# *Supplementary* for Pixel-level Semantic Correspondence through Layout-aware Representation Learning and Multi-scale Matching Integration

Yixuan Sun[1,2,*], Zhangyue Yin[3,*], Haibo Wang[3], Yan Wang[1,2], Xipeng Qiu[3], Weifeng Ge[3,†] and Wenqiang Zhang[1,2,3,†]
[1] Academy for Engineering & Technology, Fudan University, Shanghai, China
[2] Engineering Research Center of AI & Robotics, Ministry of Education, China
[3]School of Computer Science, Fudan University, Shanghai, China
{wfge, wqzhang}@fudan.edu.cn

## 1. Further Analysis of Designs in LPMFlow

### 1.1. Details of Region-based PE

We clarify that region-based position embedding (PE) is the 2D relative positional encoding [14], in which region base is explained as the procedure to separate the source and target images with 2D patch regions. Extending from the 1D sequence-relative position computations in natural language processing tasks, 2D Embedding extends the matrix of relative position into 2D format along the x & y axes in the procedure of attention. In order to be compatible with the structure designed in LARL for the joint processing of source and target features, we separately perform region-based relative position encoding on four regions (namely source to source, source to target, target to source, and target to target). The introduction of the region-based position encoding can better encode orientation and spatial layout. The efficacy of this region-based position embedding (PE) is substantiated by the ablation studies in the main paper, which demonstrates this structure can contribute to the enhancement of semantic correspondence accuracy.

Table 1. Analysis of LPMFlow Designs.

| Methods | SPair-71K $\alpha_{bbox} = 0.1$ |
|---|---|
| LPMFlow | 65.6 |
| *Analysis on Representation Learning* | |
| Isolated Dual Self-attention | 63.5 (2.1↓) |
| Bidirectional Cross-attention | 64.0 (1.6↓) |
| *Analysis on Multi-scale Integration* | |
| w/o Cross-Scale Flow Integration | 64.7 (0.9↓) |
| w/o C2F Refinement ($8\times8$) | 64.9 (0.7↓) |

### 1.2. Analysis on Representation Learning

To prove the effectiveness of our designed structure for representation learning, we compare our structure with other two structures namely dual path isolated self-attention and dual path structure with bidirectional cross-attention. Our structure demonstrates superior performance compared to the others, achieving 2.1 and 1.6 improvement (shown in Table 1). We also provide further visualization with/without LARL in Figure 3. The results show our LARL can better extract shared layout information for an image pair.

Besides, we clarify that to calculate the foreground weight, we use the averaged [cls] tokens of source and target images as global semantic tokens. We calculate the cosine distance between the global semantic and the patch token of the source image with max-min normalization to generate a weight map. We explain this as the foreground area is both the salient and shared areas for source and target images. We provide a visualization to prove our weight map can extract foreground patch tokens in Figure 1.

### 1.3. Analysis on Multi-scale Integration

The Multi-scale Integration aims to 1) generate fine-grained and multi-scale features and build up cross-scaled 4D matching tensors implemented by PFSR; 2) integrate the multi-scaled correlation and refine them into pixel-level correspondence implemented by MMFI. In these two modules, we further analyze the introduction for the correlation of asymmetric scales and the design of the coarse-to-fine matching flow refinement structure. As the results shown in Table 1, when we remove the cross-scaled $1\times2$ and $2\times1$ matching tensor, the performance declines by 0.9%. We add the ablation for coarse-to-fine refinement structure with constant $8\times8$ window size and the performance declines by 0.7%. Further visualization for the effectiveness of two modules PFSR and MMFI are also provided in Figure 3.

Table 2. Per-class level quantitative evaluation results on SPair-71k [10] benchmark, $*$ stands for the method implemented with iBOT-B backbone same with LPMFlow, the best results are in bold. TMatcher is TransforMatcher for short.

| Methods | aero. | bike | bird | boat | bott. | bus | car | cat | chai | cow | dog | hors. | mbik. | pers. | plan. | shee. | trai. | tv | all |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NC-Net[12] | 17.9 | 12.2 | 32.1 | 11.7 | 29.0 | 19.9 | 16.1 | 39.2 | 9.9 | 23.9 | 18.8 | 15.7 | 17.4 | 15.0 | 14.8 | 9.6 | 24.2 | 31.1 | 20.1 |
| SCOT [6] | 34.9 | 20.7 | 63.8 | 21.1 | 43.5 | 27.3 | 21.3 | 63.1 | 20.0 | 42.9 | 42.5 | 31.1 | 29.8 | 35 | 27.7 | 24.4 | 48.4 | 40.8 | 35.6 |
| DHPF [11] | 38.4 | 23.8 | 68.3 | 18.9 | 42.6 | 27.9 | 20.1 | 61.6 | 22.0 | 46.9 | 46.1 | 33.5 | 27.6 | 40.1 | 27.6 | 28.1 | 49.5 | 46.5 | 37.3 |
| CHM [8] | 49.6 | 29.3 | 68.7 | 29.7 | 45.3 | 48.4 | 39.5 | 64.9 | 20.3 | 60.5 | 56.1 | 46.0 | 33.8 | 44.3 | 38.9 | 31.4 | 72.2 | 55.5 | 46.3 |
| CATs [2] | 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 | 49.9 |
| MMNet[15] | 55.9 | 37.0 | 65.0 | 35.4 | 50 | 63.9 | 45.7 | 62.8 | 28.7 | 65.0 | 54.7 | 51.6 | 38.5 | 34.6 | 41.7 | 36.3 | 77.7 | 62.5 | 50.4 |
| TMatcher[5] | 59.2 | 39.3 | 73.0 | 41.2 | 52.5 | 66.3 | 55.4 | 67.1 | 26.1 | 67.1 | 56.6 | 53.2 | 45.0 | 39.9 | 42.1 | 35.3 | 75.2 | 68.6 | 53.7 |
| CATs$^*$[2] | 56.7 | 41.3 | 77.8 | 35.0 | 54.8 | 59.8 | 45.2 | 69.9 | 31.4 | 63.7 | 57.6 | 62.5 | 46.7 | 49.1 | 43.2 | 43.5 | 76.4 | 64.1 | 55.2 |
| TMatcher$^*$[5] | 57.1 | 47.4 | **83.5** | 42.3 | 56.8 | 57.0 | 55.4 | 75.3 | 34.5 | 66.1 | 64.2 | 60.2 | 52.8 | 55.2 | 40.5 | 46.0 | 75.1 | 65.8 | 57.9 |
| ACTR$^*$[13] | 65.1 | 48.5 | 82.3 | **50.4** | 55.9 | 65.3 | 63.1 | 72.8 | 35.8 | 74.1 | 70.3 | 68.9 | 58.6 | **57.1** | 46.8 | 49.5 | 84.4 | 73.3 | 62.1 |
| LPMFlow$^*$ | **71.4** | **54.8** | 83.2 | 50.3 | **57.0** | **75.4** | **68.9** | **79.3** | **41.1** | **78.4** | **74.1** | **73.7** | **58.7** | 56.9 | **48.7** | **54.7** | **87.5** | **74.6** | **65.6** |

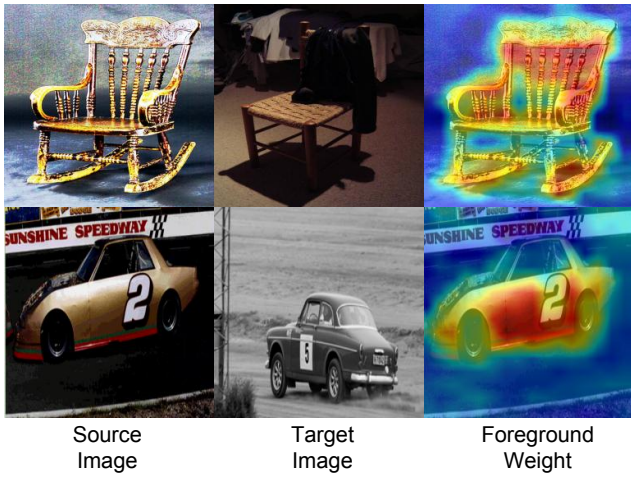


Figure 1. Visualization of generated foreground weight.

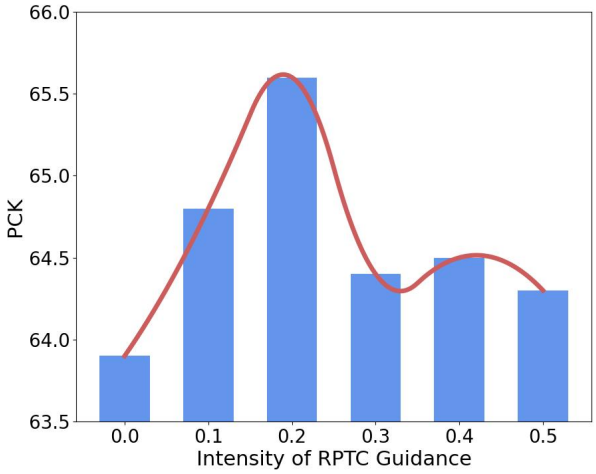


Figure 2. Performance of our method with different initial intensities of RPTC guidance.

## 2. Further Results for LPMFlow

### 2.1. Class Level Evaluation Results on SPair-71k

Table 2 shows our proposed LPMFLow outperforms all previous works with CNN-based backbones as well as CATs [2], TransforMatcher [5] and ACTR [13] in iBOT-B backbone clearly in most categories (**15 of 18**) of SPair-71K [9]. Especially in more challenging categories such as {bike, bus, car, and horse} that previous works [2, 5, 7, 13] often fail, our PCK@$\alpha_{bbox}$ 0.1 accuracy improves by {6.3%, 9.1%, 5.8% and 4.8%}. Class-level evaluation results indicate that LPMFlow can provide more accurate correspondence compatible with various object categories.

### 2.2. Evaluation of LPMFlow on MAE & iBOT-22K

We conduct experiments to investigate the impact of initial image-level feature quality on our designed pipeline. We replace the pre-trained weights with the MAE [3] method pre-trained on ImageNet 1K and iBOT [16] method pre-trained on ImageNet 22K. Results show that LPMFlow using MAE performs well with 65.9% slightly higher than that using iBOT pre-trained parameters. And LPMFlow using iBOT-22K performs at 66.4% also improves from the method using iBOT-1K backbone. This experiment proves that our LPMFlow has the potential to further improve the performance with better initial image features provided.

### 2.3. Further Ablation Analysis

We prove more ablations for LPMFlow on SPair-71K. We report the ablation results for $topK$ in LARL as 0:63.5%, 16:64.7%, 32:65.0%, 64:65.6%, 128:64.3% which shows $K$=64 is an effective setting. We also evaluate the design of the initial $\lambda$ from 0 to 0.5 for the RPTC task in Figure 2. Our method achieves the best performance when the intensity of RPTC guidance is set as 0.2. We set the decline rate for $\lambda$ as 10% until $\lambda$=0. We provide the ablation on the decline rate as 0%:64.5%, 5%:65.0%, 10%:65.6%, 15%:65.2%, 20%:65.2%, 25%:64.7%. These results prove the effectiveness of our design in decline rate.

## 3. Additional Visualization

### 3.1. Effectiveness of Three Designed Modules

We provide more visualization results in Figure 3 for LPMFlow with and without designed three modules. The challenge of objects with similar appearances such as the corners of a bus or the landing gear of an aeroplane usually leads to mismatching. As the target heatmaps in the first case show, LARL can fix incorrect representation with the consistency of the geometric layout of the components for objects in the same category. For the PFSR, the second case shows the fusion of cross-scaled matching tensors can provide more accurate matching for an object pair of different sizes. For the MMFI, the third case shows in pixel-level correspondence, that the usage of MMFI can better distinguish the matching relationship among adjacent pixels.

### 3.2. Comparison of Qualitative Results

We provide more visualization results in three gains namely warping results, dense matching flows, and matching details. We use the thin-plate splines algorithm [1] for image warping and the procedure is instructed by predicted key points. The warping results in Figure 4 demonstrate the effectiveness of our method by providing an overview of correspondence between object pairs. We provide the dense flow result for the foreground objects. Without additional supervision (sparse keypoint annotation in SPair-71K only), we acquire the foreground object using the cosine distance between averaged [cls] tokens of the image pair and the patch token similar to the implementation for generating the foreground weight map. The comparison of dense flow in Figure 5 shows except for having good correspondence in key areas, our method can also distinguish differences among nearby pixels and construct pixel-level matching. The visual comparison of matching details in Figure 6-8 also indicate the effectiveness of LPMFlow.

## References


[1] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 3, 5

[2] 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, 5, 6

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 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 *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 4, 7, 8, 9

[5] 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

[6] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10196–10205, 2020. 2

[7] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020. 2, 5, 7, 8, 9

[8] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2950, 2021. 2

[9] 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

[10] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 2

[11] 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. 2

[12] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Ncnet: Neighbourhood consensus networks for estimating image correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[13] Yixuan Sun, Dongyang Zhao, Zhangyue Yin, Yiwen Huang, Tao Gui, Wenqiang Zhang, and Weifeng Ge. Correspondence transformers with asymmetric feature learning and matching flow super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17787–17796, 2023. 2, 4, 5, 6, 7, 8, 9

[14] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 1

[15] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3354–3364, 2021. 2, 5, 6, 7, 8, 9

[16] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022. 2

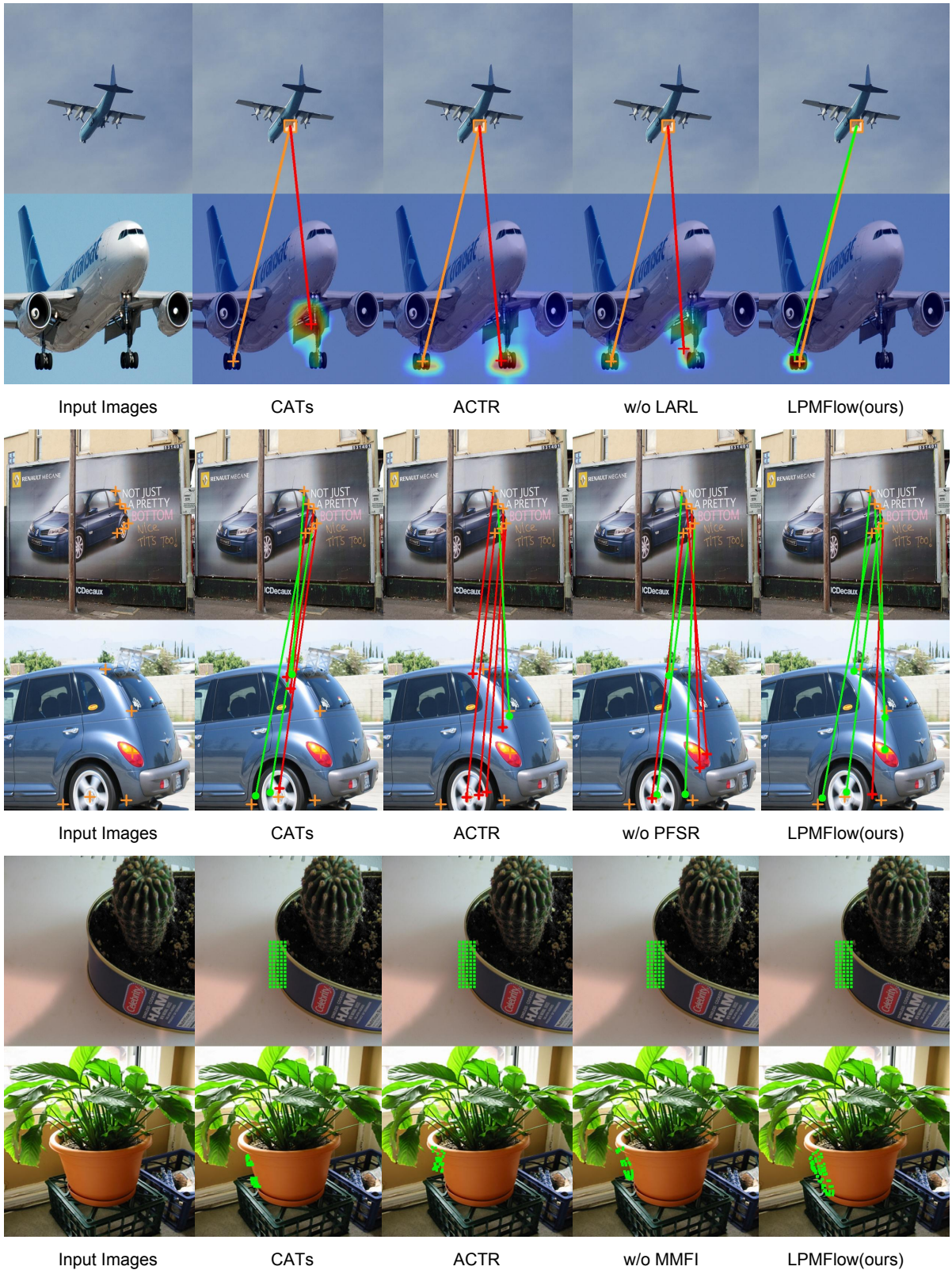


Figure 3. Visualization of the performance of three modules on three challenges. We evaluate the effectiveness of LARL for confusing regions with similar appearances in the first case. We evaluate the effectiveness of PFSR for objects with inconsistent scales in the second case. We evaluate the effectiveness of MMFI for the ability to distinguish nearby pixels. Except for the ablation models and our LPMFlow, we also visualize the results of CATs [4] and ACTR [13].

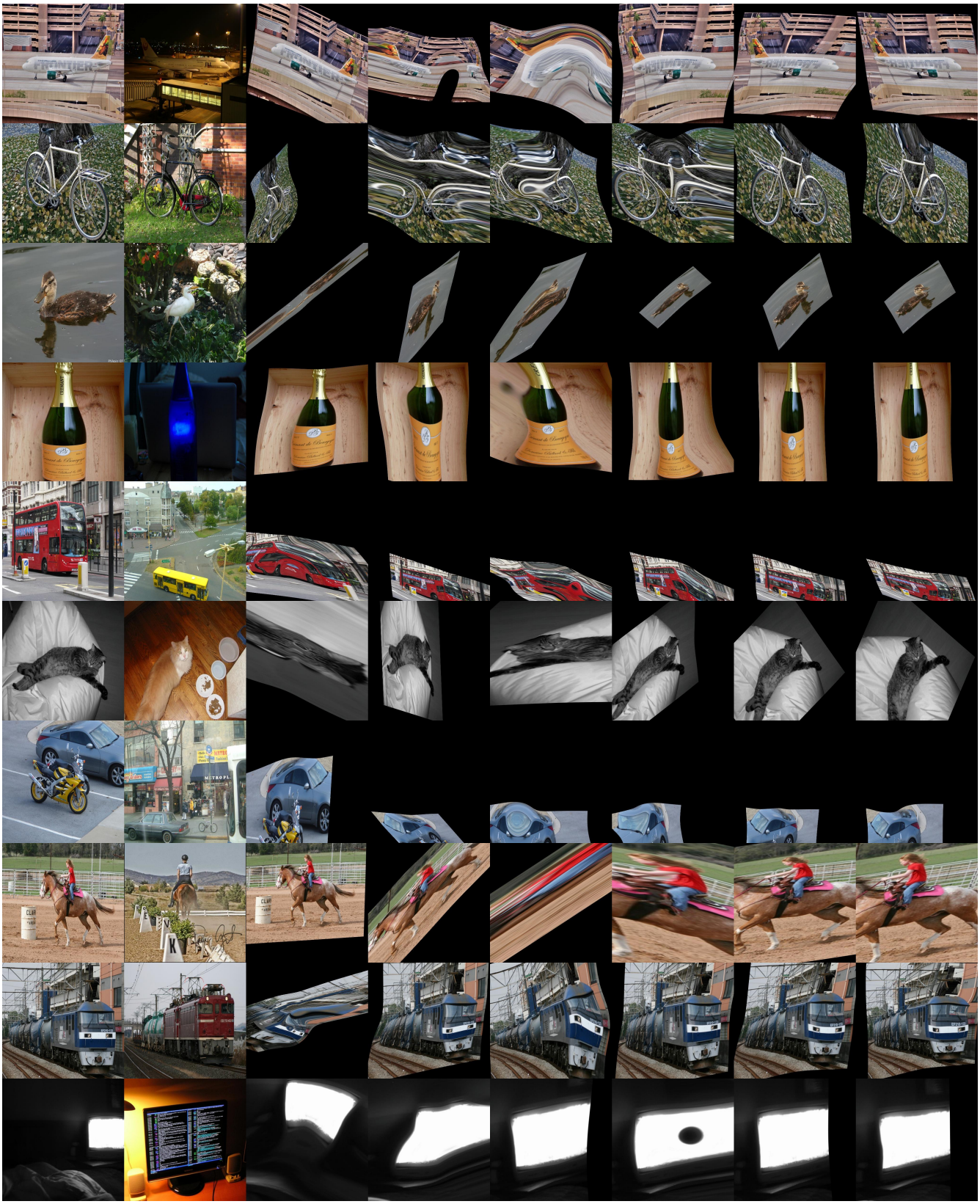


Figure 4. Visualization of dense warping result for state-of-the-art methods namely SCOT [7], CATs [2], MMNet [15] and ACTR [13] compared with our LPMFlow. Thin-plate splines algorithm [1] is used for image warping with instructed by predicted key points.

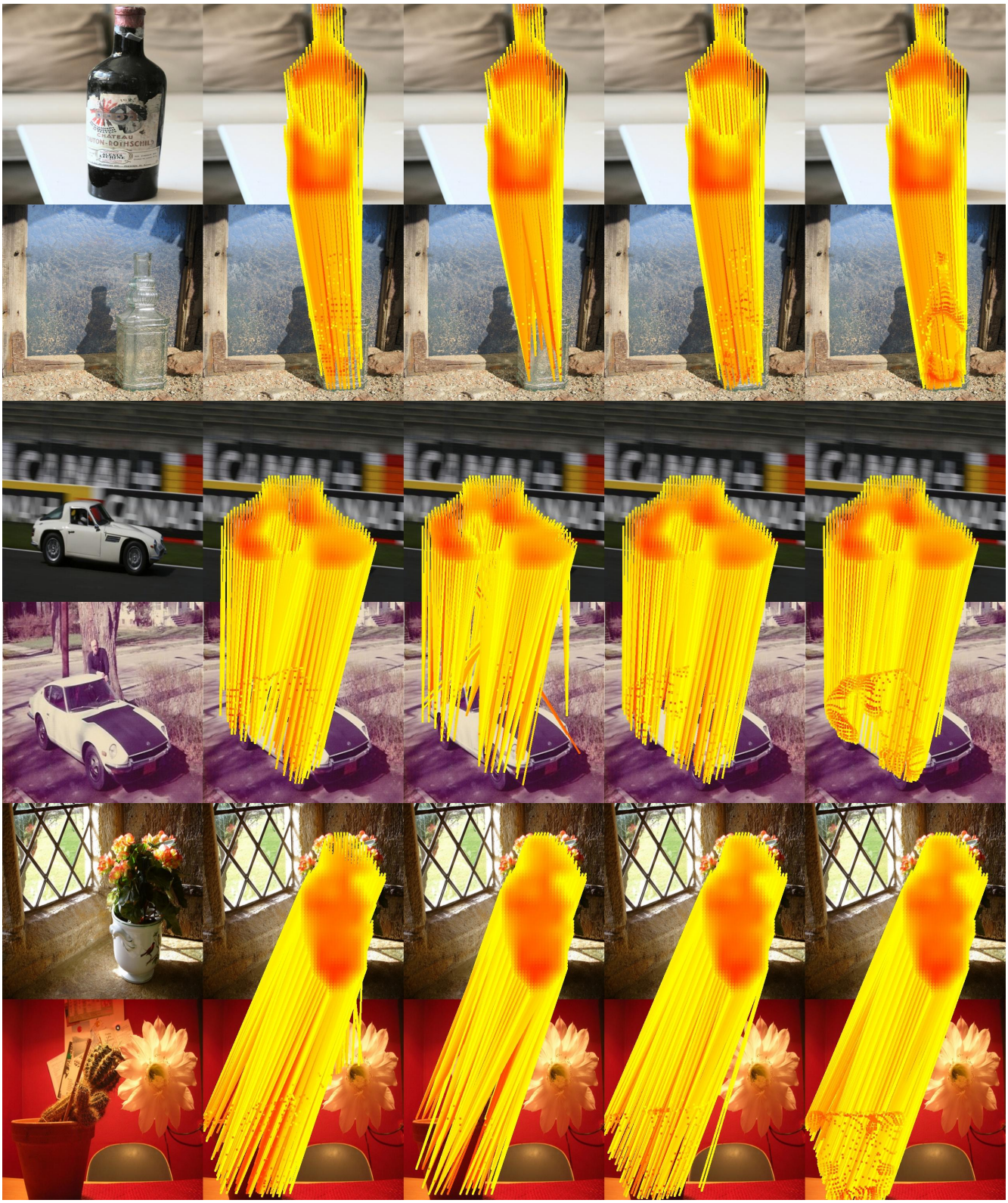


Figure 5. Visualization of dense matching flow for state-of-the-art methods namely CATs [2], MMNet [15] and ACTR [13] compared with our LPMFlow.

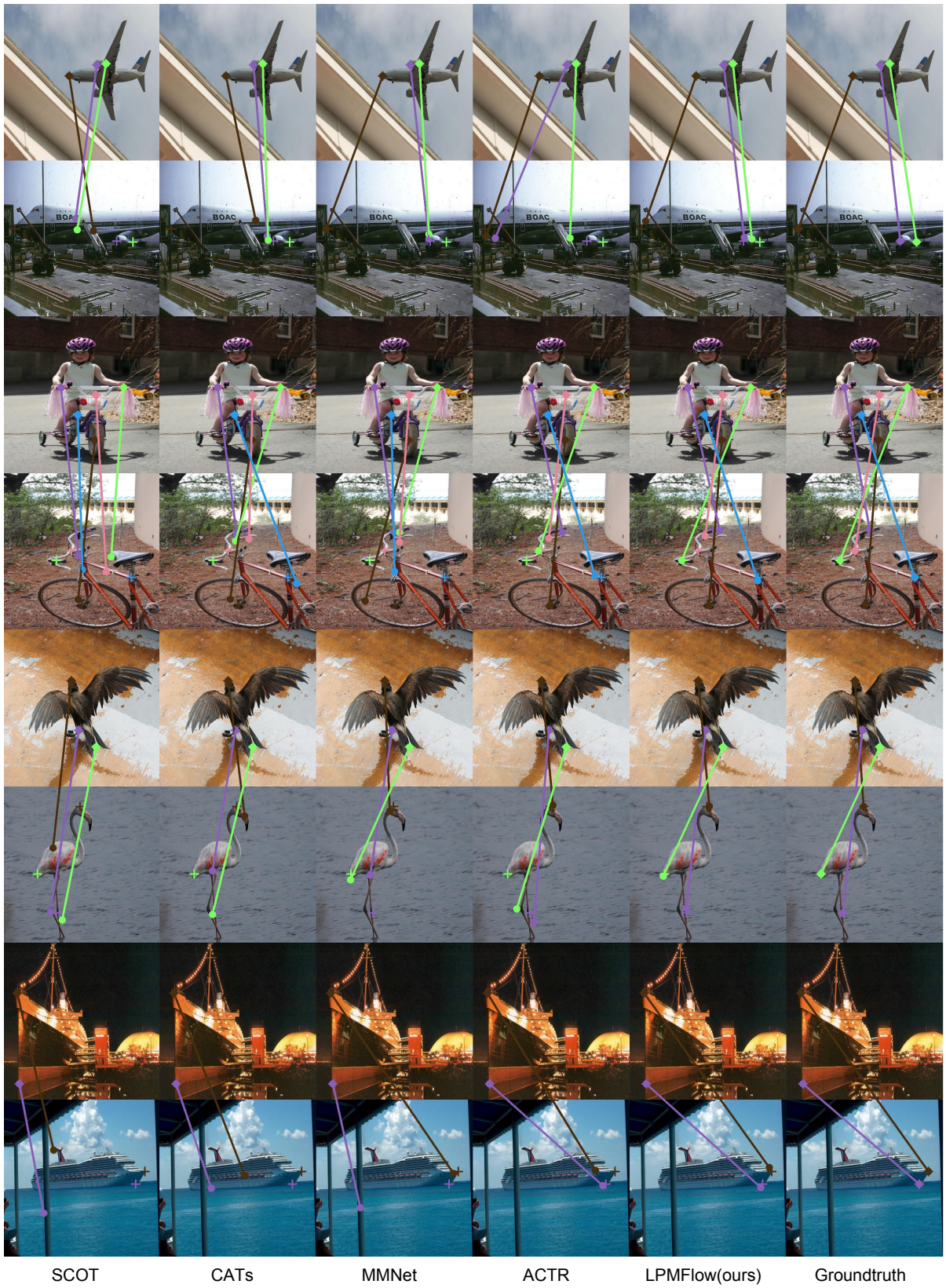


Figure 6. Visualization of matching result. The upper image is the source image the below image is the target image, and the crosses are the ground truth labels. We compare the results of SCOT [7], CATs [4], MMNet [15], ACTR [13] and our LPMFlow.

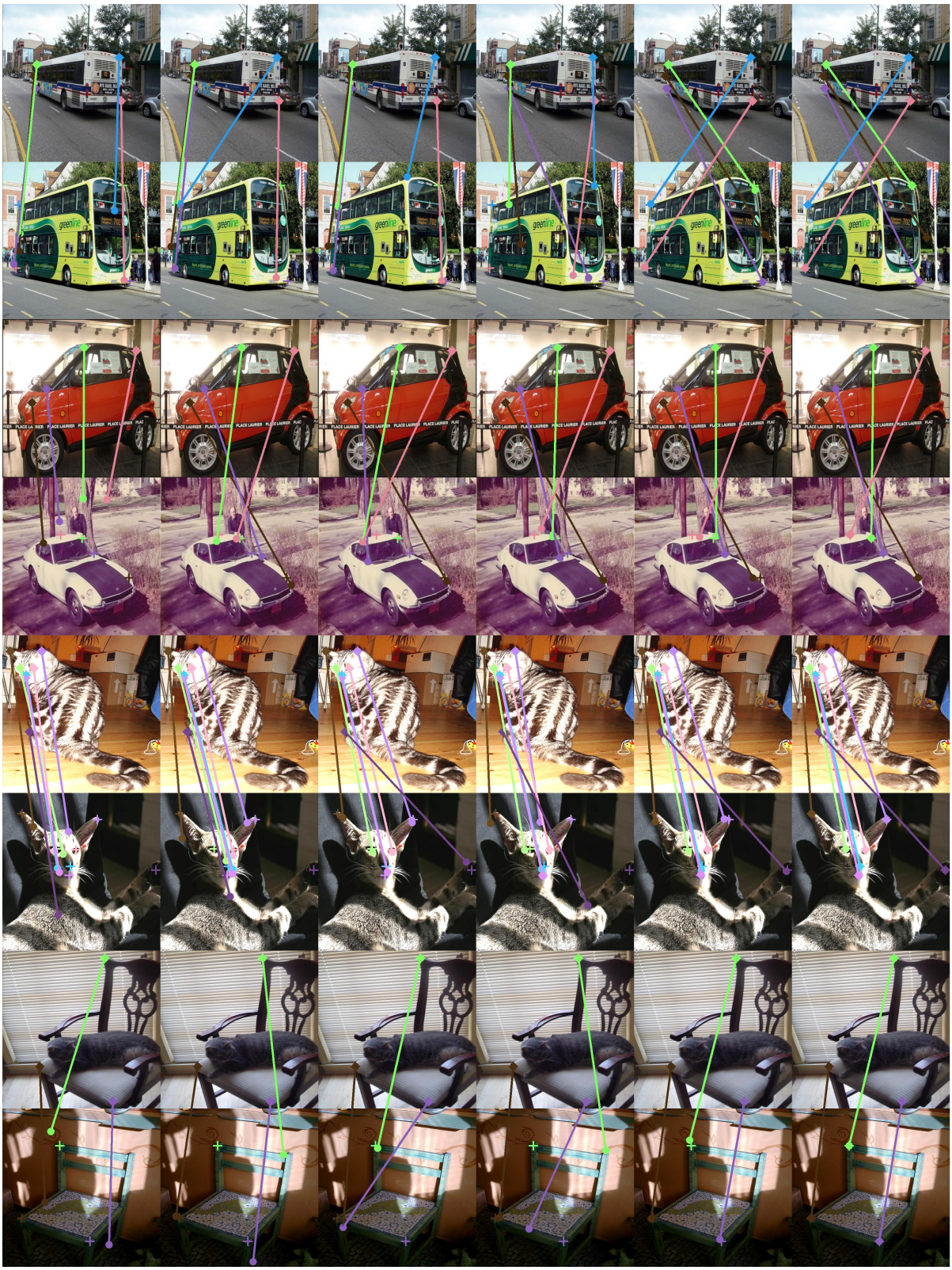


Figure 7. Visualization of matching result. The upper image is the source image the below image is the target image, and the crosses are the ground truth labels. We compare the results of SCOT [7], CATs [4], MMNet [15], ACTR [13] and our LPMFlow.

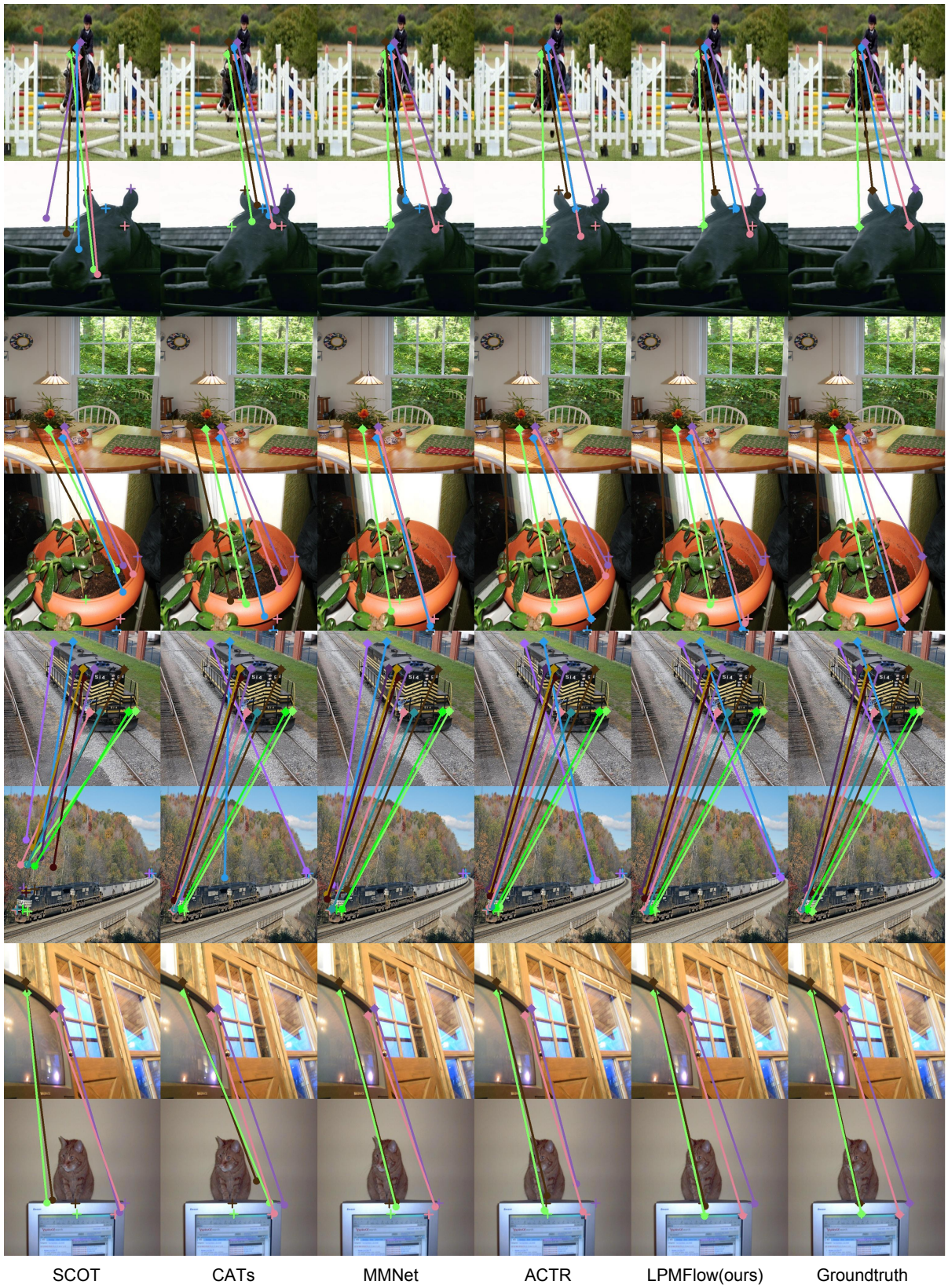


Figure 8. Visualization of matching result. The upper image is the source image the below image is the target image, and the crosses are the ground truth labels. We compare the results of SCOT [7], CATs [4], MMNet [15], ACTR [13] and our LPMFlow.