hardware + software control

Points to address

- Resolution mismatch



176 × 144



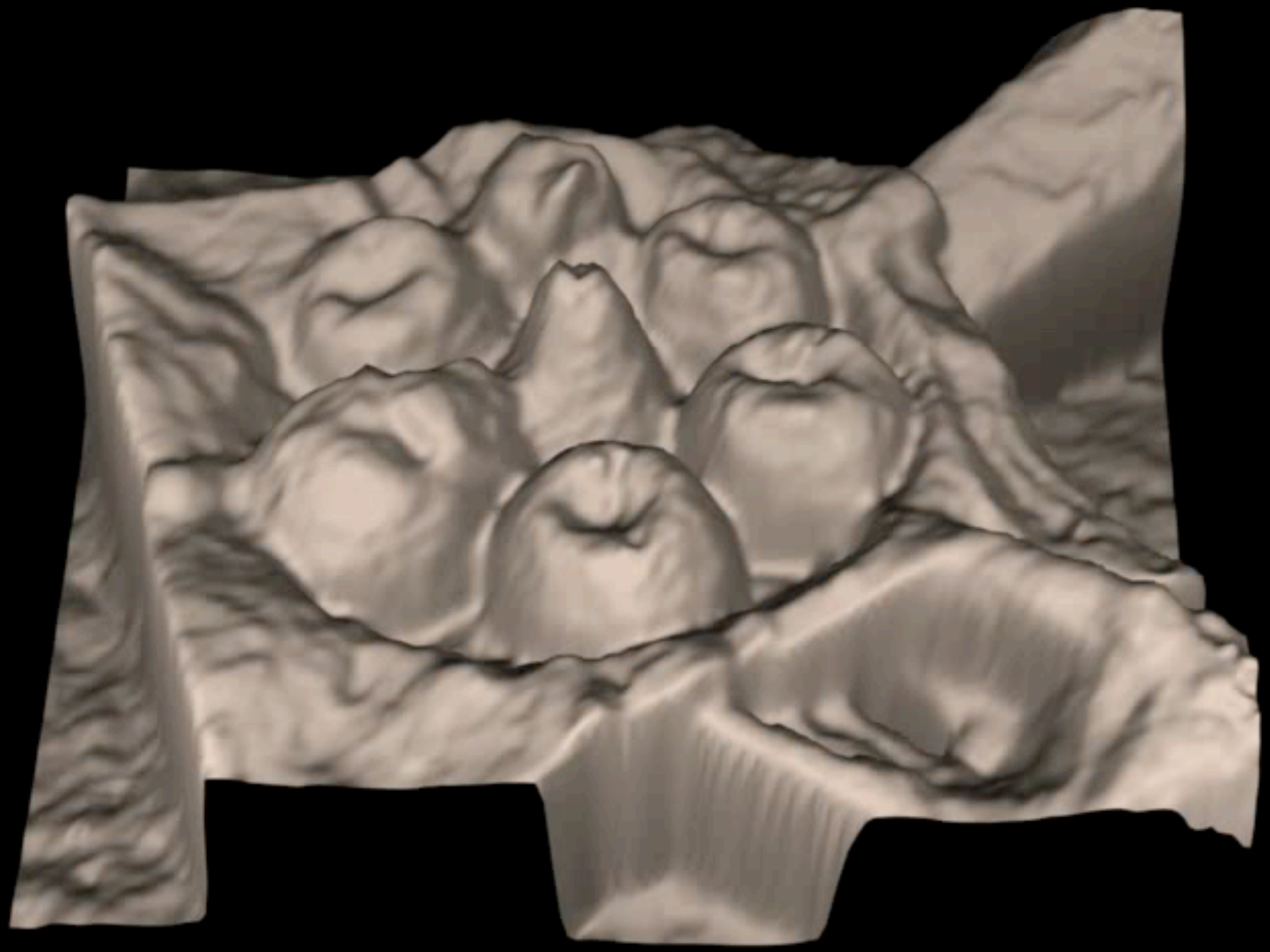
1024 × 768

Points to address

- Resolution mismatch
- Video alignment

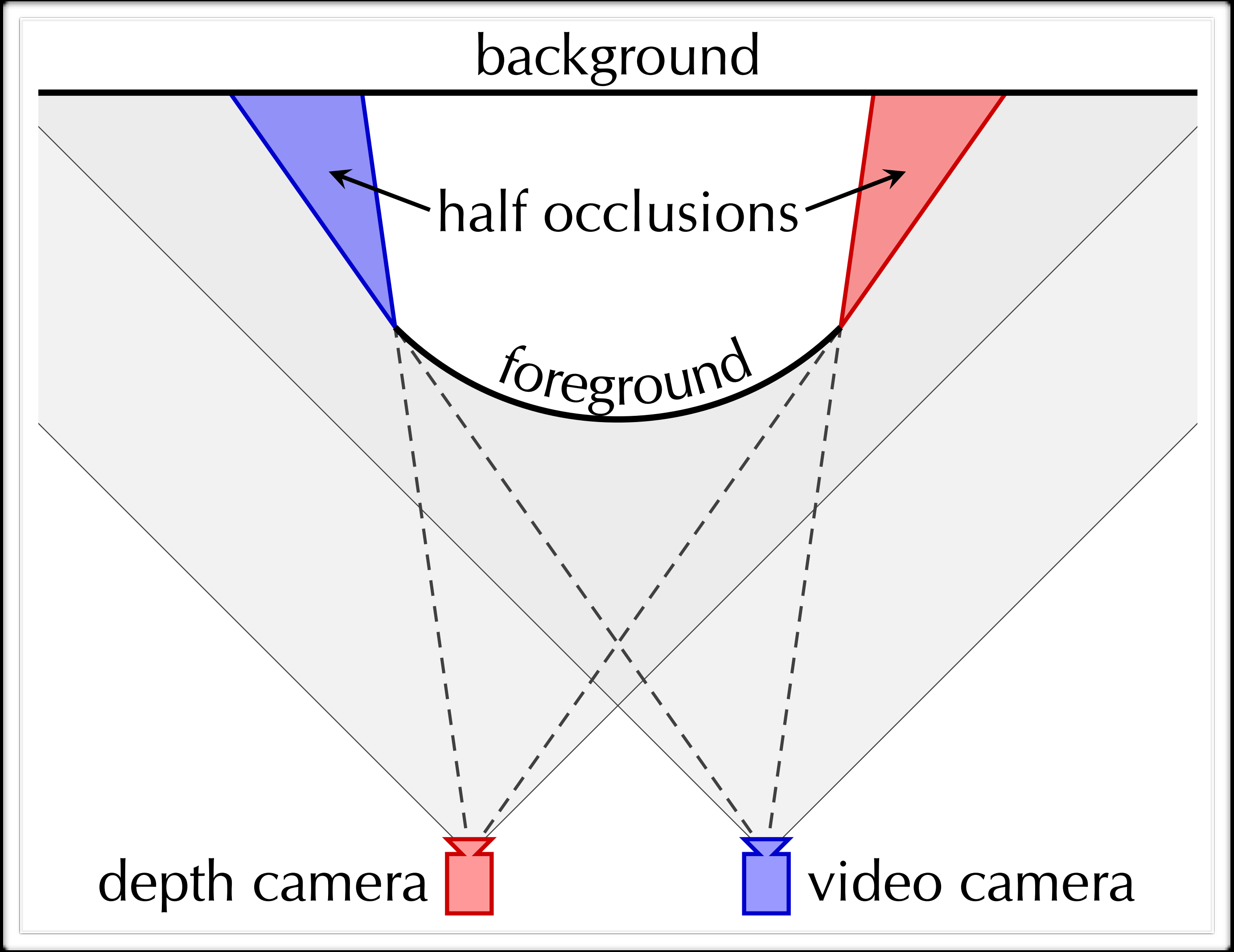
Points to address

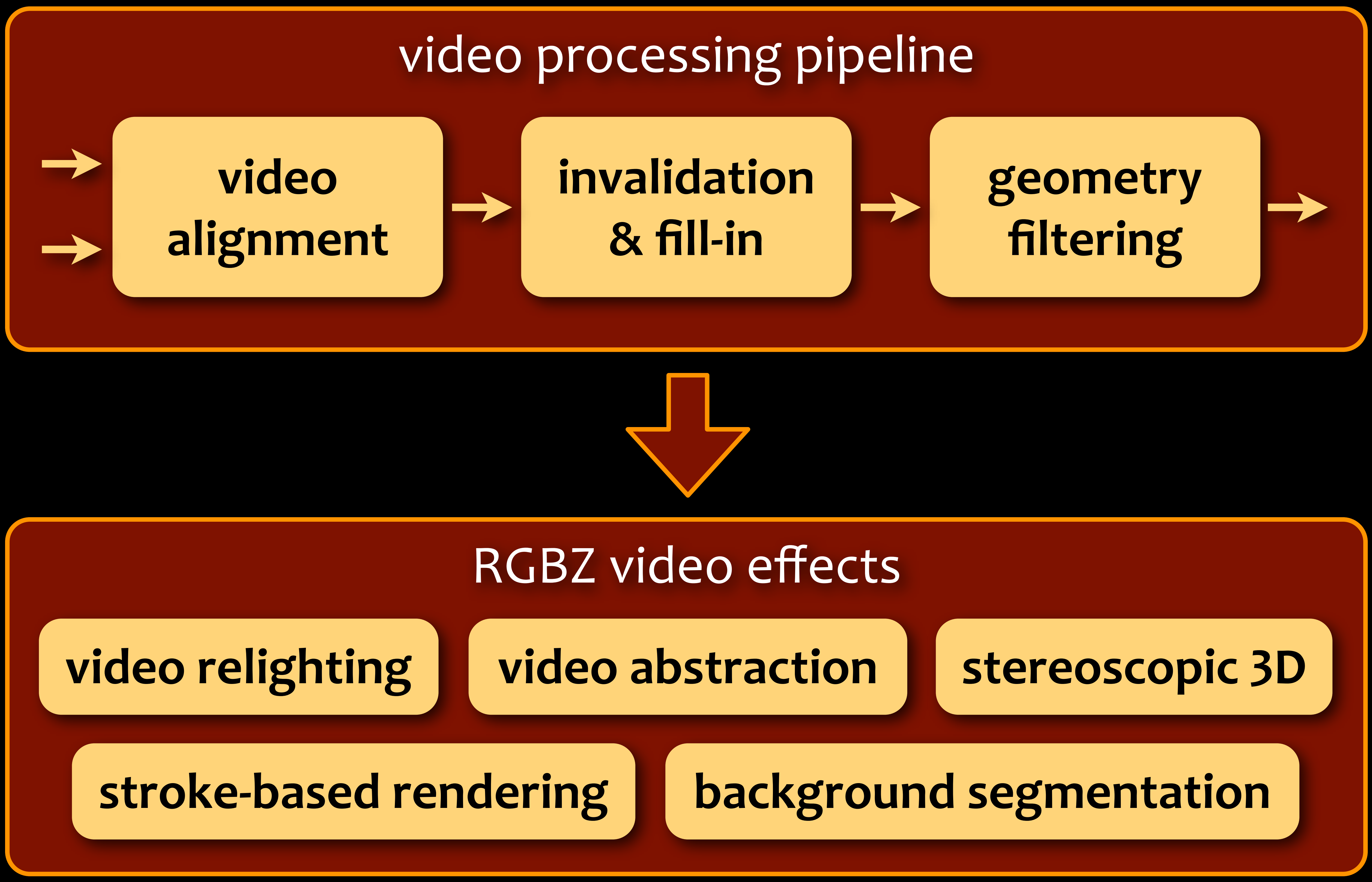
- Resolution mismatch
- Video alignment
- Noisy depth data



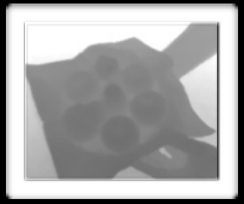
Points to address

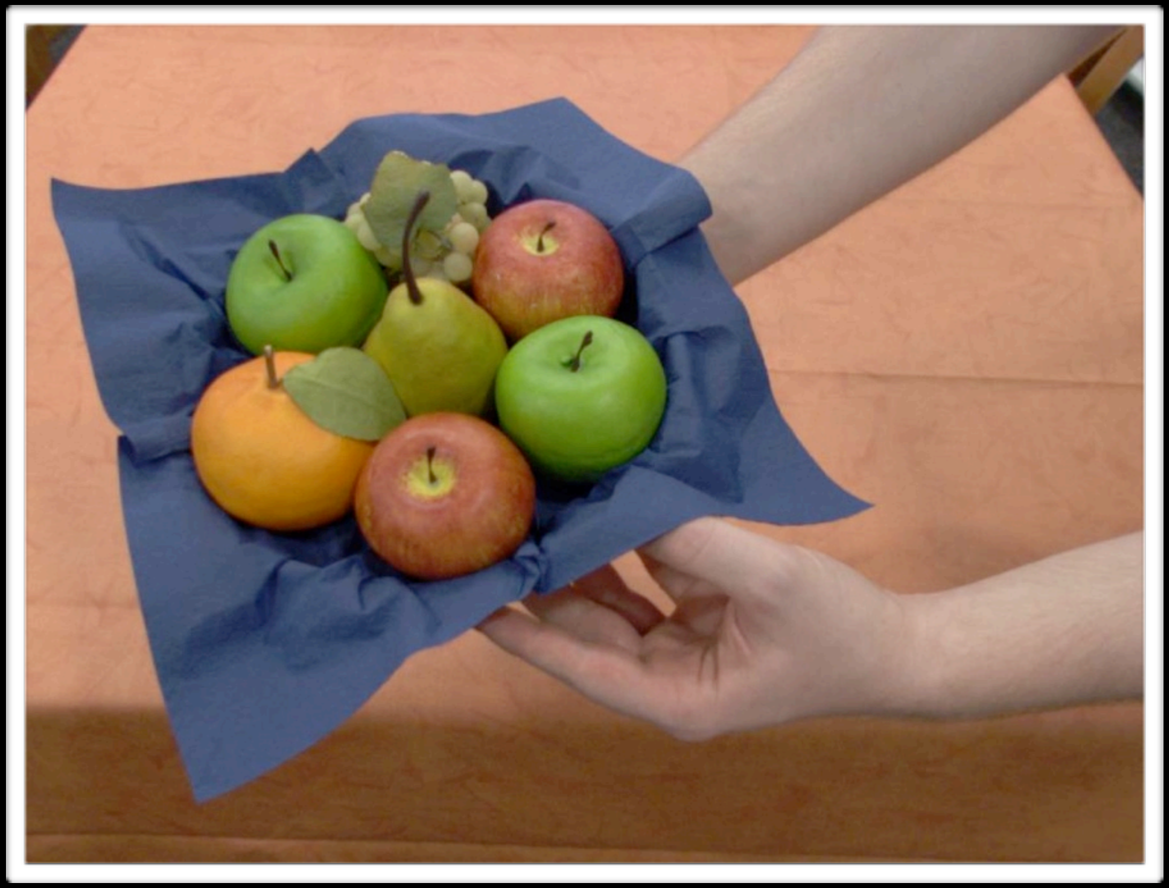
- Resolution mismatch
- Video alignment
- Noisy depth data
- Half-occlusions



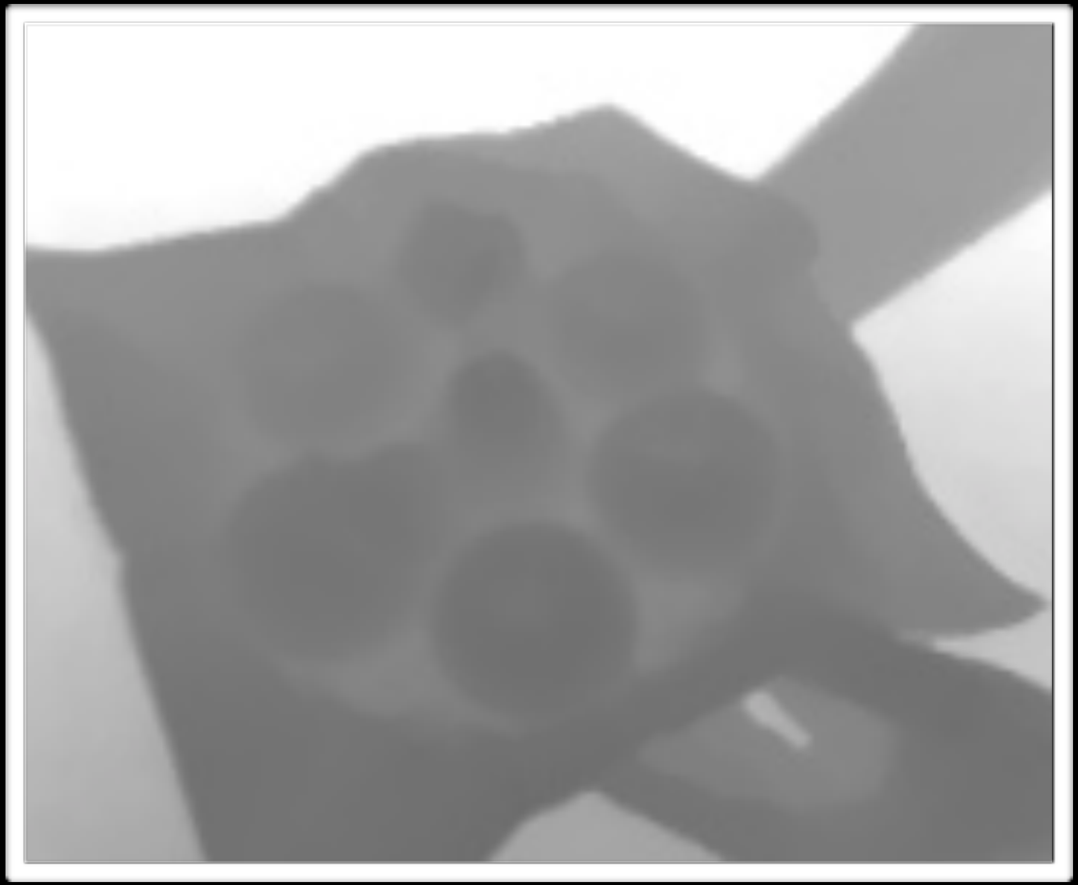

video processing pipeline
video alignment
invalidation & fill-in
geometry filtering
RGBZ video effects
video relighting
video abstraction
stereoscopic 3D
stroke-based rendering
background segmentation

Video alignment

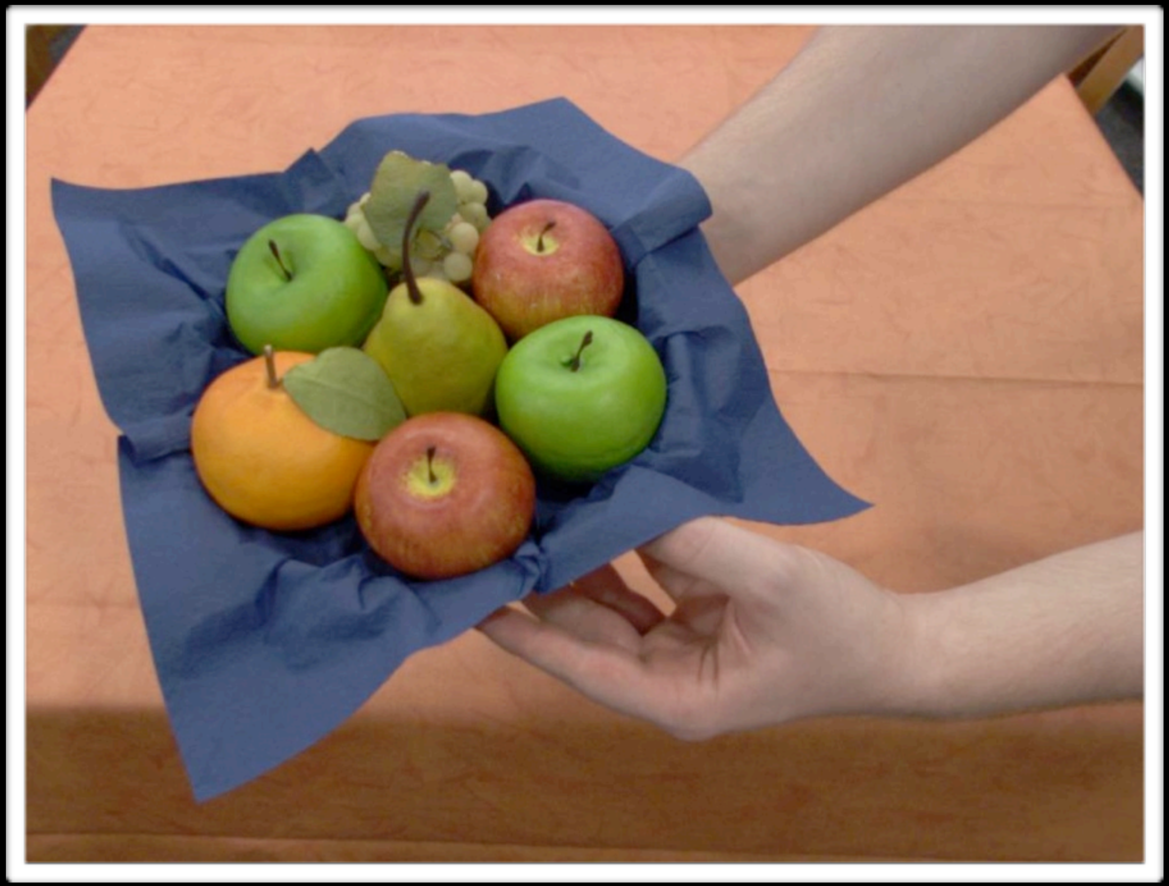

depth map


colour image

Video alignment

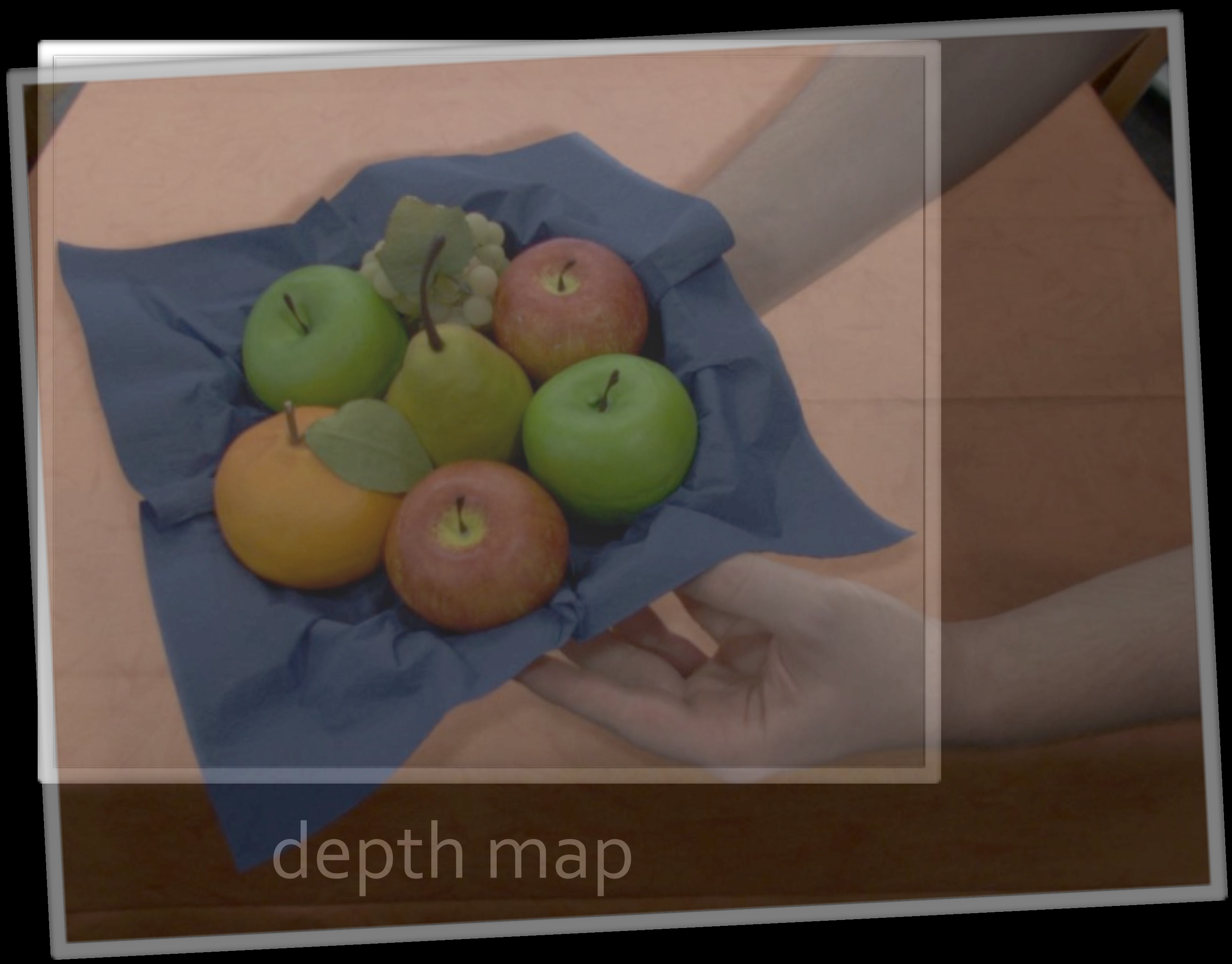


depth map



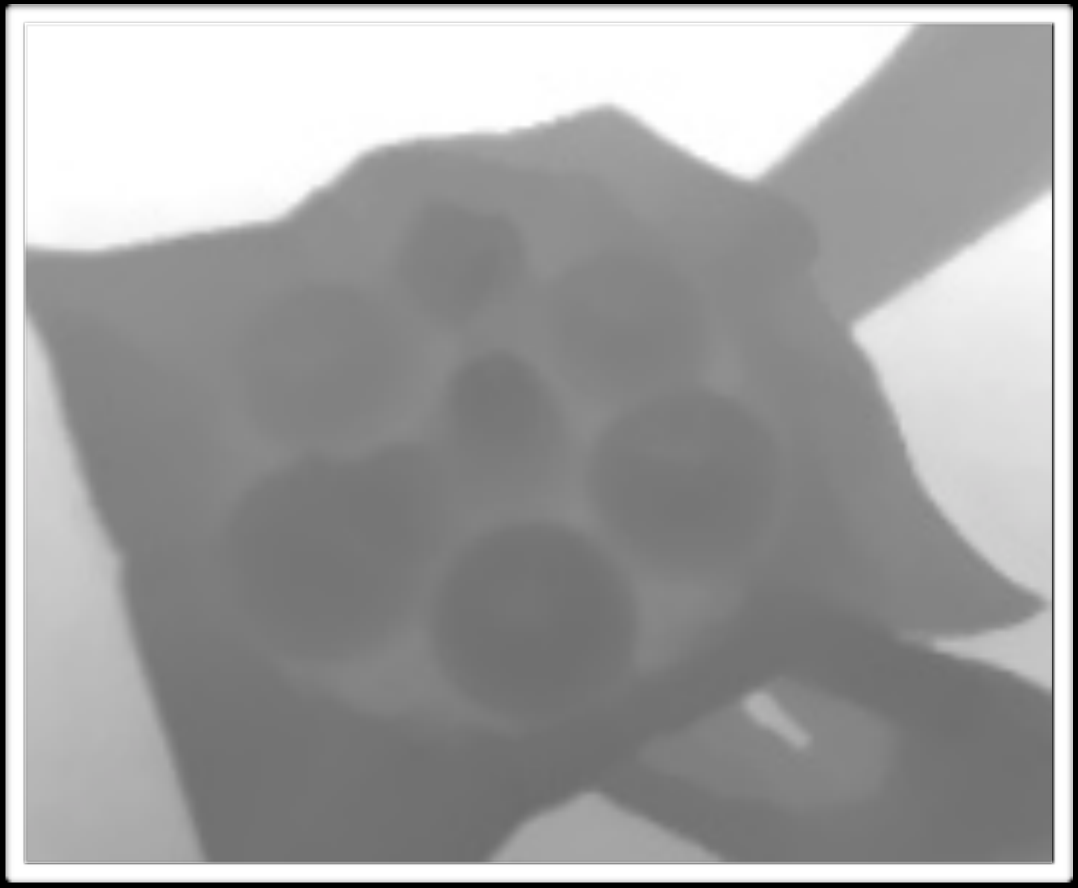
colour image

Video alignment

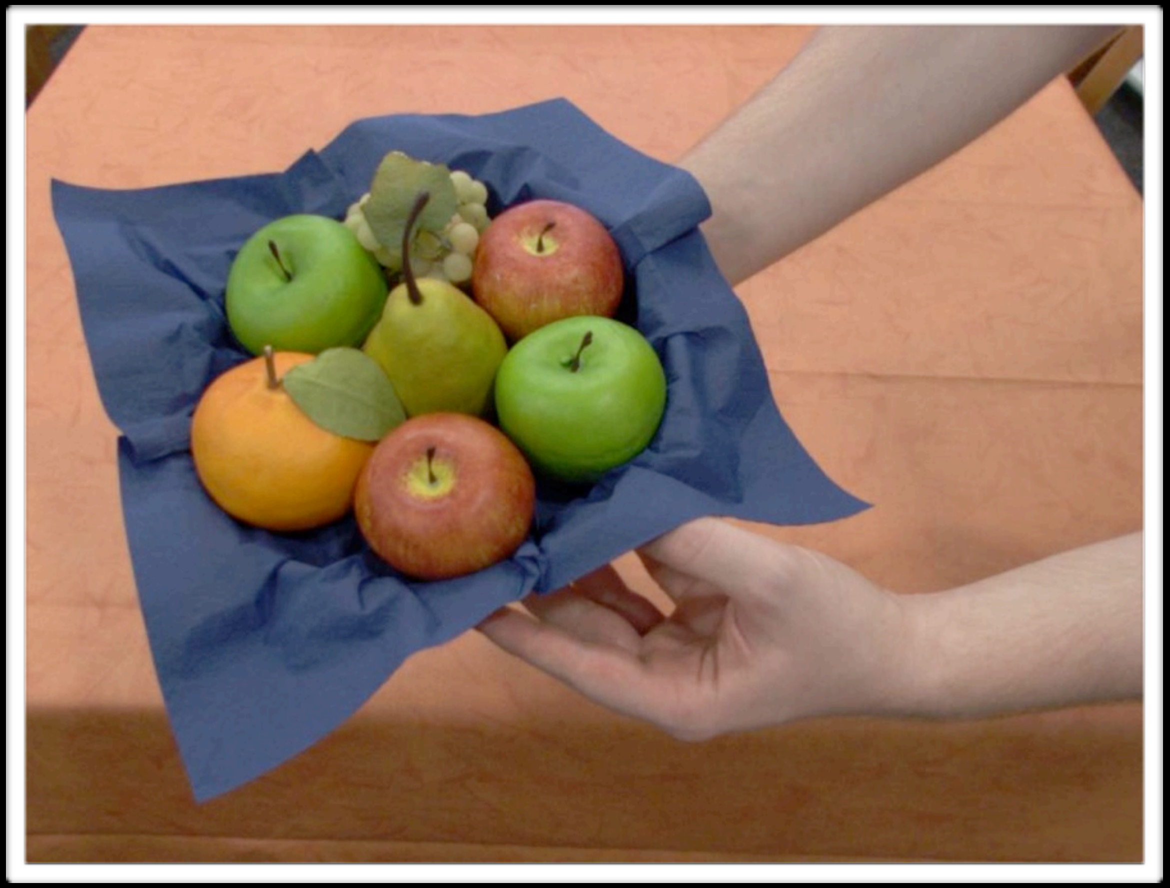


colour image

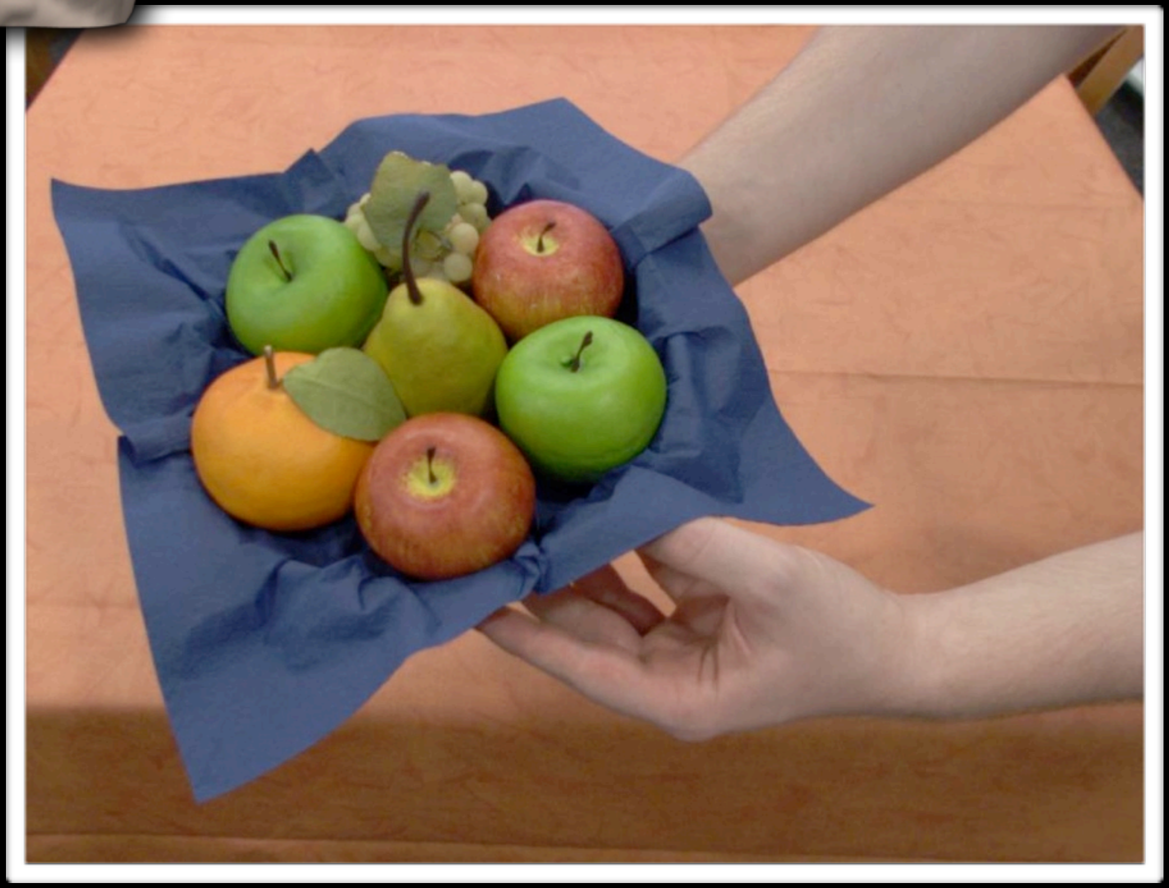
Video alignment



depth map



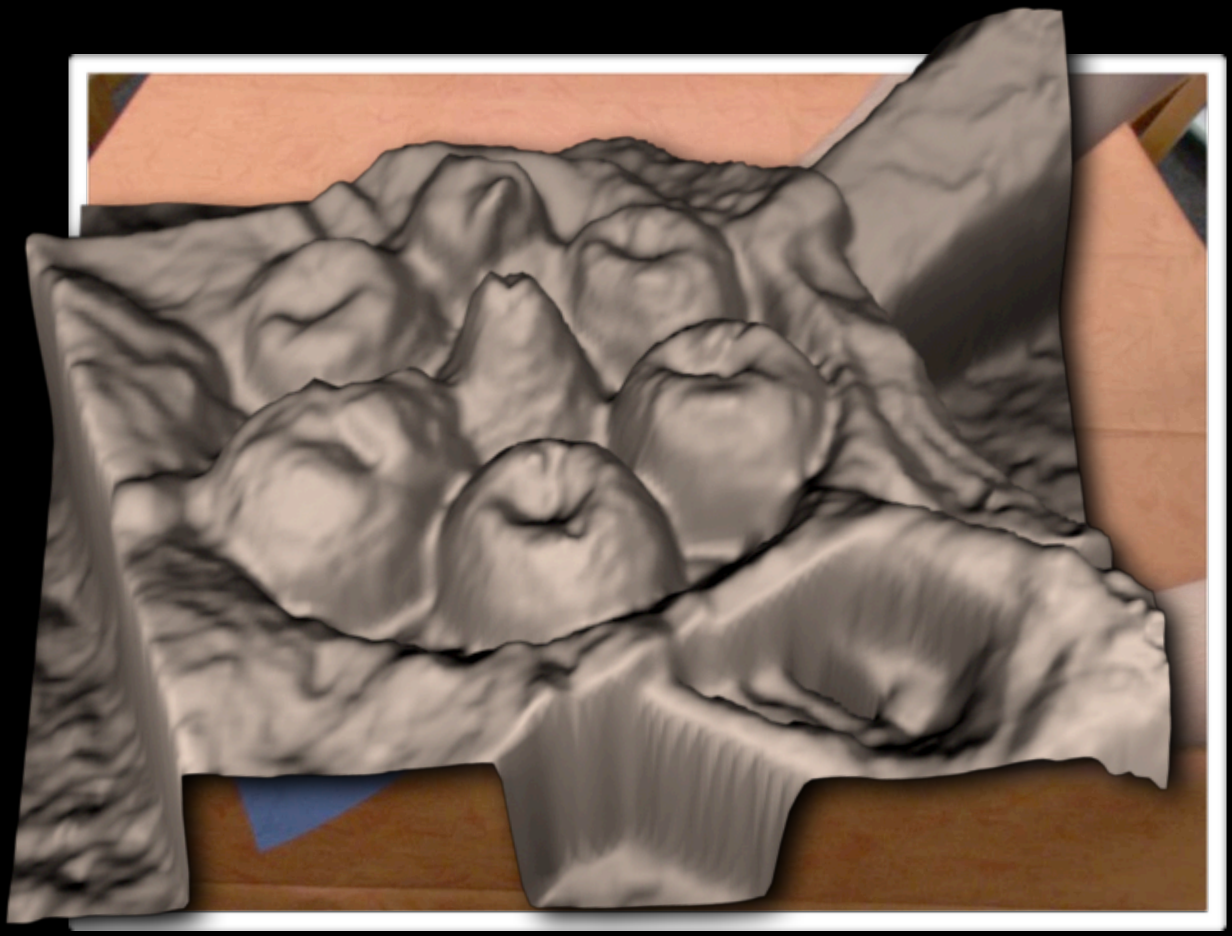
colour image



depth map

colour image

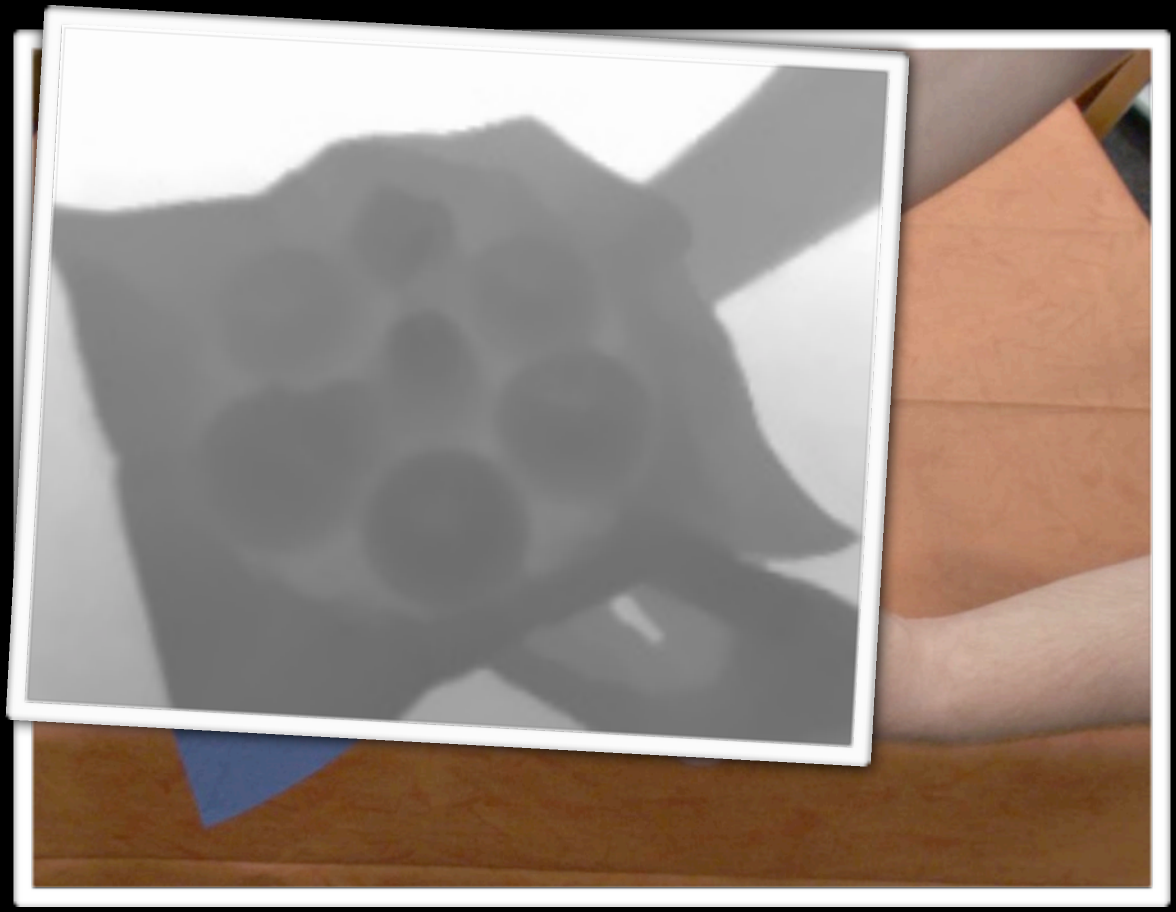
Video alignment



depth map

colour image

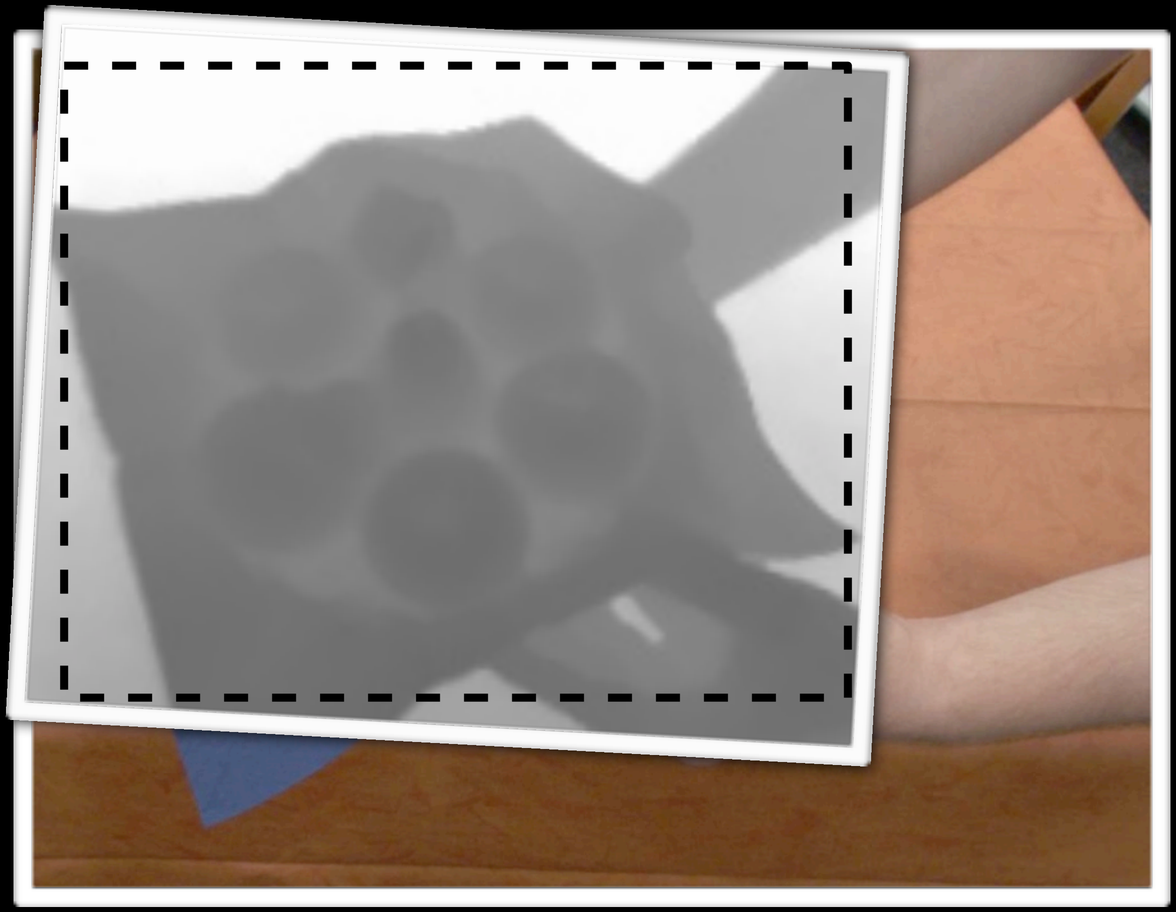
Video alignment



depth map

colour image

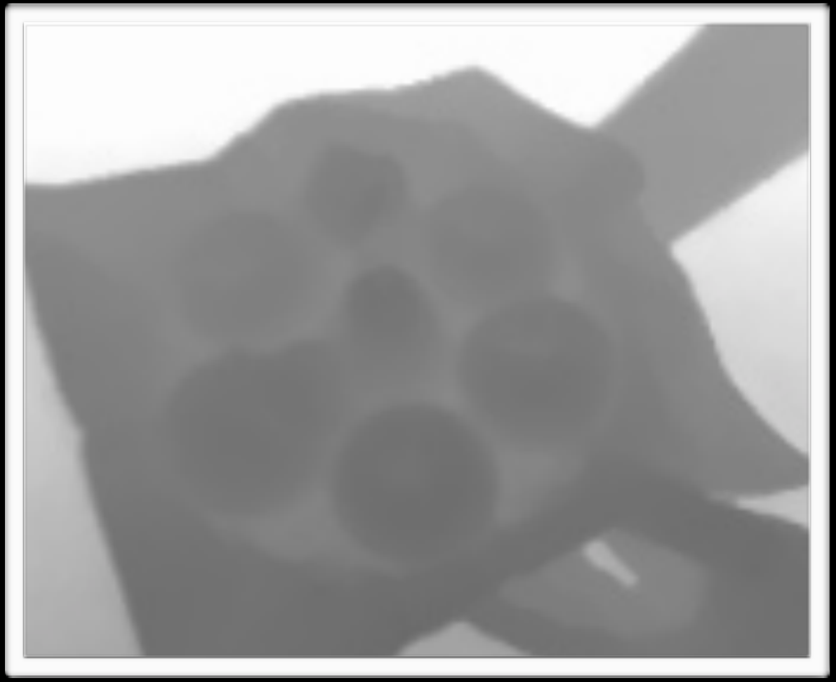
Video alignment



depth map

colour image

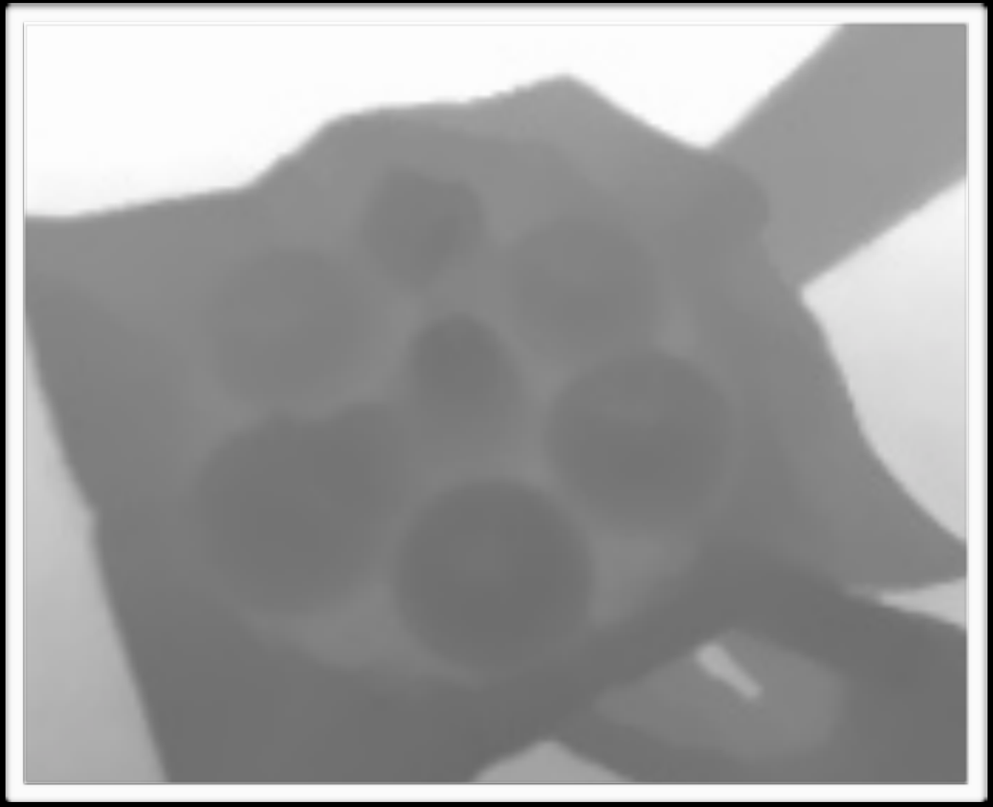
Video alignment



depth map

colour image

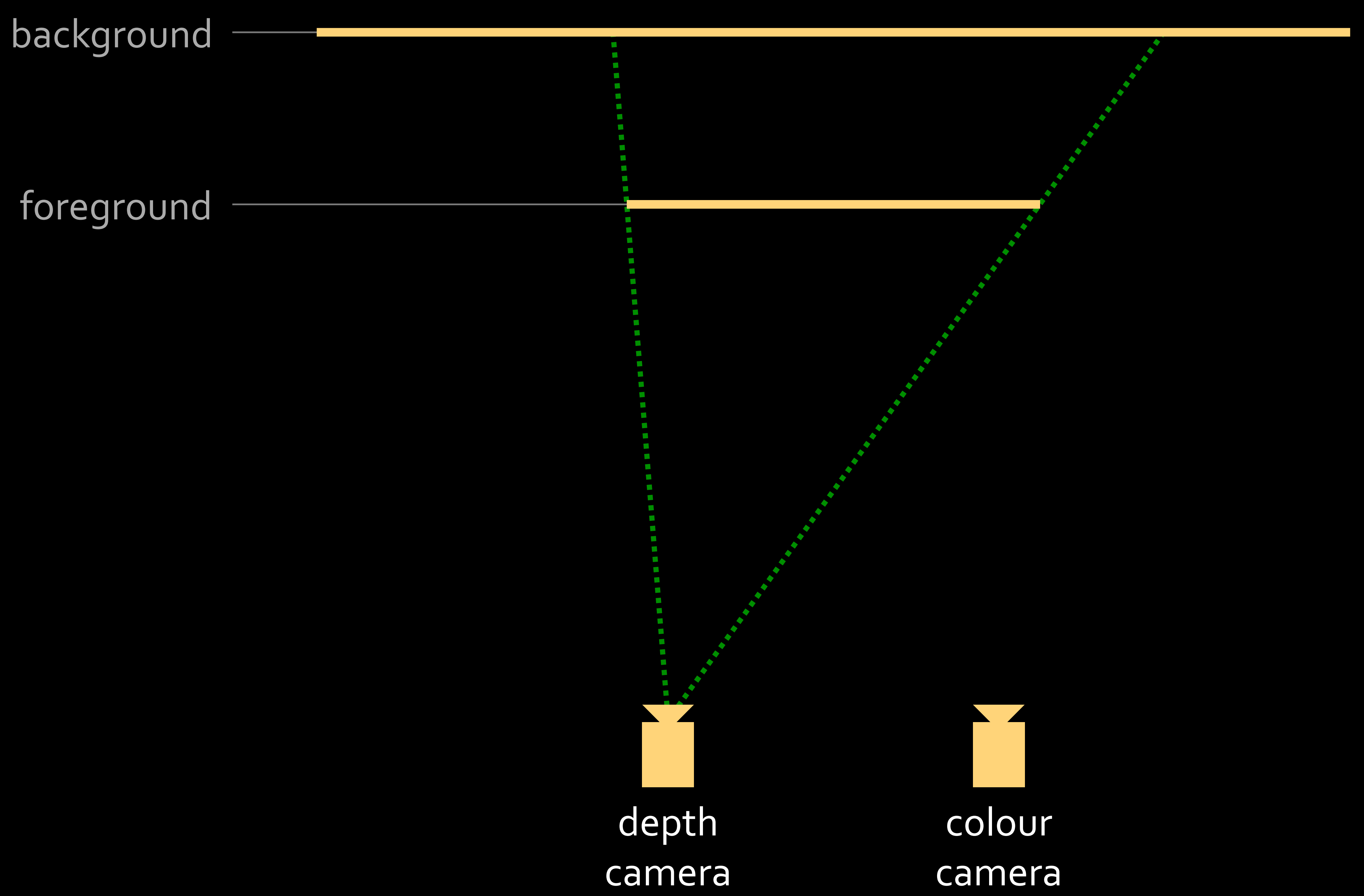
Video alignment



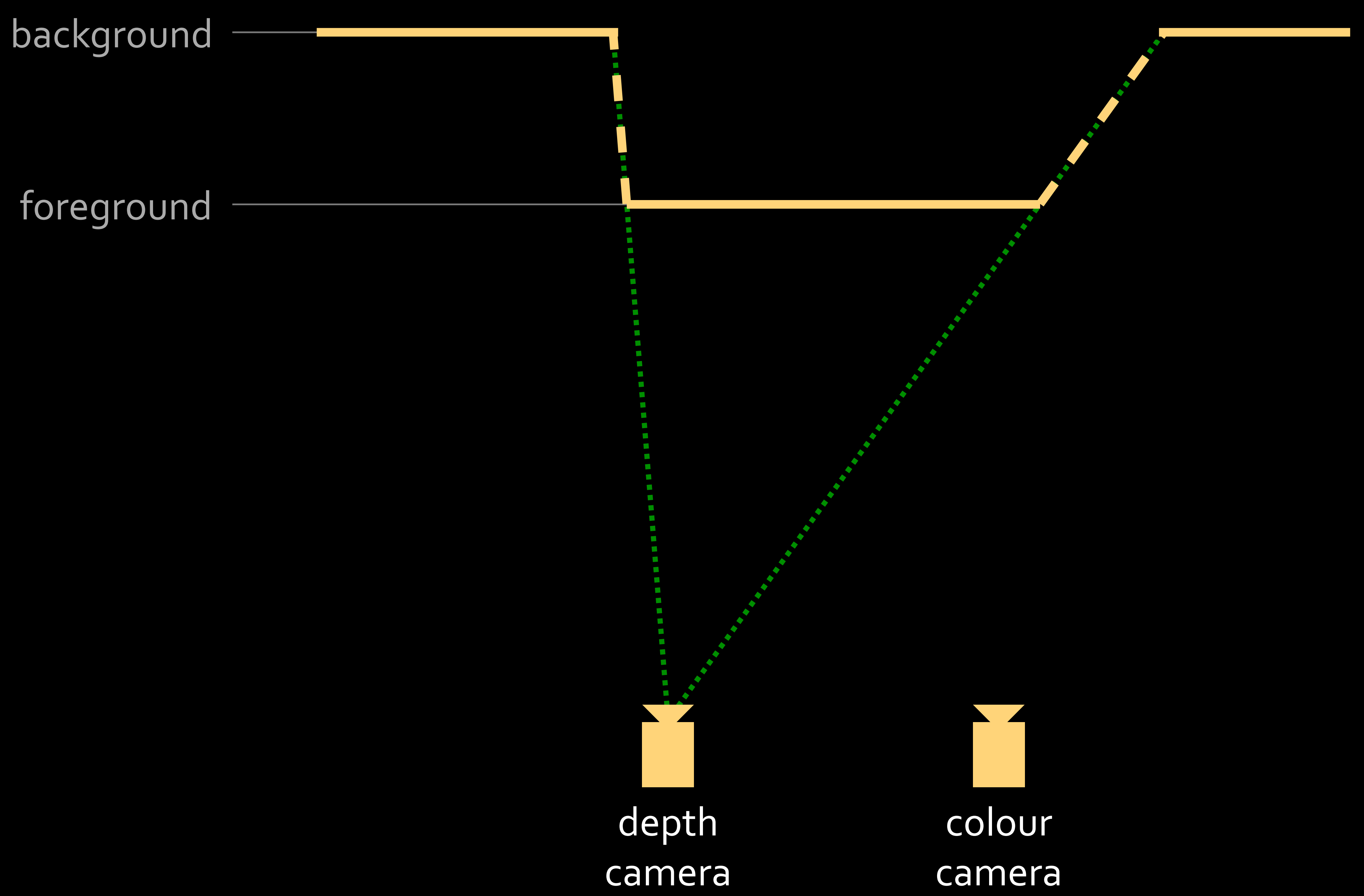
depth map

colour image

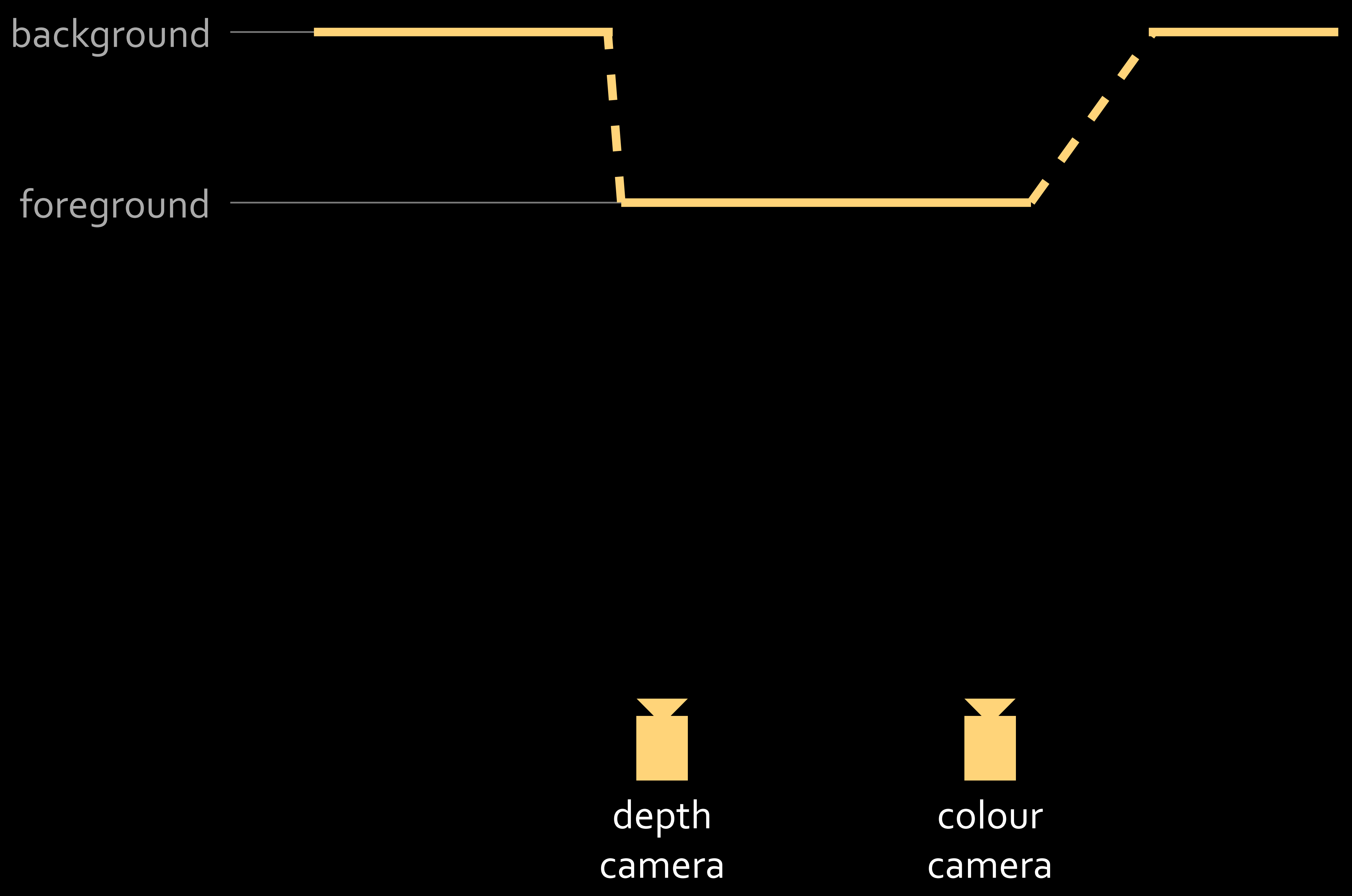
Geometry invalidation and fill-in



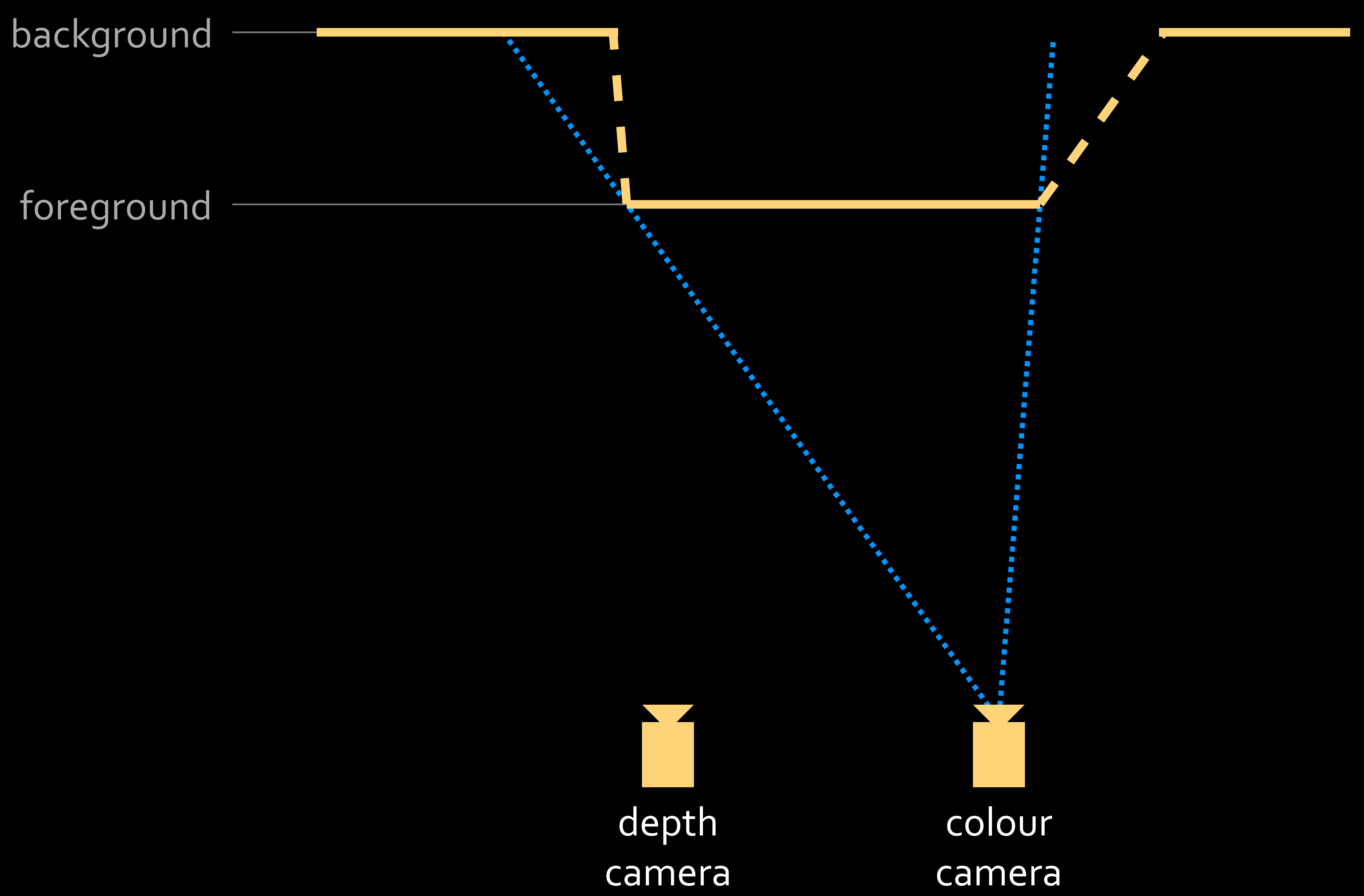
Geometry invalidation and fill-in



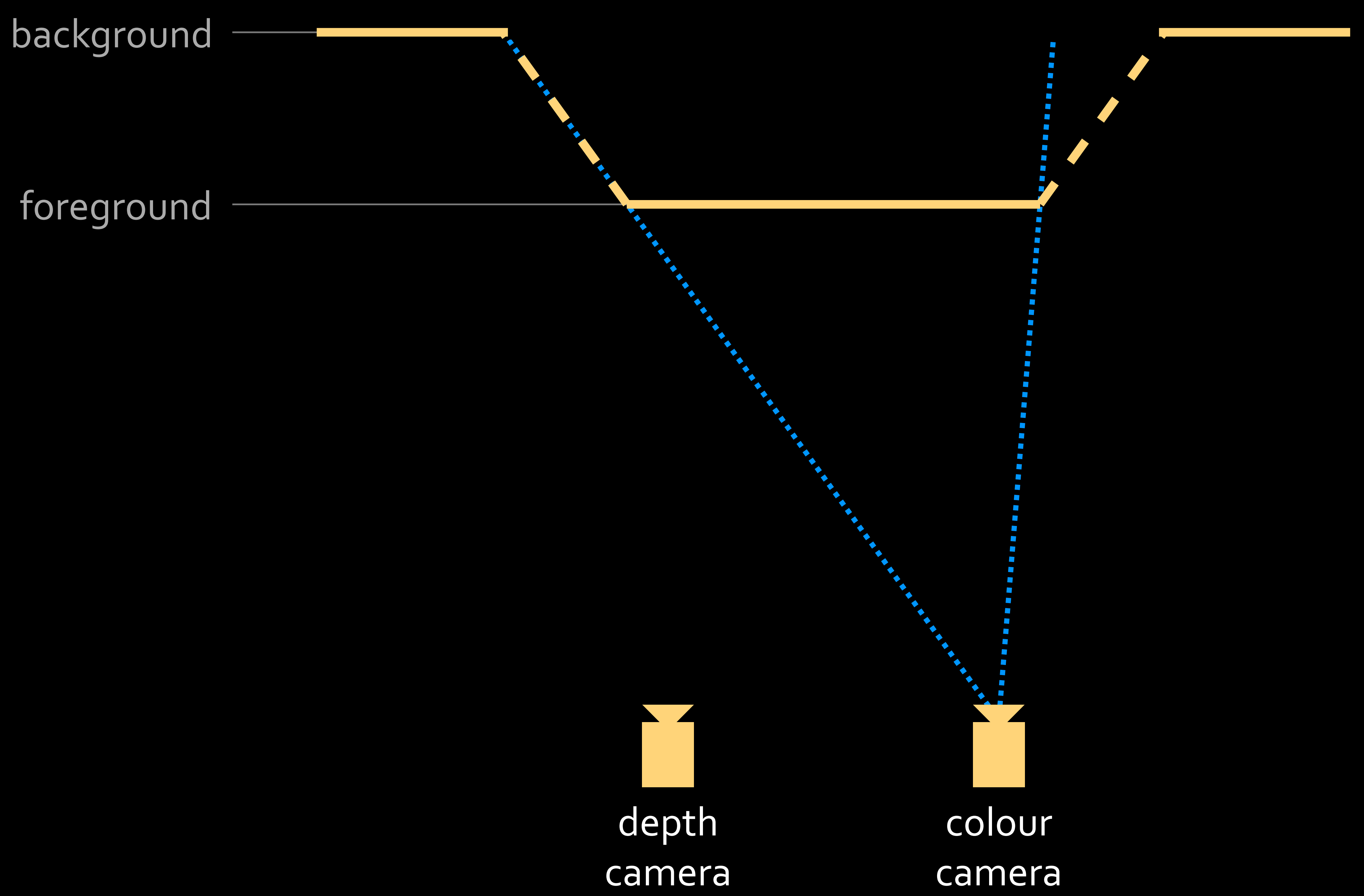
Geometry invalidation and fill-in



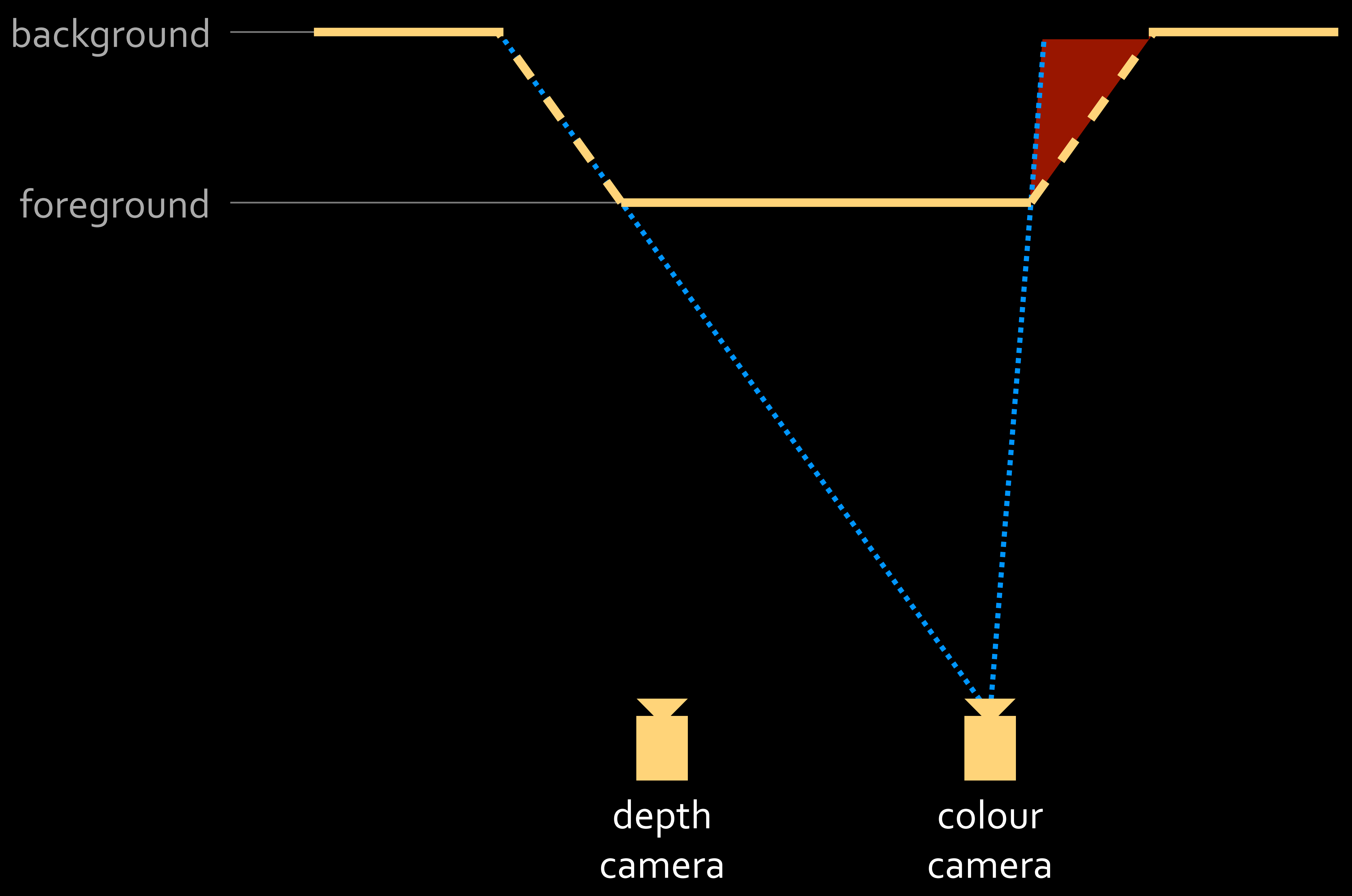
Geometry invalidation and fill-in



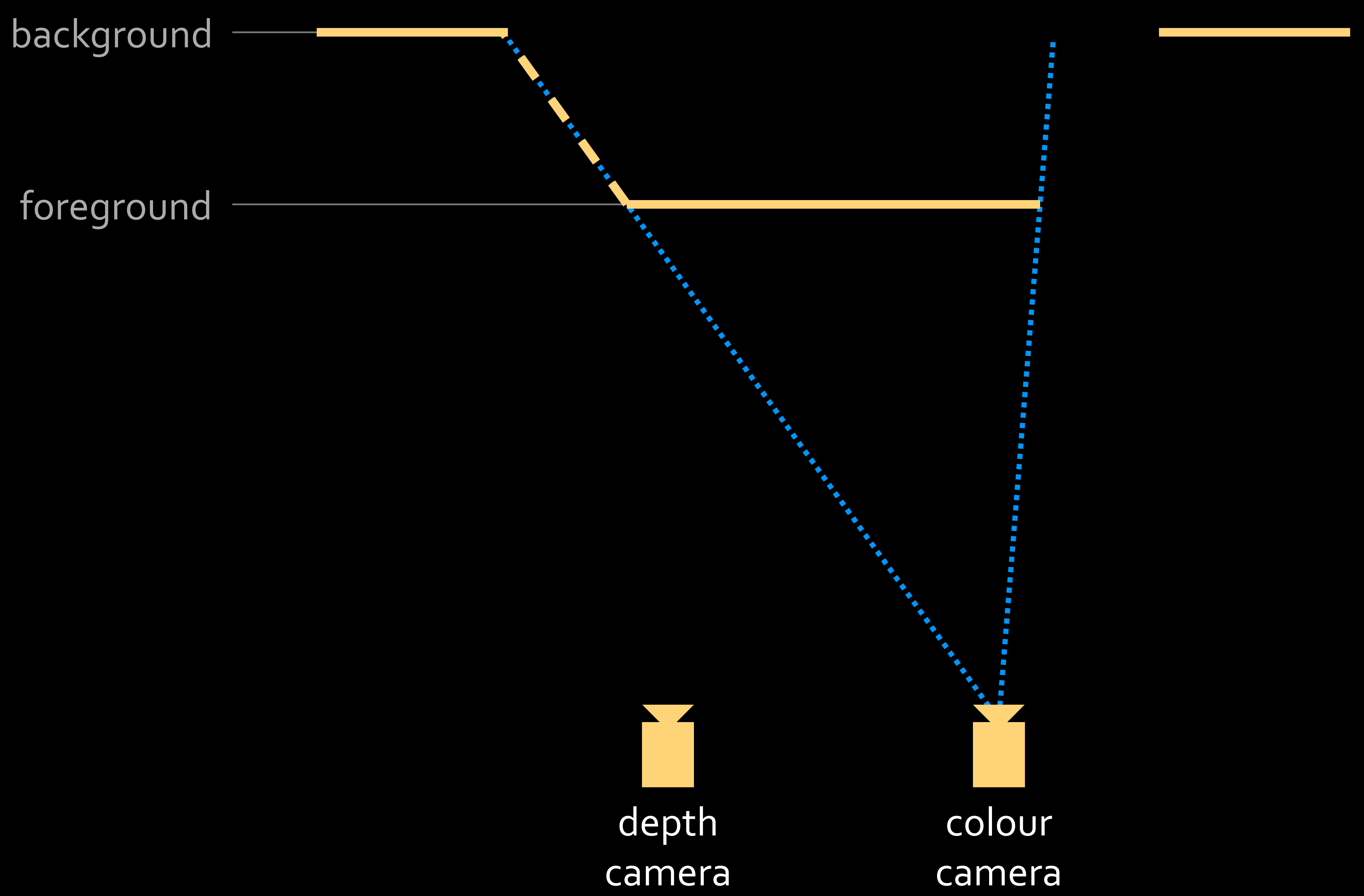
Geometry invalidation and fill-in



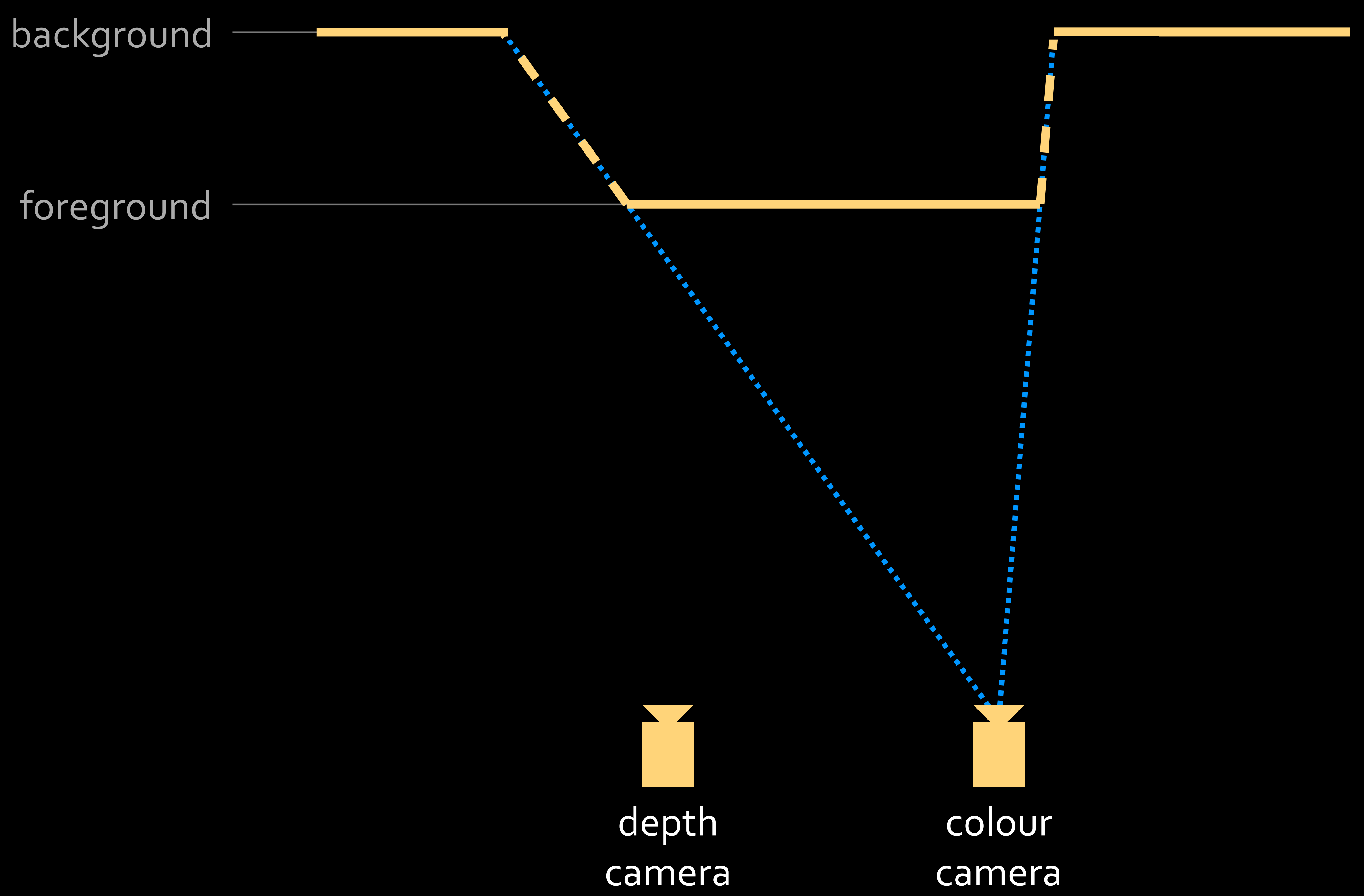
Geometry invalidation and fill-in



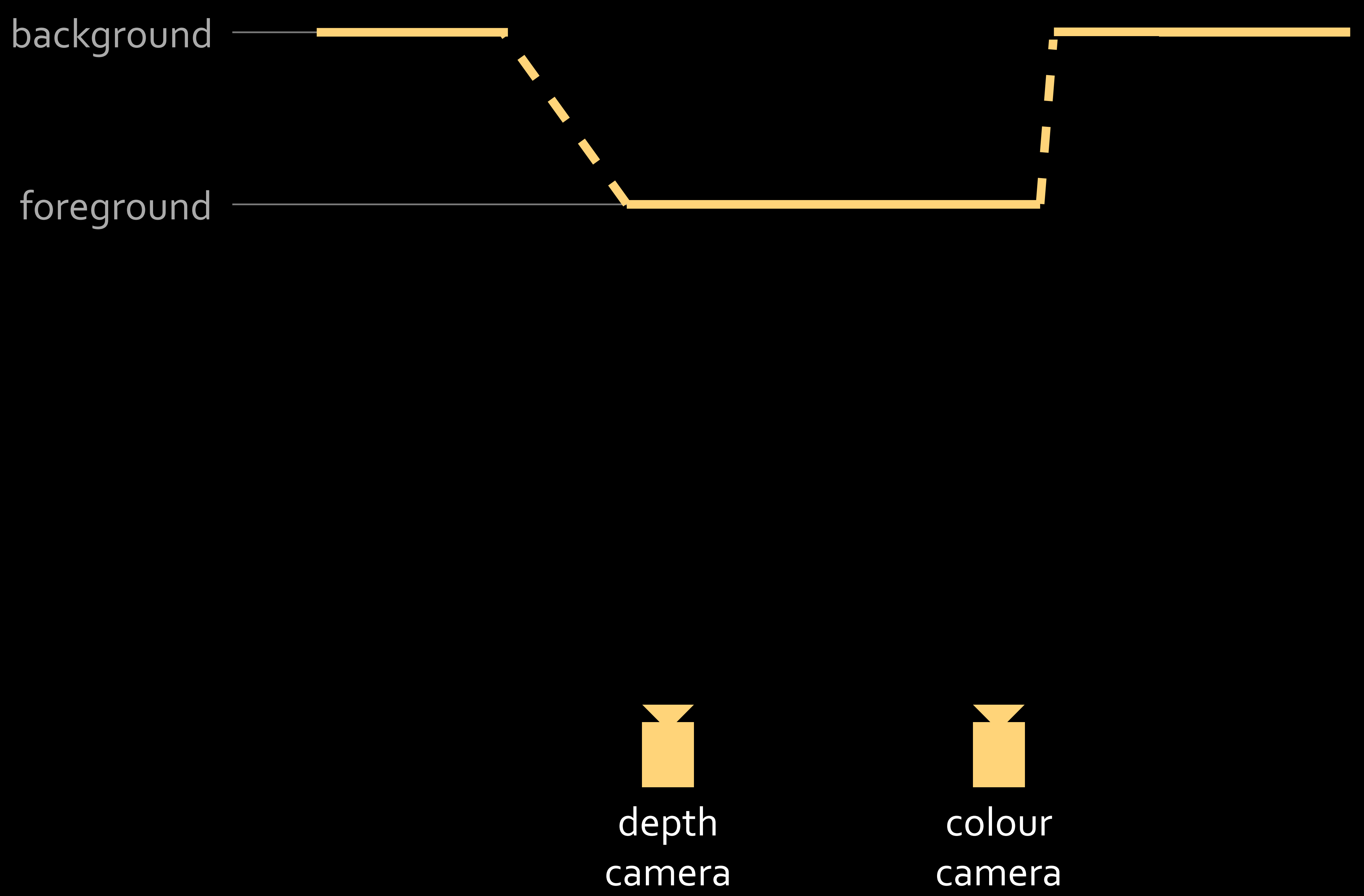
Geometry invalidation and fill-in



Geometry invalidation and fill-in



Geometry invalidation and fill-in



Geometry invalidation and fill-in

Geometry invalidation and fill-in

aligned geometry (before invalidation)

Geometry invalidation and fill-in

invalidated geometry (in orange)

Geometry invalidation and fill-in

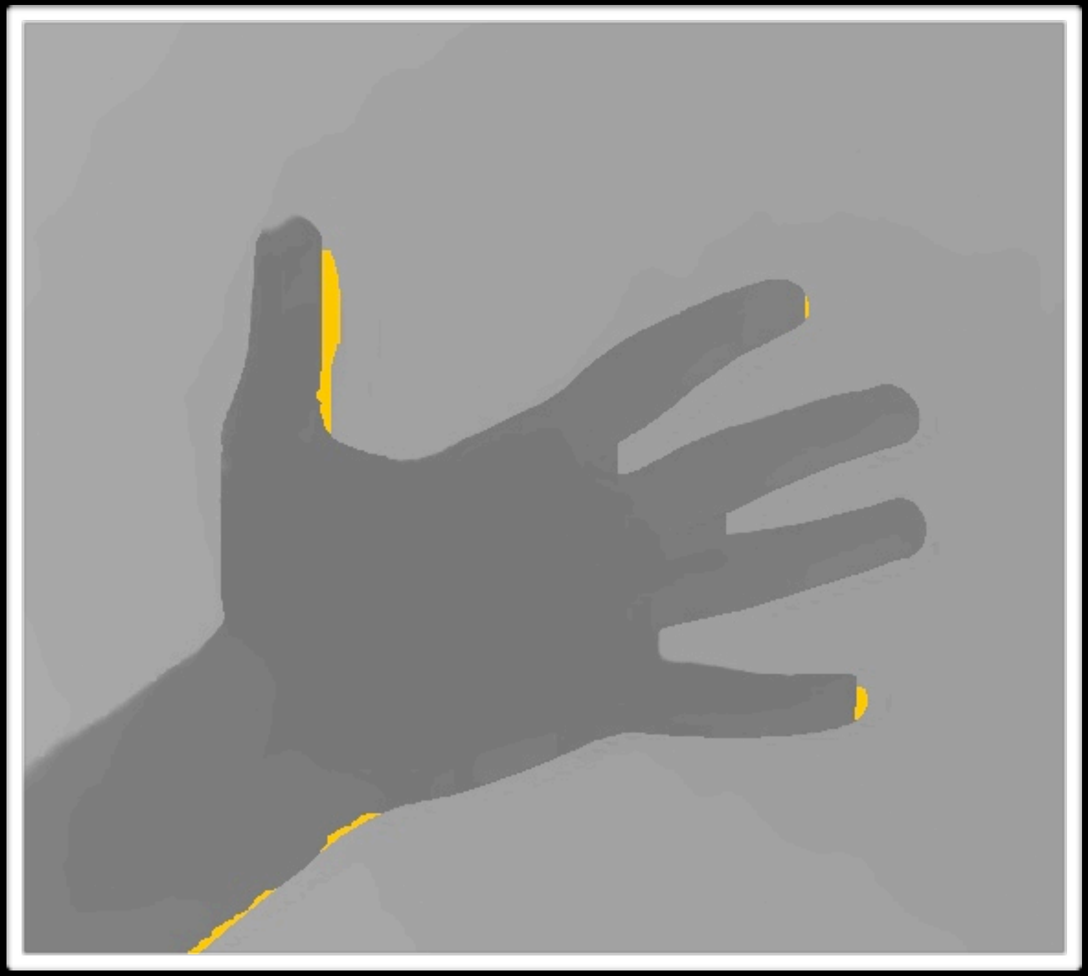
single-resolution fill-in (σ_s = 27)

70.2 ms

92.8 ms

Geometry invalidation and fill-in

single-resolution fill-in ($\sigma_s = 10$)

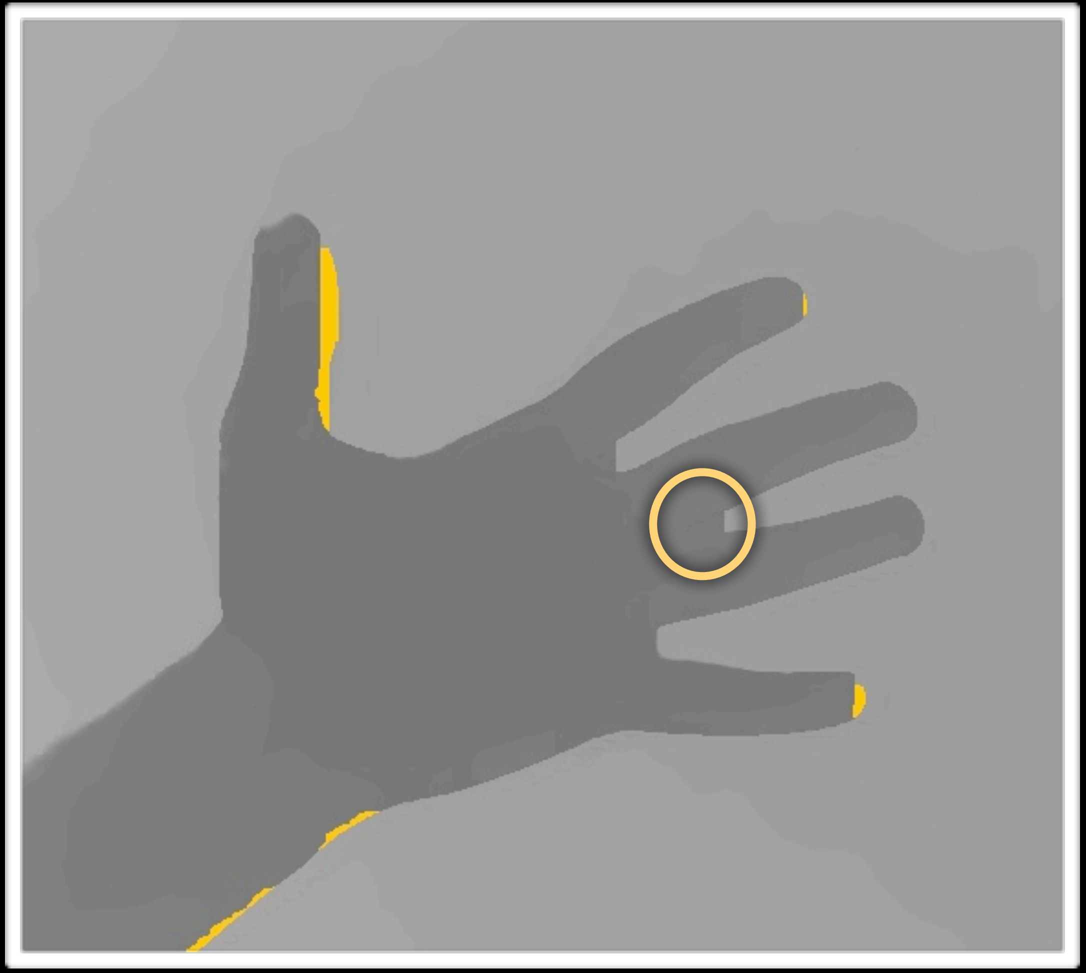


10.4 ms

14.8 ms

Geometry invalidation and fill-in

single-resolution fill-in (σ_s = 10)

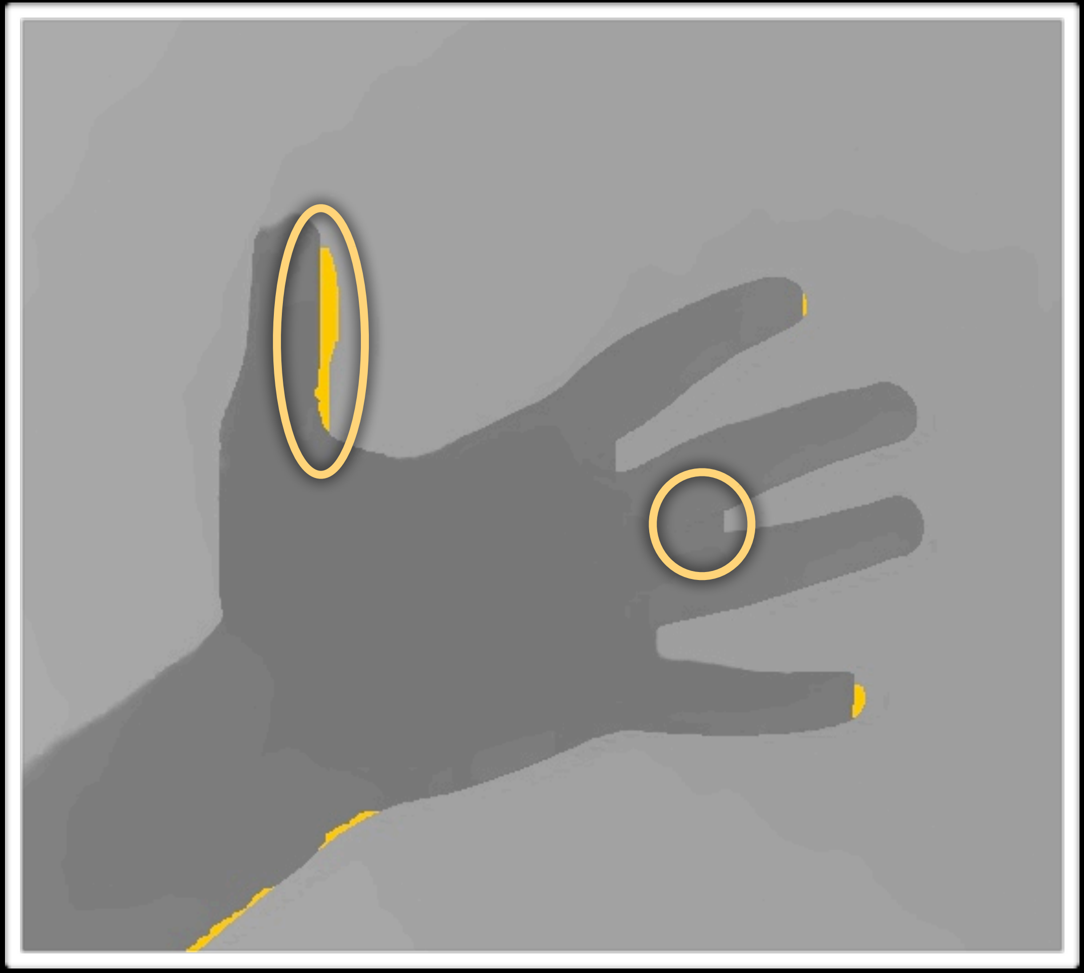


10.4 ms

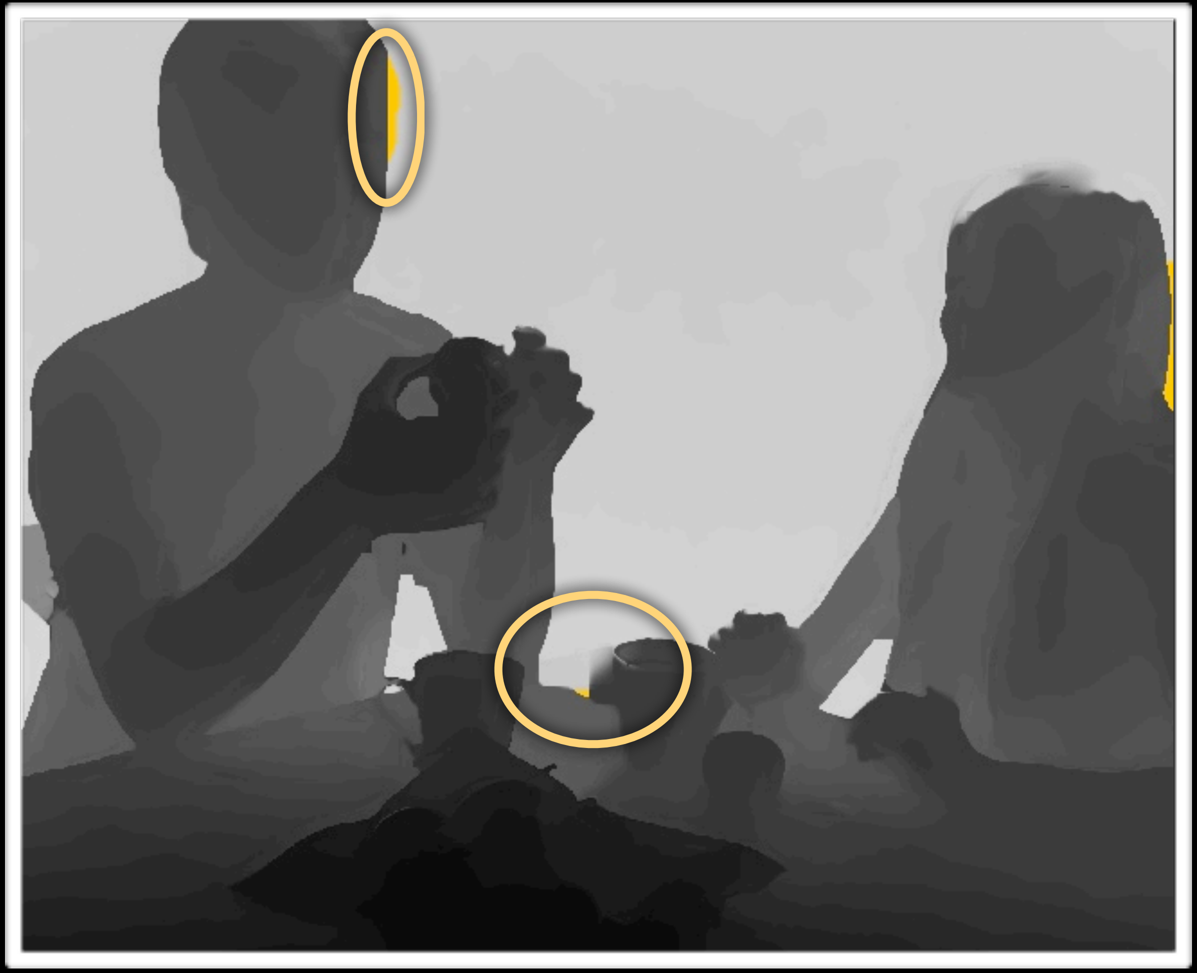
14.8 ms

Geometry invalidation and fill-in

single-resolution fill-in ($\sigma_s = 10$)



10.4 ms



14.8 ms

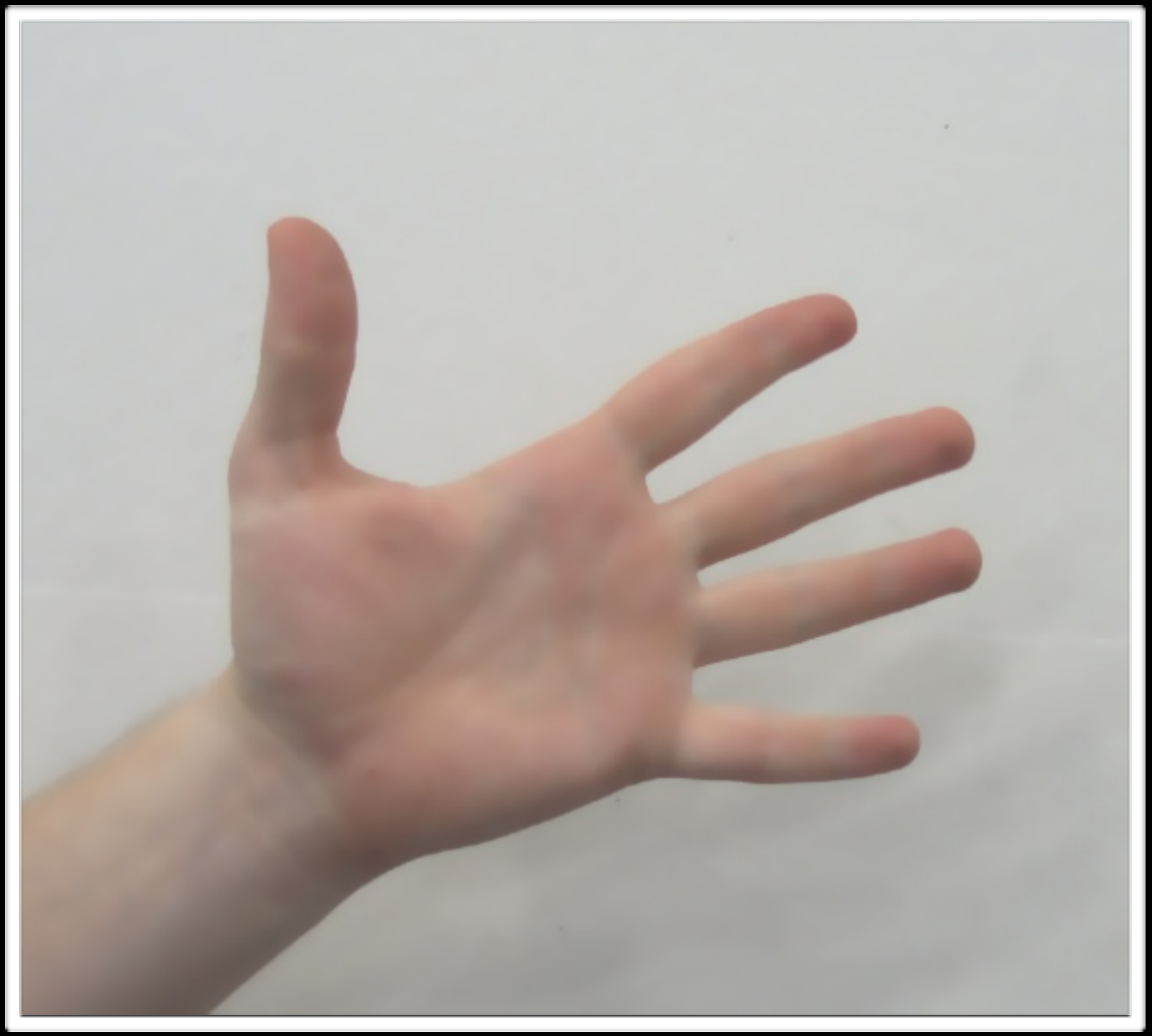
Geometry invalidation and fill-in

our multi-resolution fill-in (n = 3, g = 3, σ_s = 3)

13.7 ms

15.1 ms

Geometry invalidation and fill-in



colour image (level $k = 0$)

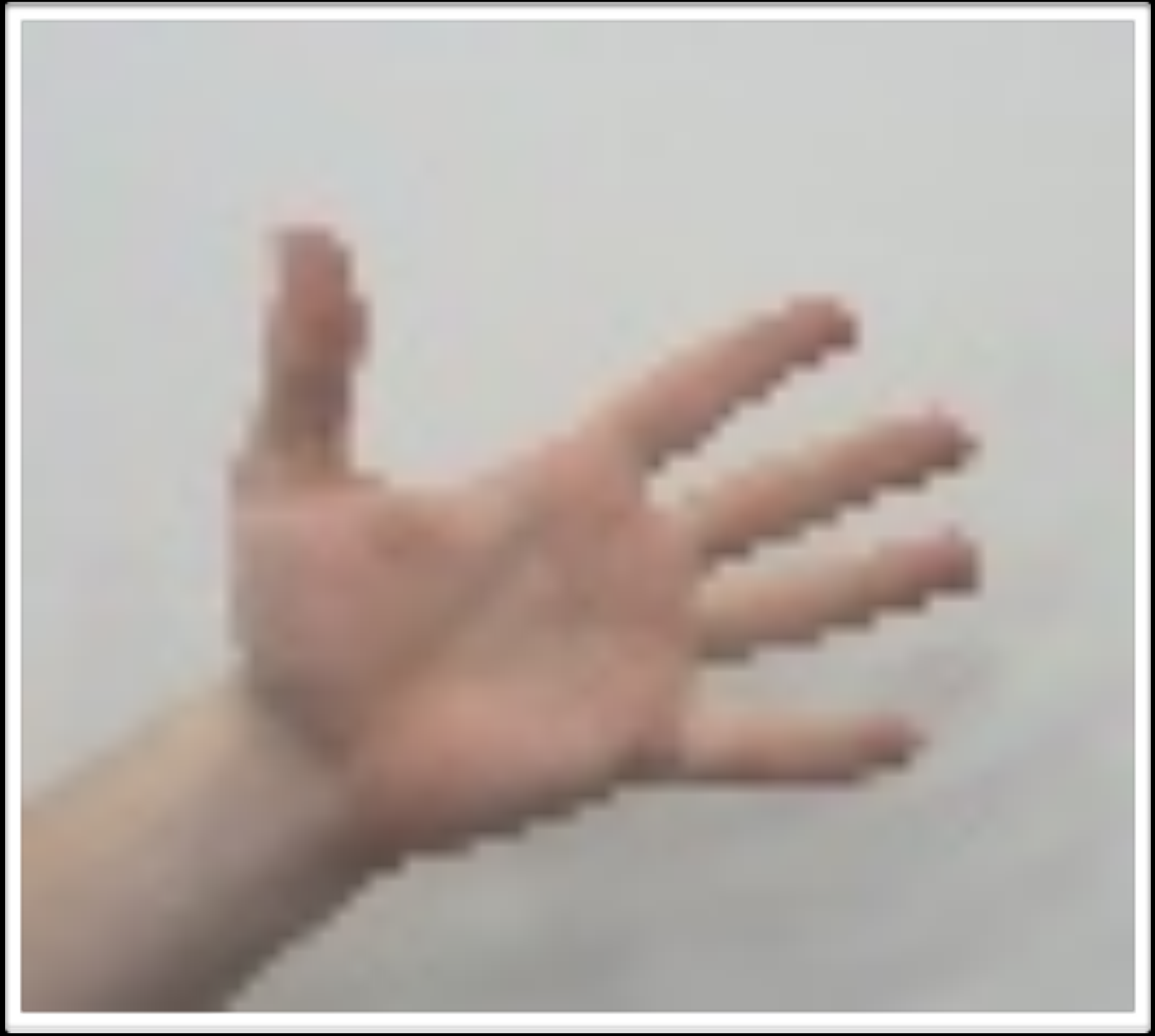
invalidated (level $k = 0$)

Geometry invalidation and fill-in

colour image (level $k = 1$)

invalidated (level $k = 1$)

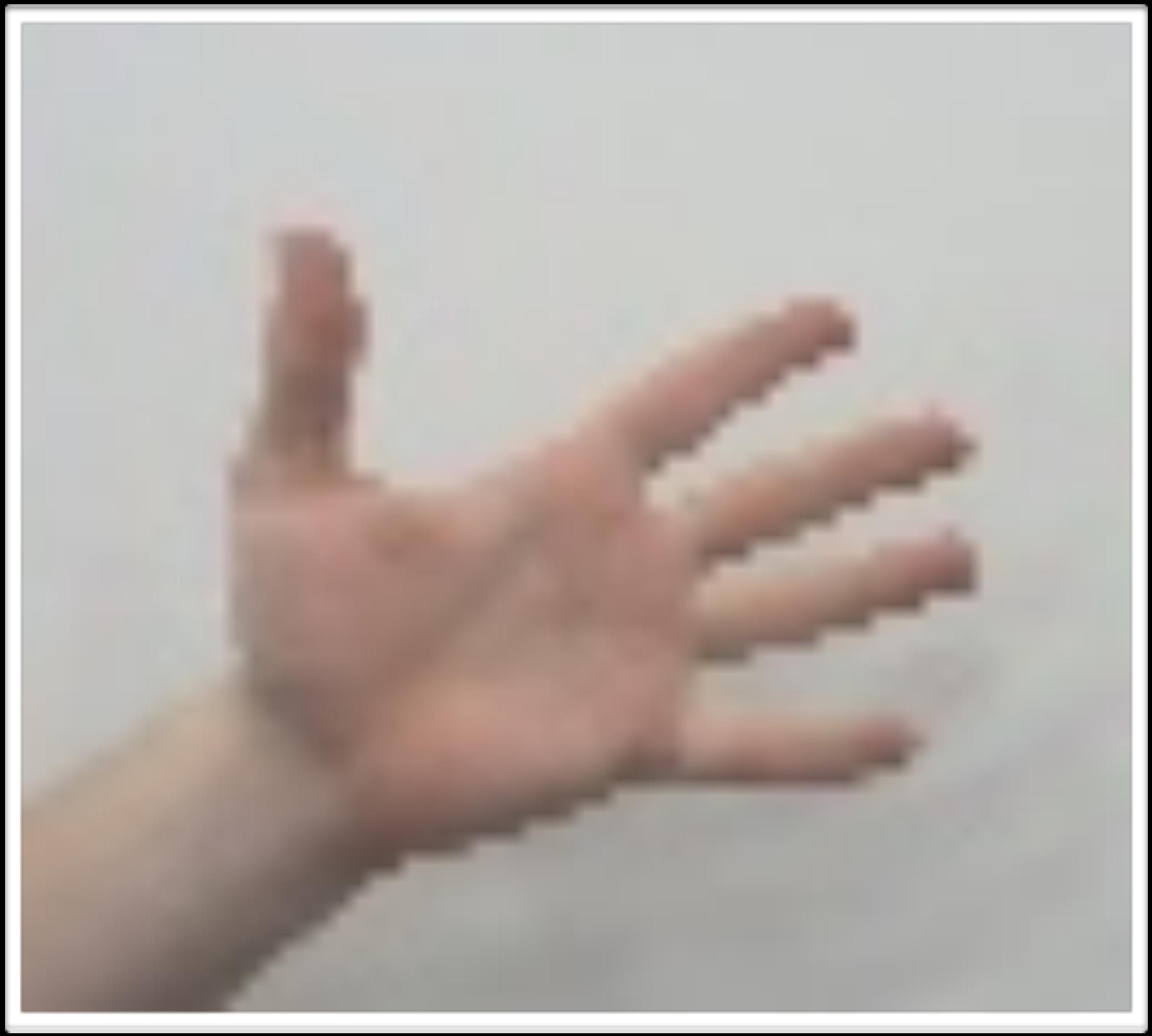
Geometry invalidation and fill-in



colour image (level k = 2)

invalidated (level k = 2)

Geometry invalidation and fill-in


colour image (level k = 2)


filled-in (level k = 2)

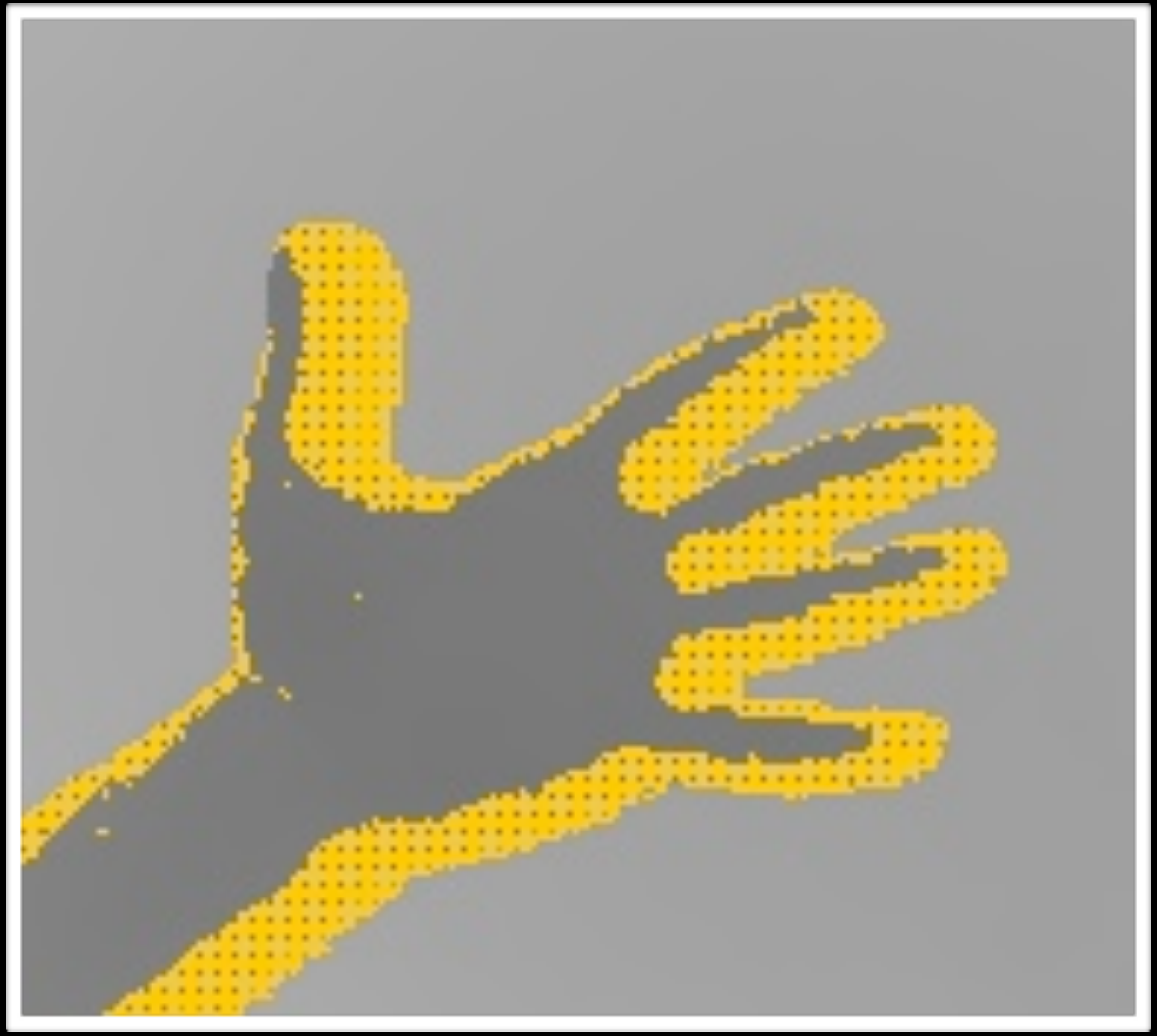
Geometry invalidation and fill-in

colour image (level $k = 1$)

invalidated (level $k = 1$)

Geometry invalidation and fill-in

colour image (level $k = 1$)



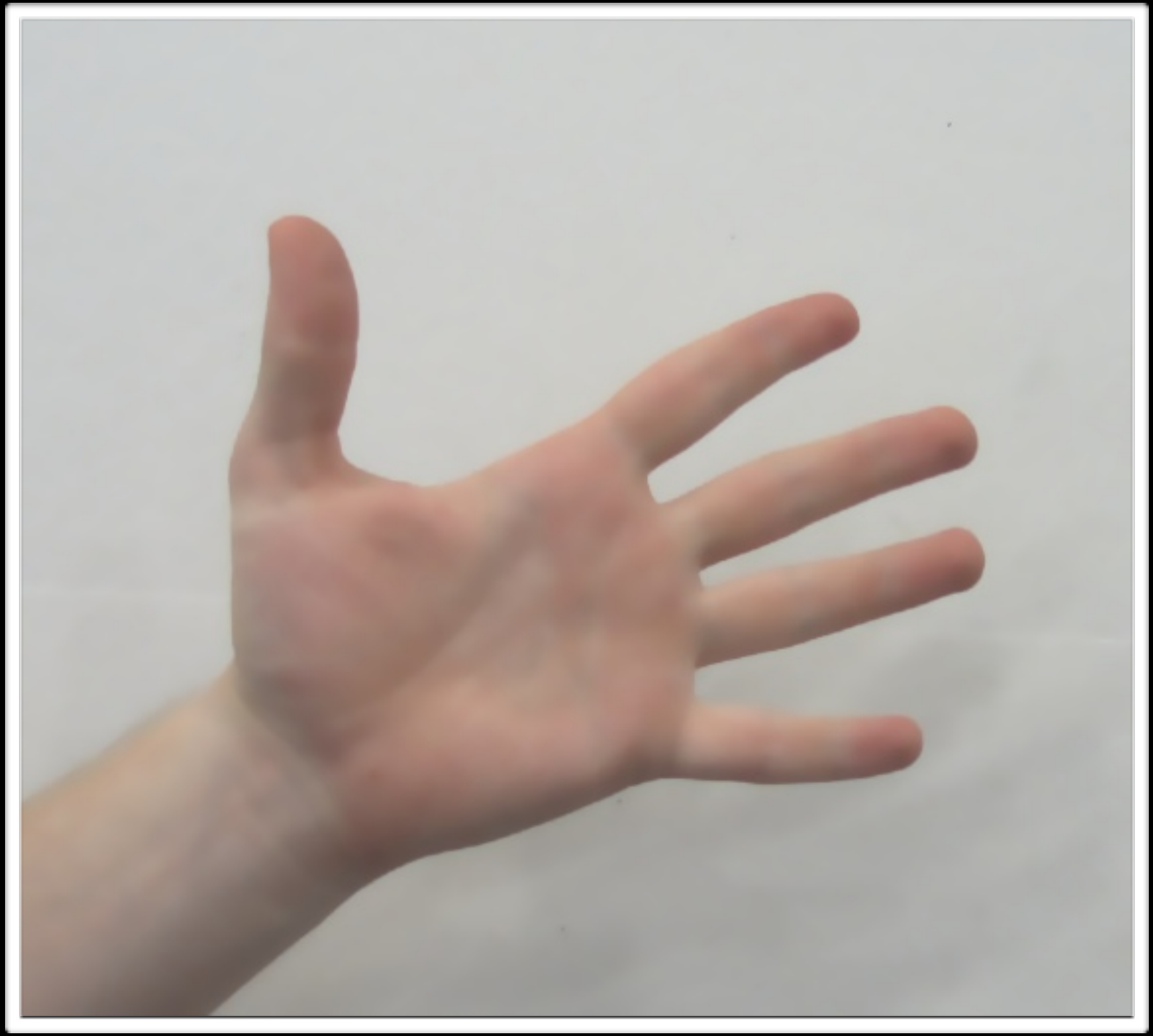
sparsely upsampled ($k = 1$)

Geometry invalidation and fill-in

colour image (level $k = 1$)

filled-in (level $k = 1$)

Geometry invalidation and fill-in



colour image (level k = 0)

filled-in (level k = 0)

Spatial-only geometry filtering

$$f_s(\mathbf{x}, t) = \frac{\sum_{\mathbf{y} \in N_{\mathbf{x}}} w(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{y}, t)}{\sum_{\mathbf{y} \in N_{\mathbf{x}}} w(\mathbf{x}, \mathbf{y})}$$

Spatial-only geometry filtering

$$f_S(\mathbf{x}, t) = \sum_{\mathbf{y} \in N_{\mathbf{x}}} w(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{y}, t)$$

Spatial-only geometry filtering

$$f_S(\mathbf{x}, t) = \sum_{\mathrm{y} \in N_{\mathrm{x}}} w_c(\mathbf{x}, \mathbf{y}) \cdot w_d(\mathbf{x}, \mathbf{y}) \cdot w_s(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{y}, t)$$

Spatial-only geometry filtering

$$f_s(\mathbf{x}, t) = \sum_{\mathrm{y} \in N_\mathrm{x}} \boxed{w_c(\mathbf{x}, \mathbf{y})} \cdot w_d(\mathbf{x}, \mathbf{y}) \cdot w_s(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{y}, t)$$

colour weight

$$\boxed{w_c(\mathbf{x}, \mathbf{y}) = \exp\left(-g_c \cdot \|\mathbf{i}(\mathbf{x}, t) - \mathbf{i}(\mathbf{y}, t)\|^2 / 2\sigma_c^2\right)}$$

Spatial-only geometry filtering

$$f_S(\mathbf{x}, t) = \sum_{y \in N_x} \boxed{w_c(\mathbf{x}, \mathbf{y})} \cdot \boxed{w_d(\mathbf{x}, \mathbf{y})} \cdot w_s(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{y}, t)$$

colour weight

$$w_c(\mathbf{x}, \mathbf{y}) = \exp\left(-g_c \cdot \|\mathbf{i}(\mathbf{x}, t) - \mathbf{i}(\mathbf{y}, t)\|^2 / 2\sigma_c^2\right)$$

distance weight

$$w_d(\mathbf{x}, \mathbf{y}) = \exp\left(-|d(\mathbf{x}, t) - d(\mathbf{y}, t)|^2 / 2\sigma_d^2\right)$$

Spatial-only geometry filtering

$$f_s(\mathbf{x}, t) = \sum_{\mathbf{y} \in N_{\mathbf{x}}} \boxed{w_c(\mathbf{x}, \mathbf{y})} \cdot \boxed{w_d(\mathbf{x}, \mathbf{y})} \cdot \boxed{w_s(\mathbf{x}, \mathbf{y})} \cdot d(\mathbf{y}, t)$$

colour weight

$$w_c(\mathbf{x}, \mathbf{y}) = \exp\left(-g_c \cdot \|\mathbf{i}(\mathbf{x}, t) - \mathbf{i}(\mathbf{y}, t)\|^2 / 2\sigma_c^2\right)$$

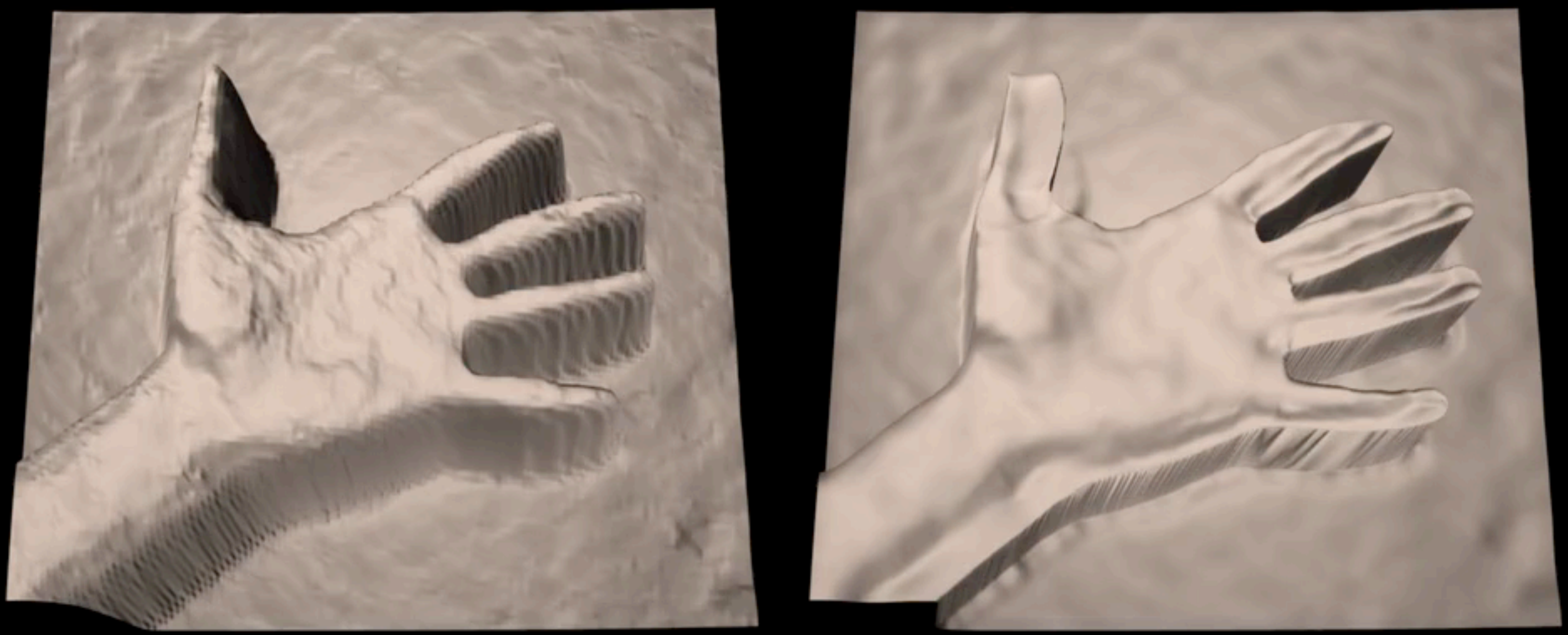
distance weight

$$w_d(\mathbf{x}, \mathbf{y}) = \exp\left(-\,|d(\mathbf{x}, t) - d(\mathbf{y}, t)|^2 / 2\sigma_d^2\right)$$

spatial weight

$$w_s(\mathbf{x}, \mathbf{y}) = \exp\left(-\,\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma_s^2\right)$$

Spatial-only filtering results



Aligned, but unfiltered

Spatially filtered

Spatiotemporal geometry filtering

$$f_{ST}(\mathbf{x}, t)$$

spatiotemporal filter

Spatiotemporal geometry filtering

$$f_{ST}(\mathbf{x}, t) = \varphi \cdot f_S(\mathbf{x}, t)$$

spatiotemporal filter

spatial filter

Spatiotemporal geometry filtering

$$f_{ST}(\mathbf{x}, t) = \varphi \cdot f_S(\mathbf{x}, t) + (1 - \varphi) \cdot f_T(\mathbf{x}, t)$$

spatiotemporal filter spatial filter temporal filter

Spatiotemporal geometry filtering

$$f_{ST}(\mathbf{x}, t) = \varphi \cdot f_S(\mathbf{x}, t) + (1 - \varphi) \cdot \boxed{f_T(\mathbf{x}, t)}$$

spatiotemporal filter   spatial filter   temporal filter

$$\boxed{f_T(\mathbf{x}, t)} = \sum_{\mathrm{y} \in N_{\mathrm{x}}} w(\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}}) \cdot f_{ST}(\bar{\mathbf{y}}, t-1)$$

Spatiotemporal geometry filtering

$$\boxed{f_{ST}(\mathbf{x}, t)} = \varphi \cdot f_S(\mathbf{x}, t) + (1 - \varphi) \cdot \boxed{f_T(\mathbf{x}, t)}$$

spatiotemporal filter

spatial filter

temporal filter

$$\boxed{f_T(\mathbf{x}, t)} = \sum_{y \in N_x} w(\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}}) \cdot \boxed{f_{ST}(\bar{\mathbf{y}}, t-1)}$$

Spatiotemporal geometry filtering

$$\boxed{f_{ST}(\mathbf{x}, t)} = \varphi \cdot f_S(\mathbf{x}, t) + (1 - \varphi) \cdot \boxed{f_T(\mathbf{x}, t)}$$

spatiotemporal filter spatial filter temporal filter

$$\boxed{f_T(\mathbf{x}, t)} = \sum_{\mathrm{y} \in N_{\mathrm{x}}} w_c(\mathbf{x}, \bar{\mathbf{y}}) \cdot w_d(\mathbf{x}, \bar{\mathbf{y}}) \cdot w_s(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \cdot w_f(\mathbf{y}, \bar{\mathbf{y}}) \cdot \boxed{f_{ST}(\bar{\mathbf{y}}, t-1)}$$

Spatiotemporal geometry filtering

$$\boxed{f_{ST}(\mathbf{x},t)} = \varphi \cdot \boxed{f_S(\mathbf{x},t)} + (1-\varphi)\cdot \boxed{f_T(\mathbf{x},t)}$$

spatiotemporal filter
spatial filter
temporal filter

$$\boxed{f_S(\mathbf{x},t)} = \sum_{\mathrm{y}\in N_\mathrm{x}} w_c(\mathbf{x},\mathbf{y})\cdot w_d(\mathbf{x},\mathbf{y})\cdot w_s(\mathbf{x},\mathbf{y})\cdot d(\mathbf{y},t)$$

$$\boxed{f_T(\mathbf{x},t)} = \sum_{\mathrm{y}\in N_\mathrm{x}} w_c(\mathbf{x},\bar{\mathbf{y}})\cdot w_d(\mathbf{x},\bar{\mathbf{y}})\cdot w_s(\bar{\mathbf{x}},\bar{\mathbf{y}})\cdot w_f(\mathbf{y},\bar{\mathbf{y}})\cdot \boxed{f_{ST}(\bar{\mathbf{y}},t-1)}$$

Spatiotemporal geometry filtering

$$\boxed{f_{ST}(\mathbf{x}, t)} = \varphi \cdot \boxed{f_S(\mathbf{x}, t)} + (1 - \varphi) \cdot \boxed{f_T(\mathbf{x}, t)}$$

spatiotemporal filter — spatial filter — temporal filter

$$\boxed{f_S(\mathbf{x}, t)} = \sum_{y \in N_x} \boxed{w_c(\mathbf{x}, \mathbf{y})} \cdot w_d(\mathbf{x}, \mathbf{y}) \cdot w_s(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{y}, t)$$

$$\boxed{f_T(\mathbf{x}, t)} = \sum_{y \in N_x} \boxed{w_c(\mathbf{x}, \bar{\mathbf{y}})} \cdot w_d(\mathbf{x}, \bar{\mathbf{y}}) \cdot w_s(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \cdot w_f(\mathbf{y}, \bar{\mathbf{y}}) \cdot \boxed{f_{ST}(\bar{\mathbf{y}}, t-1)}$$

Spatiotemporal geometry filtering

$$\boxed{f_{ST}(\mathbf{x},t)} = \varphi \cdot \boxed{f_S(\mathbf{x},t)} + (1-\varphi) \cdot \boxed{f_T(\mathbf{x},t)}$$

spatiotemporal filter — spatial filter — temporal filter

$$\boxed{f_S(\mathbf{x},t)} = \sum_{\mathbf{y}\in N_\mathbf{x}} \boxed{w_c(\mathbf{x},\mathbf{y})} \cdot \boxed{w_d(\mathbf{x},\mathbf{y})} \cdot w_s(\mathbf{x},\mathbf{y}) \cdot d(\mathbf{y},t)$$

$$\boxed{f_T(\mathbf{x},t)} = \sum_{\mathbf{y}\in N_\mathbf{x}} \boxed{w_c(\mathbf{x},\bar{\mathbf{y}})} \cdot \boxed{w_d(\mathbf{x},\bar{\mathbf{y}})} \cdot w_s(\bar{\mathbf{x}},\bar{\mathbf{y}}) \cdot w_f(\mathbf{y},\bar{\mathbf{y}}) \cdot \boxed{f_{ST}(\bar{\mathbf{y}},t-1)}$$

Spatiotemporal geometry filtering

$$f_{ST}(\mathbf{x}, t) = \varphi \cdot f_S(\mathbf{x}, t) + (1 - \varphi) \cdot f_T(\mathbf{x}, t)$$

spatiotemporal filter — spatial filter — temporal filter

$$f_S(\mathbf{x}, t) = \sum_{\mathbf{y} \in N_\mathbf{x}} w_c(\mathbf{x}, \mathbf{y}) \cdot w_d(\mathbf{x}, \mathbf{y}) \cdot w_s(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{y}, t)$$

$$f_T(\mathbf{x}, t) = \sum_{\mathbf{y} \in N_\mathbf{x}} w_c(\mathbf{x}, \bar{\mathbf{y}}) \cdot w_d(\mathbf{x}, \bar{\mathbf{y}}) \cdot w_s(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \cdot w_f(\mathbf{y}, \bar{\mathbf{y}}) \cdot f_{ST}(\bar{\mathbf{y}}, t-1)$$

Spatiotemporal geometry filtering

$$\boxed{f_{ST}(\mathbf{x}, t)} = \varphi \cdot \boxed{f_S(\mathbf{x}, t)} + (1 - \varphi) \cdot \boxed{f_T(\mathbf{x}, t)}$$

spatiotemporal filter spatial filter temporal filter

$$\boxed{f_S(\mathbf{x}, t)} = \sum_{\mathbf{y} \in N_\mathbf{x}} \boxed{w_c(\mathbf{x}, \mathbf{y})} \cdot \boxed{w_d(\mathbf{x}, \mathbf{y})} \cdot \boxed{w_s(\mathbf{x}, \mathbf{y})} \cdot d(\mathbf{y}, t)$$

$$\boxed{f_T(\mathbf{x}, t)} = \sum_{\mathbf{y} \in N_\mathbf{x}} \boxed{w_c(\mathbf{x}, \bar{\mathbf{y}})} \cdot \boxed{w_d(\mathbf{x}, \bar{\mathbf{y}})} \cdot \boxed{w_s(\bar{\mathbf{x}}, \bar{\mathbf{y}})} \cdot \boxed{w_f(\mathbf{y}, \bar{\mathbf{y}})} \cdot \boxed{f_{ST}(\bar{\mathbf{y}}, t-1)}$$

flow weight $$\boxed{w_f(\mathbf{y}, \bar{\mathbf{y}}) = \exp\left(- \|\mathbf{y} - \bar{\mathbf{y}}\|^2 / 2\sigma_f^2\right)}$$

Spatiotemporal filtering results

Aligned, but unfiltered

Spatiotemporally filtered

RGBZ video effects

- video relighting
- geometry-based video abstraction
- stroke-based video rendering
- background segmentation
- stereoscopic 3D rendering

Video relighting

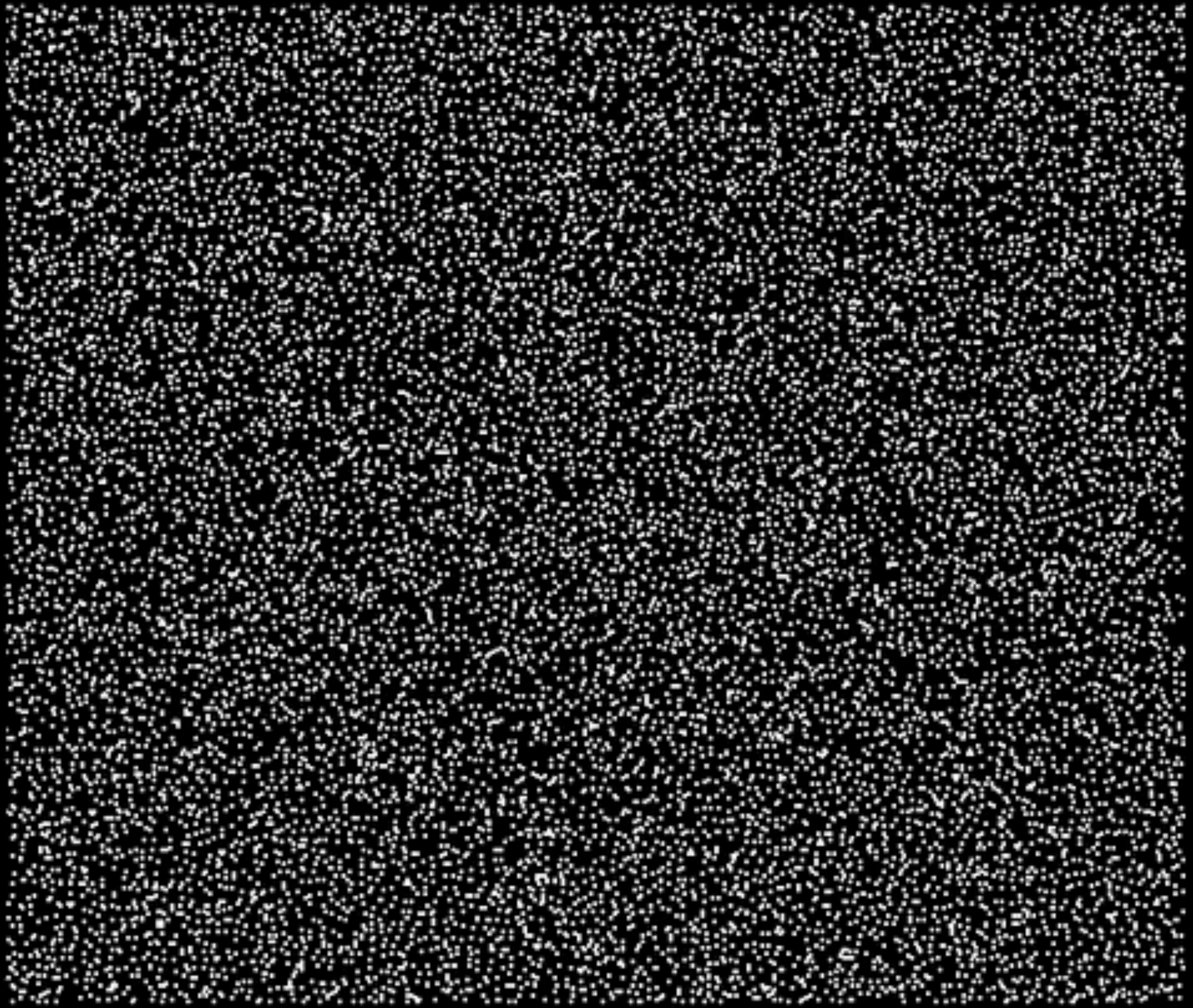
Input video

Geometry-based video abstraction

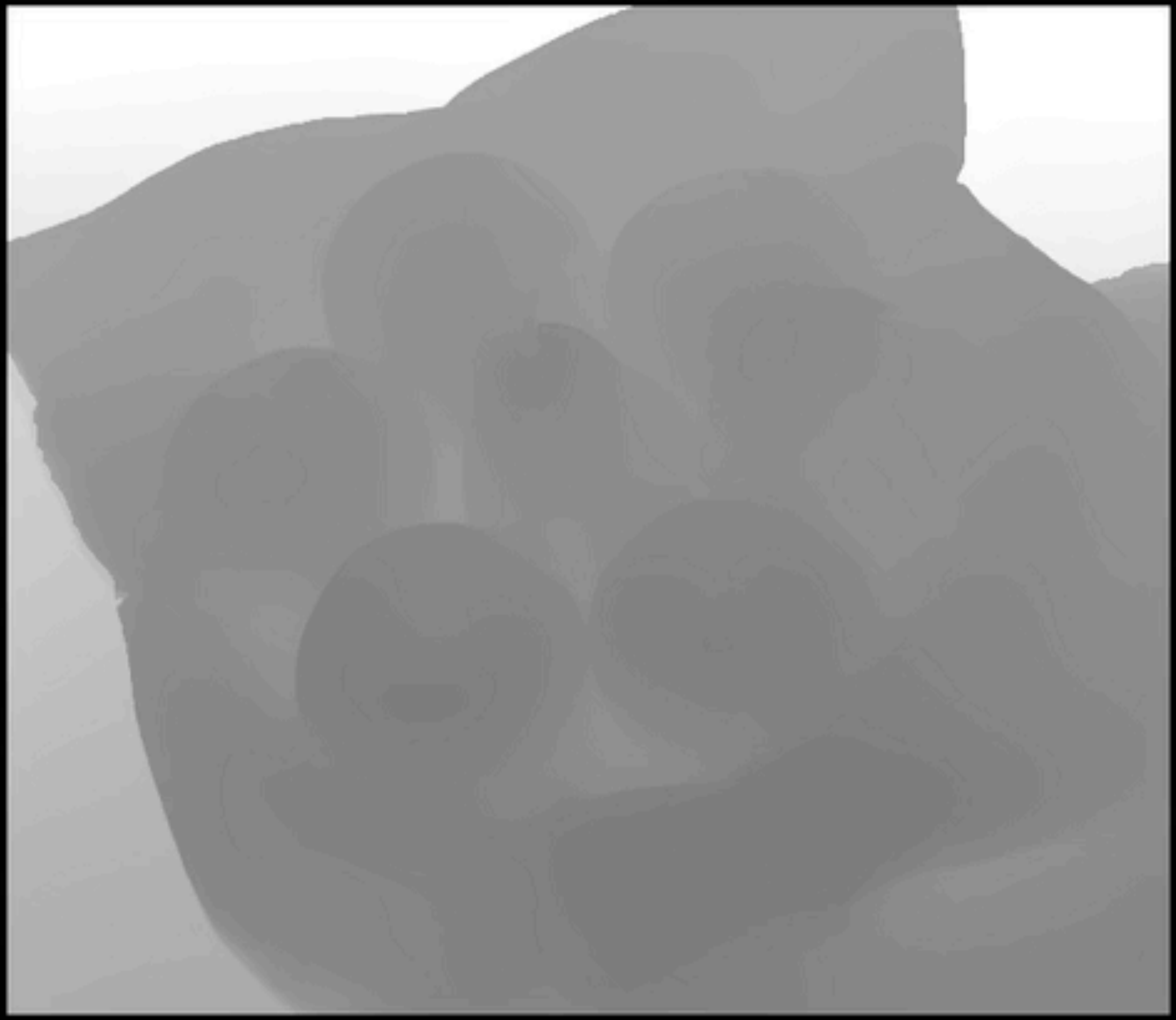
Input video

Image-based abstraction

Stroke-based video rendering



Background segmentation



Filtered distance map

Colour video

Stereoscopic 3D rendering

Synthesised right view

Synthesised stereo image

Limitations

- unreliable optical flow can lead to smearing artefacts
- assumption of coincident colour + depth edges
  - 'texture copy' artefacts in the distance map
  - edges with small colour differences not preserved well
- depth detail limited by time-of-flight camera resolution
- joint-bilateral filter not guaranteed to be optimal:
  new values are a linear combination of existing values

Future work

- improve preservation of features
  - could refine results using shape-from-shading
- formulate optical flow that respects depth discontinuities
  - would prevent ‘smearing’ artefacts in the distance map
- commodification of RGBZ video cameras and effects:
  - miniaturisation of camera hardware
  - improvements in hardware performance
  - algorithmic optimisations

Summary

- introduced a novel set of efficient and effective depth filtering and upsampling techniques for RGBZ videos:
  - a fast fill-in procedure for unreliable geometry
  - a multi-lateral spatiotemporal filtering approach
- illustrated the benefits of RGBZ video for effects
- source code and data sets are available on our project page at

http://richardt.name/rgbz-camera/

Hire me! I'm looking for a postdoc from October 2012.