Supplementary Material
Canonical Surface Mapping via Geometric Cycle Consistency

Nilesh Kulkarni Abhinav Gupta* Shubham Tulsiani*
Carnegie Mellon University Facebook AI Research
{nileshk, abhinavg}@cs.cmu.edu shubhtuls@fb.com
https://nileshkulkarni.github.io/csm/

A. Training Details

A.1. Network Archicture

We use a 5 Layer UNet [4] with 4 $\times 4$ convolution at each layer. The UNet takes input an image and learns to predict an unit-vector which parameterizes u, v. Along with that we also train the UNet to predict a segmentation of the object which is necessary for keypoint evaluations. We train our networks for over 200 epochs on all the datasets independently. We use Adam [2] for optimization of our neural network with a learning rate of 10^{-4}.

A.2. Optimization

Pose Prediction We predict N (=8) possible hypothesis for pose given an image. We initialize the poses such that they span a wide spectrum during start of the training. We add an additional loss to encourage diversity and to ensure there is no mode collapse. The diversity loss consists of two terms:

- We add an entropy term over the probabilities of hypothesis c_i which prevents mode collapse, and encourages exploration. This is equivalent to minimizing $\sum_i^N c_i \log(c_i)$
- We maximize a pair-wise distance between predicted rotations, $\text{Dist}(r_i, r_j)$ for all the predicted hypothesis of an instance. This is equivalent to minimizing $\sum_{i=1}^N \sum_{j=1, j\neq i}^N \text{Dist}(r_i, r_j)$

B. Evaluation Metrics

Keypoint Transfer AP (APK). Any keypoint transfer method given two images as input helps us infer how keypoints transfer from a source image to a target image. The method give two outputs for every keypoint a) transferred keypoint location b) confidence score. A keypoint transfer is successful if the confidence score of the method for the transfer is high and the error for the transfer is less than $d = \alpha \times \max(h, w)$, where h, w represent height and width respectively. For any method we create several confidence thresholds compute the following metrics. Let us consider we have a lot of image-pairs where we have only N_{pair} keypoint correspondences between source and target. For any given confidence threshold t following are the two cases:-

1. True Positive (TP): The confidence for the correspondence was above t, and the transfer error is less than d.
2. False Positive (FP): The confidence for the correspondence was above t, but either the given keypoint does not exist on the target image, or our transfer error is more than d.

We compute transfer precision and transfer recall as follows

$$\text{Transfer Precision} = \frac{N_{TP}}{N_{TP} + N_{FP}}$$
$$\text{Transfer Recall} = \frac{N_{TP}}{N_{\text{pair}}}$$

Here, N_{TP} represents number of True Positives and N_{FP} represents number of False Positives. We create the plots for transfer precision vs transfer recall as shown in the Figure 7 in the main manuscript. Area under such a plot represents AP and we report performance on the same in Table 1 in the main manuscript.

C. Ablations

We investigate the importance of: a) the visibility loss (-vis), b) the use of foreground pixels in $L_{\text{consistency}}$ loss (-mask). We report our quantitative evaluations in Table 1. We observe that visibility constraint is important, and the ablations show a drop in average performance across both the metrics if this loss is excluded during training. Our CSM model is trained with $L_{\text{consistency}}$ loss only on foreground pixels, and the experiments denoted by (-mask) ablate this and do not use segmentation mask while computing

the losses. We observe that using cycle and visibility loss over all the pixels in the image does not significantly affect performance. Note that the mask supervision is still critical for the reprojection loss that helps resolve degenerate solutions as described earlier, and the predicted masks are also used for correspondence transfer as in Equation 6 in the main manuscript.

Method	Birds		Cars	
	PCK	APK	PCK	APK
CSM w/ pose	56.0	30.6	51.2	21.0
CSM w/ pose - vis	57.0	31.9	42.5	12.8
CSM w/ pose - mask	53.2	27.4	51.2	21.5
CSM	48.0	22.4	40.0	11.0
CSM - vis	43.1	18.3	33.0	7.1
CSM - mask	45.1	20.0	40.0	10.9

Table 1: Ablations. The settings with (-vis) indicate results if visibility loss is not enforced. The settings with (-mask) refer to enforcing $L_{\text{consistency}}$ loss on all pixels, and not just foreground ones, though the reprojection loss still leveraged mask supervision.

D. Results on Internet Videos

In the supplementary video we show results of our method on several videos. The color map on the video sequences shows correspondence to the template shape – shown at the top right of the frame. This helps us understand and visualize intra-frame correspondences. They also show the consistency of our predictions across frames. For instance, similar colors for the tails of two birds indicates that these pixels map to similar points on the template shape. We see few snapshots from the videos in the Fig 1. It is important to note that since we are using segmentation masks from pre-trained Mask-RCNN, the failure modes of Mask-RCNN become our failure modes. We observe that false-detections and failure to detect the instance in certain frames results in absence of CSM. Furthermore, since we only train using isolated untruncated and unoccluded objects, our predictions are often inaccurate if objects overlap or are not fully visible.

It is important to note that we do not apply any smoothing or consistency across frames. Our method operates on all the frames in the video independently.

E. Additional Result Visualization

We show additional results on all the categories in Figure 2, 3, 4, 5, 6, 7

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Minh Vo N Dinesh Reddy and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicle. In *CVPR*, 2018.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

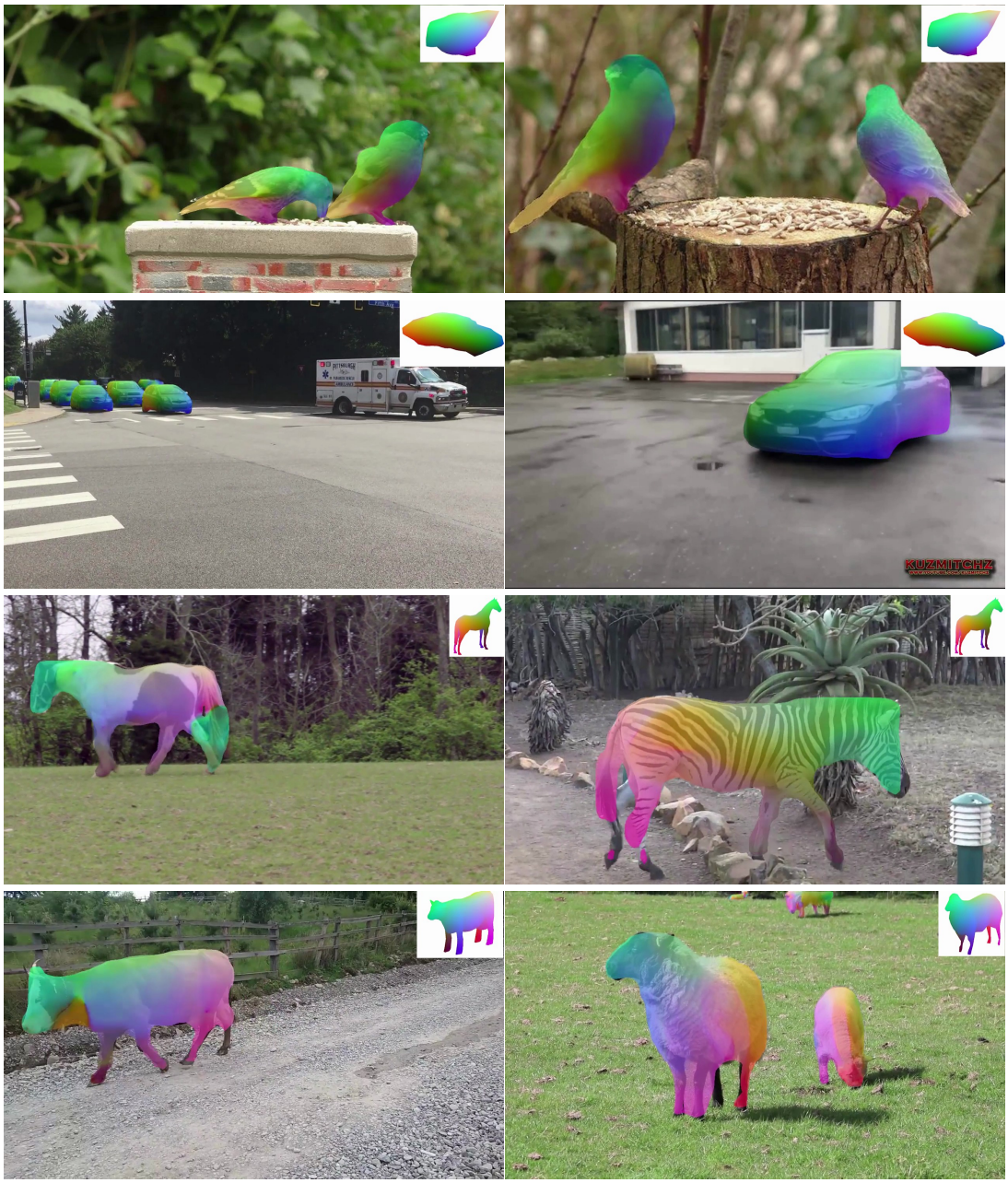

Figure 1: Snaps of a few frames from the Supplementary Video. We downloaded videos from youtube for 6 categories to show our results. We show the template shape in a canonical view on the top-right corner of the image. A few of the car videos in the qualitative results were taken from CarFusion dataset [3]

Failure Modes Our method has failure modes when the segmentation masks from Mask-RCNN [1] are incorrect. Furthmore, since our method is trained on images with a single unoccluded/untruncated object per image hence our predictions are might be inaccurate for occluded objects or partially visible objects.

Figure 2: Results of randomly sampled birds from the validation set

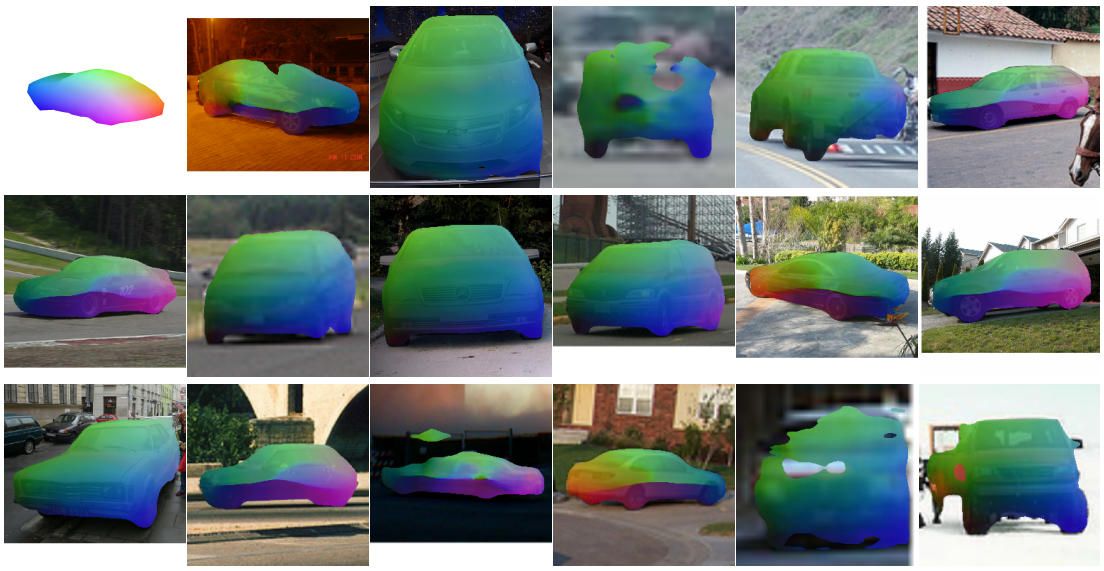

Figure 3: Results of randomly sampled cars from the validation set

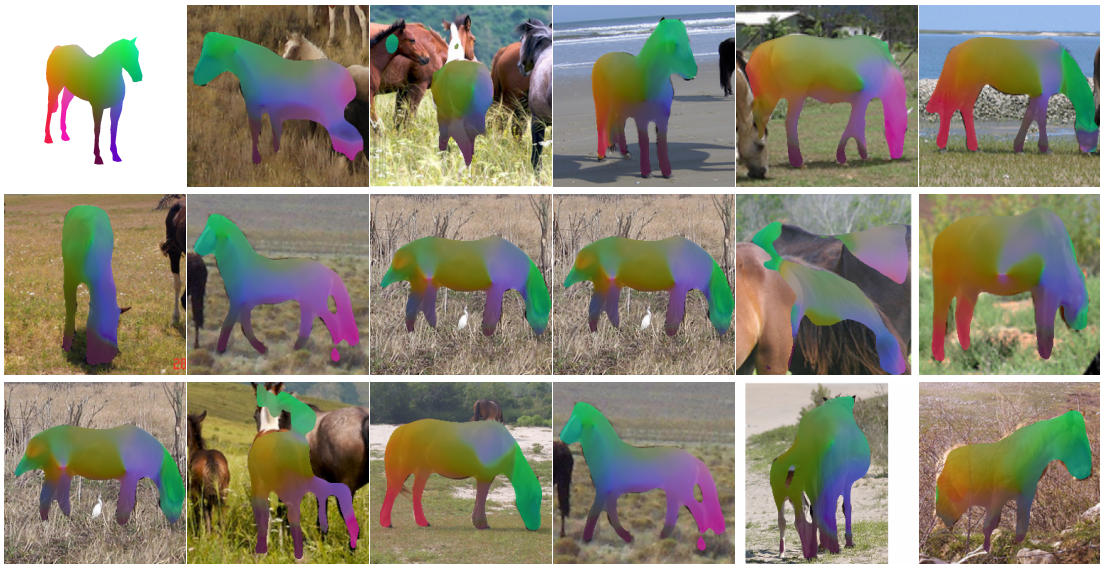

Figure 4: Results of randomly sampled horses from the validation set

Figure 5: Results of randomly sampled zebras from the validation set

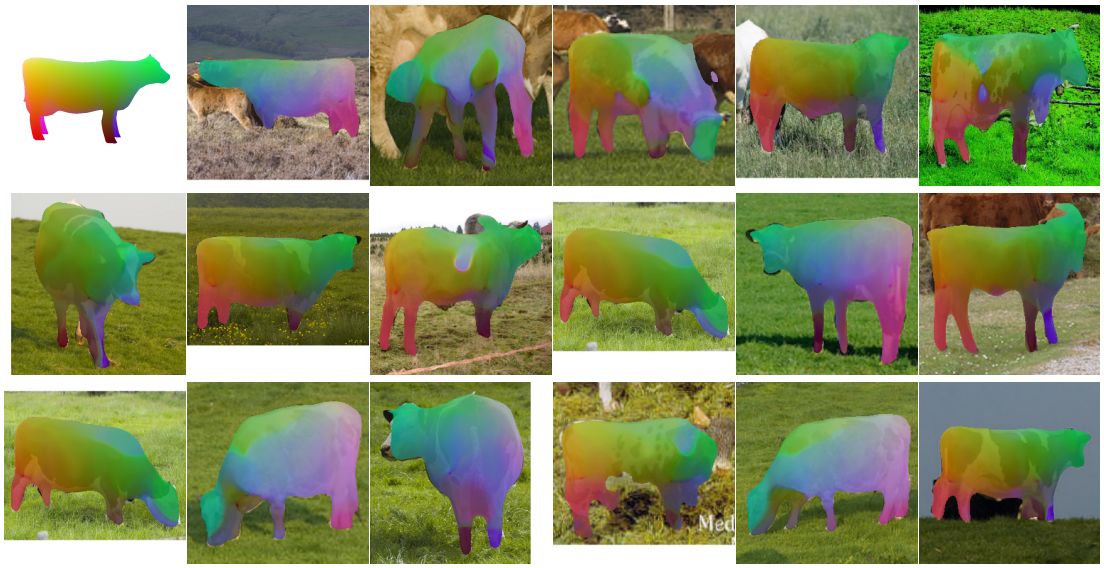

Figure 6: Results of randomly sampled cows for the validation set

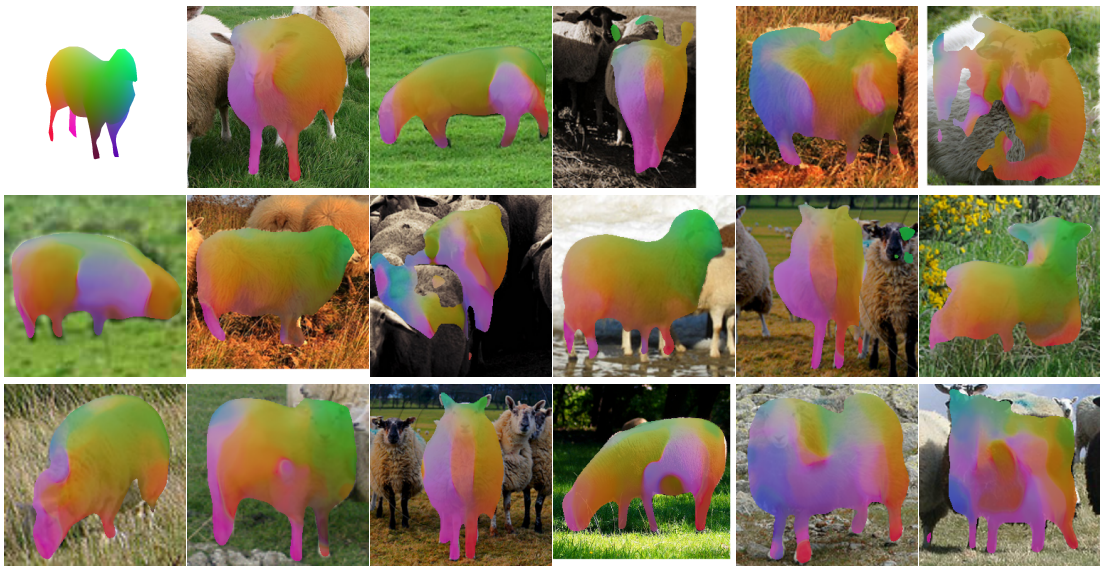

Figure 7: Results of randomly sampled sheeps from the validation set