Coarse-to-fine Animal Pose and Shape Estimation

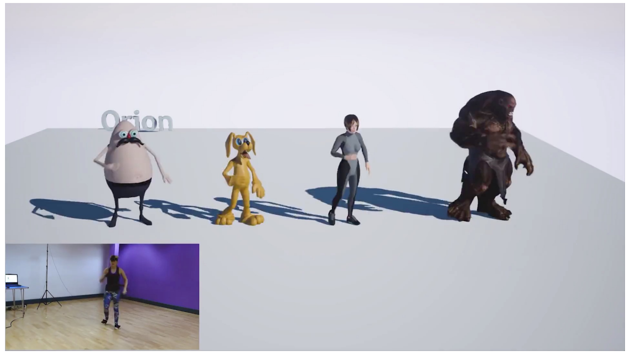
Chen Li, Gim Hee Lee

National University of Singapore

Animal pose and shape estimation

- Goal: Estimate 3D animal pose and shape from a monocular image.

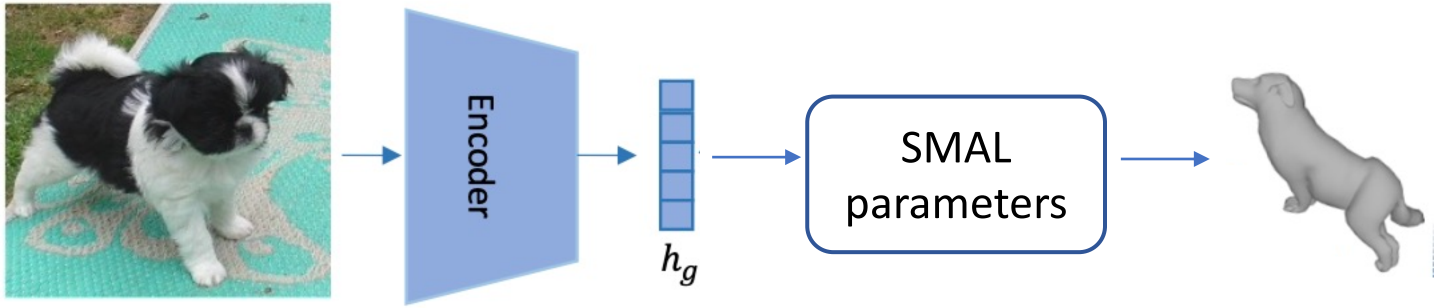
- Applications: zoology, ecology, farming and entertainment.



[1] Silvia Zuffi , et al. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. CVPR 2017
[2] Silvia Zuffi , et al. Three-D Safari: Learning to Estimate Zebra Pose, Shape, and Texture from Images "In the Wild"

Animal pose and shape estimation: Existing works

- Existing works are based on the SMAL.

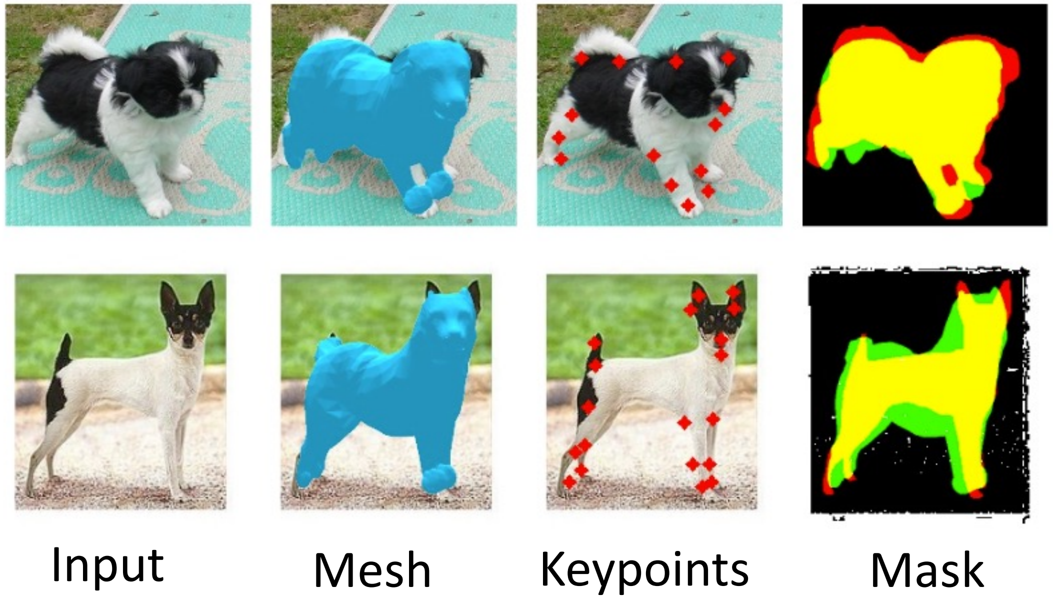


- The low dimensional parameterization of SMAL makes it easier for deep networks to learn the high dimensional 3D meshes.

- The shape space of the SMAL is learned from 41 scans of toy animals, which limits the representation capacity.

Animal pose and shape estimation: Problem

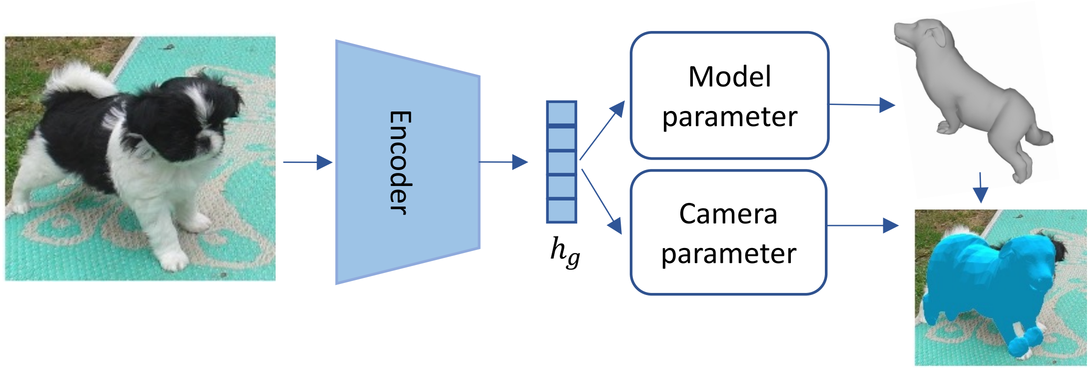
- Problem: The estimated 3D meshes do not match well with the 2D observations.



- Our solution: A two-stage approach combining parametric and non-parametric representations.

Our approach: coarse estimation stage

- Coarse estimation stage: Regress SMAL parameters from input image.



Model parameter: $\Theta' = \{\beta', \theta', \gamma'\}$

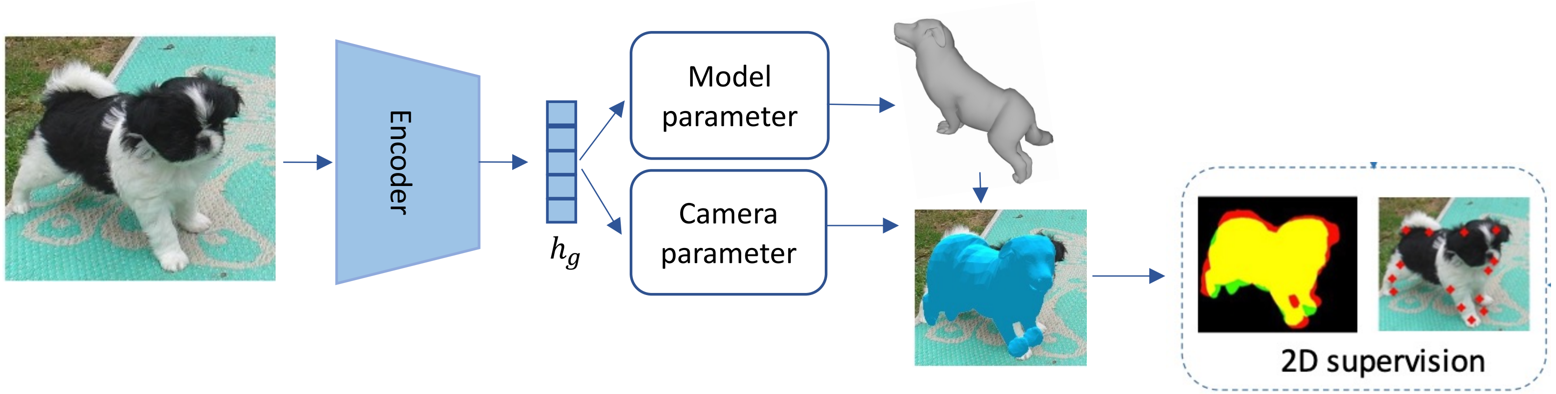
Camera parameter: f.

Mesh vertices: $V_c = \mathcal{M}(\beta', \theta', \gamma')$

Body joints: $J_{3D} = \mathcal{W} \times V_c$.

Our approach: coarse estimation stage

- Coarse estimation stage: Regress SMAL parameters from input image



Camera parameter: f

Mesh vertices: $V_c = \mathcal{M}(\beta', \theta', \gamma')$

Body joints: $J_{3D} = \mathcal{W} \times V_c$

2D keypoint based loss: $\mathcal{L}_{kp1} = \|J_{2D} - \Pi(J_{3D}, f)\|^2$

2D silhouette based loss: L1 or L2 distance

Our approach: coarse estimation stage

- The estimated shape tends to have a larger foreground area compared to GT.

- The imbalance between foreground and background pixels in the input image.

- Tversky loss: $\mathcal{T}(P, G; \alpha, \beta) = \dfrac{|PG|}{(|PG| + \alpha|P \backslash G| + \beta|G \backslash P|)},$

$|PG|$: Overlap pixels.

$|P \backslash G|$: Background pixels that are predicted as foreground, namely false positive.

$|G \backslash P|$: Foreground pixels that are predicted as background, namely false negative.

Our approach: coarse estimation stage

- The estimated shape tends to have a larger foreground area compared to GT.

- The imbalance between foreground and background pixels in the input image
- Tversky loss: $\mathcal{T}(P, G; \alpha, \beta) = \frac{|PG|}{(|PG| + \alpha|P \backslash G| + \beta|G \backslash P|)},$

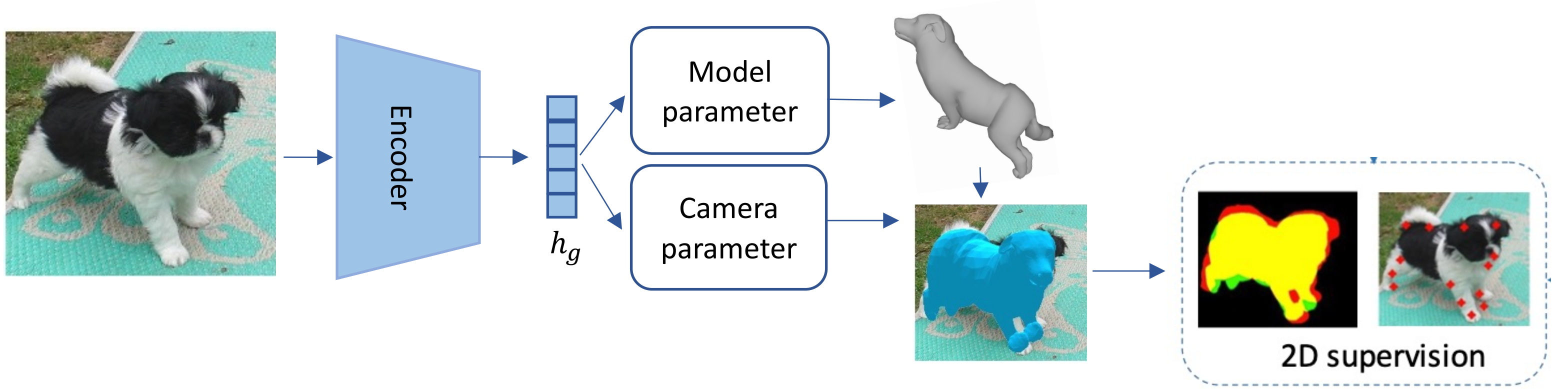
$|PG|$: True positive.

$|P \backslash G|$: False positive.

$|G \backslash \bar{P}|$: False negative.

$\alpha > \beta$ Penalize more on the false positive predictions.

Our approach: coarse estimation stage



2D supervision

Silhouette loss: $\mathcal{L}_{\text{silh1}} = 1 - \mathcal{T}(S, \mathcal{R}(V_c, f), \alpha, \beta)$

Keypoint loss: $\mathcal{L}_{\text{kp1}} = \| J_{2D} - \Pi(J_{3D}, f) \|^2$

Prior

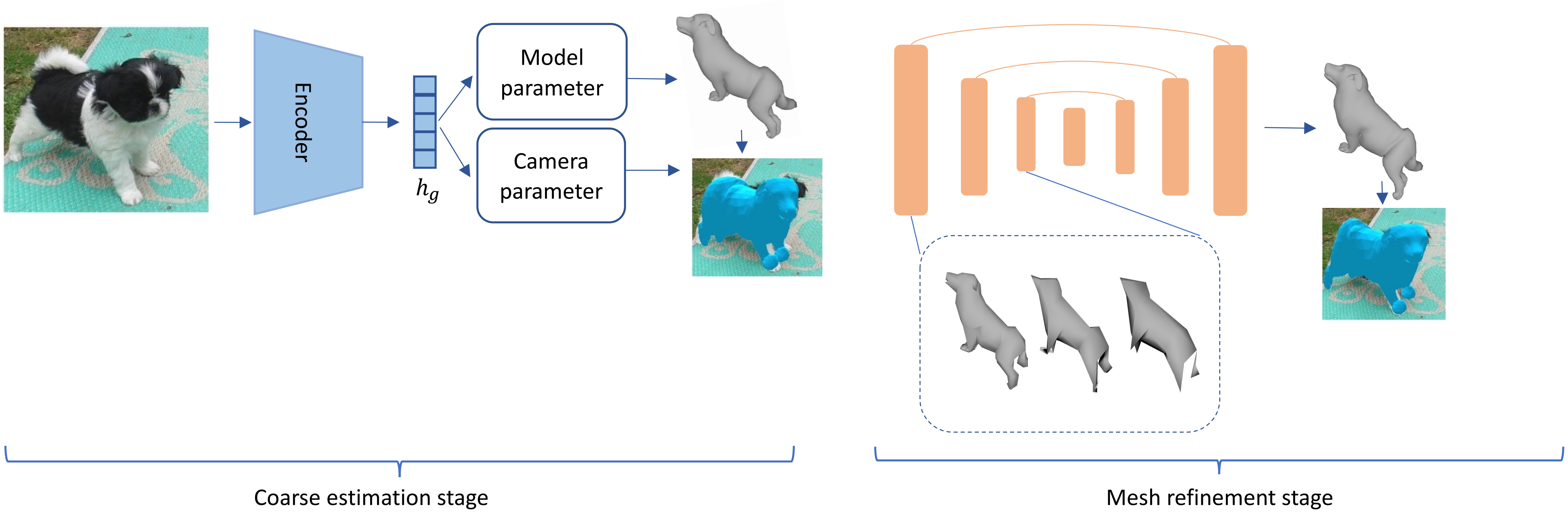
$$\mathcal{L}_\beta = (\beta' - \mu_\beta)^\top \Sigma_\beta^{-1} (\beta' - \mu_\beta)$$

$$\mathcal{L}_\theta = (\theta' - \mu_\theta)^\top \Sigma_\theta^{-1} (\theta' - \mu_\theta)$$

- Coarse stage loss function: $\mathcal{L}_{\text{st1}} = \lambda_{\text{kp1}} \mathcal{L}_{\text{kp1}} + \lambda_{\text{silh1}} \mathcal{L}_{\text{silh1}} + \lambda_\beta \mathcal{L}_\beta + \lambda_\theta \mathcal{L}_\theta$
- Pose limit constraint: enforce θ' to be in a valid range.

Our approach: Mesh refinement stage

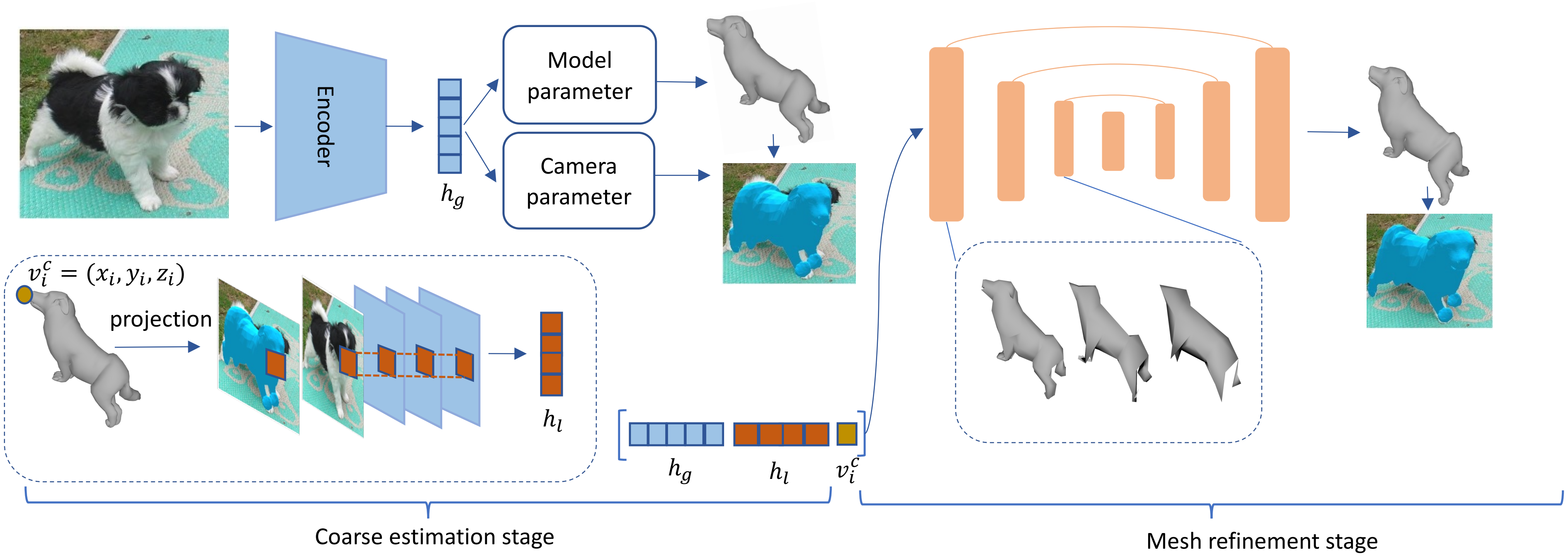
- Mesh refinement stage: Use the coarse output as an initial point, and further refine it with a MRGCN.



  - Encoder-decoder structure: Exploit features of different resolutions.
  - Skip connections: Preserve the spatial information at each resolution.

Our approach: Mesh refinement stage

- Mesh refinement stage: Refine the coarse shape with a MRGCN.



- Combination of global and local features: Capture detailed shape information.

Our approach: Mesh refinement stage

- Mesh refinement stage: Refine the coarse shape with a MRGCN.
  - Per-vertex deformation $\Delta v_i = \mathcal{F}(\mathbf{h}_i^0), \quad \text{where} \quad \mathbf{h}_i^0 = [\mathbf{h}_\mathrm{g}, \mathbf{h}_\mathrm{l}, x_i, y_i, z_i]$.

$$V_f = V_c + \Delta V, \quad \text{where} \quad \Delta V = [\Delta v_1, \Delta v_2, ..., \Delta v_C]$$

  - Laplacian regularizer to prevent large deformations:

$$\mathcal{L}_\mathrm{lap} = \sum_i \|\delta v_i^f - \delta v_i^c\|^2, \quad \text{where} \quad \delta v_i = v_i - \frac{1}{d_i} \sum_{j \in N(i)} v_j$$

  - Loss function for the refinement stage:

$$\mathcal{L}_\mathrm{st2} = \lambda_\mathrm{kp2} \mathcal{L}_\mathrm{kp2} + \lambda_\mathrm{silh2} \mathcal{L}_\mathrm{silh2} + \lambda_\mathrm{lap} \mathcal{L}_\mathrm{lap}$$

Our approach: Experiments

- Training details:
  - Train the coarse estimation part with $\mathcal{L}_{\mathrm{st1}}$ (without $\mathcal{L}_{\mathrm{silh1}}$) for 200 epochs.
  - Train the mesh refinement part with $\mathcal{L}_{\mathrm{kp2}}$ for 10 epochs.
  - Train the whole network with all losses for 200 epochs.

The silhouette loss can lead the network to unsatisfactory local minima if applied too early [3, 4].

[3] Benjamin Biggs, et al. Who left the dogs out?: 3D animal reconstruction with expectation maximization in the loop. ECCV 2020.

[4] Benjamin Biggs, et al. Creatures great and SMAL: Recovering the shape and motion of animals from video. ACCV 2018.

Our approach: Quantitative Results

- Evaluation matric: IOU for shape and PCK for pose.

- Results on the StanfordExtra dataset.

Method	Keypoints	Silhouette	IOU	PCK@0.15				
				Avg	Legs	Tail	Ears	Face
3D-M [32]	Pred	Pred	69.9	69.7	68.3	68.0	57.8	93.7
3D-M	GT	GT	71.0	75.6	74.2	**89.5**	60.7	**98.6**
3D-M	GT	Pred	70.7	75.5	74.1	88.1	60.2	98.7
3D-M	Pred	GT	70.5	70.3	69.0	69.4	58.5	94.0
CGAS [3]	CGAS	Pred	63.5	28.6	30.7	34.5	25.9	24.1
CGAS	CGAS	GT	64.2	28.2	30.1	33.4	26.3	24.5
WLDO [2]	-	-	74.2	78.8	76.4	63.9	78.1	92.1
Ours-coarse	-	-	72.5	77.0	75.9	55.3	76.1	89.8
Ours	-	-	**81.6**	**83.4**	**81.9**	63.7	**84.4**	94.4

Our approach: Quantitative Results

- Results on the Animal Pose dataset

Method	Keypoints	Silhouette	IOU	PCK@0.15				
				Avg	Legs	Tail	Ears	Face
3D-M [32]	Pred	Pred	64.9	59.2	55.7	56.9	61.3	86.7
WLDO [2]	-	-	67.5	67.6	60.4	**62.7**	86.0	86.7
Ours-coarse	-	-	67.5	62.0	57.1	45.1	75.8	78.9
Ours	-	-	**75.7**	**67.8**	**62.2**	45.1	**86.6**	**87.8**

- Results on the BADJA dataset

Method	IOU	PCK@0.15				
		Avg	Legs	Tail	Ears	Face
WLDO [2]	65.0	48.6	40.4	**78.2**	55.2	76.5
Ours-coarse	59.6	42.5	33.7	57.5	63.4	**79.2**
Ours	**72.0**	**54.1**	**47.6**	76.1	**66.2**	74.4

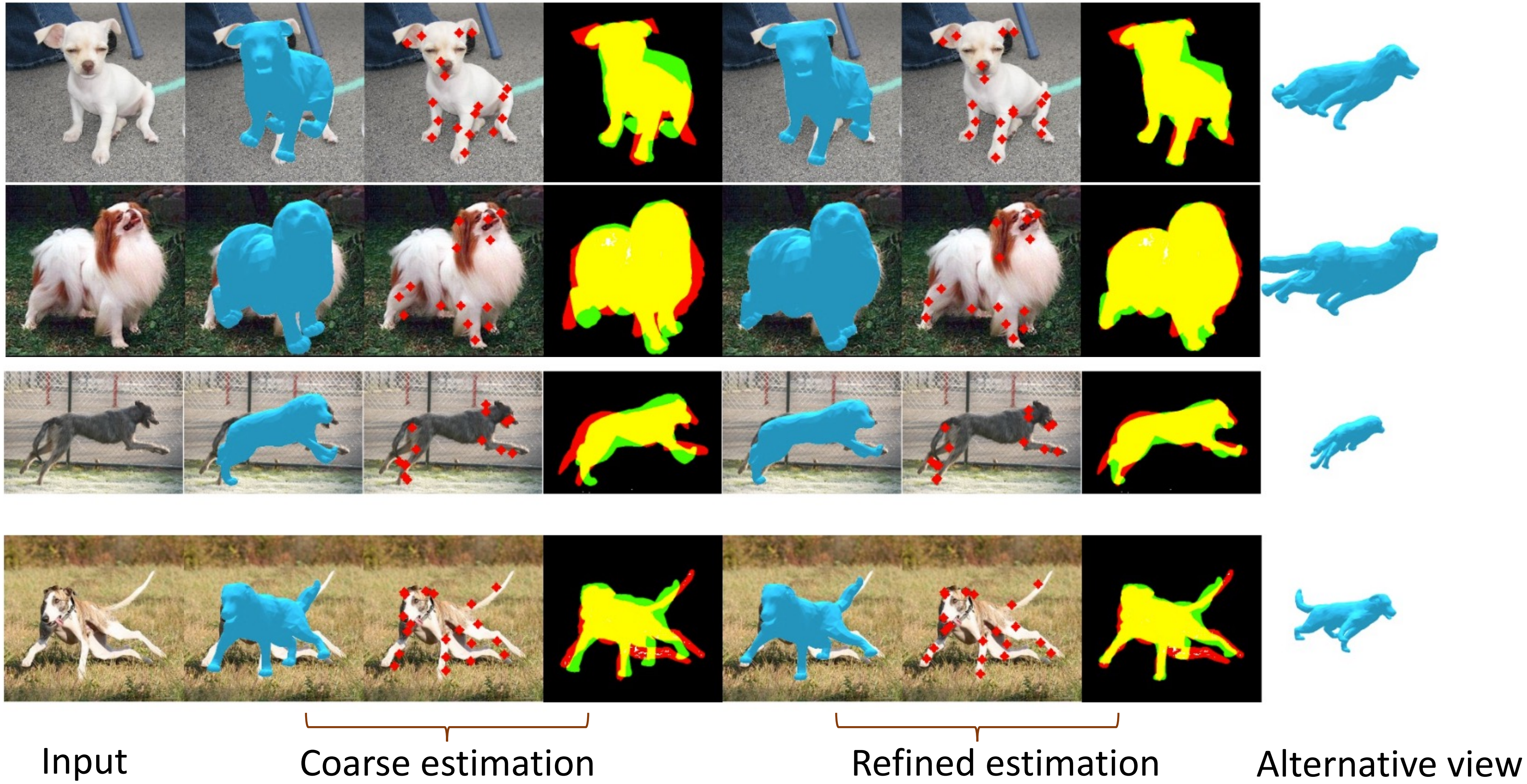
Our approach: Ablation studies

- Removing each component from the full model to evaluate the corresponding contribution.

Method	IOU	PCK@0.15				
		Avg	Legs	Tail	Ears	Face
Full	**81.6**	**83.4**	**81.9**	63.7	**84.4**	**94.4**
-MR	72.5	77.0	75.9	55.3	76.1	89.8
-LF	73.3	76.9	76.1	57.0	75.1	89.1
-ED	79.4	80.2	79.2	59.3	79.9	91.5
-TL	81.1	82.5	81.0	**65.2**	82.9	92.8

The performance drops when each component is removed from the full model.

Our approach: Qualitative results



Conclusion and future work

- We propose a coarse-to-fine approach, which combines SMAL-based and vertex-based representations.
- We design an encoder-decoder structured mesh refinement GCN, which combines image-level and vertex-level features to recover detailed shapes.

- Failure cases: camera looking at the back of the animal and extreme animal poses.

Thank you !