Using primitives (at last!)

D.A. Forsyth and Vaibhav Vavilala, UIUC
Anand Bhattad, UIUC->TTIC->JHU

Points

- Generators are DUMB
- There is no “right” geometric representation
  - some are good for one thing, some for another
  - Q: how to make them live together in ways that are productive
- Bias and Variance are the key ideas
- Primitive representations offer refined control

Generators are DUMB

- Diffusion generators are DUMB:
  - They don’t understand scale or physics
  - They get geometry wrong
  - They hallucinate
- Consequence:
  - a prodigious need to control generators

Diffusion applications:
make commercial art cheaper

and political discourse nastier

Applications: make commercial art cheaper

and political discourse nastier

Scale relative to politician

horrid geometry

Are generated and real images different?

- Procedure:
  - build reduced geometric representation of image
    - lines; perspective field; object+shadow
      - using existing tools
  - train a classifier to distinguish between real and generated
  - prequalify images so that at test, classifier sees only “good” generated images.
- Result
  - Near perfect classification

- Conclusion:
  - generators are bad at lines; perspective; shadows

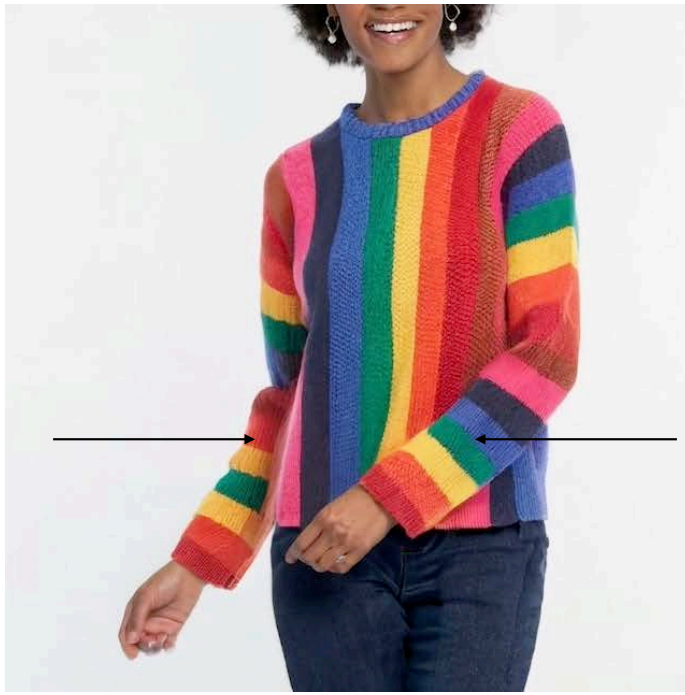
Sarkar, 24: Shadows Don’t Lie and Lines Can’t Bend! Generative Models don’t know Projective Geometry...for now

Synthesizing clothing

Q: Are these models wearing a real garment?

Q: Are these models wearing the same garment?

Zhu et al 23: Tryondiffusion: A tale of two unets



Zhu et al 23: Tryondiffusion: A tale of two unets

Big questions

- Q: why do we put up with this?
  - generators know stuff that is unexpected, cf Anand's papers
- Q: where is this coming from?
  - (guess) the denoiser architecture
- Q: what do we do about this?
  - get better at measuring performance; current situation is laughable

Visual representation

- What do representations do?
  - answer queries (cf Malcolm Sabin)
  - support interaction

- Latent
  - no obvious meaning
  - example: style code in styleGAN, etc.
  - requires a decoder to map to “meaningful” variables
  - characteristic property: very good at managing detail and complexity

- Explicit
  - has an intrinsic meaning
    - albedo, depth, normal, roughness, point-sets, geometry, etc…
  - example: geometric primitives
  - characteristic property: compress and simplify with loss of detail

Visual representation, II

- Bias-variance tradeoffs
- Bias
  - model makes errors because it can't represent some things correctly
  - a characteristic property of many kinds of low parameter parametric model
    - eg explicit physical models, meshes, etc.
- Variance
  - model makes errors because you can't estimate the model correctly
  - less scary than it used to be
    - networks trained with huge quantities of data should have variance problems, but don't

Argument

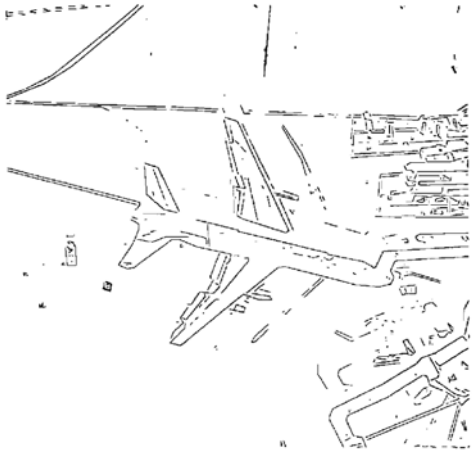
- Much modern 3D vision
  - is about turning sets of pictures into other sets of pictures that must be “right”
  - this should use only latent representations
    - it’s really mostly plenoptic function interpolation

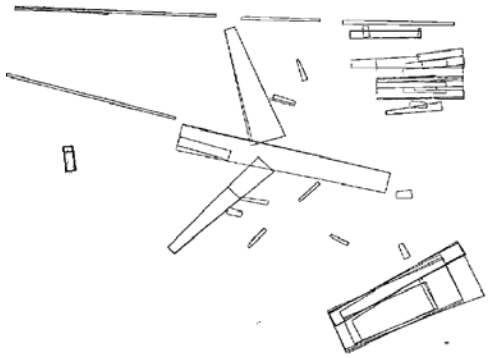
- Other applications demand simplified representations
  - interaction with human (eg scene editing)
  - manipulation (eg appropriate contact points for families of shape)
  - navigation (eg simplified representation OK, because you don’t go too close)
  - legacy (eg a bazillion GPUs want mesh based content for games)

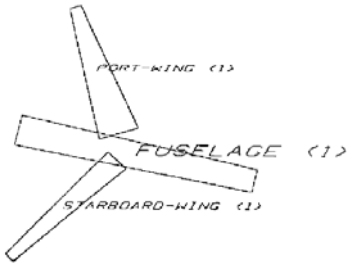
- Unfashionable hypothesis:
  - geometric primitives will come back

Primitives (ancient)

- Traditional idea, back to at least Binford 71
- Objects are a composite of primitive shapes
- Two issues:
  - What are the primitives?
  - Given some input, parse into the set of primitives?
- Traditional literature:
  - Construct a set of primitives, using
    - geometric insight, guessing, etc
  - Now infer presence of primitive from local image properties, edges, etc.
  - Major problem:
    - objects aren't precisely primitives, so….
  - Could almost be made to work







Brooks, 1981, Symbolic reasoning among 3-dimensional models and 2-dimensional images

Ioffe+Forsyth 01

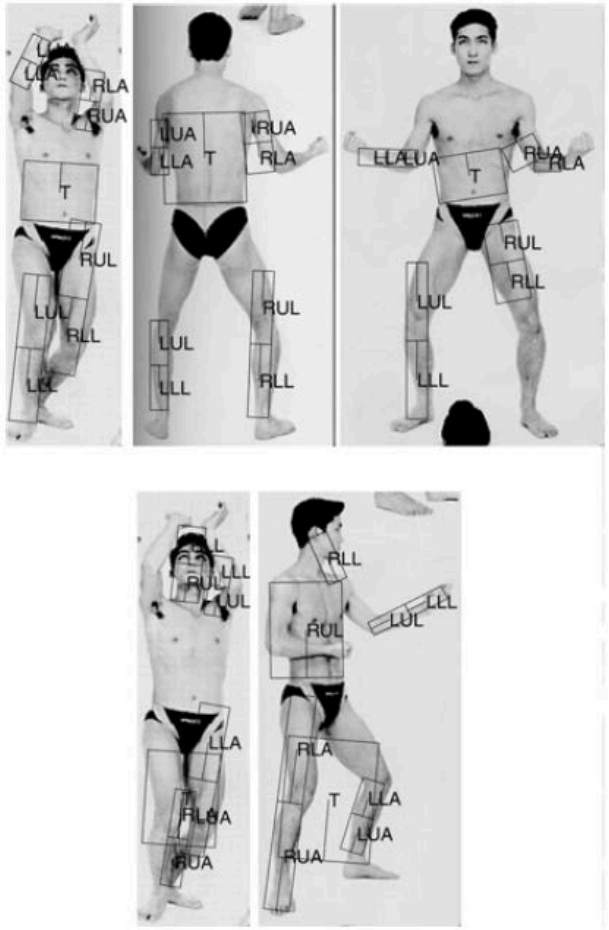


Figure 15. Examples of representative assemblies found for images of people. Each representative assembly is the highest-likelihood sample from a set of samples with overlapping torsos. We use representatives to count people and argue that using representatives does not change the count. Often (top) representatives can also be used to infer configurations of people, although (bottom) that is not always the case.

Obstacle: Primitive fitting is hard

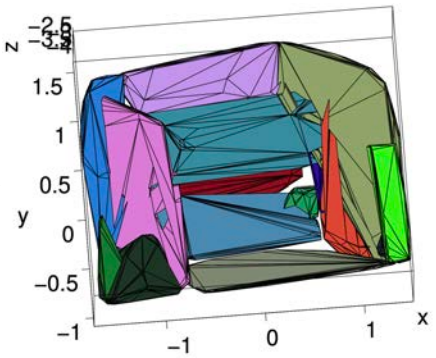
- Balance simplification against representation quality
- Usually different numbers per scene
  - Incremental RANSAC is really clumsy and doesn't work great

Primitives (modern)

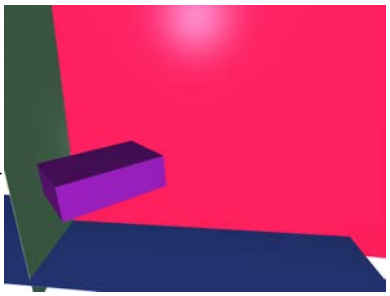
Tulsiani et al. 2017

Deng et al. 2020

Jiang 2014



Kluger et al. 2021

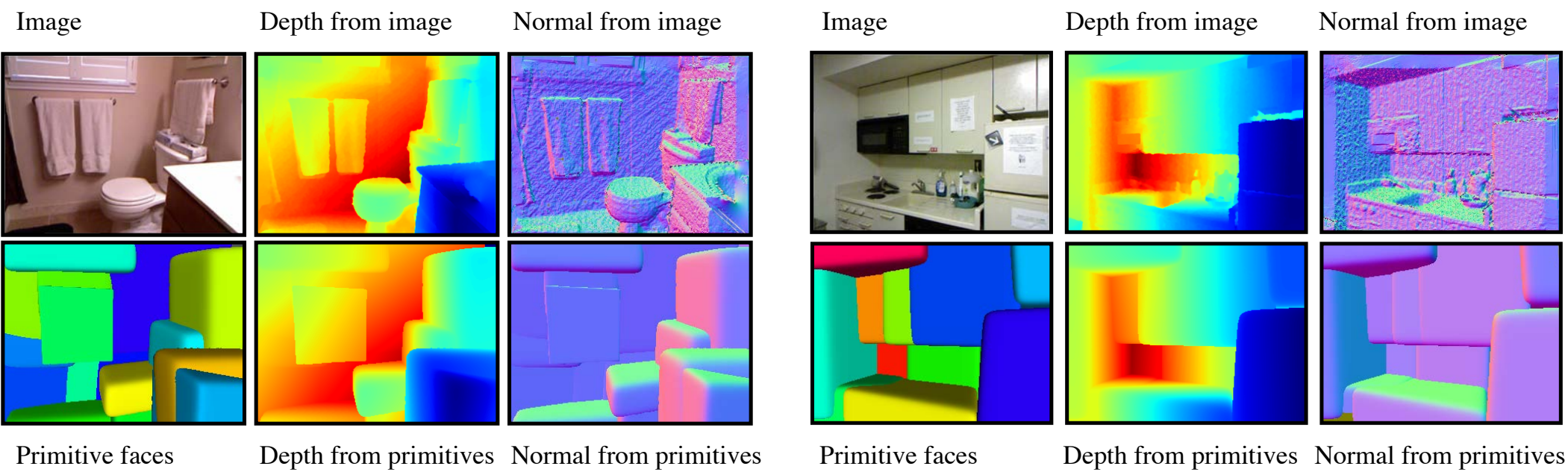


Fitting primitives to scenes

- Fit to depth map
  - Losses:
    - approximate depth; spread out; all points encoded; etc
    - cf Deng et al 20, CVXNet
    - segmentation loss if labelled images are available

- Fitting:
  - train network to accept image, produce fixed number of primitives
    - these primitives should minimize loss
  - this gives a great start point
    - polish by descent on loss

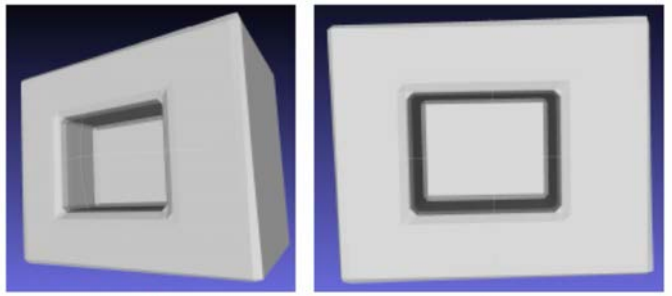
Map indoor scenes to primitives



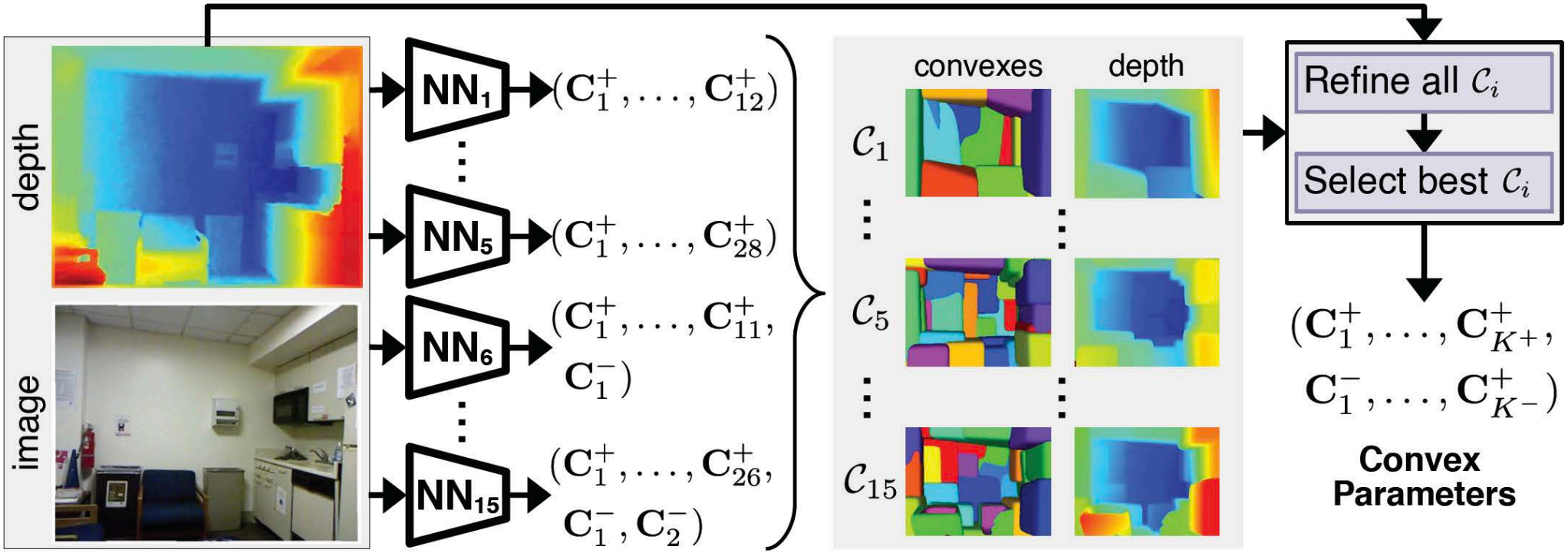
Vavilala, Forsyth, 23: Convex decomposition of indoor scenes

But…

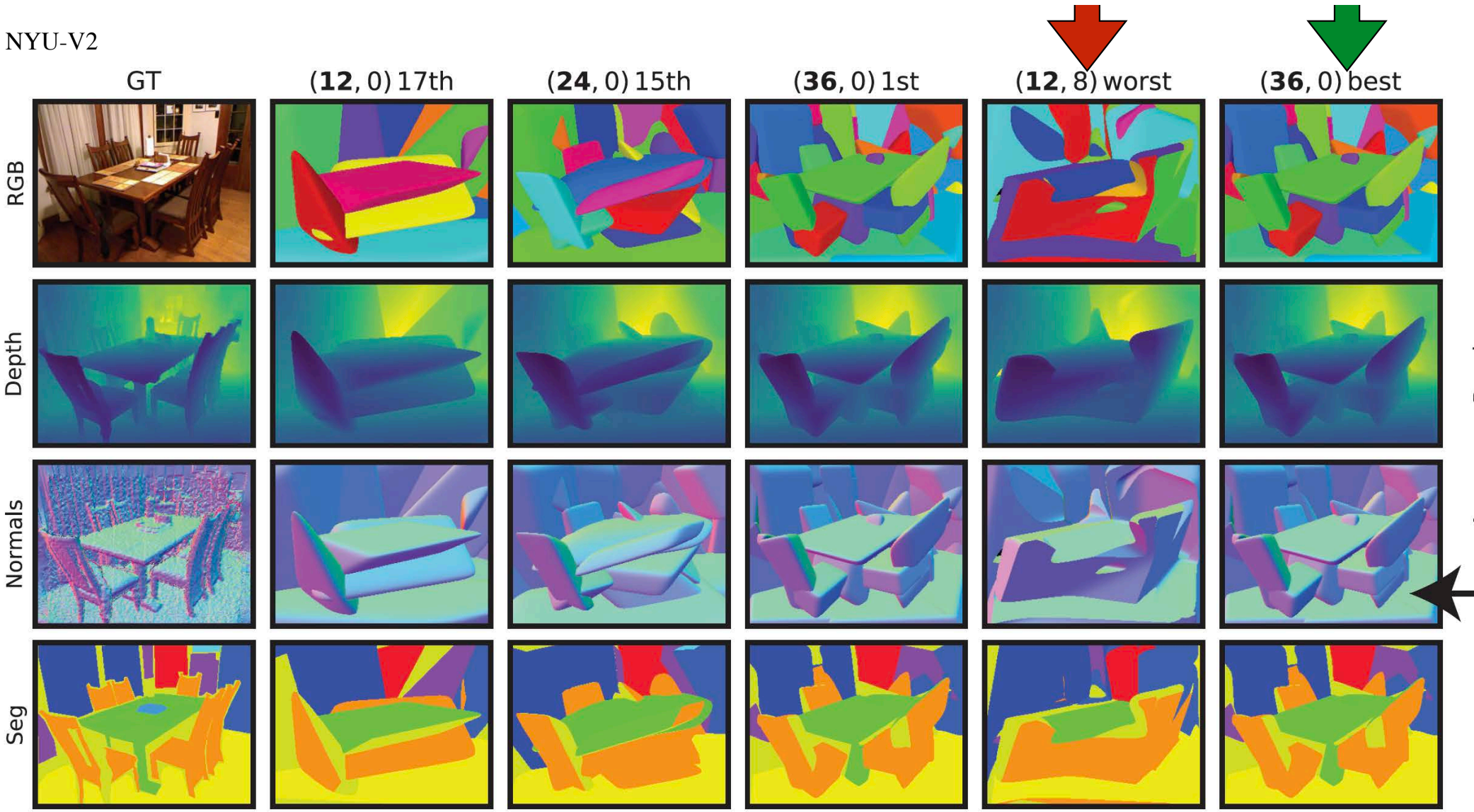
- Fixed number of primitives
  - ensemble, choose best (which you can do at test time)
- Subtraction
  - as in CSG
    - “negative primitives”
  - massively increases geometric complexity
  - easily included
    - now ensemble with different numbers of negatives
- Key: efficient evaluation of loss



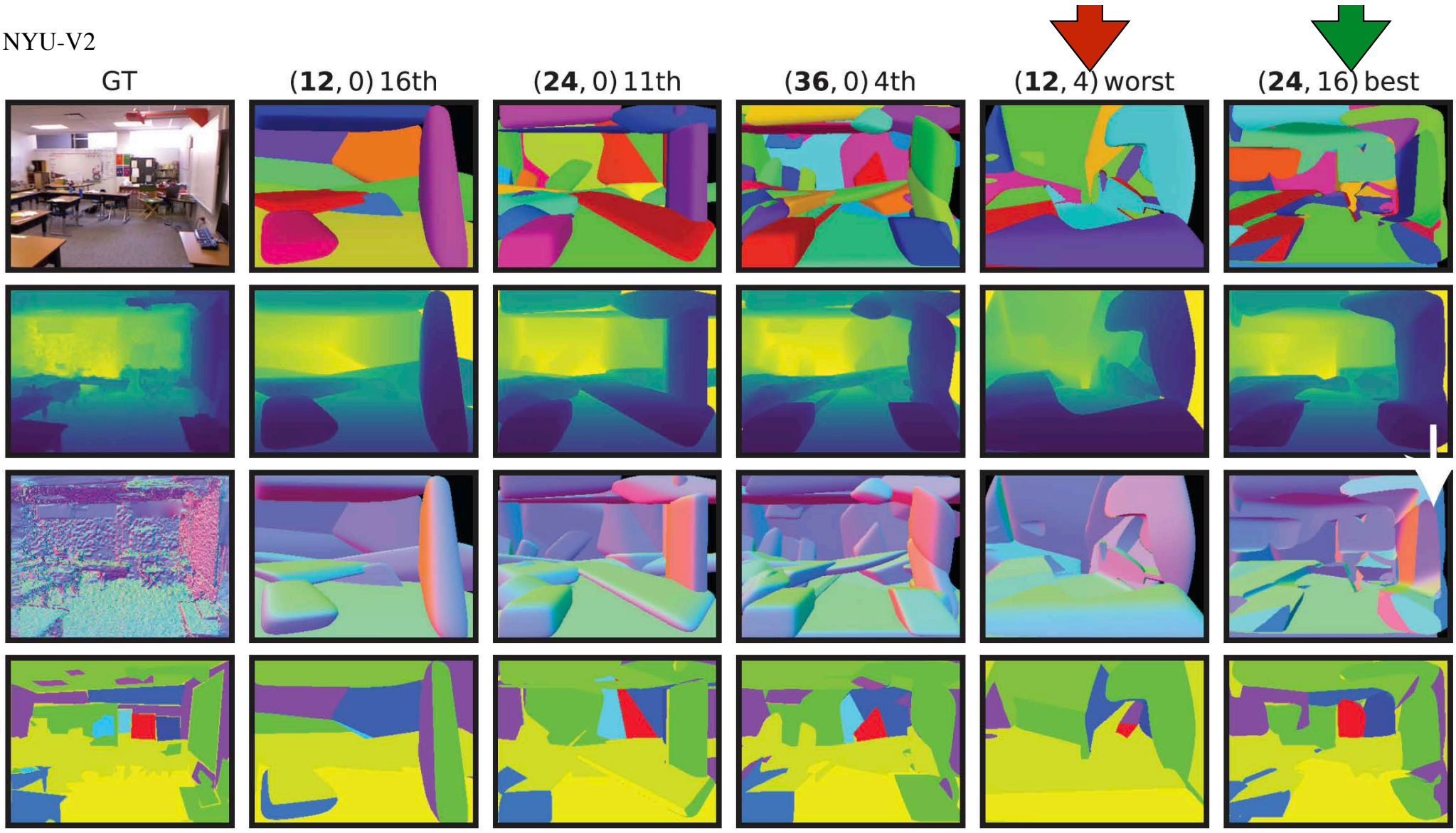
Ensembling to choose pos/neg numbers



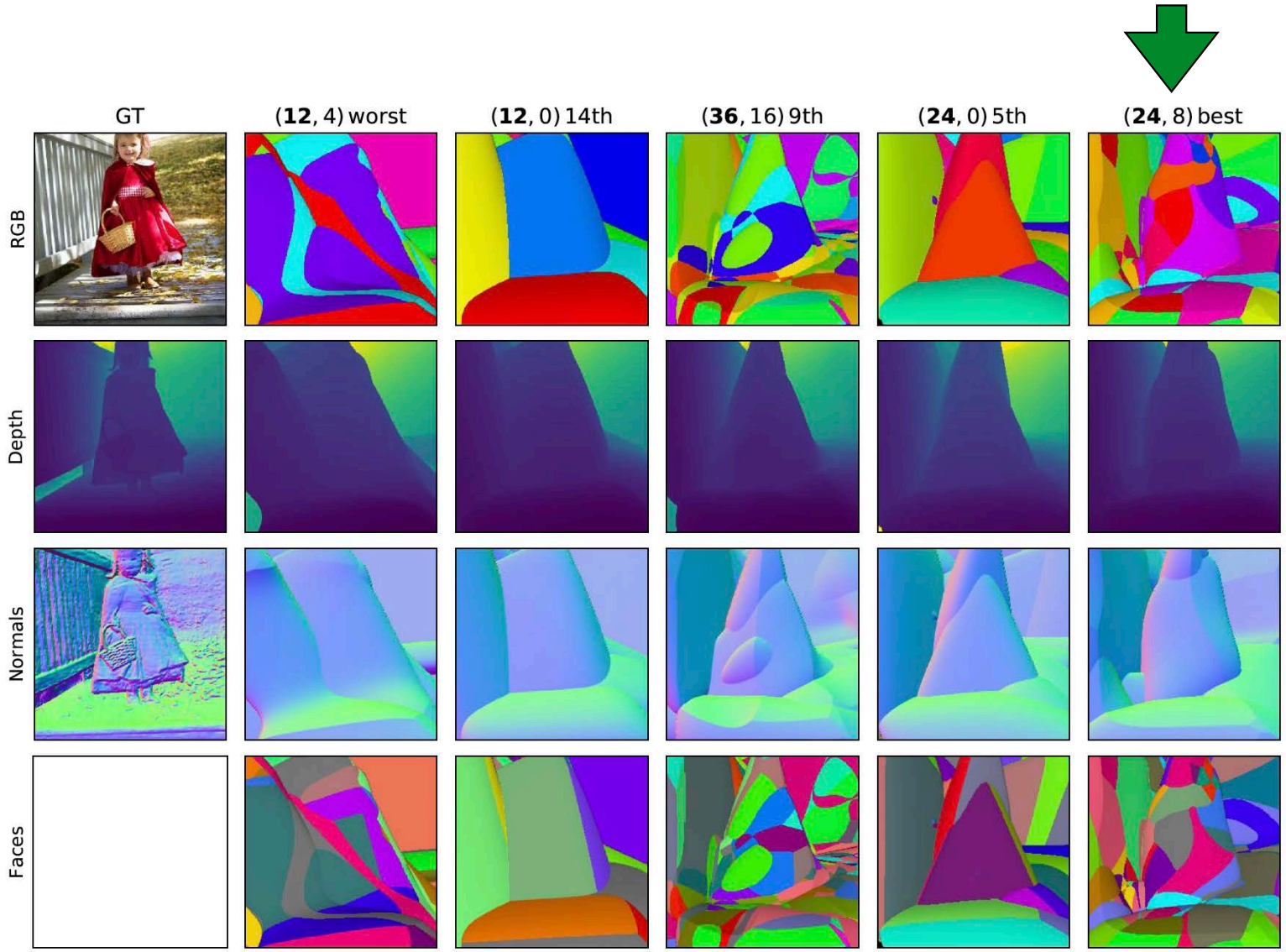
Vavilala, Forsyth, in review, Arxiv, 25


NYU-V2
GT
(12, 0) 17th
(24, 0) 15th
(36, 0) 1st
(12, 8) worst
(36, 0) best
RGB
Depth
Normals
Seg

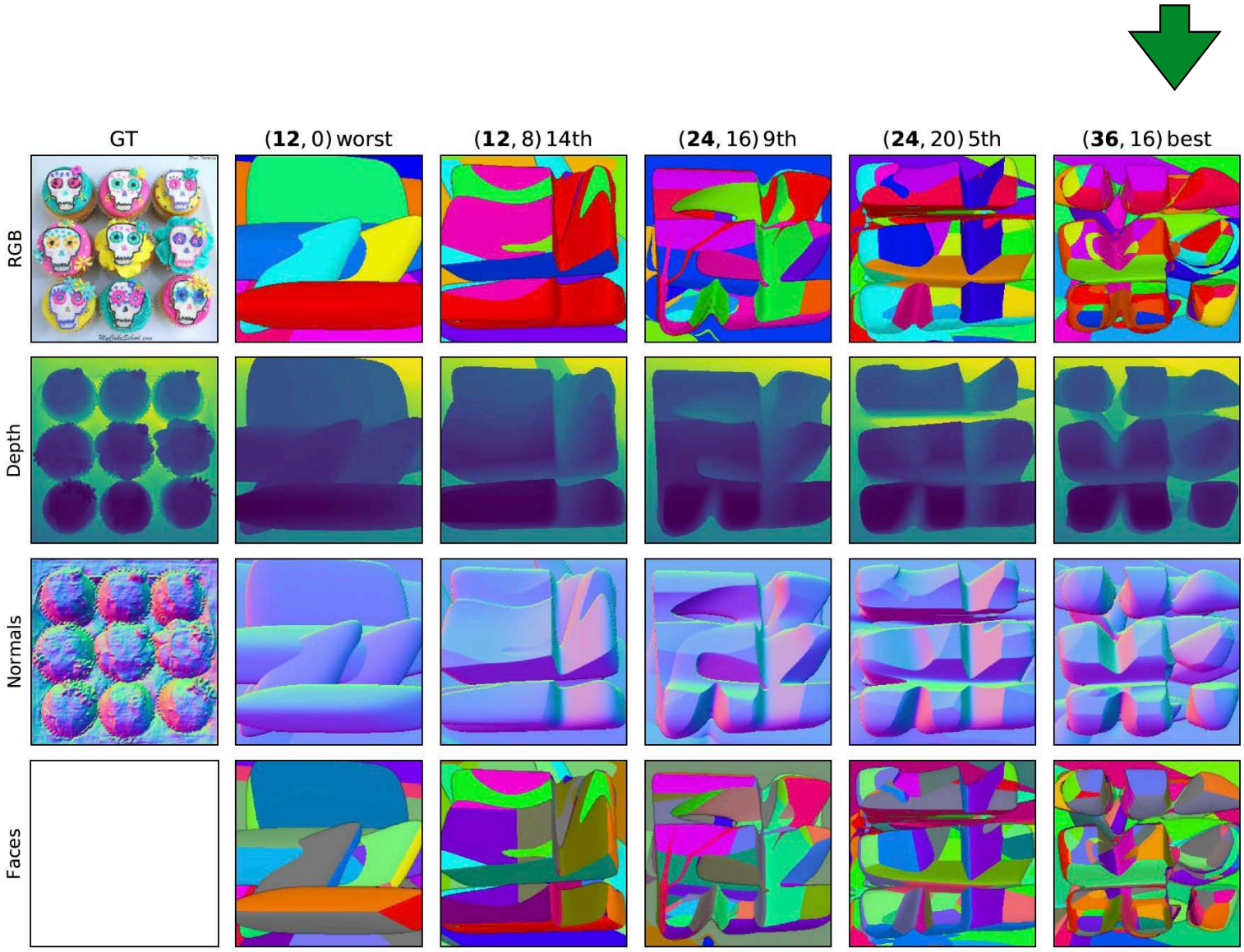
NYU-V2


GT
(12, 0) 16th
(24, 0) 11th
(36, 0) 4th
(12, 4) worst
(24, 16) best
RGB
Depth
Normals
Seg

LAION


GT
(12, 4) worst
(12, 0) 14th
(36, 16) 9th
(24, 0) 5th
(24, 8) best
RGB
Depth
Normals
Faces

LAION


GT
(12, 0) worst
(12, 8) 14th
(24, 16) 9th
(24, 20) 5th
(36, 16) best
RGB
Depth
Normals
Faces

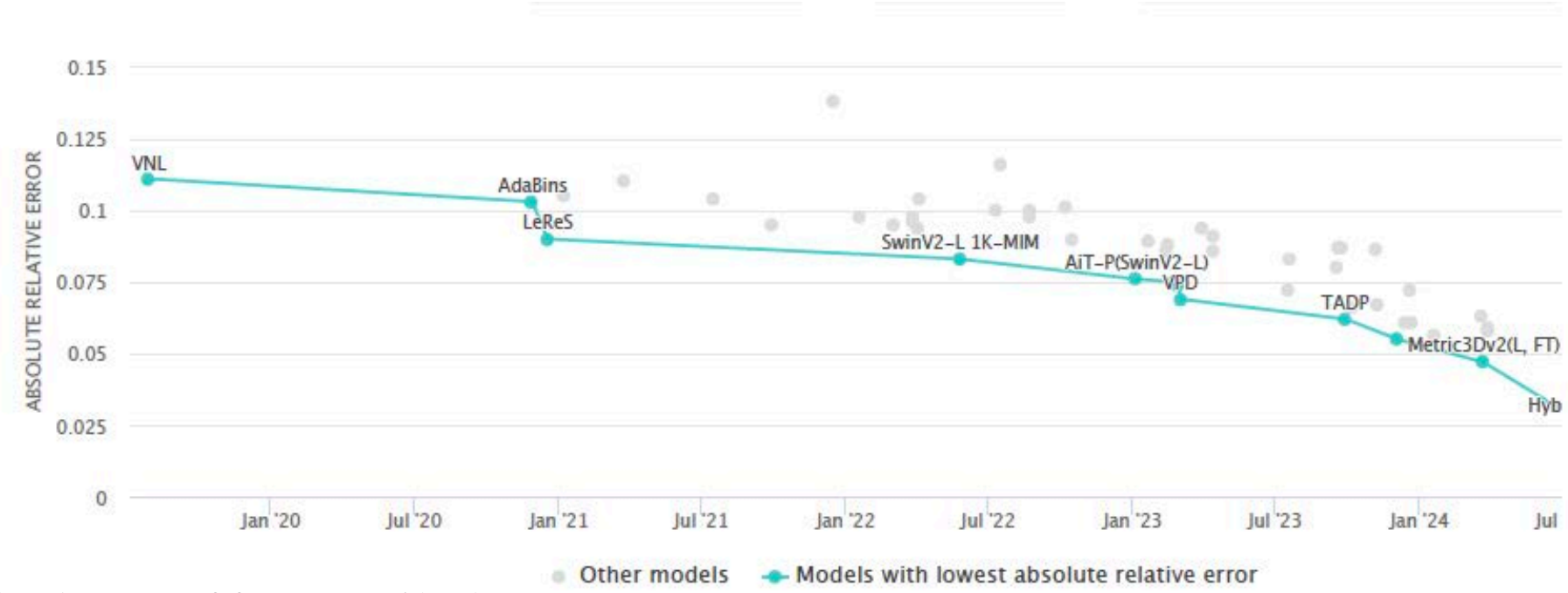
What are primitives good for?

- Interaction
  - following slides - requires accuracy, fair segmentation
- Planning
  - maybe - requires accuracy
- Inductive bias on generators
  - maybe - we can certainly fit on a very large scale

V. Vavilala and others, Generative Blocks World: Moving Things Around in Pictures, 2025, in review

cf Laconic, which we just heard about

Accuracy - single image depth prediction



AbsRel on NYU V2 from paperswithcode

Accuracy

AbsRel of primitive depths against ground truth (NYUV2) or monocular depth prediction (LAION)

Scale of error is comparable - the loss in depth accuracy passing to primitives is similar to that using monocular depth estimation

Dataset	faces	AbsRel↓
NYUv2	6	0.0417
LAION	6	0.0193
LAION	12	0.0178

Table 2: Depth metrics for NYUv2 data and LAION data compared, when ensembling pos + neg $R \to S$. The much larger data volume in LAION significantly improves the generalization ability of our procedure. Further, increasing the representational power of our method by removing the symmetric normal constraint (for parallelepipeds, top two rows) and increasing the number of faces per polytope to $f = 12$ (bottom row) yields even better primitive decompositions.

Interaction with scenes

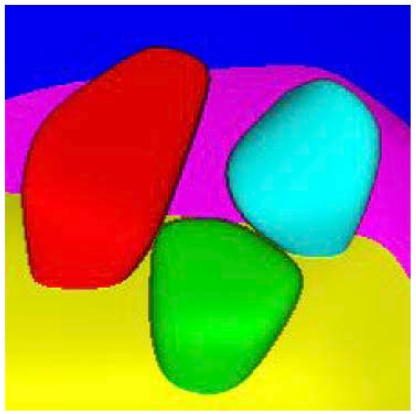
- Strategy:
  - image to primitives
  - move primitives, camera, etc
  - transfer texture from input using simple ray-tracing argument
    - there will be holes, signal problems, etc.
  - use conditional image generator to produce results
    - must fill holes, fix signal problems, etc.
    - OFF THE SHELF depth conditioned generator (no finetuning)

Abdelrahman Eldesokey and Peter Wonka. 2024. Build-a-scene: Interactive 3d layout control for diffusion-based image generation.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. 2024. Stable Flow: Vital Layers for Training-Free Image Editing.

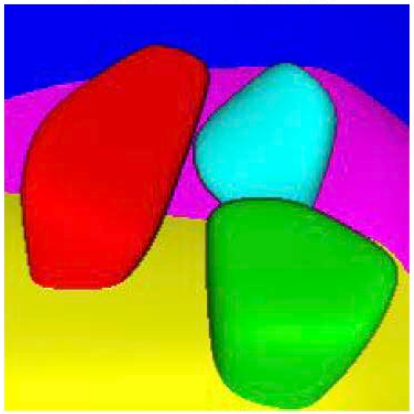
Shariq Farooq Bhat, Niloy J. Mitra, and Peter Wonka. 2023. LooseControl: Lifting ControlNet for Generalized Depth Conditioning.

Source Primitives



Move

Modified Primitives

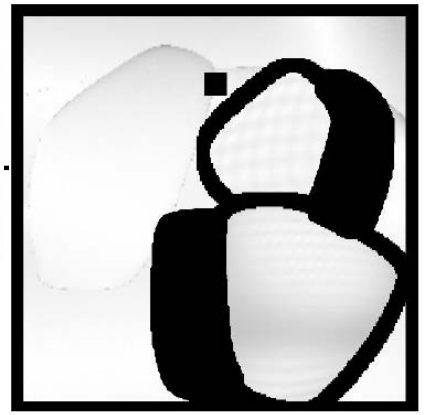


Generate

"Red, green, and blue couches"

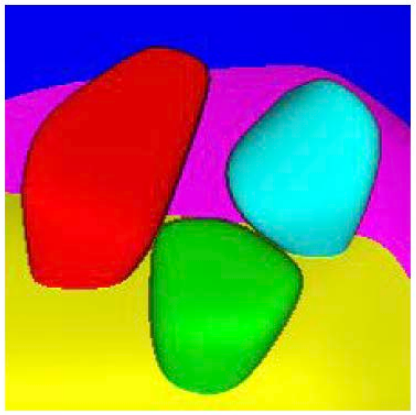
Hint image

Confidence Mask



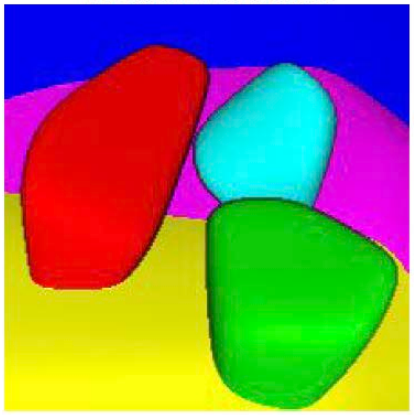
Edited result

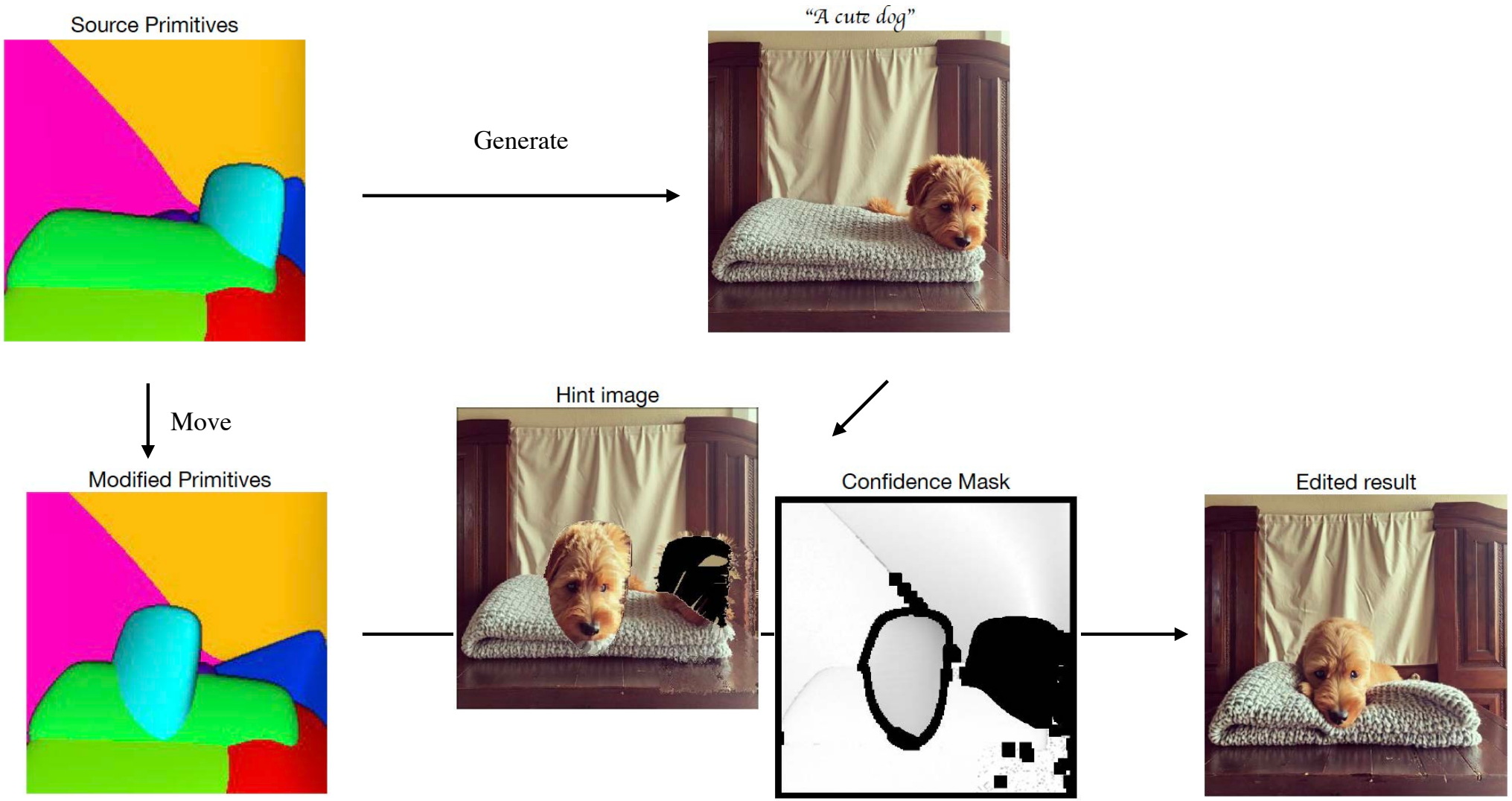
Source Primitives



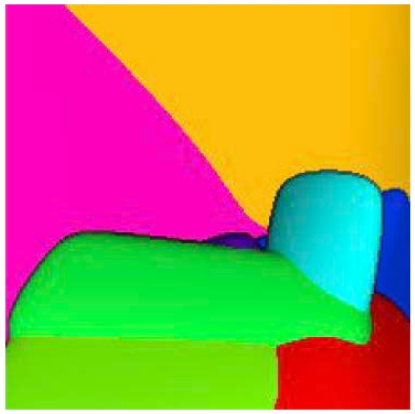
Move

Modified Primitives



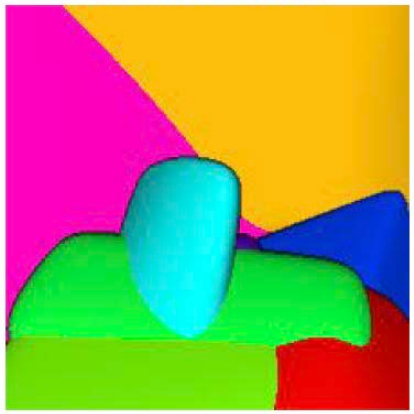

Source Primitives
Generate
"A cute dog"
Move
Modified Primitives
Hint image
Confidence Mask
Edited result

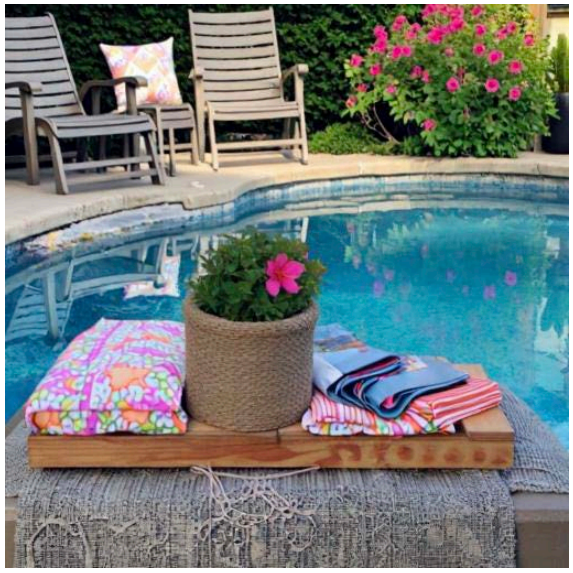
Source Primitives



Move

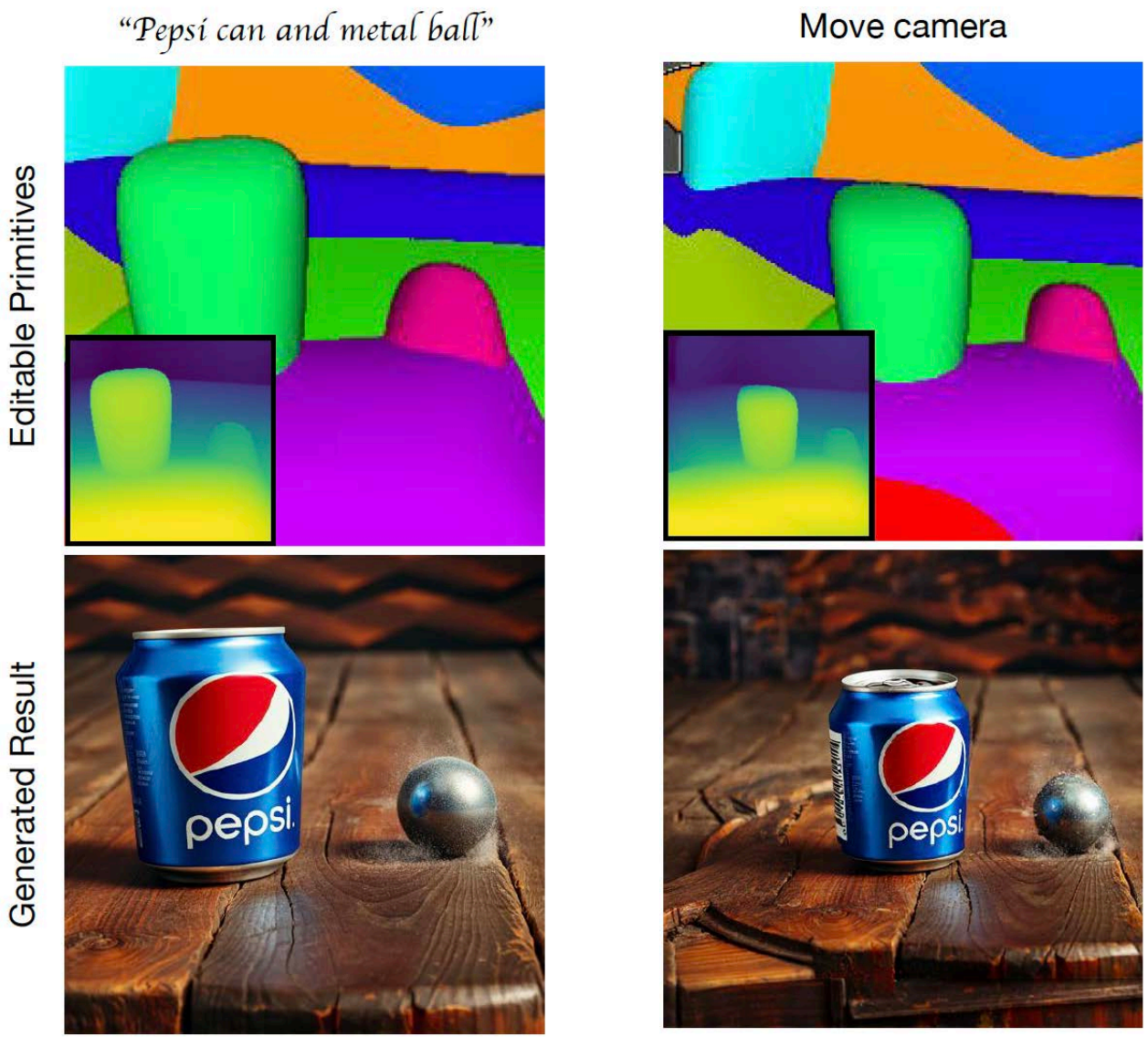
Modified Primitives





Camera moves are hard

- Routine methods:
  - holes and signal issues
- Diffusion methods
  - hallucinate

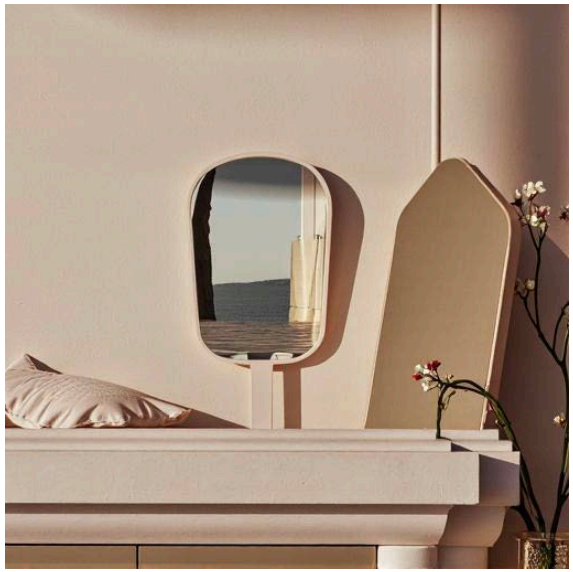

"Pepsi can and metal ball"
Move camera
Editable Primitives
Generated Result
pepsi
pepsi

Simple pan

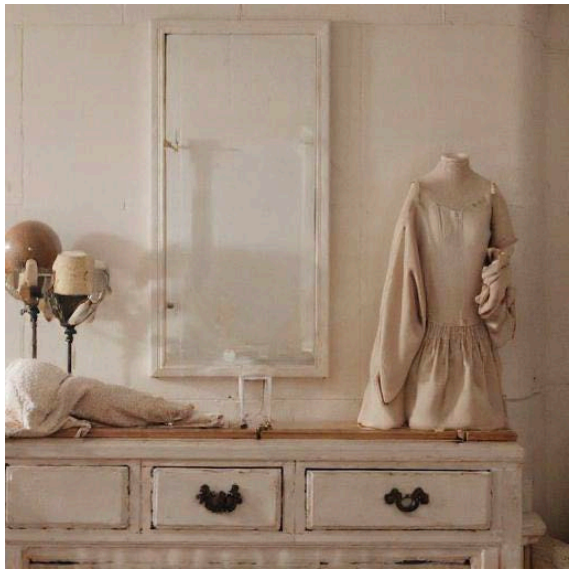
Initial

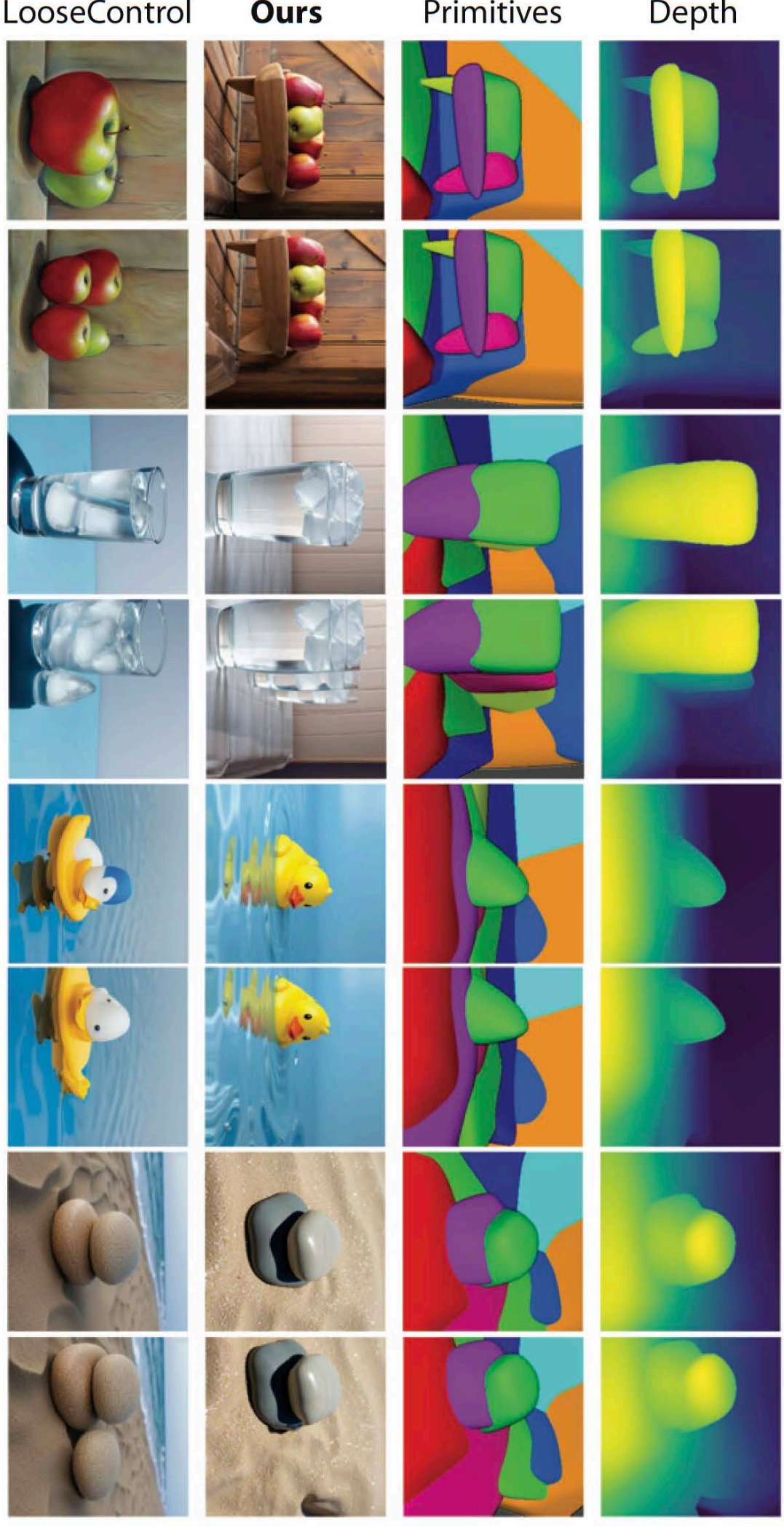
Our move

Omit texture, keep text prompt



StableFlow (keep key-value pairs)




LooseControl
Ours
Primitives
Depth

Evaluate camera moves

- Image 1 -> Move camera -> Image 2
- Image 2 -> Move camera back -> Image 1(ish)

- Compare Image 1 to Image 1(ish)
  - eg PSNR, SSIM

Method	PSNR ↑	SSIM ↑
Ours (FLUX)	**18.7**	**0.874**
LC [Bhat et al. 2023]	6.65	0.670

Points

- Generators are DUMB
- There is no “right” geometric representation
  - some are good for one thing, some for another
  - Q: how to make them live together in ways that are productive
- Bias and Variance are the key ideas
- Primitive representations offer refined control