Supplementary Material for
Structured Set Matching Networks for One-Shot Part Labeling

Jonghyun Choi*** Jayant Krishnamurthy*† Aniruddha Kembhavi* Ali Farhadi*‡

Allen Institute for Artificial Intelligence* Semantic Machines† University of Washington‡

jonghyunc@allenai.org jayant@semanticmachines.com anik@allenai.org ali@cs.uw.edu

1. Datasets

1.1. DiPART Dataset

1.1.1 Curation

We obtain line-drawing images from image search engines such as Google Image Search and Bing Image Search. We then obtain the list of part names for each object category by asking a set of annotators. We encourage the annotators to list a set of parts that are visually distinctive from one another and unique to that object category, if possible. We then choose the top ten parts. For each image and part, multiple annotators mark a point on the specified part, and the final point is chosen during a second round of annotations to choose the most representative point among the number of annotations.

1.1.2 Statistics

Figure 1 shows the number of images per category in the DiPART data set, sorted in descending order. Although the number of images per category is not uniform, it is reasonably well-balanced across categories.

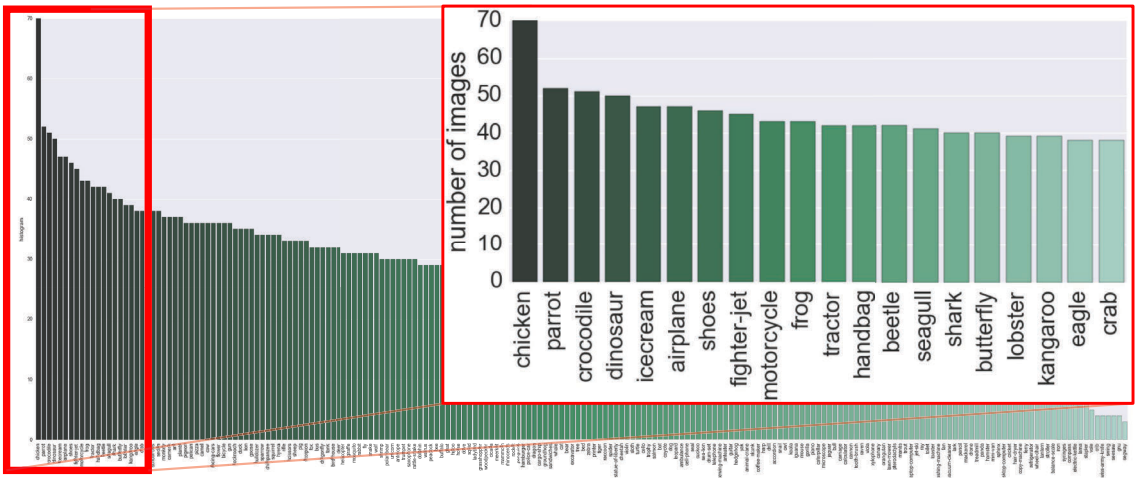

Figure 1. Number of images per category in the DiPART dataset (please zoom in to see category names. Inset shows the largest categories).

* indicates equal contribution. Majority of the work has been done while JK is in AI2.

1.1.3 Split

The dataset is split into train, validation and test set with 140, 30 and 30 categories, respectively. As per the one-shot setting, no object category is repeated among the three sets. This results in 3,507, 681 and 733 images for each set. The number of part labeling examples is much larger however, since all pairs of images within a category can be chosen as data points; 101,670 training, 21,262 validation, and 20,110 test examples (pairs).

Although the object categories do not overlap between the training, validation and testing sets, there are common part names overlapping among these sets, *e.g.*, animal categories typically have parts such as *leg* and *eye*. 51% of the part names in the validation set and 54% in the test set appear in the training set. As noted in the paper, our analysis shows that despite this overlap, learning part appearances provides very little benefit to the model. We conjecture that this is because part appearances do not always transfer across categories; *e.g.*, a *head* of an elephant and a giraffe look significantly different. The DiPART dataset will be publicly released.

Coherence and Incoherence of Part Locations. The location of a part is generally not coherent across the images of a category. Figures 2 and 3 show the spatial distributions of part locations for several categories. (These figures are similar to Figure 3-(a) in the main paper) The distribution of part locations is calculated using all of the images within the category, then overlaid on a single image for context. Figure 2 shows categories where parts can appear in many different locations. In contrast, Figure 3 shows categories where parts tend to appear in the same location, as indicated by the tight distributions. These figures demonstrate that part locations may be highly variable within a category, which suggests that simply using the spatial location of a part is insufficient to predict its label.

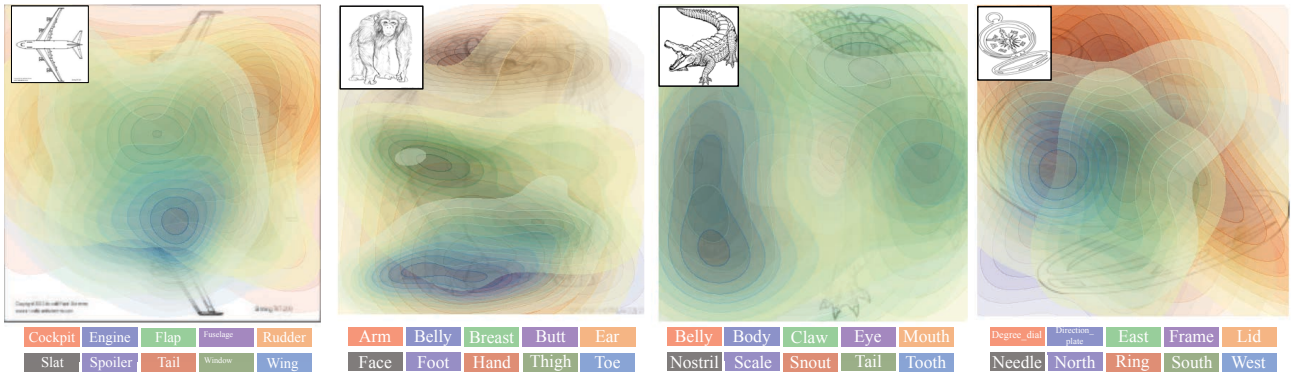

Figure 2. The spatial distribution of part locations for several categories with relatively incoherent distributions.

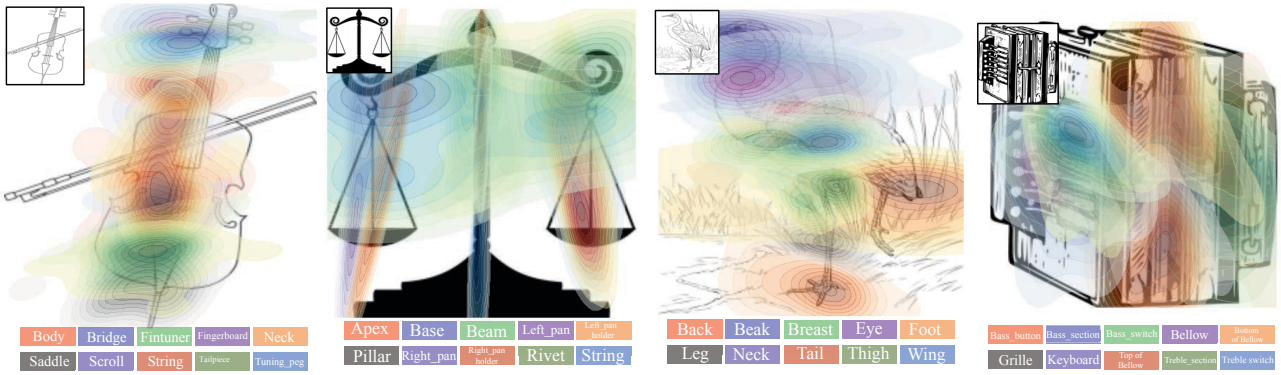

Figure 3. The spatial distribution of part locations for several categories with relatively coherent distributions.

Coherence and Incoherence of Part Appearances. Figures 4 and 5 show average image patches for several parts from the dataset. (These figures are similar to Figure 3-(b) in the main paper) Figure 4 shows parts whose appearances are relatively consistent across images, while Figure 5 shows parts whose appearances are diverse. Parts with consistent appearances are easier to match using visual appearance similarity. Parts with inconsistent appearances pose more challenges to the model, requiring other cues including global consistency constraints.

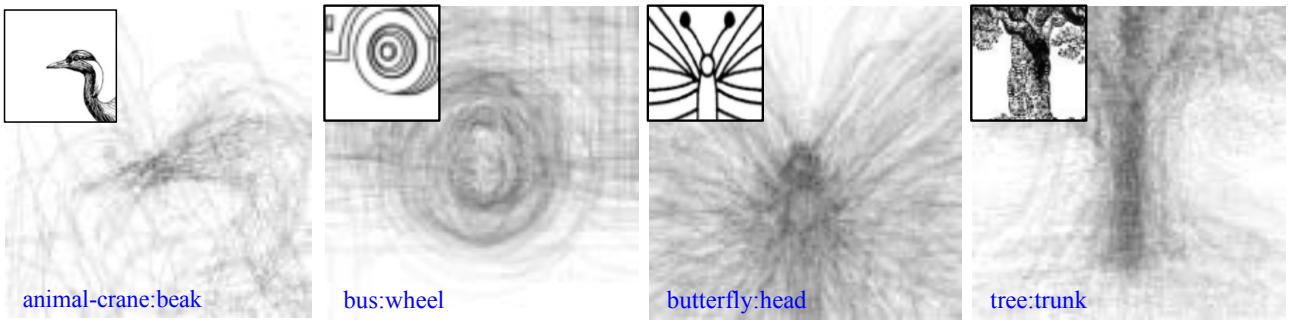

Figure 4. Average images of part patches for parts with relatively coherent appearances.

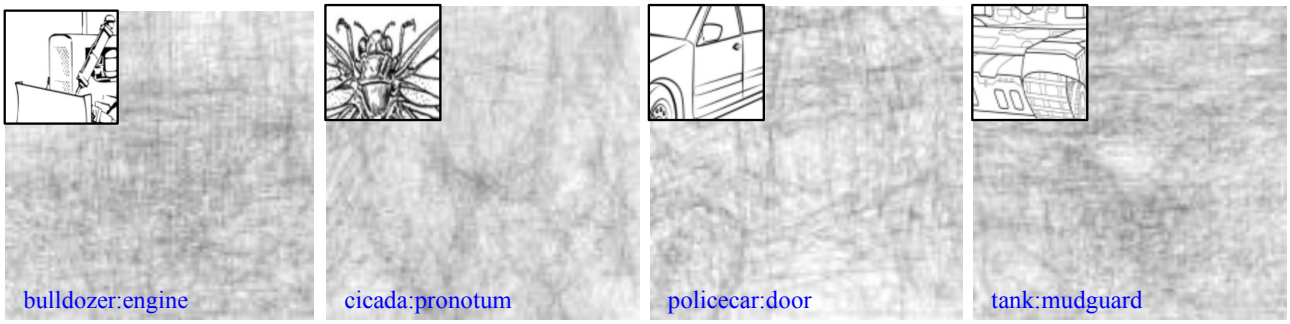

Figure 5. Average images of part patches for parts with incoherent appearances.

1.2. Pascal Part Matching (PPM) Dataset

The Pascal Part Matching (PPM) dataset is a subset of Pascal Part dataset [1]. The Pascal Part dataset consists of images taken from the PASCAL VOC 2010 dataset with segmentation masks for each part of the object. To adapt Pascal Part towards our task, we chose categories that have more than ten parts. Although the original Pascal Part dataset provides object bounding box and the segmentation mask of each part, for setup of the part labeling task, we remove the object bounding box and convert the segmentation mask into a point supervision by taking a center of mass of the mask as shown in Figure 6. The PPM dataset will also be publicly released.

The PPM dataset consists of 6,936 images with ten parts in six categories; 5,390 images (74,660 pairs) in four categories (*dog*, *person*, *sheep* and *cow*) in train set and 1,546 images (18,120 pairs) in two categories (*cat* and *horse*) in test set.

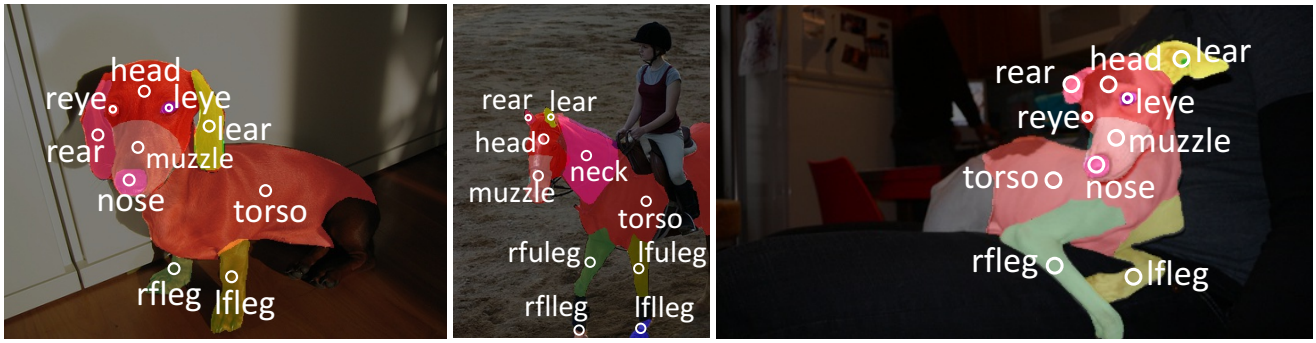

Figure 6. Visualization of the point annotation of Pascal-Part-Matching Dataset overlaid with segmentation mask of Pascal Part dataset.

1.3. Cross-DiPART-PPM Dataset

For the cross domain matching experiments, we find all overlapping categories and part names between DiPART and PPM to make Cross-DiPART-PPM. It consists of 22,969 image-to-diagram pairs in five categories with four parts (18,489 pairs of

four categories (*dog*, *cat*, *sheep*, *cow*) in train set and 4,480 pairs of one category, *horse*, in test set).

2. Additional Results on DiPART Dataset

2.1. Ablation Study

We perform an ablation study of our SSMN model and summarize in Table 1.

Dataset	SSMN	SSMN$-f_{gc}$	SSMN$-f_a$	SSMN$-f_p$
DiPART	58.1%	44.7%	57.3%	56.6%

Table 1. Ablation Study. '$-$' denotes 'without'

2.2. Effect of Global Consistency for Drastic Pose Change

To measure the effect of the global consistency term f_{gc} quantitatively, we randomly sampled 500 pairs in the test split of the DiPART dataset and annotated 1) drastic pose difference or 2) small pose difference and found: Drastic: 130 pairs, Small: 370 pairs. Figure 7 shows an example of matching of a pair in small and drastic pose difference by both SSMN and SSMN-f_{gc}.

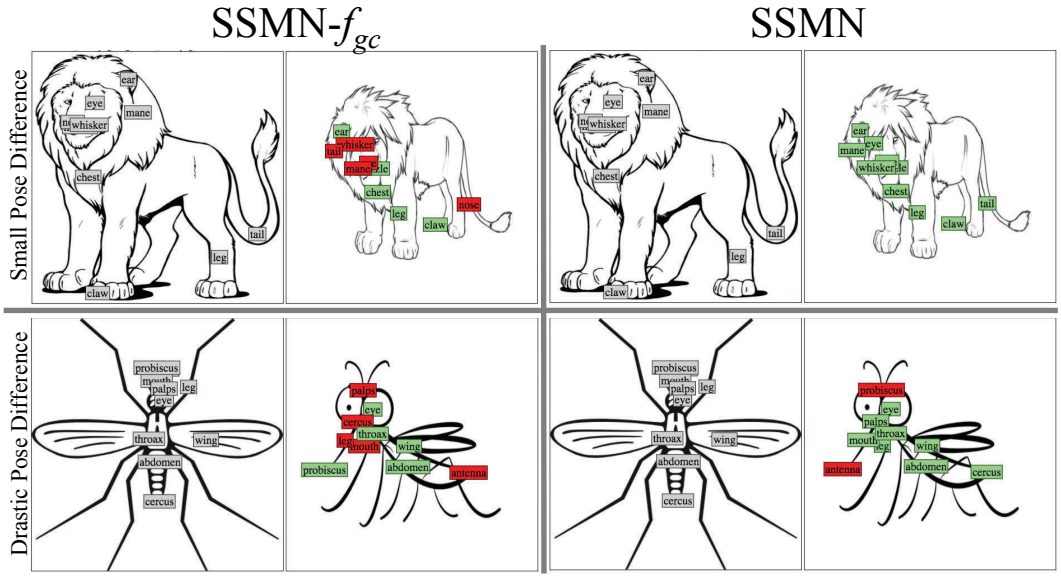

Figure 7. A Matching Example of Pairs in Small (Upper Row) and Drastic Pose Difference (Lower Row)

Table 2 shows the accuracy. Global consistency helps significantly in both cases, with a slight increase in relative performance for Small Pose difference (13.7% vs. 11.9%).

DiPART subset (500 pairs)	SSMN	SSMN$-f_{gc}$	Δ
All	59.0%	45.7%	13.3%
Small Pose Difference	60.6%	46.9%	13.7%
Drastic Pose Difference	54.2%	42.3%	11.9%

Table 2. Matching Accuracy on Different Level of Pose Difference

3. Additional Qualitative Results

3.1. Diagram to Diagram Part Labeling with DiPART dataset

Figures 8 and 9 show additional qualitative one-shot part labeling results of the SSMN on the test set of the DiPART dataset. Figure 8 shows examples where the SSMN successfully predicts the correct part labels, despite variations in pose and appearance between the source and target images.

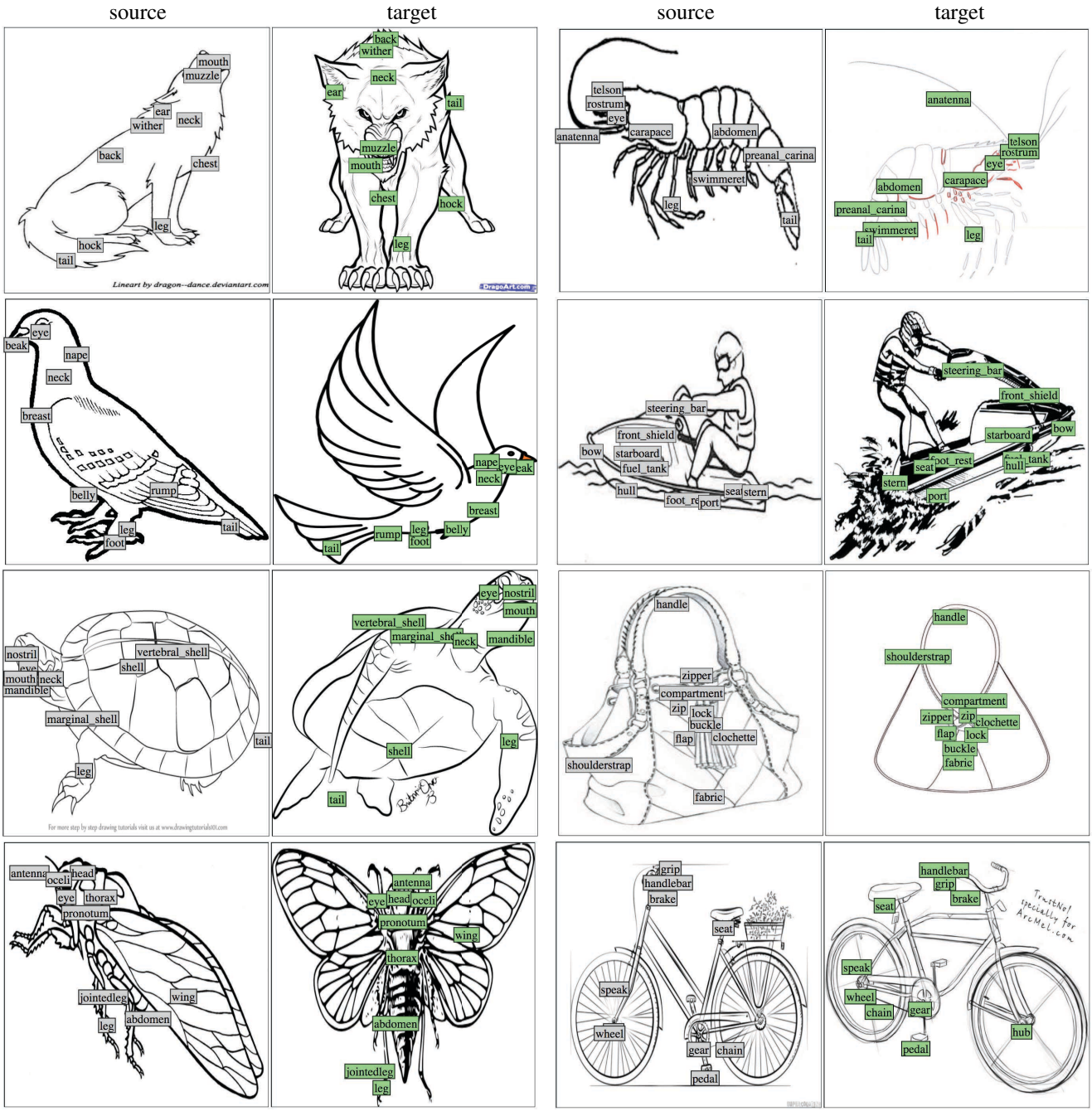

Figure 8. Test set examples where the SSMN predicts the correct labeling of the target image. A green box indicates a correctly predicted label and a red box indicates an incorrectly predicted one.

Figure 9 shows examples where the SSMN fails to predict the correct part labels. These failures occur for various reasons: a part's appearance may be drastically different in the source and the target (*e.g.*, the "sidemirror" of the *minivan*); the two parts may be nearly indistinguishable owing to their proximity and lack of texture (*e.g.*, the "arm" and "elbow" of the *koala*); or two parts may be located very close to each other (*e.g.*, the "battery release button" and "battery" of the *drill*).

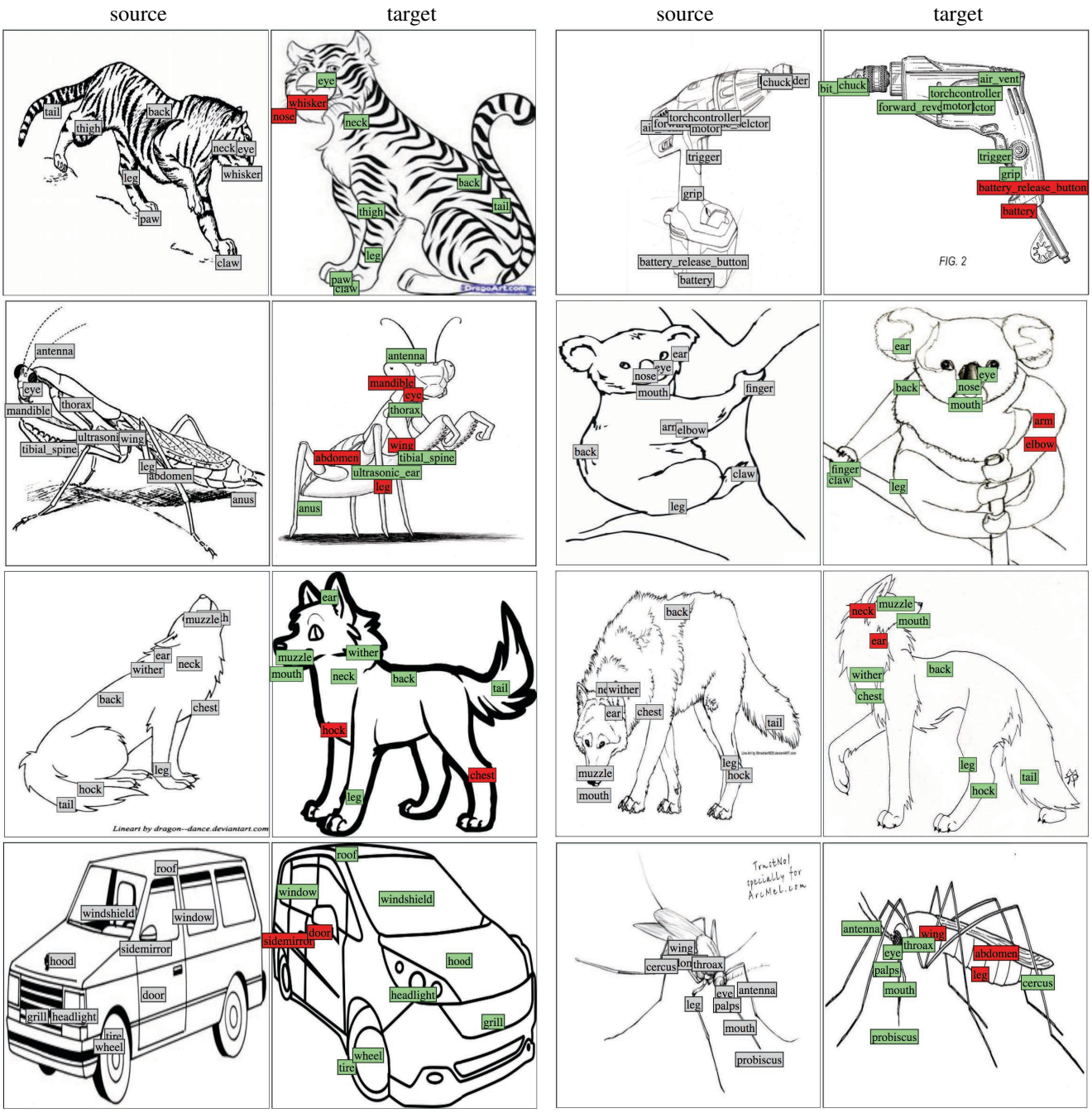

Figure 9. Test set examples where the SSMN predicts an incorrect labeling of the target image. A green box indicates a correctly predicted label and a red box indicates an incorrectly predicted one.

3.2. Image to Image Part Labeling with Pascal Part Matching (PPM) Dataset

Figures 10 and 11 show additional qualitative one-shot part labeling results of the SSMN on the test set of Pascal Part Matching (PPM) dataset. Figure 10 shows examples where the SSMN successfully predicts the correct part labels, despite variations in pose and appearance between the source and the target images.

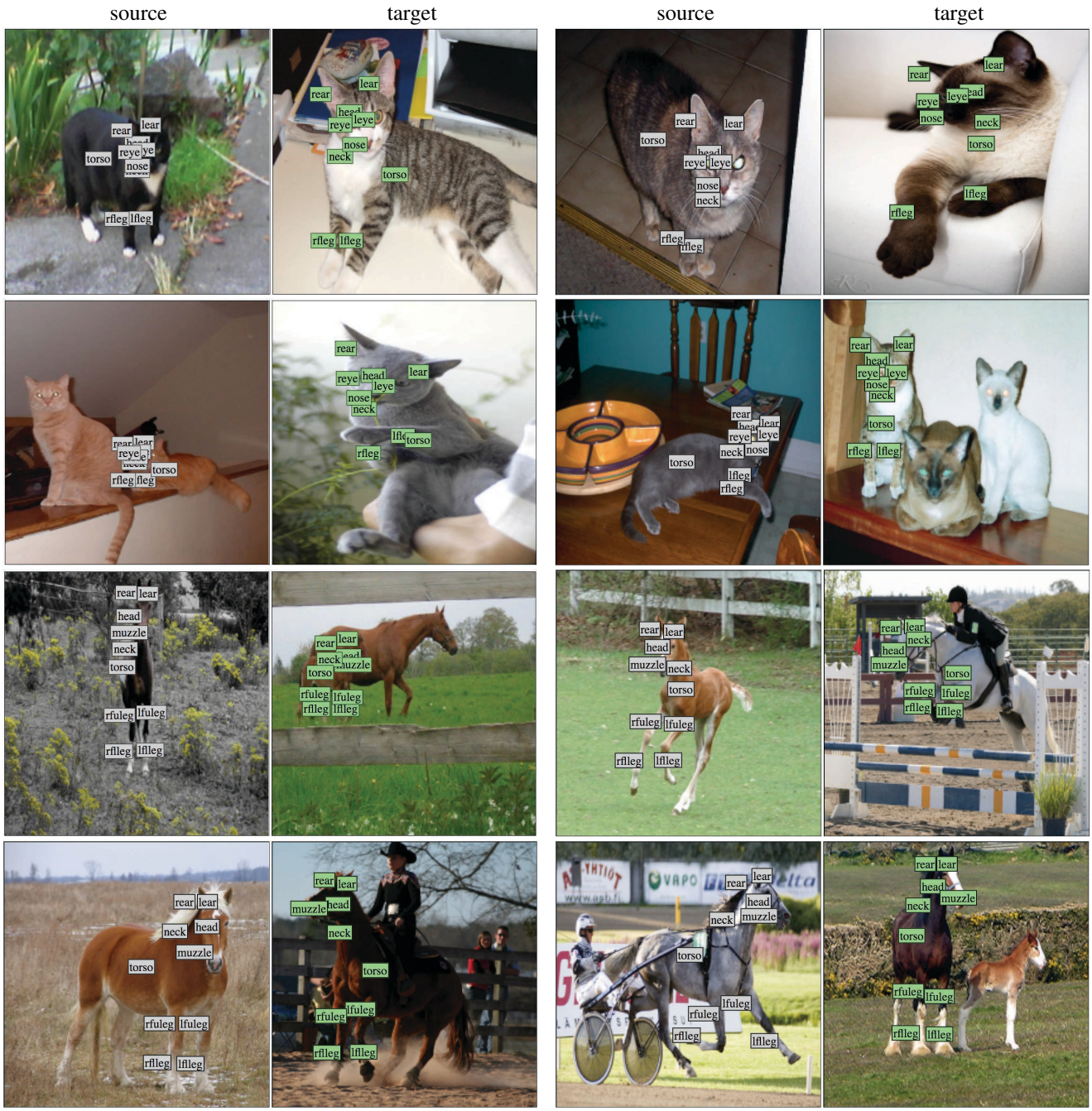

Figure 10. Test set examples where the SSMN predicts the correct labeling of the target image. A green box indicates a correctly predicted label and a red box indicates an incorrectly predicted one.

Figure 11 shows examples where the SSMN fails to predict the correct part labels. These failures occur for a number of reasons: significant changes in pose of the non-rigid object (*e.g.*, a cat lying on the floor with mingled legs and a cat standing up shown in the left pair of the second row); occlusion by other objects (*e.g.*, the standing horse and the horse in the arms of a person shown in the left pair of the fourth row); a very different part appearances between the source and the target (*e.g.*, the ears ("lear" and "rear") of the gray standing cat and the ears of the black lying cat shown in the right pair of the first row); or two parts have nearly the same appearance and are located very close to each other (*e.g.*, the "lfleg" (left front leg) and "rfleg" (right front leg) of the cat).

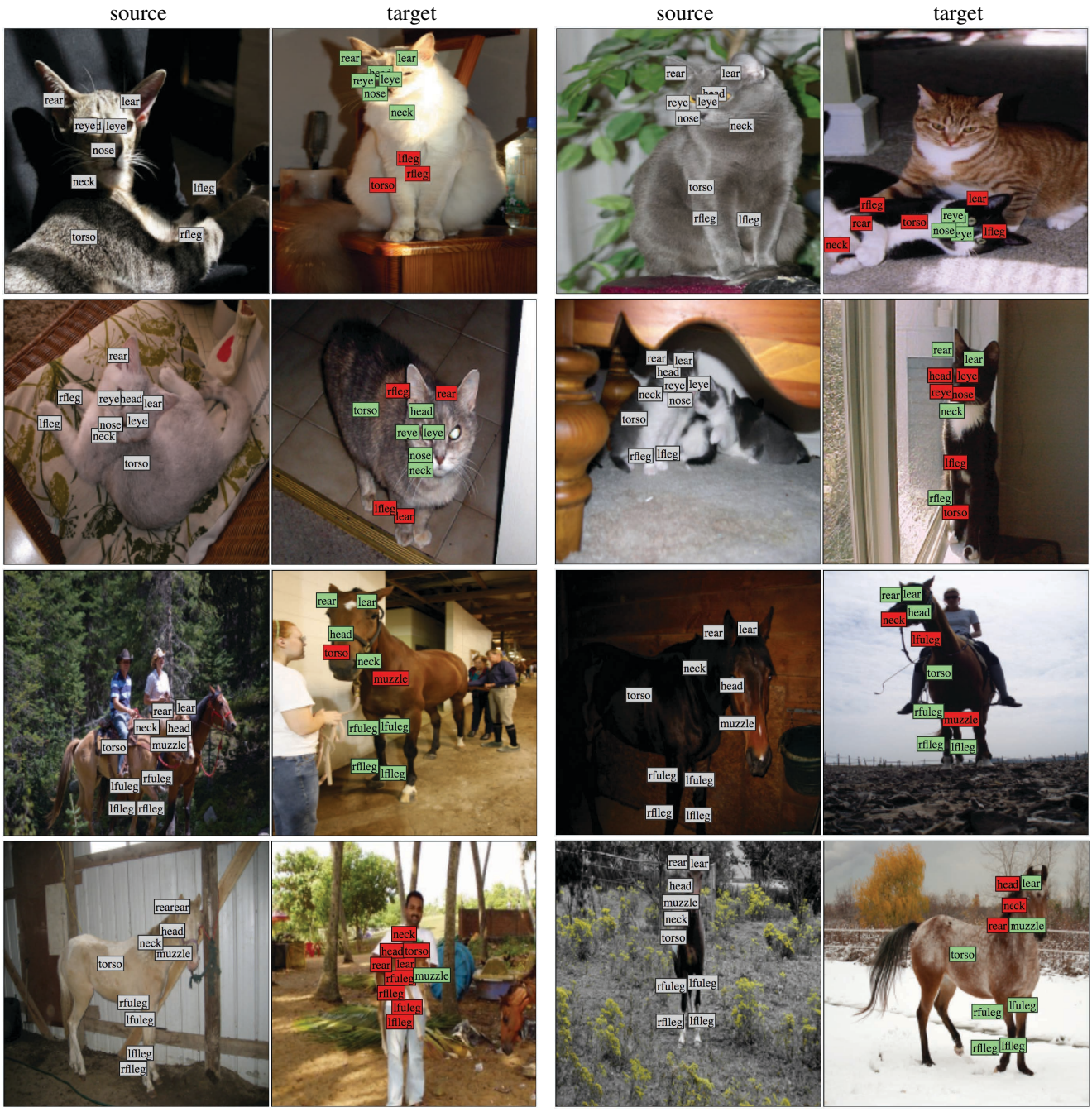

Figure 11. Test set examples where the SSMN predicts an incorrect labeling of the target image. A green box indicates a correctly predicted label and a red box indicates an incorrectly predicted one.

3.3. Image to Diagram Part Labeling with Cross-DiPART-PPM Dataset

Figures 12 and 13 show additional qualitative one-shot part labeling results of the SSMN on the test set of the Cross-DiPART-PPM dataset. Figure 12 shows examples where the SSMN successfully predicts the correct part labels, despite the drastic appearance variation across the domains (image to diagram) and pose variation between the source and the target images.

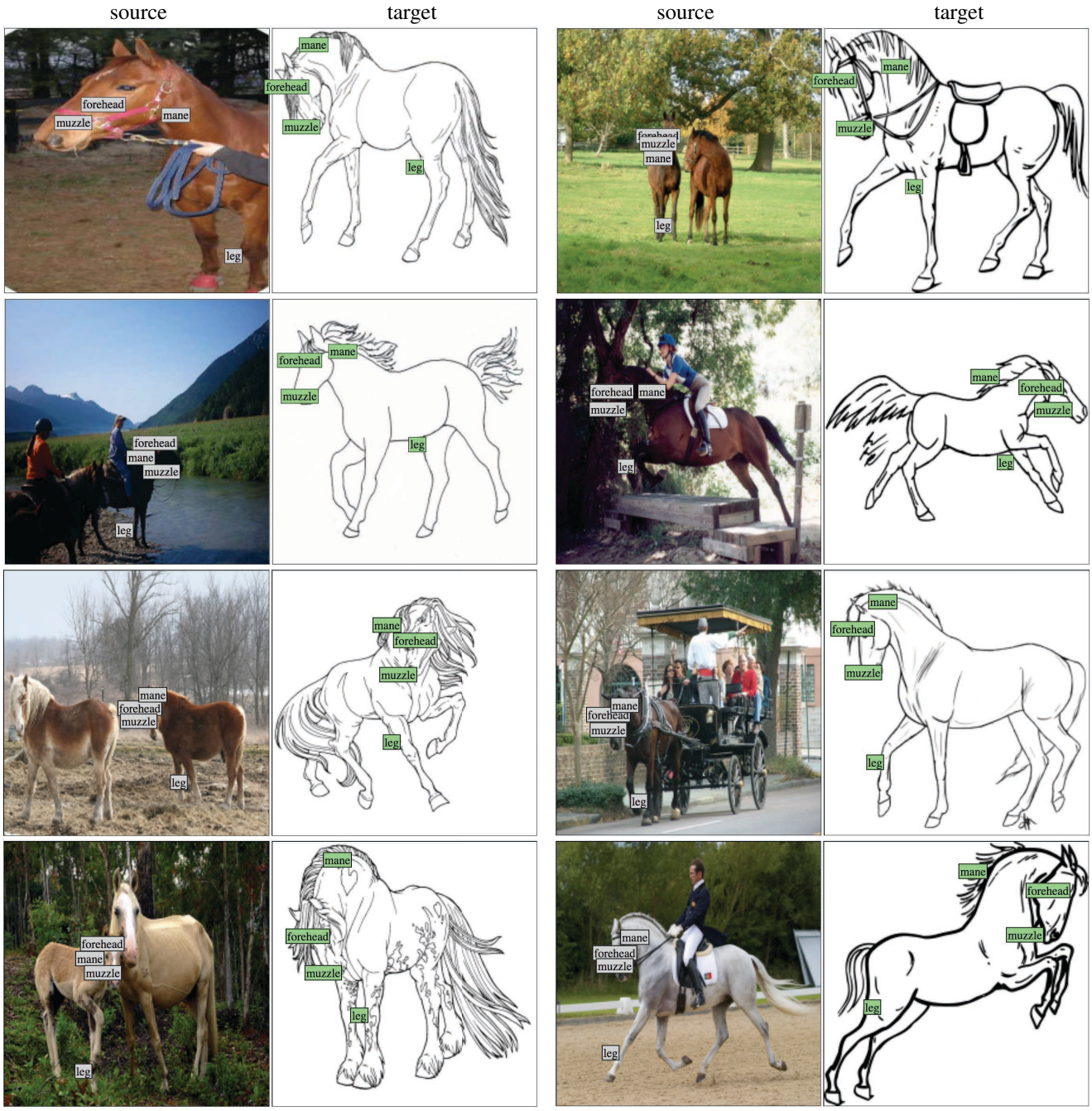

Figure 12. Test set examples where the SSMN predicts the correct labeling of the target image. A green box indicates a correctly predicted label and a red box indicates an incorrectly predicted one.

Figure 13 shows examples where the SSMN fails to predict the correct part labels. These failures occur for a set of reasons: two parts have nearly the same appearance and are located very close to each other (*e.g.*, the "forehead" and "mane" of a *horse*); confused with nearby part appearance (*e.g.*, "mane" is confused with tail nearby a "leg"); annotation difficulties (*e.g.*, "mane" of the source image in the example shown in the left pair of the second row is the position of a neck, which is very close to the location of the "muzzle" annotation in the target image).

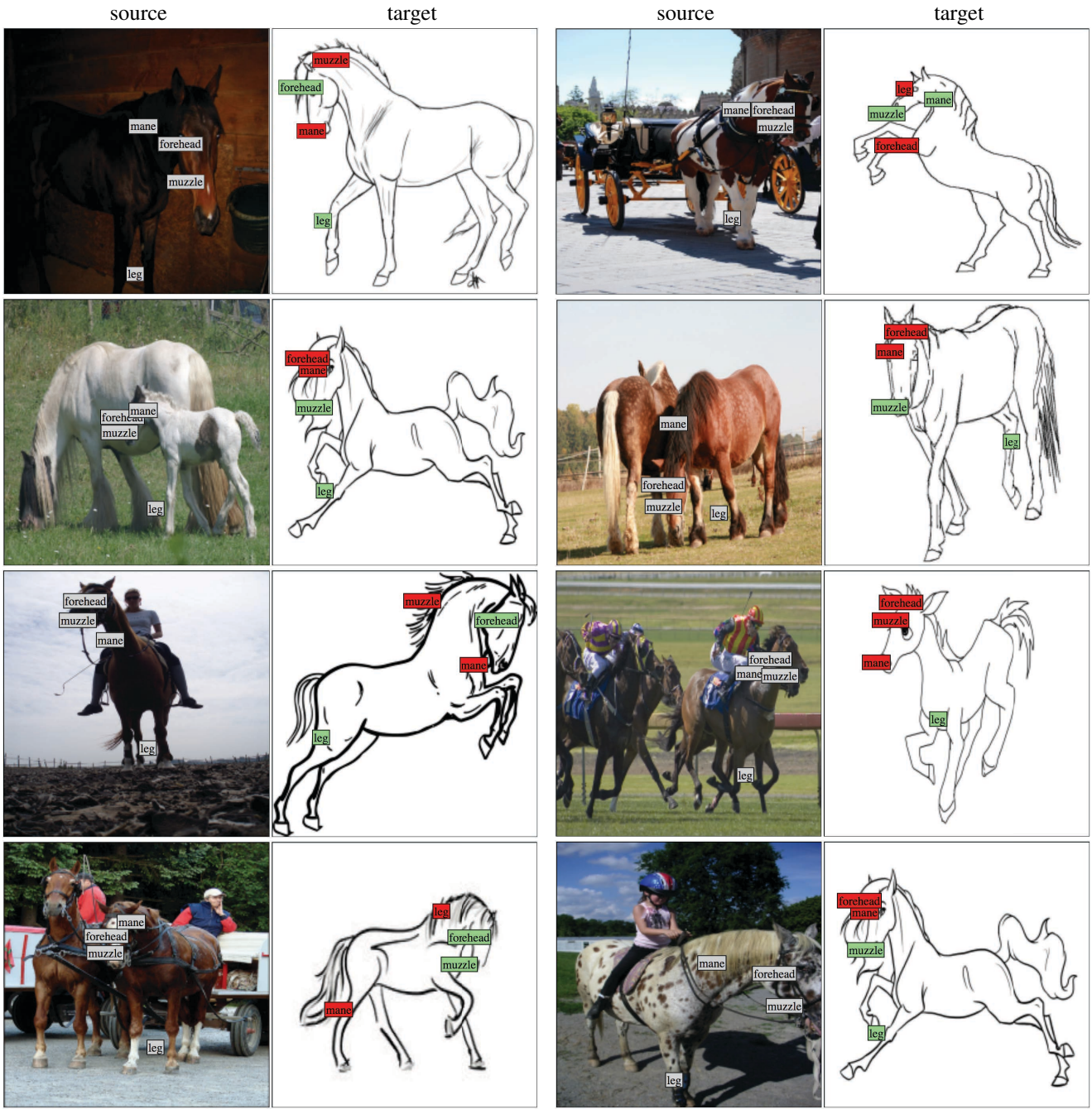

Figure 13. Test set examples where the SSMN predicts an incorrect labeling of the target image. A green box indicates a correctly predicted label and a red box indicates an incorrectly predicted one.

References

[1] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. In *CVPR*, 2014. 3