Composite Shape Modeling via Latent Space Factorization

Anastasia Dubrovina, Fei Xia, Raphae Groscot, Panos Achlioptas, Mira Shalah, and Leonidas Guibas

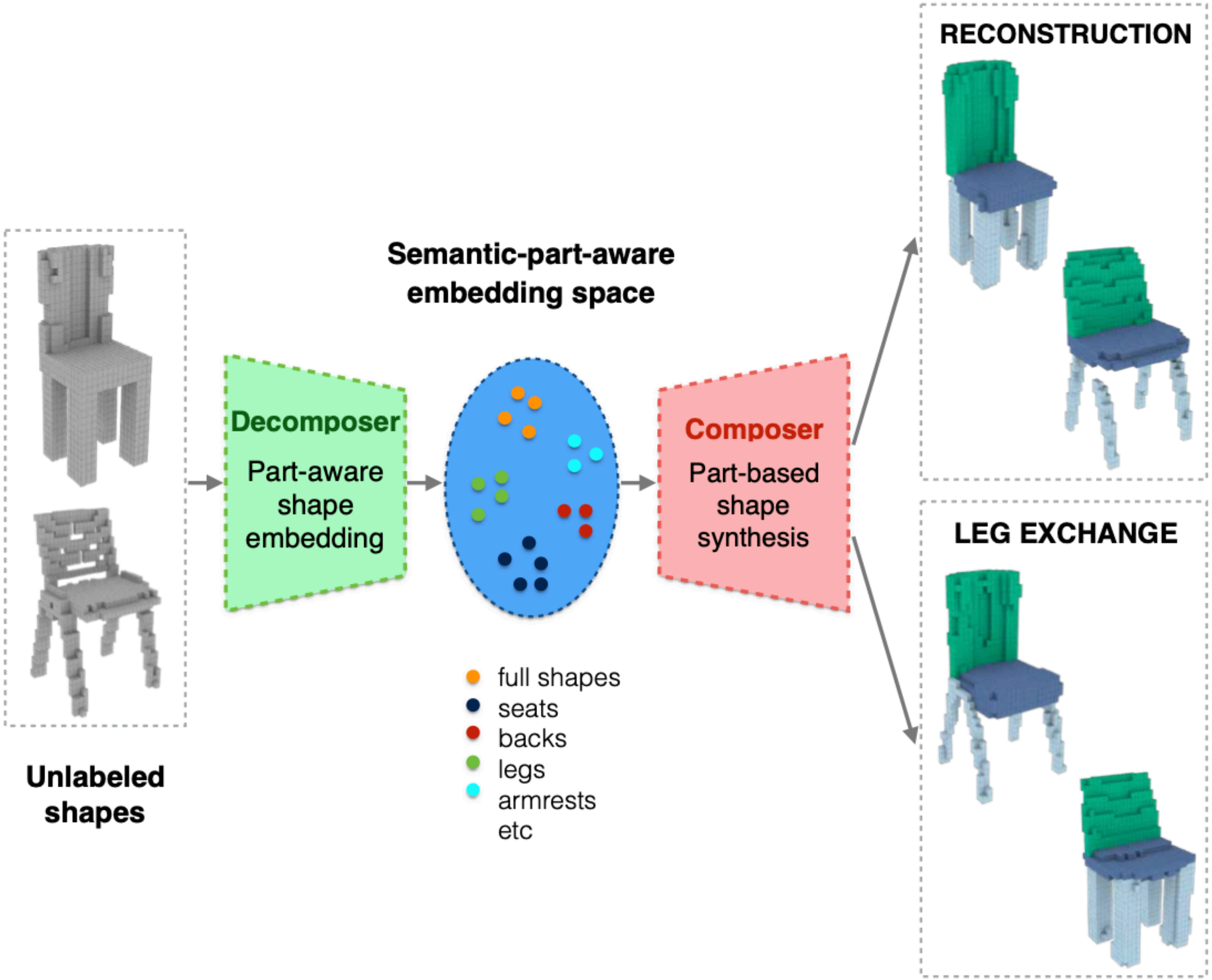
CS348N Student Presentation by Jean Betterton and Boxiao Pan

Introduction

- **The Problem:**
  - learn shape modeling / synthesis in a structure-aware manner
    - work hierarchically with semantic shape parts
- **Vanilla VAE Shortcomings:**
  - Latent spaces of the VAEs correspond to complete shapes
    - entangled latent factors corresponding to different semantic parts
    - difficult to do part-level shape manipulation
      - single part replacement
      - part interpolation
      - part-level shape synthesis.

Introduction

- **Proposed Approach**
    - Auto-encoder-based pipeline
    - Produces a factorized latent space
        - Factorization reflects the semantic part structure of the shapes
        - Compactly encodes geometry
        - Different part encodings lie in separate linear subspaces
        - Shape composition by summing up part embedding coordinates
        - uses data to learn the factorization
    - Operates on unlabeled input shapes
        - infers the shape’s semantic structure
        - compactly encodes its geometry



Unlabeled shapes

Decomposer

Part-aware shape embedding

Semantic-part-aware embedding space

● full shapes
● seats
● backs
● legs
● armrests
etc

Composer

Part-based shape synthesis

RECONSTRUCTION

LEG EXCHANGE

Introduction

- **Network Overview**
  - **The Decomposer**
    - input occupancy grid -> factorized latent space
  - **The Composer**
    - set of part-embedding coordinates -> semantically and geometrically plausible shape w/ part labels
  - **3D Spatial Transformer Network (STN)**
    - Learns and applies part transformations in-network to create coherent hierarchical shape
  - **Cycle consistency constraint**
    - to learn part-based shape manipulation
      - part replacement
      - part interpolation
      - shape synthesis

Main Contributions

- **Novel Latent Space Factorization**
    - Enables shape structure manipulation using linear operations directly in the learned latent space
- **In-network 3D STN**
    - The application of a 3D STN to perform in-network affine shape deformation, used in end-to-end training
- **Cycle Consistency Loss**
    - Improves shape synthesis and reconstruction quality

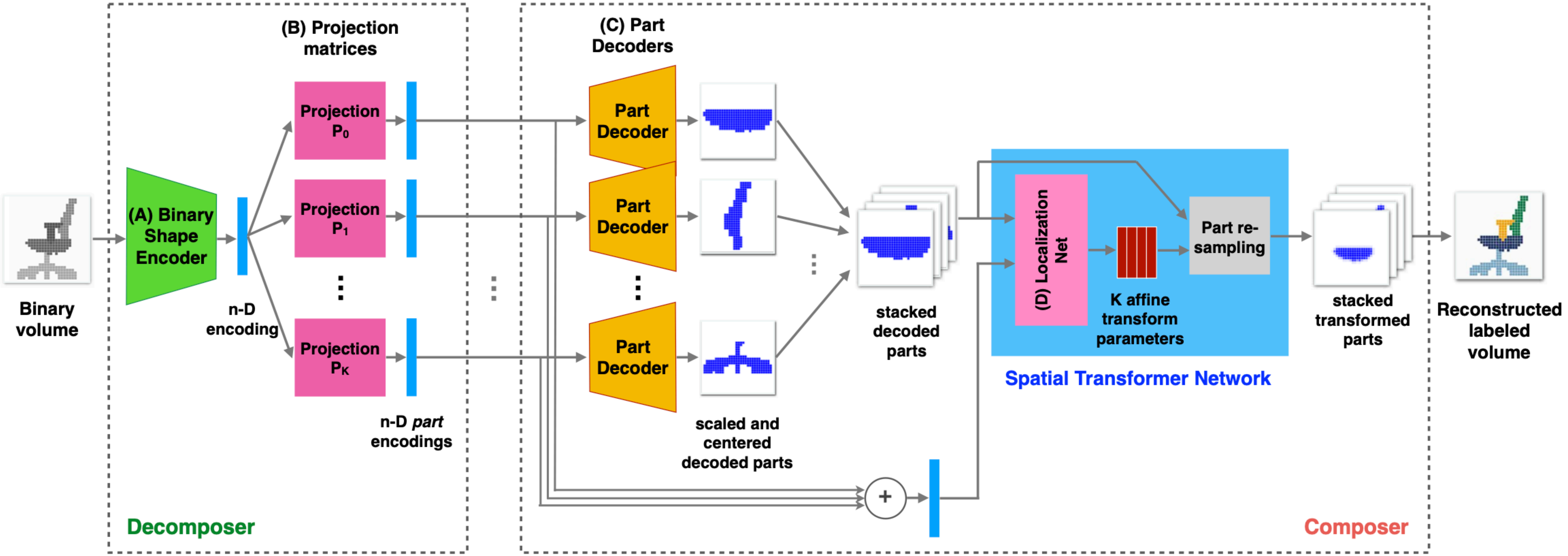


Figure 2: The proposed Decomposer-Composer architecture.

Decomposer Network

- Unlabeled shapes -> factorized embedding space
    - Hierarchical structure
- Has to satisfy two properties
    - 1) Factorization consistency
    - 2) Can combine embeddings of different shape components

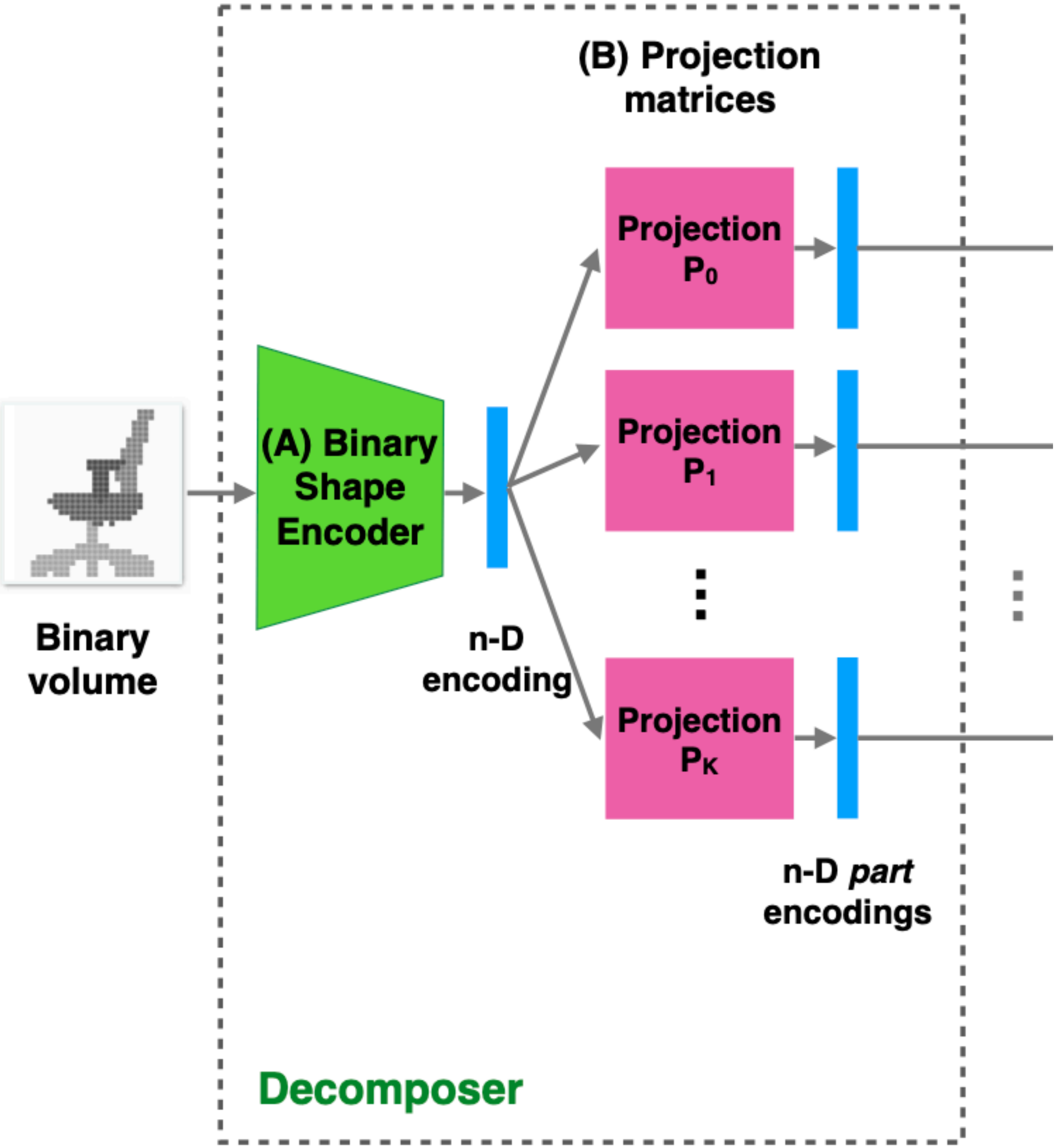
Decomposer Network

- Model embedding space as a direct sum of subspaces

$$V = V_1 \oplus ... \oplus V_k$$

- Could split embedding into K equal sized coordinate groups
  - constrains dimensionality of part embeddings
    - limited capacity, suboptimal
- Learned factorization of the embedding space

$$\begin{aligned} &(1)\ \ P_i^2 = P_i, \forall i, \\ &(2)\ \ P_i P_j = \mathit{0} \text{ whenever } i \neq j, \\ &(3)\ \ P_1 + ... + P_K = I, \end{aligned} \qquad (1)$$


(B) Projection matrices
Projection P_0
Projection P_1
Projection P_K
(A) Binary Shape Encoder
Binary volume
n-D encoding
n-D *part* encodings
Decomposer

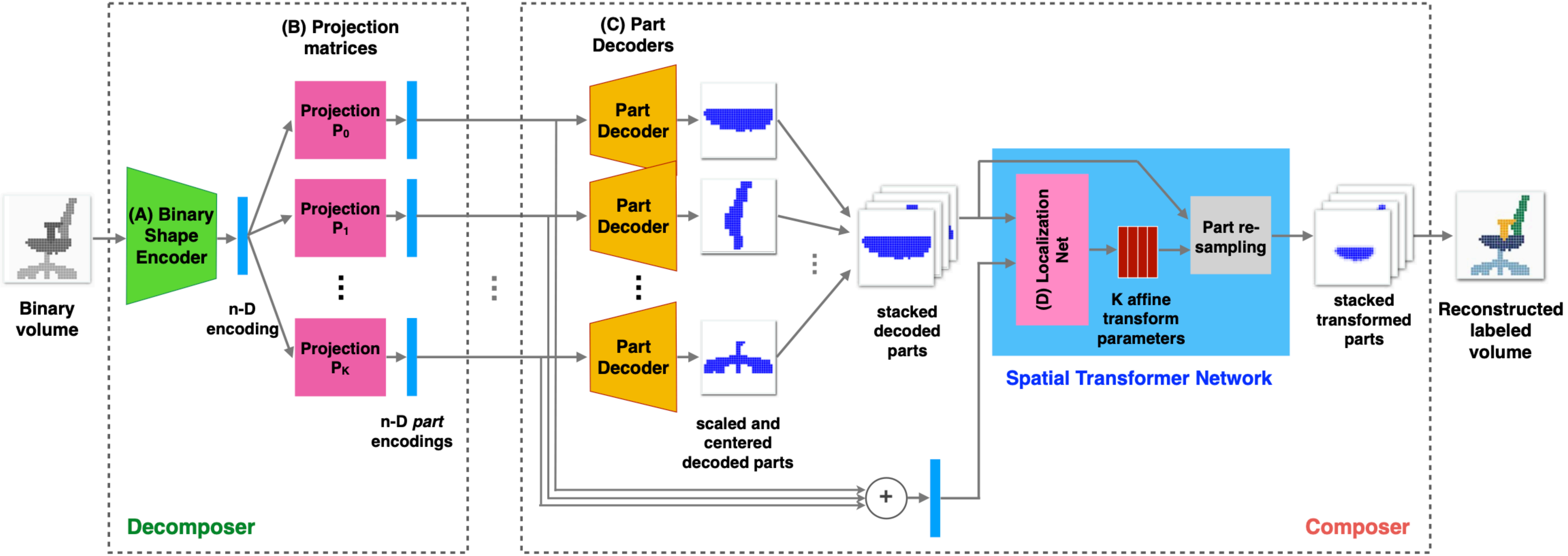
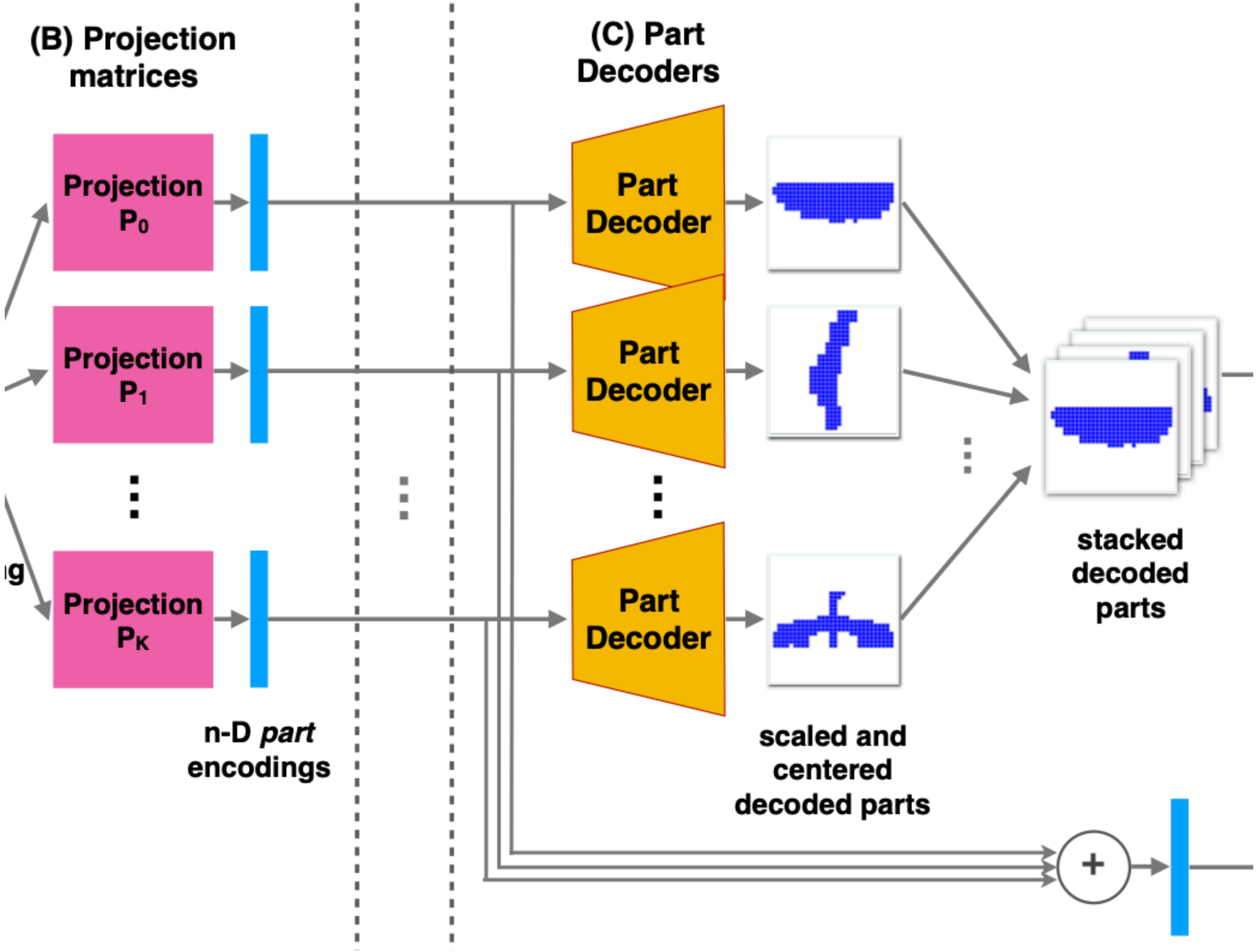


Figure 2: The proposed Decomposer-Composer architecture.

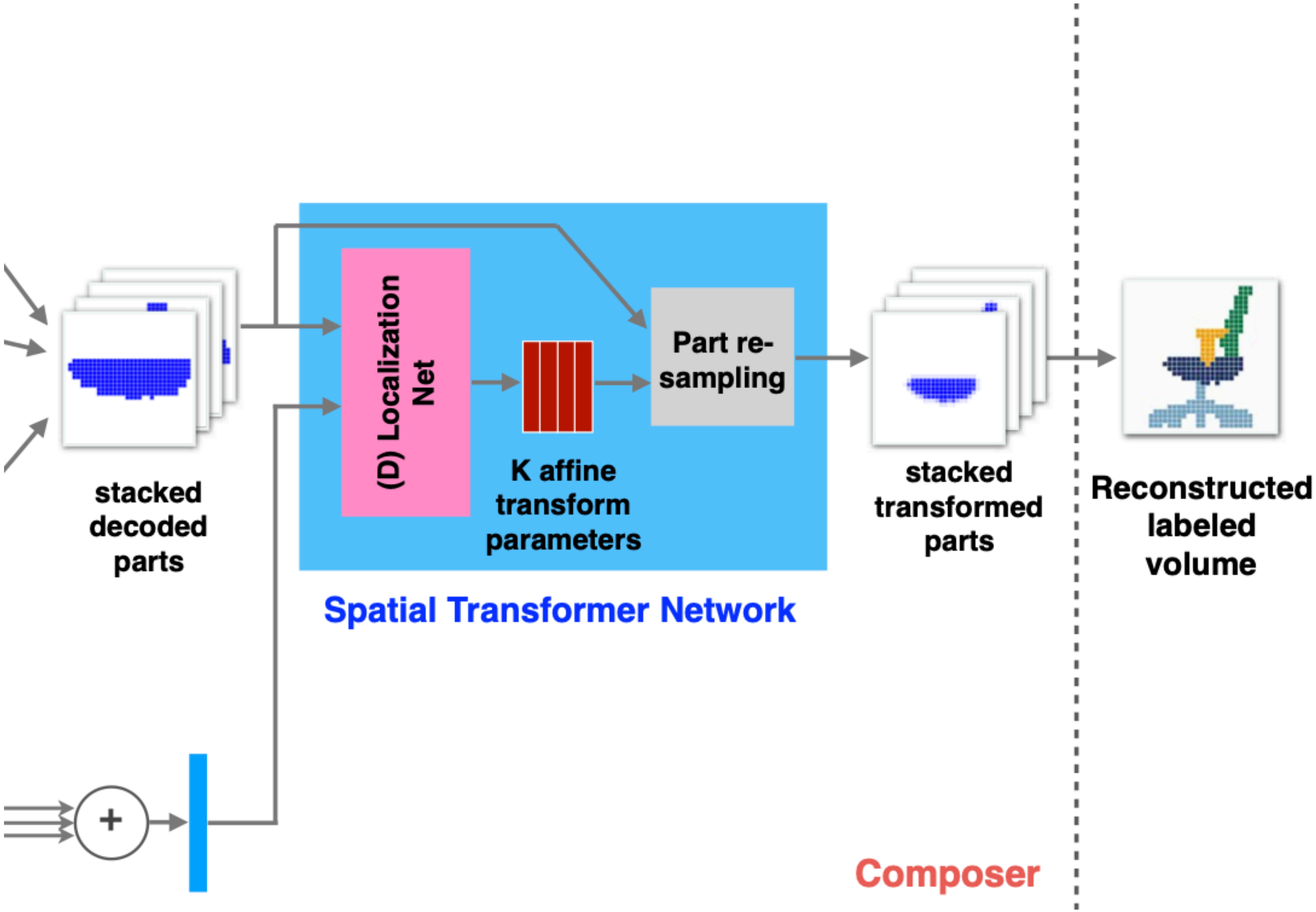
Composer Network

- Sets of part embeddings -> shapes with semantic part labels
- Could use single decoder
  - fails with thin shapes and fine details
- Instead each part’s decoder generates scaled and centered shape parts


(B) Projection matrices
Projection P0
Projection P1
Projection PK
n-D part encodings
(C) Part Decoders
Part Decoder
Part Decoder
Part Decoder
scaled and centered decoded parts
stacked decoded parts
+
g

Composer Network

- Produce per-part parameters to combine the parts into a complete shape
  - per-part affine transformations and translations
    - simplifying assumption
- Uses 3D spatial transformer network (STN)
  - localization net -> 12-D affine transformations / translations
  - re-sampling unit -> which transforms and places part components
  - inputs
    - scaled / centered parts
    - the sum of part encodings



stacked decoded parts

(D) Localization Net

K affine transform parameters

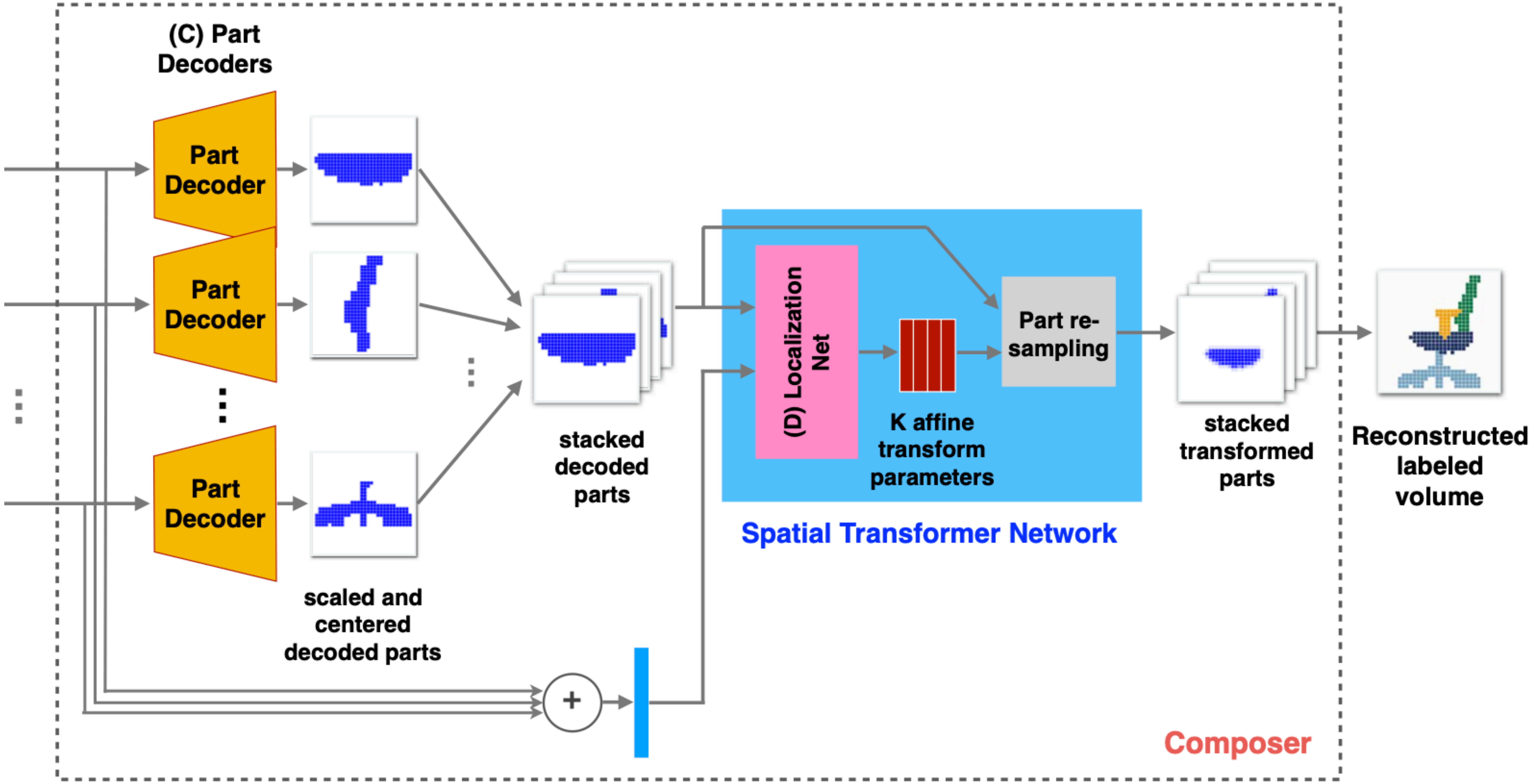
Part re-sampling

Spatial Transformer Network

stacked transformed parts

Reconstructed labeled volume

Composer


(C) Part Decoders
Part Decoder
Part Decoder
Part Decoder
scaled and centered decoded parts
stacked decoded parts
+
(D) Localization Net
K affine transform parameters
Part re-sampling
Spatial Transformer Network
stacked transformed parts
Reconstructed labeled volume
Composer

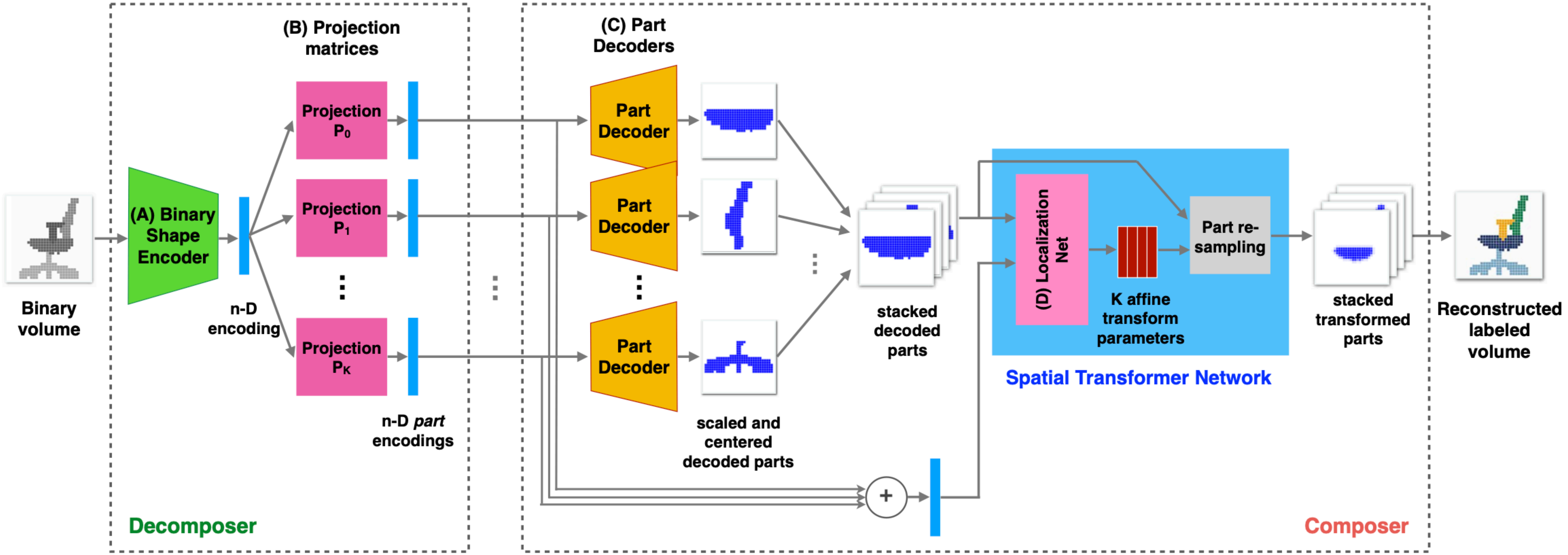


Figure 2: The proposed Decomposer-Composer architecture.

Cycle Consistency

- **Problem:** No training data for synthesized composite shapes!
- **Solution:** Use a cycle consistency constraint
    - 1) Batch of M training shapes
    - 2) K semantic part encodings per shape (w/ Decomposer)
    - 3) randomly mix the part encodings within the batch
        - M new encoding sets w/ one embedding coordinate per part
    - 4) reconstruct the shapes with Composer.
    - 5) Reverse engineer!
    - 6) Compare final output to original

Cycle Consistency

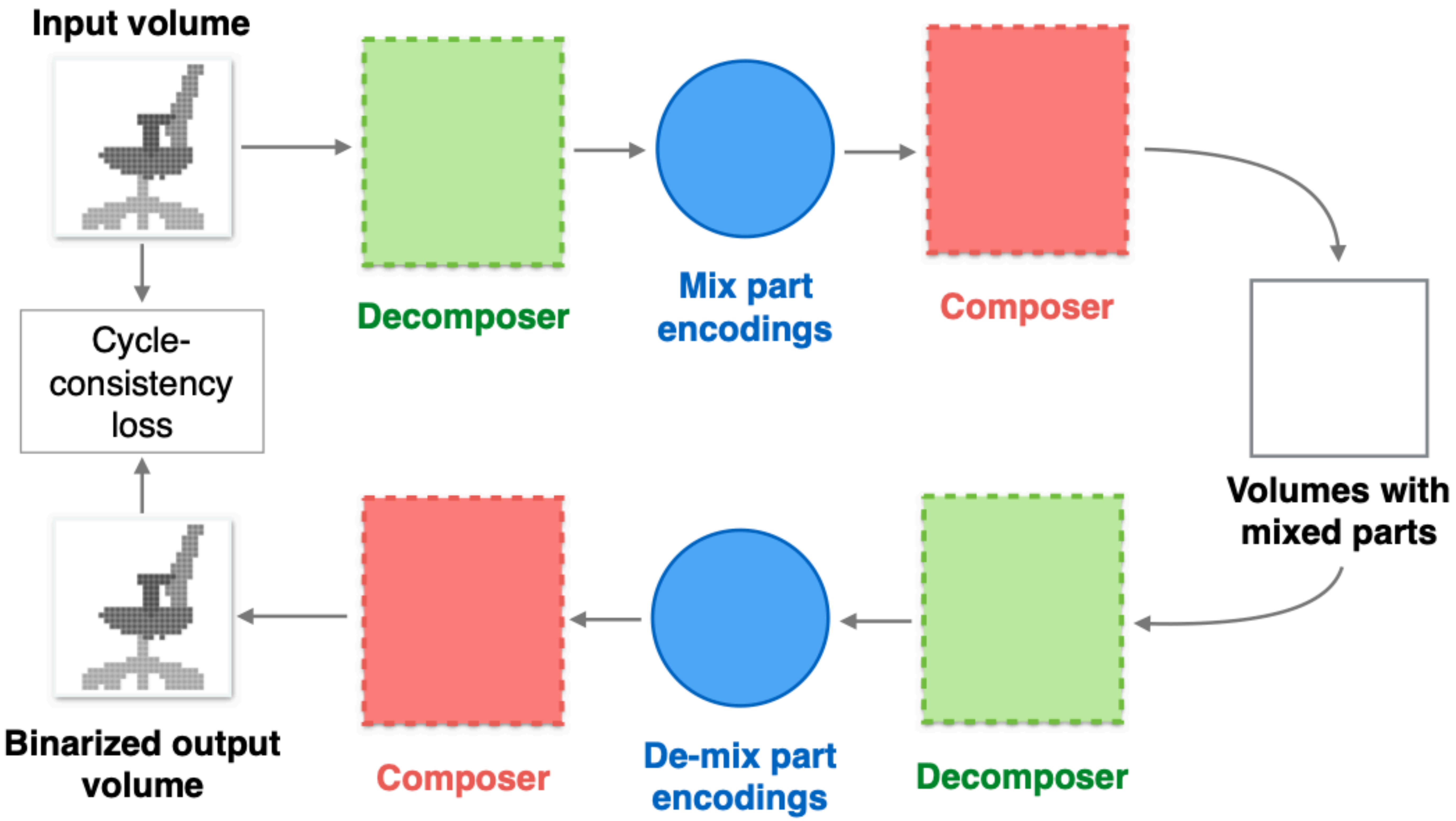


Figure 3: Schematic description of the cycle consistency constraint. See Section 3.3 for details.

Loss Function

$$L = w_{\mathrm{PI}}\mathcal{L}_{\mathrm{PI}} + w_{\mathrm{part}}\mathcal{L}_{\mathrm{part}} + w_{\mathrm{trans}}\mathcal{L}_{\mathrm{trans}} + w_{\mathrm{cycle}}\mathcal{L}_{\mathrm{cycle}}. \quad (2)$$

- ***L*PI:** Deviation of the predicted projection matrices from projection constraints
- ***L*part**: Reconstructed centered and scaled part volumes vs GT
- ***L*trans:** Regression loss between the predicted and the ground truth transformation parameter vectors
- ***L*cycle:** Cycle consistency loss
- wPI = 0.1, wpart = 100, wtrans = 0.1, wcycle = 0.1 in experiments

Interesting Training Details

- The network was trained on each shape category separately
- Training over **500 epochs**
  - **150 epochs** Essential to pretrain the binary shape encoder, projection layer, and part decoder parameters separately
    - Use ***L*****PI** and ***L*****part**, ignore ***L*****trans** and ***L*****cycle**
  - **100 epochs** Train the parameters of the spatial transformer network keeping the rest of the parameters fixed.
  - **250 epochs** Train everything together for fine tuning

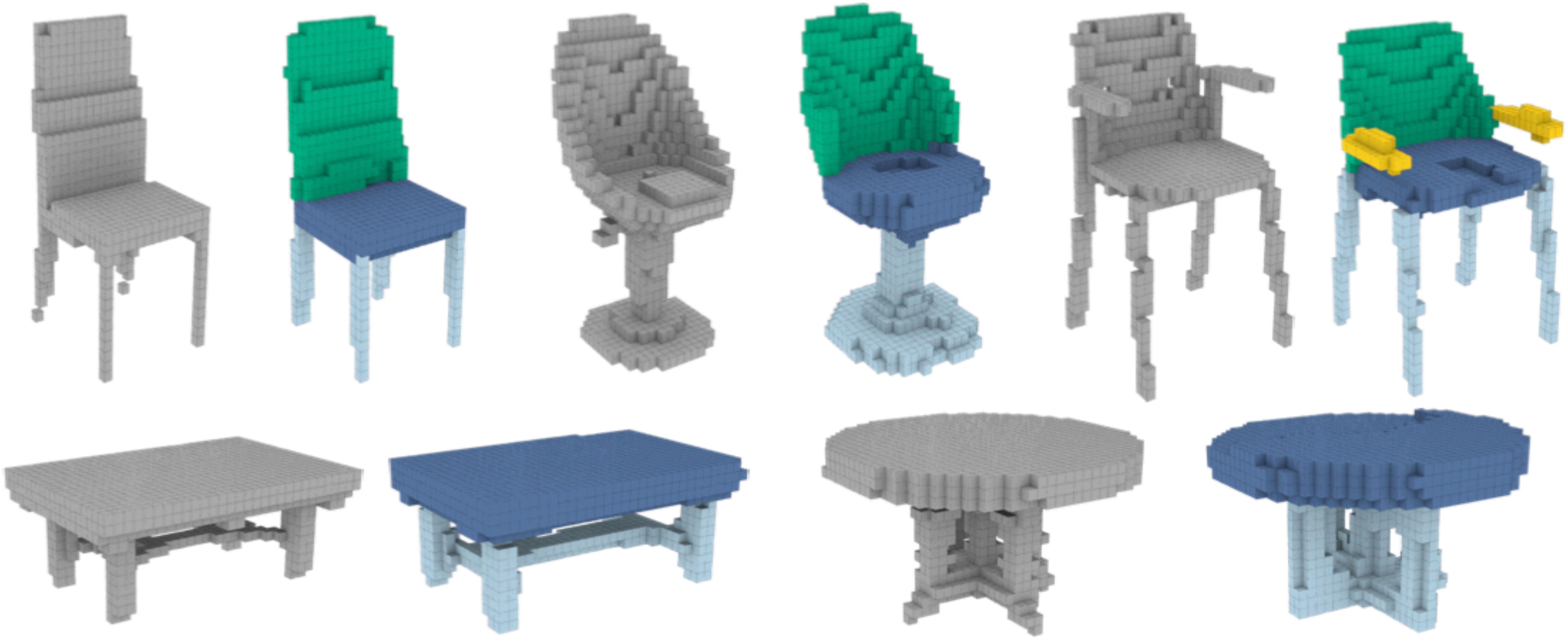


Figure 4: Reconstruction results of the proposed pipeline, for chair and table shapes. Gray shapes are the input test shapes; the results are colored according to the part label.

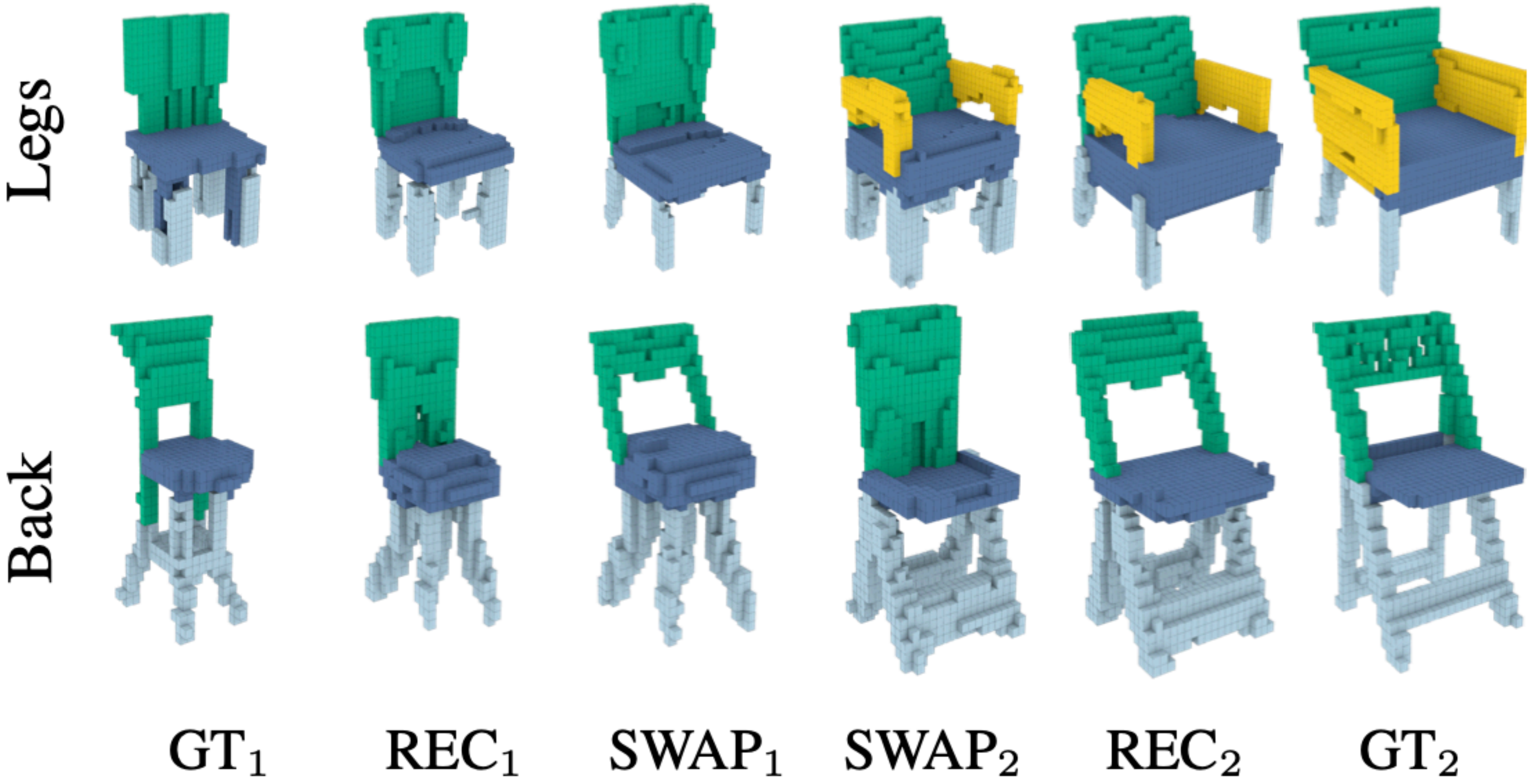


Figure 5: Single part exchange experiment. $GT_{1/2}$ denote ground truth shapes, $REC_{1/2}$ - reconstruction results, $SWAP_{1/2}$ - part exchange results. Unlabeled shapes were used as an input.

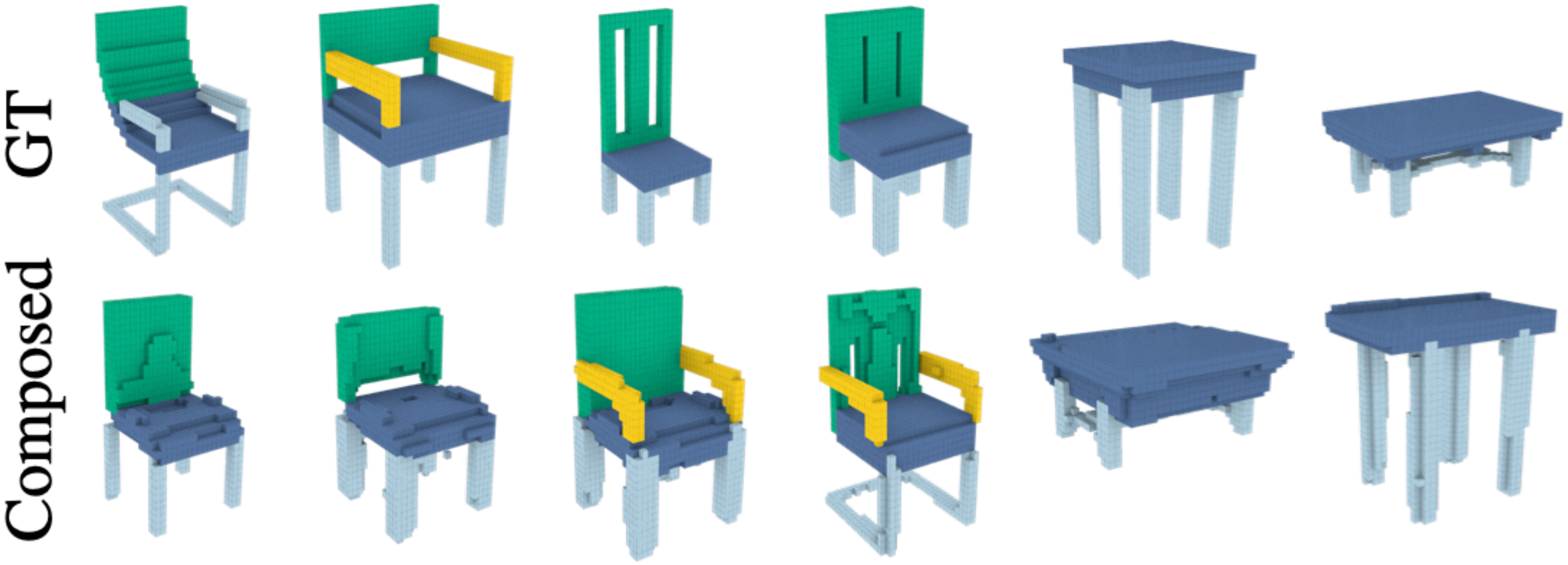


Figure 6: Shape composition by random part assembly. The top row shows the ground truth (GT) shapes, and the bottom row - shapes assembled using the proposed approach (see Section 4.2). Unlabeled shapes were used as an input.

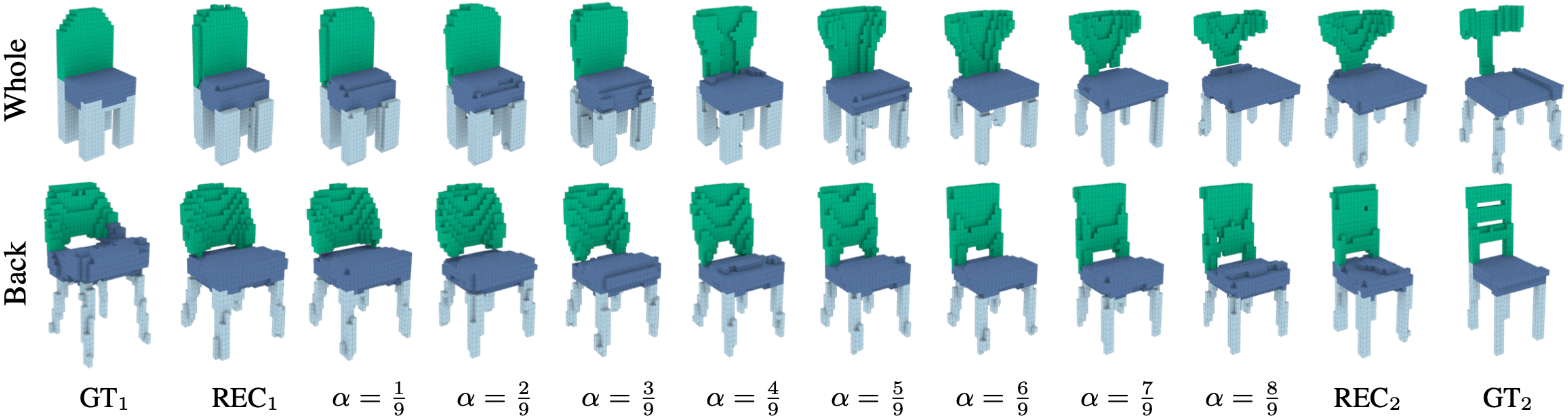


Figure 7: Example of a whole (top) and partial (bottom) shape interpolation. GT$_{1/2}$ denote original models, REC$_{1/2}$ - their reconstructions, and linear interpolation results are in the middle. Unlabeled shapes were used as an input.

Metric Method	mIoU	mIoU (parts)	Connectivity			Classifier accuracy			Symmetry score		
	Rec.	Rec.	Rec.	Swap	Mix	Rec.	Swap	Mix	Rec.	Swap	Mix
Our method	0.64	0.65	**0.82**	**0.71**	**0.65**	**0.95**	0.89	**0.83**	0.95	0.95	0.95
W/o cycle loss	0.63	0.66	0.74	0.62	0.54	0.93	0.84	0.80	0.96	0.96	0.95
Fixed projection	0.63	0.65	0.72	0.61	0.58	0.94	0.86	0.77	0.94	0.95	0.95
Composer w/o STN	**0.75**	**0.8**	0.69	0.48	0.23	**0.95**	**0.9**	0.71	0.95	0.91	0.85
Naive placement	-	-	-	0.68	0.62	0.61	0.47	0.21	-	**0.96**	**0.96**
ComplementMe	-	-	-	**0.71**	0.47	-	0.66	0.43	-	0.66	0.43
Segmentation+STN	-	-	-	0.41	0.64	-	0.64	0.36	-	0.77	0.77

Table 1: Ablation study results. The evaluation metrics are mean Intersection over Union (*mIoU*), per-part mean IoU (*mIoU (parts)*), shape *connectivity* measure, binary shape *classifier accuracy*, and shape *symmetry score*. Rec., Swap and Mix stand for the shape reconstruction, part exchange and random part assembly experiment results, respectively (see Section 4.2). See Section 4.4 for a detailed description of the compared methods and the evaluation metrics.

Main Contributions

- **Novel Latent Space Factorization**
    - Enables shape structure manipulation using linear operations directly in the learned latent space
- **In-network 3D STN**
    - The application of a 3D STN to perform in-network affine shape deformation, used in end-to-end training
- **Cycle Consistency Loss**
    - Improves shape synthesis and reconstruction quality

Summary Strengths and Limitations

- **Strengths**
    - Structure-aware 3D shape modeling
    - Generate a factorized latent shape representation
        - Different semantic part embedding coordinates lie in separate linear subspaces
    - Allows shape manipulation via part embedding coordinates
        - exchange / interpolate parts between shapes
        - synthesize novel shapes
- **Limitations**
    - Memory constraints limit resolution of voxel representations
    - Simplifying assumptions on affine transformations