Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning

Marat Saidov, AMI171
September, 22th

01. SHAPENET

Dataset’s components
And a crucial property

02. ARCHITECTURE & LOSSES

Visually interpretable objective functions

03. REGULARIZATION

Keypoints modeling is unstable

04. APPLICATIONS IN CV

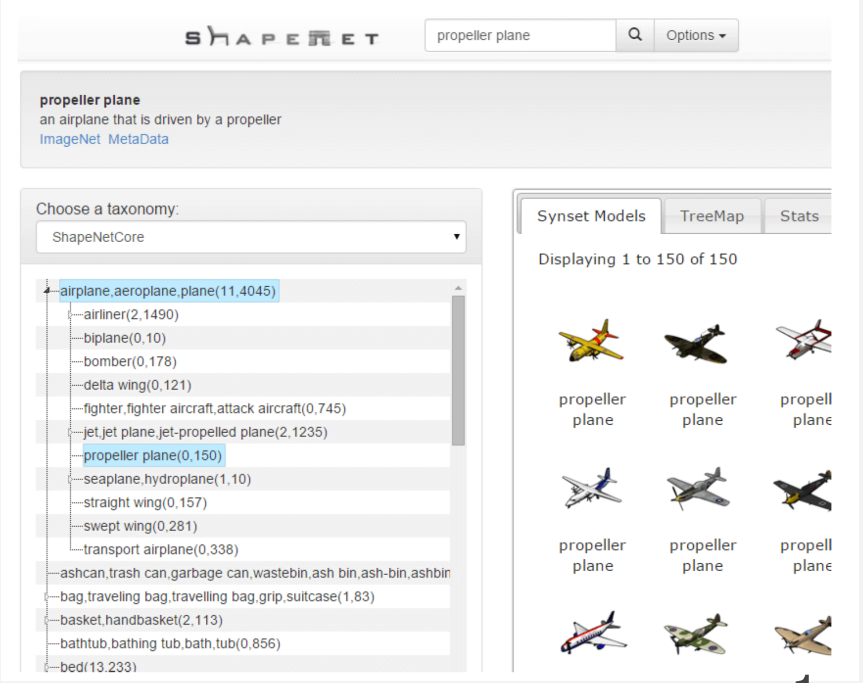
Pose2Seg: human instance segmentation

05. VISUALIZATION

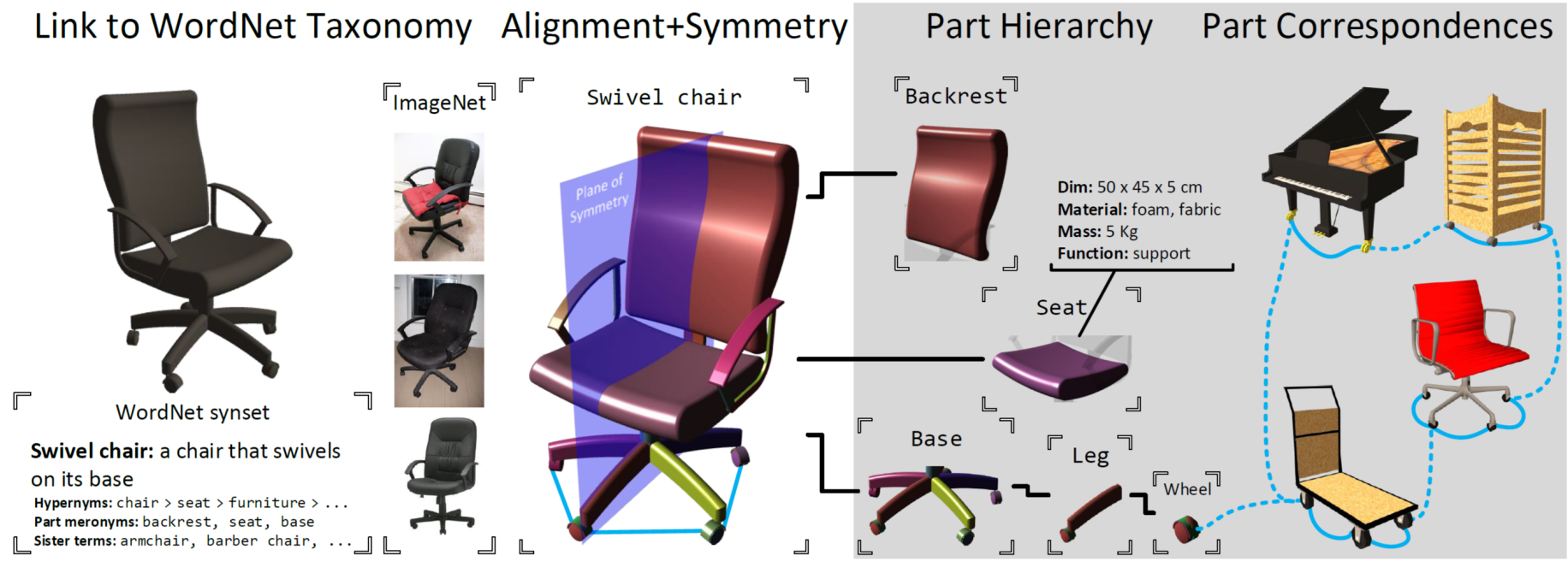
Ablation studies and exceptional cases

WELL-STRUCTURED 3D-MODELS DATASET

- Salient physical attributes: planes of symmetry, parts decompositions;
- Interfaces: textual keywords, taxonomy traversal, image & shape similarity search;



PARTS DECOMPOSITIONS



RIGID ALIGNMENTS

An algorithm of rigid alignments

- Bottom-up manner: gradually elevating from low-level categories;
- Low-level shape is a random variable;
- Tree-structured set of variables is a Markov Random Field (MRF);
- Discretized set of rigid transformations;
- MRF measures a consistency over all transformations;
- Insignificant human supervision;

GENERAL VIEW

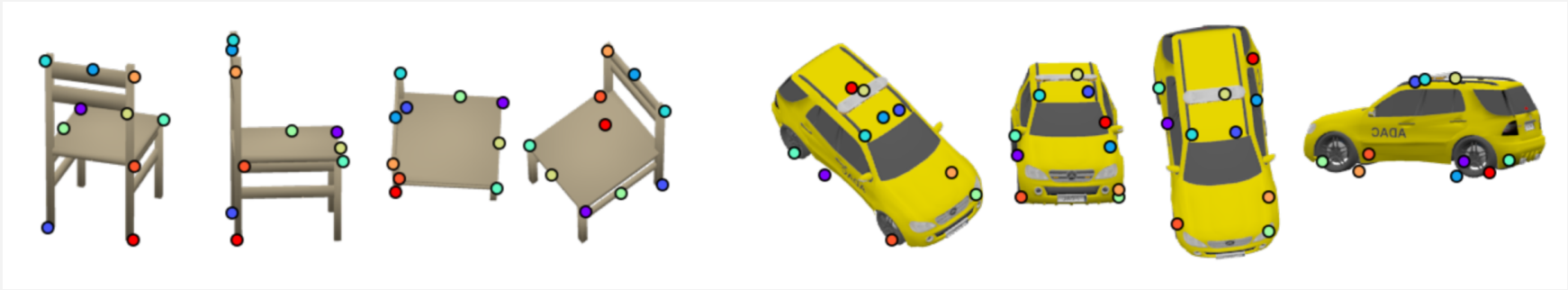
Learns set of **category-specific** 3D keypoints which could be optimized for
a downstream task. It does not obtain ground-truth keypoint annotations.

GENERAL VIEW

3D Keypoint = Pixel Coordinate + Associated Depth

Modeling an ordered list of 3D keypoints for a single image:

- Occluded points;
- Generalization over different views;



GENERAL VIEW

Informative properties of keypoints:

- Distinctiveness / diversity;
- Ease of detection;

Restriction on a downstream task: an objective function should be differentiable w.r.t keypoints values;

OBJECTIVE FUNCTION

KeypointNet has N heads that extract N keypoints **(fixed before training).** Each head tends to predict points with the same semantic interpretation.

Keypoints are optimized directly on a downstream task! At training time it was chosen a **relative pose estimation** task.

RIGID TRANSFORMATIONS

Rigid transformation preserves the Euclidian distance between every pair of points.

Let (I, I') be a pair of different views on the same object. A transformation T is rigid if it satisfies the following matrix form:

$$T = \begin{pmatrix} R^{3\times3} & t^{3\times1} \\ 0 & 1 \end{pmatrix}$$

R – rotation matrix, t – translation vector.

OBJECTIVE FUNCTION

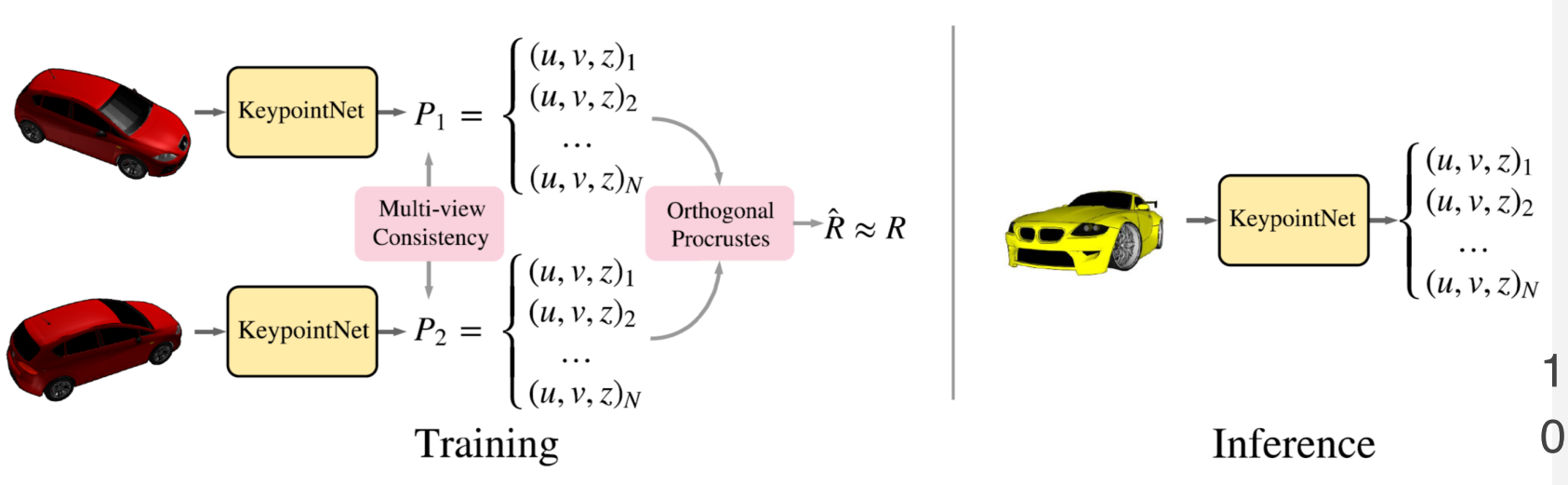
Relative pose estimation: considering two views (I, I') predict two lists of keypoints: P_1 and P_2. They should correspond to the same semantic positions.

An objective $\mathrm{O}(P_1, P_2)$ consists of two crucial components:

- **Relative pose estimation loss** measures the difference between a ground truth rotation R and rotation $\hat{R}$ recovered from P_1 and P_2 using **orthogonal procrustes**;
- **Multi-view consistency loss** measures the discrepancy between two sets of points under a ground truth transformation;

ARCHITECTURE

These two terms allow to predict meaningful locations. However, they could be suboptimal.



ARCHITECTURE

Thus, there is a learning transformation $f_\theta(I)$ that maps an image I to a list of 3D points $P = (p_1, \ldots, p_N)$, $p_i \equiv (u_i, v_i, z_i)$.

Objective function $O(f_\theta(I), f_\theta(I'))$ that is differentiable w.r.t θ.

MULTI-VIEW CONSISTENCY

Perspective projection operation. We have a viewpoint with a focal length f. Denote $[x, y, z]$ as a 3D coordinate, $[u, v]$ – pixel coordinate.

$\pi : \mathbb{R}^4 \rightarrow \mathbb{R}^4$

$$\pi\left(\left[x, y, z, 1\right]^T\right) = \left[\frac{fx}{z}, \frac{fy}{z}, z, 1\right]^T = [u, v, z, 1]^T$$

MULTI-VIEW CONSISTENCY

Multi-view consistency is a pixel-wise loss. We optimize a predictive ability of a model for a direct transformation T and the reciprocal one T^{-1}:

$$[\hat{u}, \hat{v}, \hat{z}, 1]^T \sim \pi T \pi^{-1}\left([u, v, z, 1]^T\right)$$

$$[\hat{u}', \hat{v}', \hat{z}', 1]^T \sim \pi T \pi^{-1}([u', v', z', 1]^T)$$

$\hat{u}$ – a projection of u to the second view, $\hat{u}'$ – a projection of u' to the first view.

MULTI-VIEW CONSISTENCY

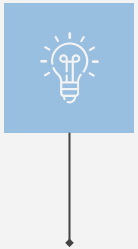
Multi-view consistency loss:

$$L_{con} = \frac{1}{2N} \sum_{i=1}^{N} \left\| \left[u_i, v_i, {u_i}', {v_i}' \right]^T - \left[\hat{{u'}}_i, \hat{{v'}}_i, \hat{u}_i, \hat{v}_i \right]^T \right\|$$

Depth is measured with different units and never directly observed.
However, its knowledge is crucial to project between 3D and pixel views.

MULTI-VIEW CONSISTENCY

An optimal set of keypoints with multi-view consistency leads to a collapse into one single location.



Natural way of encouragement their diversity is the usage of a downstream task.

RELATIVE POSE ESTIMATION

Relative pose estimation loss measures the misfit between an estimated rotation $\hat{R}$ between two sets of keypoints (calculated by **Procrustes' alignment**) and the ground truth R.

$$L_{pose} = 2arcsin\left(\frac{1}{2\sqrt{2}}\left\|\hat{R} - R\right\|_F\right)$$

ORTHOGONAL PROCRUSTES PROBLEM

How to estimate $\hat{R}$ properply?

Let $X \in \mathbb{R}^{3\times N}$and $X' \in \mathbb{R}^{3\times N}$ denote two sets of unprojected 3D keypoints.

Take mean-subtracted versions of X and X': $\tilde{X}$ and $\tilde{X}'$.

Apply SVD to their multiplication: $U\Sigma V^T = SVD(\tilde{X}\tilde{X}'^T)$.

ORTHOGONAL PROCRUSTES PROBLEM

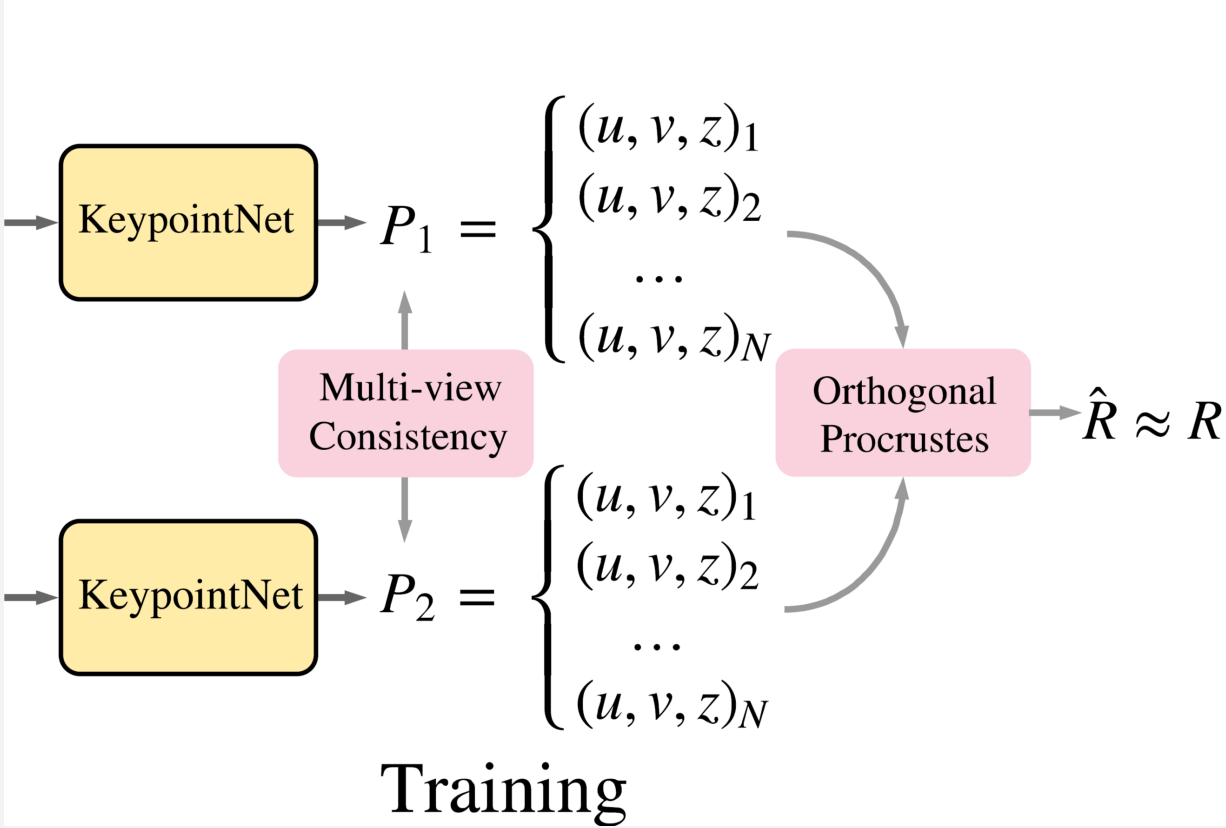
The optimal least-squares solution for estimated rotation is:

$\hat{R} = Vdiag(1,1,\ldots,\det(VU^T))U^T$

Adding small Gaussian noise to make sure that $\tilde{X}\tilde{X}'^T$ is invertible.

To minimize relative pose estimation loss we can backpropagate through SVD.

ORTHOGONAL PROCRUSTES PROBLEM

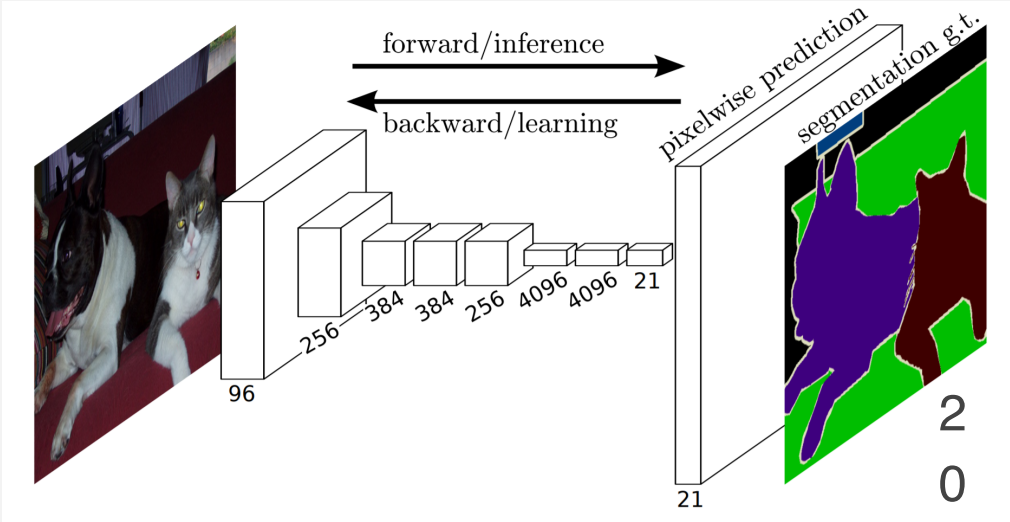


ARCHITECTURE

Equivariance for translations – keypoints should shift if the image shifts.

- Each head predicts not $[x, y, z]$ but a probability distribution map $g_i(u, v)$.
- The expected value of this distribution is a predicted keypoint:

$$[u_i, v_i]^T = \sum_{u,v} [u \bullet g_i(u, v), v \bullet g_i(u, v)]^T$$



ARCHITECTURE

To predict depth it is necessary to learn another transformation $d_i(u, v)$:

$$z_i = \sum_{u,v} d_i(u, v) g_i(u, v)$$

ARCHITECTURE

- 13 layers with 3x3 kernels with the following dilation rates:

1	1	2	4	8	16	1	2	4	8	16	1	1

- Last layer has $2N$ output channels splitted between g_i and d_i.
- LeakyReLU and BN for all layers except the last layer;
- Spatial SoftMax for distributions;

SYMMETRICAL OBJECTS

Symmetrical parts of an object should not be perceived identically.

Understanding global context: e.g., knowing car orientation helps to differ left wheels from the right wheels.

Using consistent orientations from ShapeNet.

SYMMETRICAL OBJECTS

Additional half-sized network to predict a binary flag: supervised from ShapeNet.

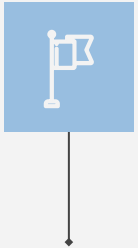
Normalizing and object and training to predict $[-1,0,0]$ or $[1,0,0]$ (left or right directions).

Use such predictions to aid KeypointNet main model.

REGULARIZATION

Add an artificial way of keypoints diversity: **separation loss**.

$$L_{sep} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i \neq j}^{N} \max\left(0, \delta^2 - \left\| X_i - X_j \right\|^2\right)$$



Loss is computed between 3D points to let keypoints have the same pixel location with a different depth.

Silhoutte consistency: the probabilities $g_i(u, v)$ are allowed to be non-zero only inside of an object silhouette.

Let $b(u, v) \in \{0,1\}$ be a binary segmentation mask.

$$L_{obj} = \frac{1}{N} \sum_{i=1}^{N} -log \sum_{u,v} b(u, v) g_i(u, v)$$

$L_{obj} = 0$ if the probability mass lies within a silhouette.

REGULARIZATION

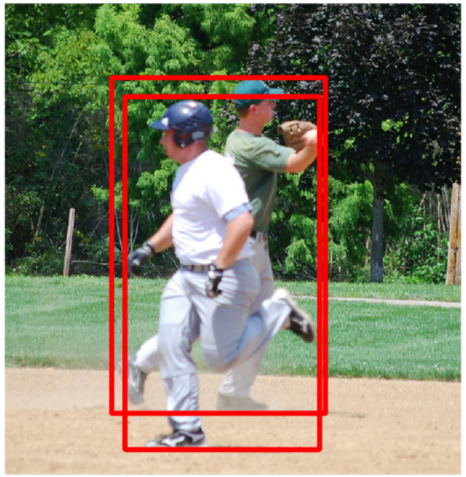
Silhoutte consistency: the probabilities $g_i(u, v)$ are allowed to be non-zero only inside of an object silhouette.

By penalizing the variance of a pixel-wise distribution the keypoints are stimulated to lay within a silhouette as well:

$$L_{var} = \frac{1}{N} \sum_{i=1}^{N} \sum_{u,v} g_i(u, v) \left\| [u, v]^T - [u_i, v_i]^T \right\|^2$$

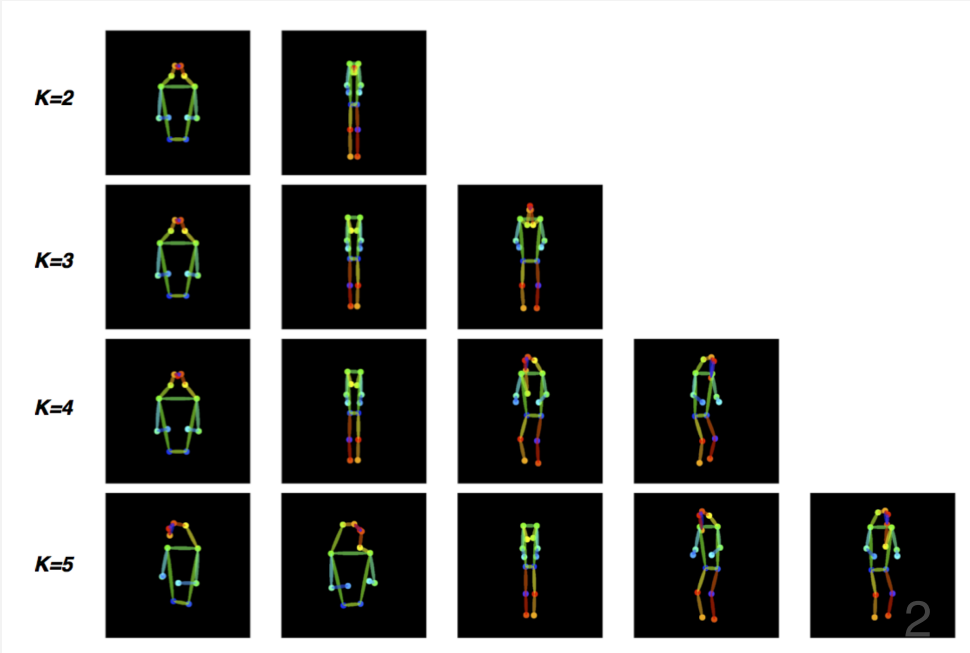
POSE2SEG: OVERVIEW

- Previously: object detection ⇨ instance segmentation;
- Problem: heavily occluded people are unrecognizable;
- Better solution: pose estimation ⇨ instance segmentation;

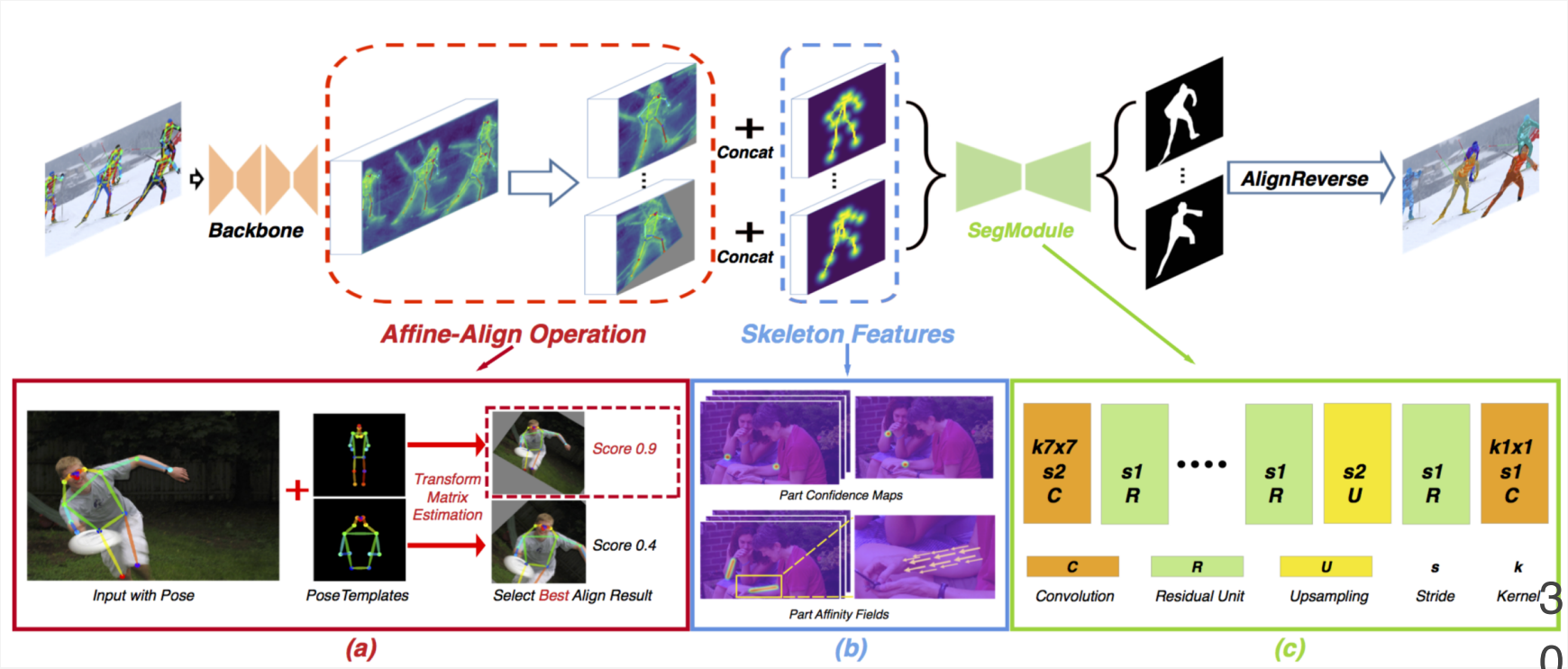


POSE2SEG: OVERVIEW

1. Clusterization of main human poses based on K-means;
2. Choose the best affine transformation H among pose templates and a given pose;
3. Apply transformation H to an input image;



POSE2SEG: OVERVIEW



05 VISUALIZATION

QUESTIONS

- Describe a perspective projection operation, rigid transformations and a mult-view consistency loss.
- Formulate the Orthogonal Procrustes Problem and describe how it is used in KeypointNet.
- Write out a silhouette consistency loss. How does it help in training?

REFERENCES

- **Suwajanakorn et. al (2018)**

Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning

- **Zhang et. al (2018)**

Pose2Seg: Detection Free Human Instance Segmentation

- **Chang et. al (2015)**

ShapeNet: An Information-Rich 3D Model Repository

- **Hejrati et al (2012)**

Analyzing 3D Objects in Cluttered Images