# Weakly Supervised Learning of Semantic Correspondence through Cascaded Online Correspondence Refinement
# ———*Supplementary Material*———

Yiwen Huang[1,*], Yixuan Sun[2,*], Chenghang Lai[1], Qing Xu[3],
Xiaomei Wang[3], Xuli Shen[1] and Weifeng Ge[1,†]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Academy of Engineering & Technology, Fudan University, Shanghai, China
[3]UniDT Technology, Shanghai, China
wfge@fudan.edu.cn

## 1. Details on One-to-one Greedy Selection

In section 3.1 of our main paper, we introduce pseudo matching pair selection module, where an algorithm named one-to-one greedy selection method is designed to generate final pseudo matchings. In this algorithm, we need to extract pixel-level corresponding relationship between pixel $\mathbf{P}_s^i$ and $\mathbf{P}_t^i$ based on neighborhood hough matching consistency scores $\mathbf{C} \in \mathbb{R}^{h_s \times w_s \times h_t \times w_t}$, in which $\mathbf{P}_s^i$ are token out from input source image $\mathbf{I}_s$ at pixel-level coordinate $(h_s^i, w_s^i)$, and similarly, $\mathbf{P}_t^i$ are token out from input target image $\mathbf{I}_t$ at pixel-level coordinate $(h_t^i, w_t^i)$, in addition, $h_s \times w_s, h_t \times w_t$ are resolution (counted by pixel) of input source image $\mathbf{I}_s$ and input target image $\mathbf{I}_t$, respectively.

Therefore, a natural idea occurs that we can simply take out pixel pair $(\mathbf{P}_s^i, \mathbf{P}_t^i)$ with the highest matching score from 4D matrix $\mathbf{C}$. Then pixel pairs with lower scores can be taken out accordingly. Besides, in order to remove background information, both pixels in the pixel pair $(\mathbf{P}_s^i, \mathbf{P}_t^i)$ are required to be located in the fore-ground region of corresponding saliency map $\mathbf{S}_s \in \mathbb{R}^{h_s \times w_s}$ or $\mathbf{S}_t \in \mathbb{R}^{h_t \times w_t}$. In summary, such procedures can be assembled as pseudo code in our Algorithm 1. In addition, since region-level corresponding relationship is expanded from its centric pixel pair, Algorithm 1 is appropriate for all stages in our pseudo-label generation framework.

## 2. Further Overview for SC-ImageNet

In section 3.5 of our main paper, we discuss the procedure of building SC-ImageNet, where some categories of ImageNet [2] are abandoned due to their inherent conflict with the single-instance semantic correspondence task. For example, some image categories are not object-centric like 'lakeside' and 'seacoast'. In addition, even if a category is object-centric, instance in images labeled as such category may also lack salient and unique feature points, like various kinds of balls. Besides, in some classes of ImageNet, images contain too complex clutters, where 'coral reef' and 'sea anemone' are typical cases. Furthermore, since our task is single-instance semantic correspondence, categories mainly composed of multiple instances should also be removed, like 'conch' and 'goldfish'. In summary, Figure 4 shows typical cases of abandoned categories. Moreover, we also provide more visualization of our automatically labeled SC-ImageNet in Figure 5 ∼ Figure 7.

**Algorithm 1** One-to-one Greedy Selection

```
Input: 4D Matching Score Matrix $\mathbf{C} \in \mathbb{R}^{h_s \times w_s \times h_t \times w_t}$,
Input: Saliency Map for Source Image $\mathbf{S}_s \in \mathbb{R}^{h_s \times w_s}$,
Input: Saliency Map for Target Image $\mathbf{S}_t \in \mathbb{R}^{h_t \times w_t}$,
Input: Saliency Threshold $T \in [0, 1]$.
Initialization: Empty pseudo-label set $\mathcal{M} = \{\}$,
Initialization: Source pixel set $\mathcal{P}_s = \{(h_s^1, w_s^1), ...\}$,
Initialization: Target pixel set $\mathcal{P}_t = \{(h_t^1, w_t^1), ...\}$.
Remove background pixels $(h_s^i, w_s^i)$ from $\mathcal{P}_s$, where $\mathbf{S}_s(h_s^i, w_s^i) < T$.
Remove background pixels $(h_t^i, w_t^i)$ from $\mathcal{P}_t$, where $\mathbf{S}_t(h_t^i, w_t^i) < T$.
repeat
    Find the corresponding pixel pair $(\mathbf{P}_s^i, \mathbf{P}_t^i)$ from $\mathcal{P}_s, \mathcal{P}_t$: $[(h_s^i, w_s^i), (h_t^i, w_t^i)]$, where $(h_s^i, w_s^i, h_t^i, w_t^i) = \texttt{argmax}_{\{(a,b,x,y)|(a,b)\in\mathcal{P}_s,(x,y)\in\mathcal{P}_t\}} \mathbf{C}(a, b, x, y)$
    Take out the pixel $\mathbf{P}_s^i$ from $\mathcal{P}_s$: $(h_s^i, w_s^i)$
    Take out the pixel $\mathbf{P}_t^i$ from $\mathcal{P}_t$: $(h_t^i, w_t^i)$
    Put corresponding pixel pair $(\mathbf{P}_s^i, \mathbf{P}_t^i)$ into $\mathcal{M}$.
until $|\mathcal{P}_s||\mathcal{P}_t| = 0$
Output: Pseudo-labels $\mathcal{M} = \{(\mathbf{P}_s^1, \mathbf{P}_t^1), (\mathbf{P}_s^2, \mathbf{P}_t^2), ...\}$.
```

[*] : Equal Contribution
[†] : Corresponding Author

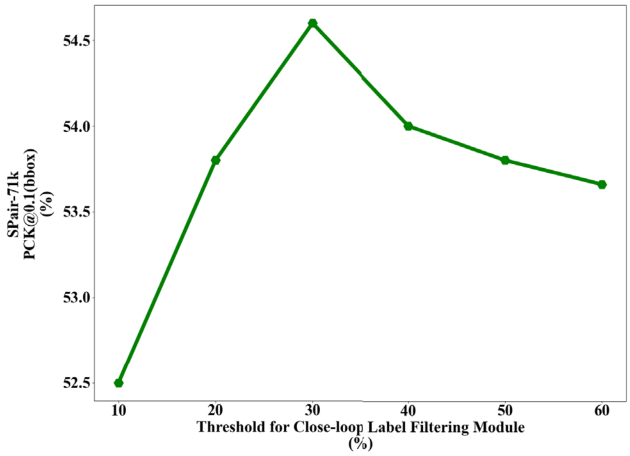

Figure 1. **Ablations on close-loop label filtering.** CATs [1] models are pre-trained with variously sized SC-ImageNet, where the number of training iterations is ***kept the same***.

## 3. Ablations on Close-loop Label Filtering

We investigate the best size of our SC-ImageNet as well as the most appropriate threshold for close-loop label filtering module (introduced in section 3.4 of our main paper) by conducting ablation experiments using single-instance images from ImageNet [2] with top-(10, 20, ..., 60)% consistency score as pre-training dataset. Then pre-trained model will be fine-tuned on SPair-71k [8] for evaluation. Note that we ensure all models are pre-trained with ***the same number of iterations*** by adjusting the number of pre-training epoch. Figure 1 shows that a limited number of highly consistent images are not enough for pre-training a precise semantic correspondence model, while too many inconsistent images are also harmful for pre-training. Finally, 113,516 highly consistent single-instance images with top-30% consistency score are selected to build our SC-ImageNet.

## 4. Comparison with Semi-supervised Methods

In section 4.3 of our main paper, we demonstrate the effectiveness of pre-training with our SC-ImageNet by conducting experiments with state-of-the-art fully-supervised semantic correspondence algorithms such as CATs [1] and TransforMatcher [7]. Nevertheless, our pre-training strategy with additional automatically labeled images could be seen as a semi-supervised approach, which is not sufficiently discussed in our main paper. As a consequence, in this section, our pre-training setting is compared with SOTA semi-supervised semantic correspondence methods, including SemiMatch [6] as well as SCorrSAN [5]. Both algorithms propose a semi-supervised strategy for semantic correspondence with additional key-point pairs, instead of our additional image pairs, provided by disagreement of views from different data augmentation settings and teacher-student model, respectively. Therefore, following [10], SemiMatch and SCorrSAN can be considered as semi-supervised approaches based on pseudo-labels, which is also appropriate for our pre-training setting.

| Method | PF-P. | | PF-W. | | Spair-71k |
|---|---|---|---|---|---|
| | 0.10 | 0.15 | 0.10 | 0.15 | 0.10 |
| CATs[1] | 92.6 | 96.4 | 79.2 | 90.3 | 49.9 |
| CATs[1]+SemiMatch[6] | **93.5** | **96.6** | 82.1 | 92.1 | 50.7 |
| CATs[1]+Ours | 92.9 | **96.6** | **84.0** | **94.5** | **54.6** |

Table 1. **Quantitative evaluation of CATs [1] and CATs based semi-supervised approaches on PF-PASCAL (PF-P.), PF-Willow (PF-W.) [3] and SPair-71k [8].** The best results in bold. +Ours demonstrates the model of CATs is firstly pre-trained with our SC-ImageNet, then fine-tuned on target dataset. Note that the input resolution of original algorithms is not modified.

| Method | PF-P. | | PF-W. | | Spair-71k |
|---|---|---|---|---|---|
| | 0.10 | 0.15 | 0.10 | 0.15 | 0.10 |
| SCorrSAN[5] | 93.3 | 96.6 | 80.0 | 89.8 | 55.3 |
| SCorrSAN[5]+Ours | **93.4** | **96.8** | **82.3** | **91.6** | **58.9** |

Table 2. **Quantitative evaluation of SCorrSAN [5] and SCorrSAN based semi-supervised approaches on PF-PASCAL (PF-P.), PF-Willow (PF-W.) [3] and SPair-71k [8].** The best results in bold. +Ours demonstrates the model of SCorrSAN is firstly pre-trained with our SC-ImageNet, then fine-tuned on target dataset. Note that the input resolution of original algorithms is not modified.

**Comparison between Additional Image Pairs and Additional Key-point Pairs.** For fair comparison, CATs [1] is selected as our baseline, where SemiMatch [6] and our pre-training strategy with SC-ImageNet are independently implemented. As shown in Table 1, for the most challenging SPair-71k with large-scale variation, our approach outperforms SemiMatch by 3.9% PCK@0.1. It demonstrates the effectiveness of pre-training on our SC-ImageNet containing a large number of additional image pairs. To show the effectiveness and robustness of our framework in detail, we compare per-class accuracy in Table 3 and our approach outperforms original CATs and SemiMatch on 13 of the 18 classes. Our qualitative results are shown in Figure 2. Furthermore, pre-training on our SC-ImageNet can also significantly improve the generalization power of semantic correspondence algorithms, which is proven through the best performance in PF-WILLOW for all PCKs by 1.9% and 2.4%, respectively compared to the same baseline model, CATs, trained with additional key-point pairs given by SemiMatch. However, our method do not perform well on PF-PASCAL at PCK@0.10, due to sparse key-point annotation in image pairs from PF-PASCAL and our automatically labeled SC-ImageNet, which results in less information on neighboring key-point [6]. Finally, we can draw the conclusion that semi-supervised semantic correspondence method with additional image pairs are better than approaches with additional key-point pairs in generalization ability, such as deal-

| Method | aero. | bike | bird | boat | bott. | bus | car | cat | chai. | cow | dog | hors. | mbik. | pers. | plan. | shee. | trai. | tv | all |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CATs[1] | 52.0 | 34.7 | 72.2 | 34.3 | **49.9** | 57.5 | 43.6 | 66.5 | **24.4** | 63.2 | 56.5 | 52.0 | 42.6 | **41.7** | 43.0 | 33.6 | 72.6 | 58.0 | 49.9 |
| CATs[1]+SemiMatch[6] | 53.6 | 37.0 | **74.6** | 32.3 | 47.5 | 57.7 | 42.4 | 67.4 | 23.7 | 64.2 | 57.3 | 51.7 | 43.8 | 40.4 | **45.3** | 33.1 | 74.1 | 65.9 | 50.7 |
| CATs[1]+Ours | **58.8** | **44.6** | 71.7 | **41.0** | 49.2 | **71.2** | **46.4** | **73.0** | 23.3 | **69.8** | **58.3** | **59.8** | **55.8** | 39.2 | 33.0 | **41.6** | **75.8** | **73.9** | **54.5** |

Table 3. **Per-class quantitative evaluation of CATs [1] and CATs based semi-supervised approaches on SPair-71k dataset. [8]** The best results are in bold. +Ours demonstrates the model of CATs is firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Note that the input resolution of original algorithms is not modified.

| Method | aero. | bike | bird | boat | bott. | bus | car | cat | chai. | cow | dog | hors. | mbik. | pers. | plan. | shee. | trai. | tv | all |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SCorrSAN[5] | 57.1 | 40.3 | **78.3** | 38.1 | **51.8** | 57.8 | 47.1 | 67.9 | 25.2 | 71.3 | 63.9 | 49.3 | 45.3 | 49.8 | **48.8** | 40.3 | 77.7 | 69.7 | 55.3 |
| SCorrSAN[5]+Ours | **64.5** | **47.5** | 78.0 | **39.9** | 49.2 | **65.1** | **49.0** | **74.0** | **29.4** | **75.4** | **64.3** | **60.9** | **54.8** | **52.0** | 48.1 | **47.4** | **87.0** | **75.4** | **58.9** |

Table 4. **Per-class quantitative evaluation of SCorrSAN [5] and SCorrSAN based semi-supervised approaches on SPair-71k dataset. [8]** The best results are in bold. +Ours demonstrates the model of SCorrSAN is firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Note that the input resolution of original algorithms is not modified.

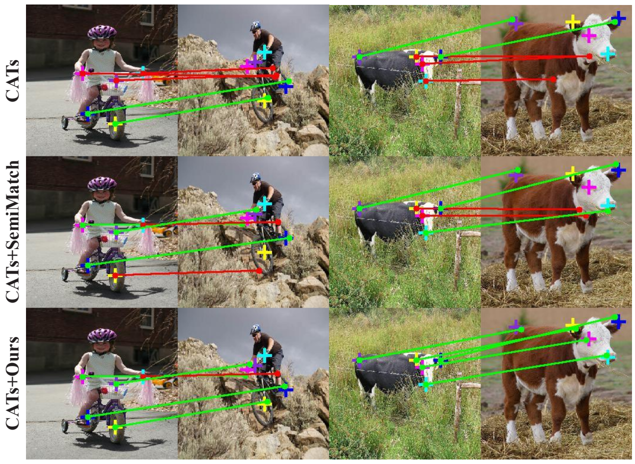

Figure 2. **Qualitative results of CATs [1] and CATs based semi-supervised approaches on SPair-71k [8].** SemiMatch is proposed in [6]. +Ours demonstrates the model of CATs is firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Colored cross means corresponding key-point (ground-truth). Green dots and lines mean correct matches, while red dots and lines mean incorrect matches (measured by PCK@0.1 as in Table 1).

ing with matching scenes with obvious variation, while approaches with additional key-point pairs show more expertise in strict matching criterion.

**Cooperation between Additional Image Pairs and Additional Key-point Pairs.** In this part, SCorrSAN [5], an originally semi-supervised semantic correspondence method, is our baseline. As shown in Table 2, for PF-PASCAL, SCorrSAN pre-trained with our SC-ImageNet gets slightly better performance. At the same time, for SPair-71k, our approach gets 3.6% higher PCK@0.1. Both results indicate that cooperation between additional image pairs and additional key-point pairs is effective in improving semantic correspondence methods. To show the effectiveness and robustness of such cooperation in detail, we compare per-class accuracy in Table 4 and our approach provides more accurate corresponding key-point prediction on 15 of the 18 classes. Our qualitative results are shown in Figure 3. In addition, based on the experimental results of PF-WILLOW, generalization power of SCorrSAN is also improved by pre-training with our SC-ImageNet, for getting better performance than SCorrSAN with only additional key-point pairs on all PCKs by 2.3% and 1.8%, respectively. As a consequence, for semi-supervised semantic correspondence algorithms, cooperation between additional image pairs and additional key-point pairs is beneficial for their effectiveness, robustness and generalization ability. In other words, additional image pairs are compatible with additional key-point pairs, which has shown promising matching power.

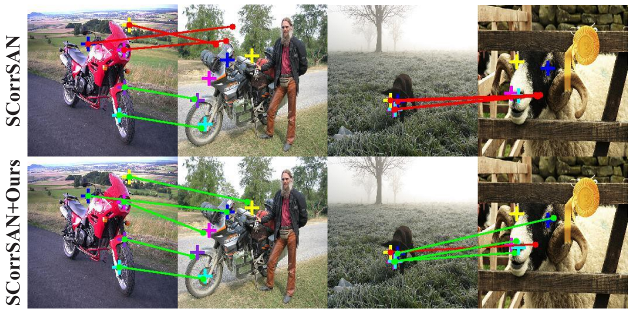

Figure 3. **Qualitative results of SCorrSAN [5] and SCorrSAN based semi-supervised approaches on SPair-71k [8].** +Ours demonstrates the model of SCorrSAN is firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Colored cross indicates corresponding key-point (ground-truth). Green dots and lines mean correct matches, while red dots and lines mean incorrect matches (measured by PCK@0.1 as in Table 2).

## 5. Additional Qualitative Results

As shown in Figure 8 $\sim$ Figure 10, we qualitatively compare state-of-the-art fully-supervised semantic correspondence models, including CATs [1], TransforMatcher [7], DHPF [9] as well as VAT [4] , with pre-training on our SC-ImageNet and accordant models with original training settings. On the basis of our qualitative results, models pre-trained on SC-ImageNet are seen to establish more accurate correspondences in the challenging SPair-71k [8], which contains obvious intraclass variations, scale difference, occlusion and truncation.

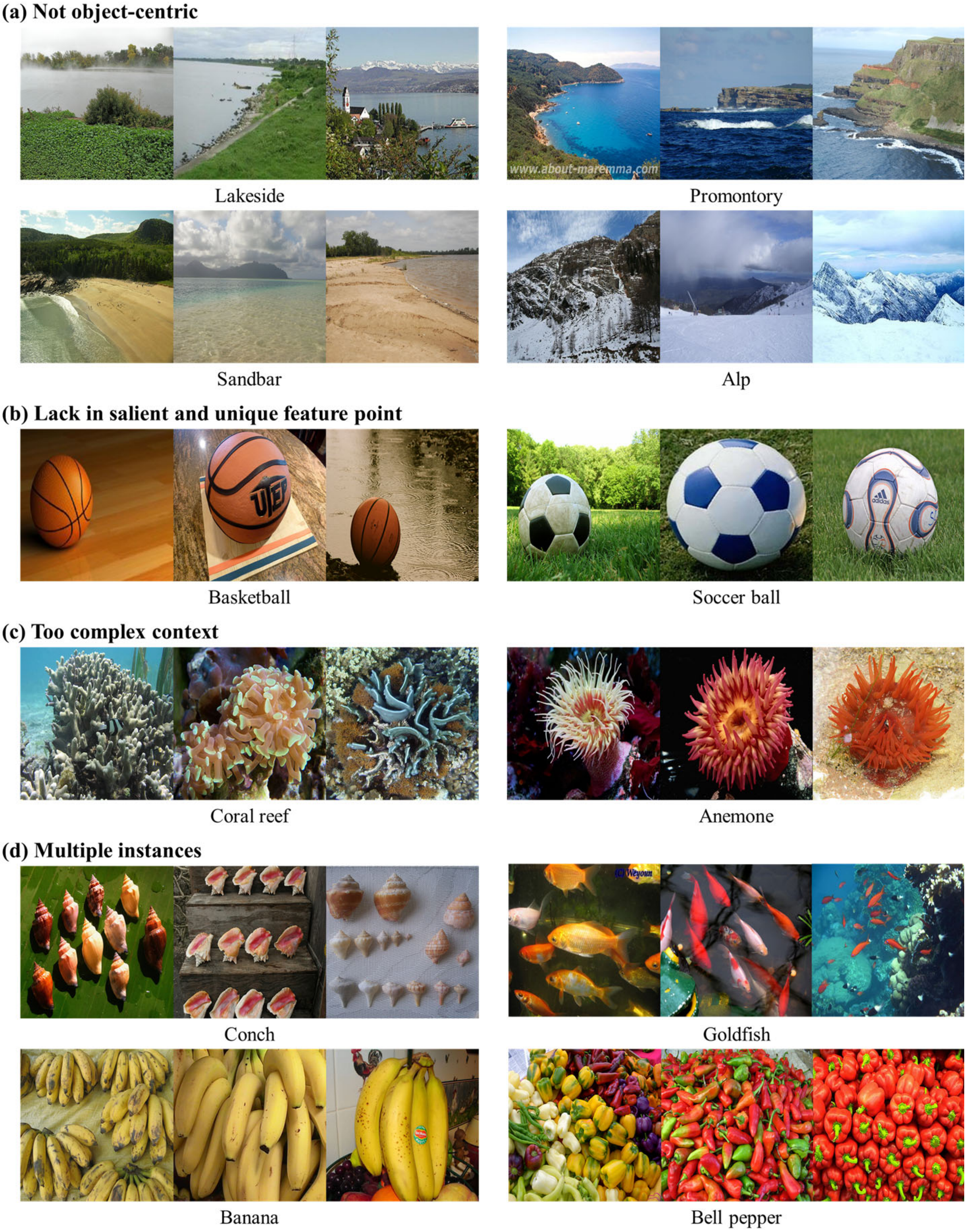

Figure 4. **Typical cases of abandoned categories.** All images are selected from ImageNet [2]. **(a) Not object-centric**: categories are not object-centric. **(b) Lack in salient and unique feature point**: there are not enough salient and unique feature points in instance. **(c) Too complex context**: images contain too complex clutters. **(d) Multiple instances**: categories are mainly composed of multiple instance images, which is conflict with single-instance semantic correspondence task.

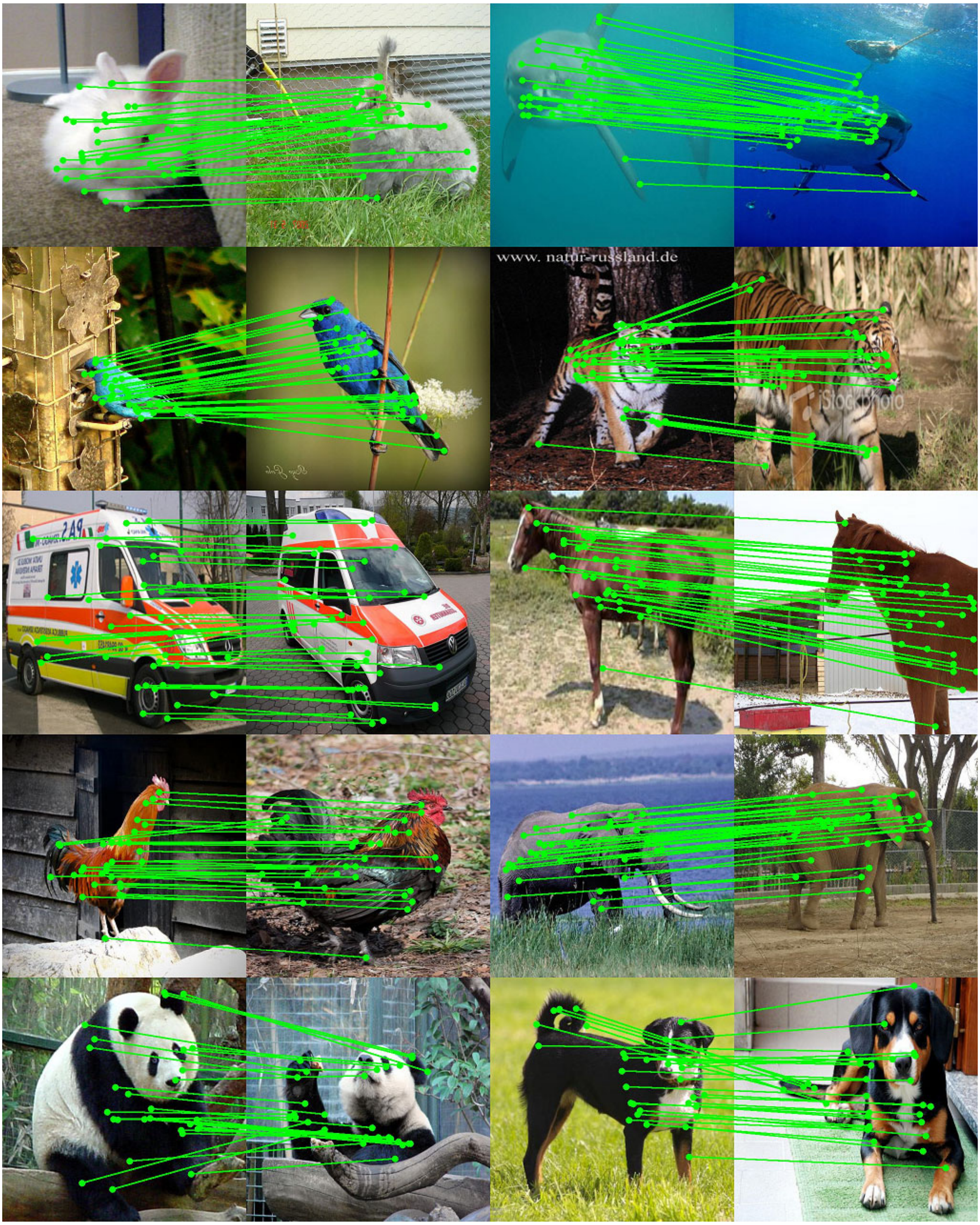

Figure 5. **Visualization of our SC-ImageNet (Part 1 of 3).** All images are selected from ImageNet [2]. Green dots and lines are our automatically generated pseudo-labels.

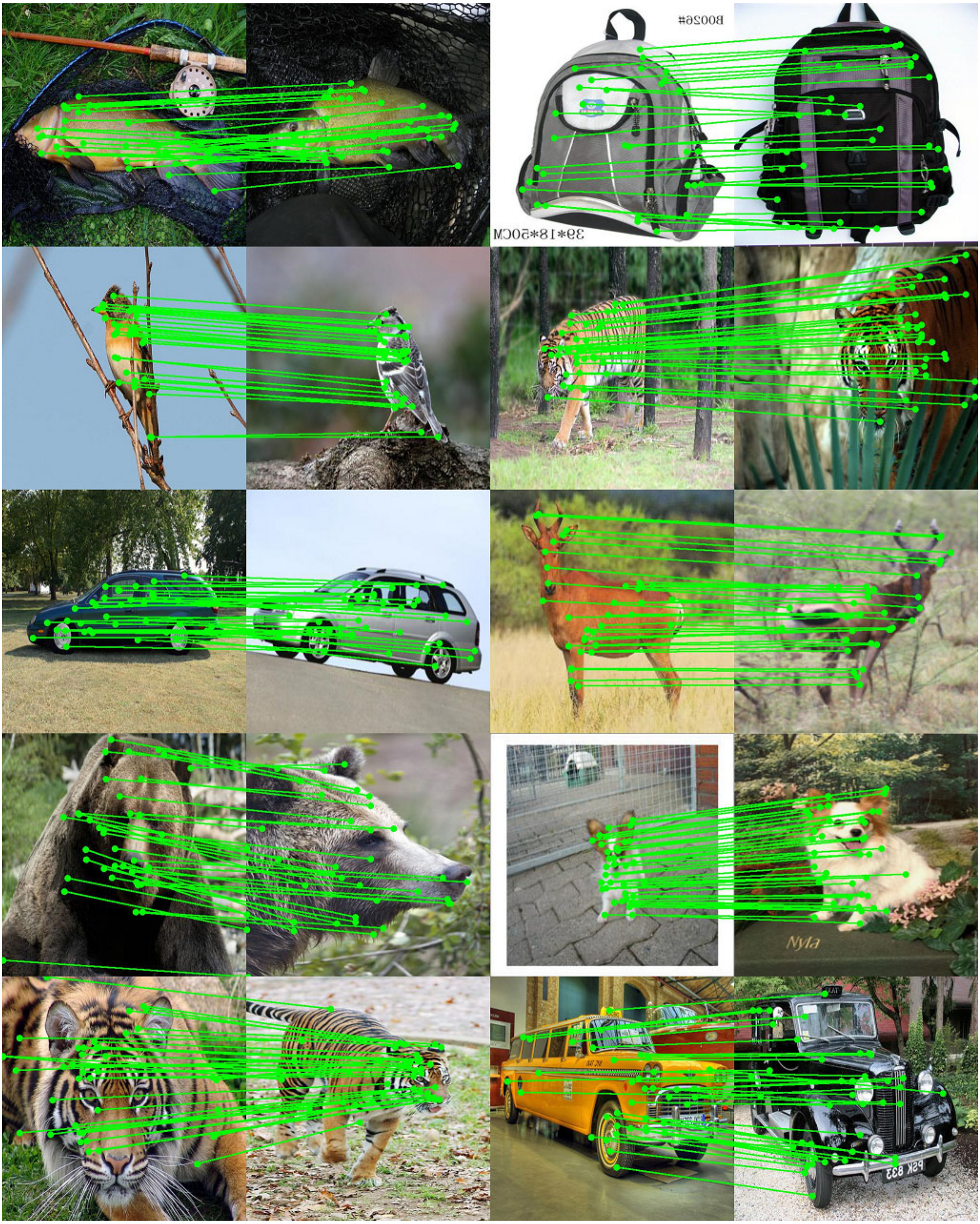

Figure 6. **Visualization of our SC-ImageNet (Part 2 of 3).** All images are selected from ImageNet [2]. Green dots and lines are our automatically generated pseudo-labels.

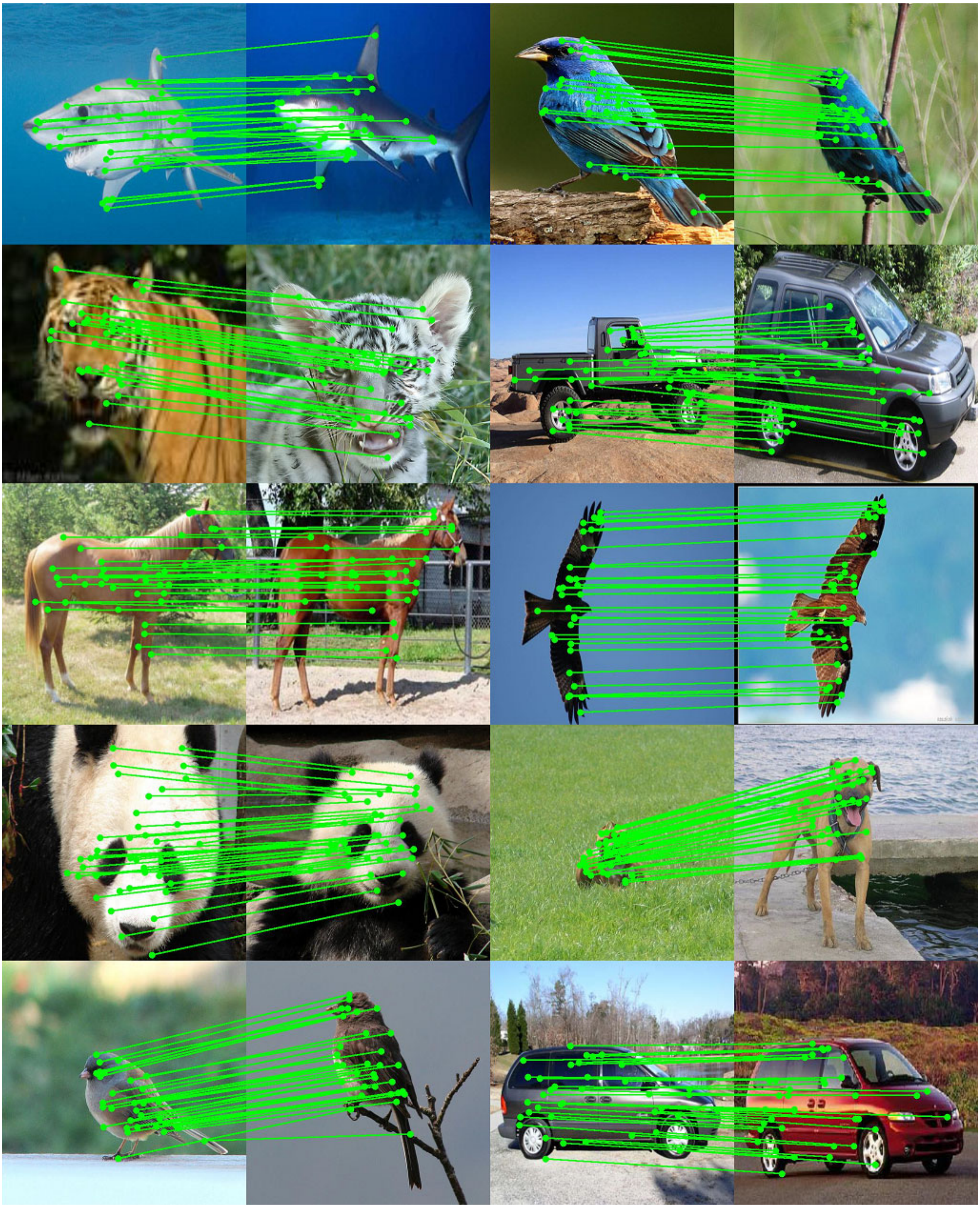

Figure 7. **Visualization of our SC-ImageNet (Part 3 of 3).** All images are selected from ImageNet [2]. Green dots and lines are our automatically generated pseudo-labels.

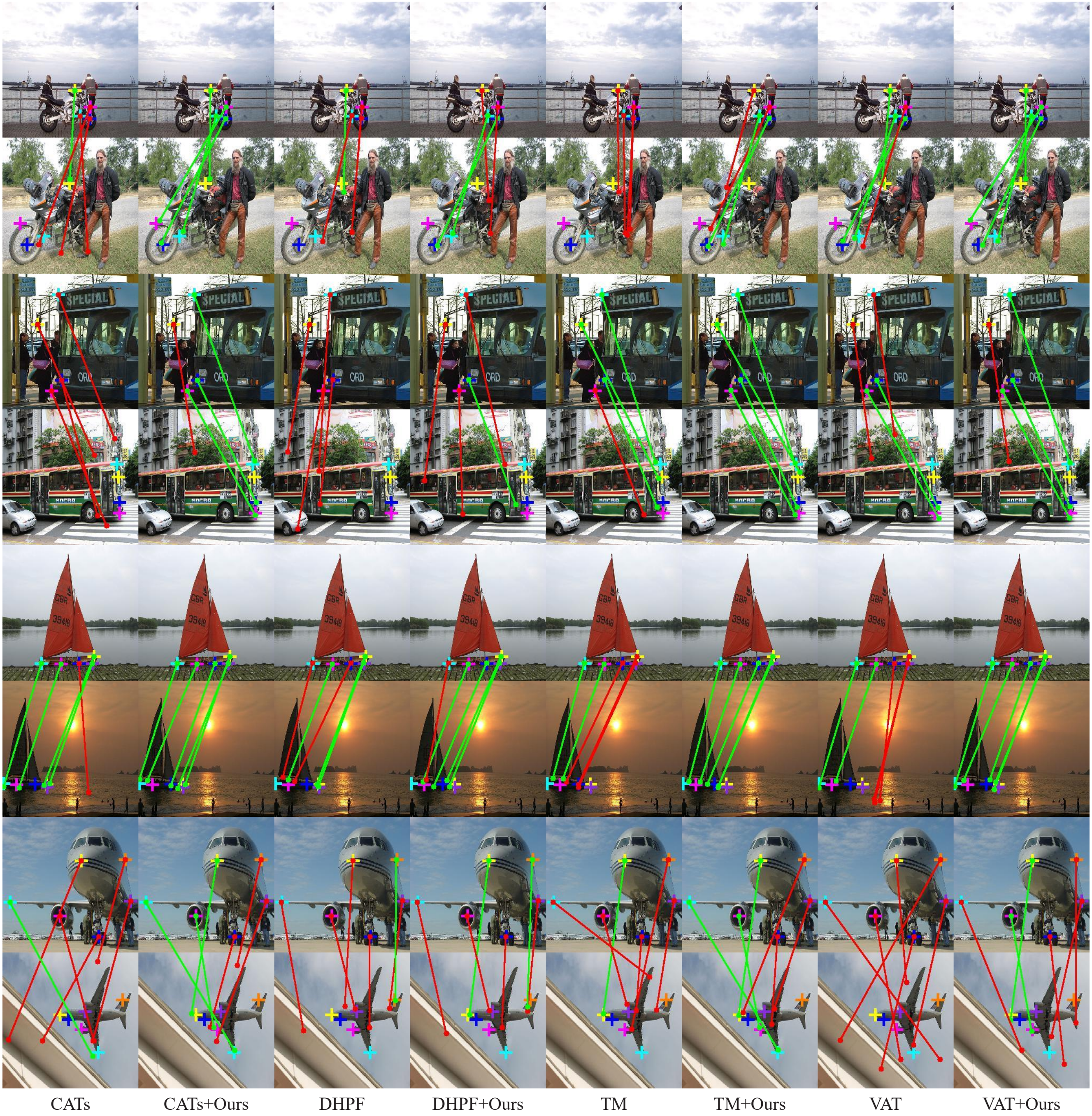

Figure 8. **Additional qualitative results on SPair-71k [8] (Part 1 of 3).** TM means TransforMatcher [7]. Other methods are proposed in CATs[1], DHPF[9] and VAT[4], respectively. +Ours demonstrates semantic correspondence models are firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Colored cross indicates corresponding key-point (ground-truth). Green dots and lines mean correct matches, while red dots and lines mean incorrect matches (measured by PCK@0.1 as in Table 2 of our main paper).

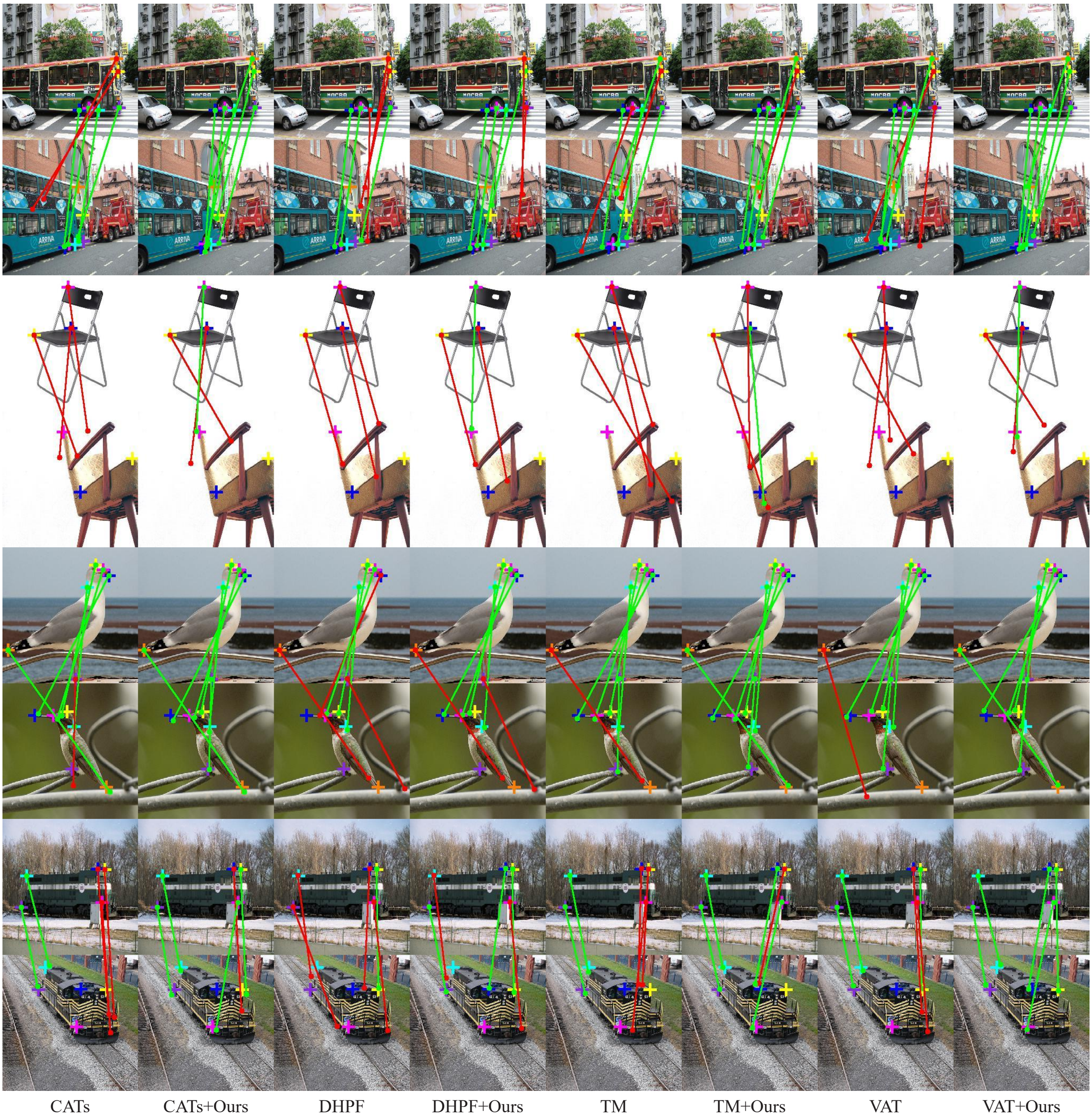

Figure 9. **Additional qualitative results on SPair-71k [8] (Part 2 of 3).** TM means TransforMatcher [7]. Other methods are proposed in CATs[1], DHPF[9] and VAT[4], respectively. +Ours demonstrates semantic correspondence models are firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Colored cross indicates corresponding key-point (ground-truth). Green dots and lines mean correct matches, while red dots and lines mean incorrect matches (measured by PCK@0.1 as in Table 2 of our main paper).

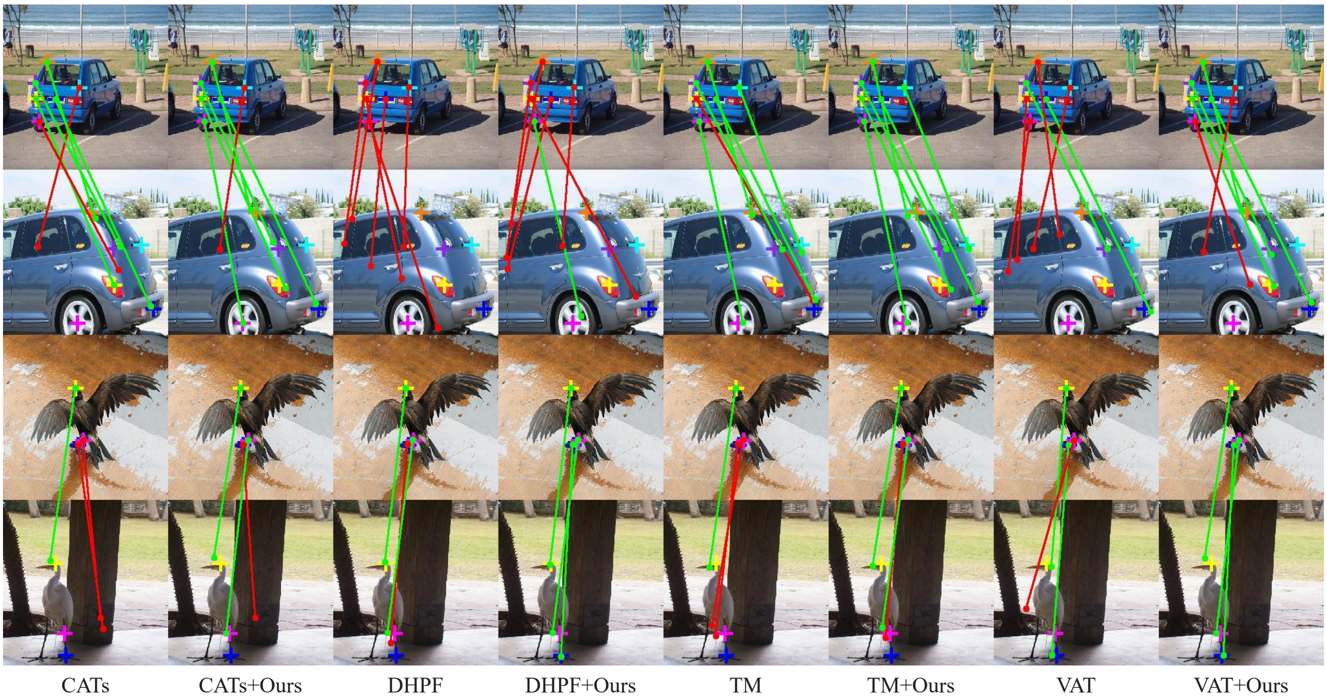

Figure 10. **Additional qualitative results on SPair-71k [8] (Part 3 of 3).** TM means TransforMatcher [7]. Other methods are proposed in CATs[1], DHPF[9] and VAT[4], respectively. +Ours demonstrates semantic correspondence models are firstly pre-trained with our SC-ImageNet, then fine-tuned on SPair-71k. Colored cross indicates corresponding key-point (ground-truth). Green dots and lines mean correct matches, while red dots and lines mean incorrect matches (measured by PCK@0.1 as in Table 2 of our main paper).

## References

[1] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021. 2, 3, 8, 9, 10

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 1, 2, 4, 5, 6, 7

[3] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1711–1725, 2017. 2

[4] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 108–126. Springer, 2022. 3, 8, 9, 10

[5] Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 267–284. Springer, 2022. 2, 3

[6] Jiwon Kim, Kwangrok Ryoo, Junyoung Seo, Gyuseong Lee, Daehwan Kim, Hansang Cho, and Seungryong Kim. Semi-supervised learning of semantic correspondence with pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19699–19709, 2022. 2, 3

[7] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8707, 2022. 2, 3, 8, 9, 10

[8] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3395–3404, 2019. 2, 3, 8, 9, 10

[9] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020. 3, 8, 9, 10

[10] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2