Supplementary Material of MATCHA🍵

Fei Xue[1*] Sven Elflein[2,3,4] Laura Leal-Taixé[3] Qunjie Zhou[3†]
[1]University of Cambridge [2]University of Toronto [3]NVIDIA [4]Vector Institute

In this supplementary document, we describe model training details in Appendix A, and provide more evaluation results in Appendix B. In addition, we present qualitative visualization for geometric, semantic and temporal matches across different methods in Appendix C.

A. Supervision and Training Details

Geometric matching supervision. We train geometric descriptors with ground-truth geometric correspondences as previous local features [2, 12, 14]. We leverage the dual-softmax loss function proposed in [12] which employs the negative log-likelihood loss over matching probabilities from mutual directions. Given an image pair I^a and I^b with M ground-truth geometric correspondences, we first subsample sparse geometric descriptors X_g^a and $X_g^b \in R^{M\times D_g}$ located at keypoints with ground truth annotations from the extracted dense geometric descriptors F_g^a and F_g^b. We then compute the similarity matrix $S \in R^{M\times M}$ from two sets of sparse descriptors, *i.e.*, $S = X_g^a (X_g^b)^T$, and compute the geometric loss defined as:

$$L_{geo} = -\sum_i \texttt{log}(\texttt{softmax}_r(S)_{ii}) - \sum_i \texttt{log}(\texttt{softmax}_r(S^T)_{ii}), \quad (1)$$

where we apply `softmax` from both matching directions over the similarity matrix.

Semantic matching supervision. Similar to geometric matching supervision, we train semantic descriptors with ground-truth semantic matches which are sparsely annotated by human. Thus, we subsample sparse semantic descriptors $X_s^a \in R^{M\times D_s}$ and $X_s^b \in R^{M\times D_s}$ at keypoint locations with ground truth. We adopt the commonly used the CLIP contrastive loss [13] f_{cl} defined as:

$$f_{cl} = f_{ce}(\tau X_s^a (X_s^b)^T, \mathcal{O}) + f_{ce}(\tau X_s^b (X_s^a)^T, \mathcal{O}), \quad (2)$$

where f_{ce} is the CrossEntropy loss and τ is the scale parameter. $\mathcal{O} = (0, 1, ..., M-1)^T$ is the ground-truth labels with M classes. The sparse contrastive loss, however, only minimizes the distances between positive pairs and ignores the distances between negative pairs. To compensate for that, an additional dense semantic flow loss [5] is adopted as

$$L_{flow} = \sum_i ||(p_i^a - (\hat{p}_i^a + \epsilon))||_2 + \sum_i ||(p_i^b - (\hat{p}_i^b + \epsilon))||_2 \ , \quad (3)$$

where ϵ is the Gaussian noise with mean of 0 and standard variance of 25, $\hat{p}_i^a$ is the ground-truth correspondence, and $p_i^a = \sum_q m_i(q) q$ is the predicted correspondence. p_i^a is the average of all positions $q = (u, v)$ of F_s^b weighted by matching probability $m_i(q)$ between descriptor $X_{s,i}^a$ at the index of i and $F_{s,q}^b$. The matching probability $m_i(q)$ is the normalized similarity value between $X_{s,i}^a$ and $F_{s,q}^b$ and is computed as:

$$m_i(q) = \frac{exp(\frac{X_{s,i}^a (F_{s,q}^b)^T}{\beta})}{\sum_q^{'} exp(\frac{X_{s,i}^a (F_{s,q'}^b)^T}{\beta})} \ , \quad (4)$$

where β is the temperature.

The optical flow loss enforces semantic descriptors to maximizes the distances between negative pairs while minimizing the distances between positive pairs. The total loss for semantic matching is the combination of sparse contrastive loss L_{cl} and dense flow loss L_{flow}:

$$L_{sem} = w_{cl} L_{cl} + w_{flow} L_{flow}, \quad (5)$$

where w_{cl} and w_{flow} are weights balancing the two losses.

Training data. We train our model using both geometric and semantic datasets, balancing samples across each dataset to ensure even representation. For geometric matching supervision, we use the ScanNet [1] and MegaDepth [7] datasets adopting dataset splits used in [15] leading to approximate 15k indoor sequences from ScanNet and 441 outdoor sequences from MegaDepth. We use the ground-truth poses and depth maps to generate correspondences for training. For semantic matching supervision, we use PF-PASCAL [4], SPair-71k [9], and AP-10k [18] as in [19].

*This work was done when Fei Xue was an intern at NVIDIA.
†Project leader.

PF-PASCAL includes 2941 training pairs from 20 object categories. SPair-71k offers 53k training pairs across 18 categories with high intra-class variation. AP-10k provides 10k images across 23 categories, with an additional 261k pairs generated for semantic training.

Training schema. To properly train MATCHA, we adopt a multi-stage training schema. Empirically, we found geometric descriptors require more iterations to be trained properly. This is likely to be caused by the imbalanced number of available annotated data, *i.e.*, we have more geometric samples than semantic samples. Training too long on limited semantic matching correspondences harms generalization. Therefore, to compensate the data imbalance, we 1) first train the model purely on geometric matching with frozen semantic features using L_{geo}, and 2) next jointly train both geometric and semantic descriptors on geometric and semantic matching using a weighted combination of both supervisions as:

$$L_{total} = L_{geo} + w_{sem} L_{sem}. \tag{6}$$

Implementation details. MATCHA is implemented on PyTorch [11] with 8 blocks consisting of both self and cross attention layers. The hidden size of the self and cross attention layer is 512 and the number of head is 8. The dimension of final geometric and semantic descriptors is 256 and 768. The patch size p used to patchify geometric and semantic features is set to 2 for both geometric and semantic features. In the training process, hyper-parameters of τ (Eq. 2), β (Eq. 4), w_{cl} (Eq. 5), w_{flow} (Eq. 5), and w_{sem} (Eq. 6) are set to 0.02, 14.3, 1.0, 1.0, 0.1, respectively.

We train MATCHA using AdamW [8] optimizer with weight decay of 1×10^{-3} and initial learning rate of 1×10^{-4} on 4 H100 GPUs for 220k iterations in total with 150k iterations at the first stage. The learning rate is reduced to 5×10^{-5} and 2×10^{-5} after 100k and 150k iterations. The batch size is set to 24 and 48 for the first and second stage training, respectively. All images are sized to 512×512 in the training process.

B. Additional Evaluations

Temporal matching. We provide additional ablation study to understand the performance of different types of features on temporal matching. Specifically, we consider the geometric (geo) and semantic (sem) of descriptors of MATCHA-Light and DIFT [16] models and their unified feature version, *i.e.*, DIFT.Uni, MATCHA-Light.Uni. We also consider the feature models that combine DINOv2 [10], *i.e.*, DIFT.Uni + DINOv2 and MATCHA.

As shown in Tab. 1, low-level geometric features are more important to temporal matching than high-level semantic features, *i.e.*, DIFT (geo) *vs* DIFT (sem), and MATCHA-Light (geo) *vs* MATCHA-Light (sem). We also notice that geometric supervision leads to improved temporal matching, *i.e.*, MATCHA-Light (geo) *vs* DIFT (geo). In contrast, adding semantic supervision produces degraded temporal matching accuracy, *i.e.*, MATCHA-Light (sem) *vs* DIFT (sem), which shows that sparse semantic correspondence supervision across instances leads to decreased capability in establishing matches between the same instance.

Method	Supervision	TAPVid-Davis [3] PCK$_{@0.01/0.05/0.1}$
DIFT (geo) [16]	✗	75.6 / 82.6 / 86.9
DIFT (sem) [16]	✗	71.9 / 81.4 / 86.4
MATCHA-Light (geo)	GM+SM	75.7 / 82.8 / 87.0
MATCHA-Light (sem)	GM+SM	64.9 / 77.9 / 84.3
DINOv2 [10]	✗	83.2 / 89.7 / 92.0
DIFT.Uni [16]	✗	79.7 / 86.7 / 90.5
DIFT.Uni + DINOv2 [10, 16]	✗	<u>86.4</u> / <u>91.6</u> / <u>93.5</u>
MATCHA-Light.Uni	GM+SM	78.7 / 86.3 / 90.2
MATCHA	GM+SM	**87.8 / 93.5 / 95.5**

Table 1. **Ablation Study on Temporal Matching.** We report the Percentage of Correct Keypoints (PCK) under different thresholds. The **best** and <u>second-best</u> results are highlighted.

The combination of geometric and semantic features contains the properties of the both features, giving better temporal matching accuracy *i.e.*, DIFT.Uni *vs* DIFT (geo), and MATCHA-Light.Uni *vs* MATCHA-Light (geo). As discussed in the main paper, DINOv2 benefiting from its large-scale learning on single object-centric data, is able to well handle large viewpoint and scale changes, especially for single-object dominant scenes, leading to surprisingly superior temporal matching performance. By combining with DINOv2 features, both DIFT.Uni and MATCHA have significant improvement in temporal matching, *i.e.*, MATCHA *vs* MATCHA-Light.Uni, DIFT.Uni + DINOv2 *vs* DIFT.Uni, validating object-level semantic representation learned by DINOv2 is complementary to semantic features extracted from stable diffusion models.

Ablation on obtaining a unified feature. In the main paper, we adopt a simple concatenation-based merging mechanism to obtain a unified feature. To further validate this design choice, we provide additional ablation study focusing on comparing different ways of unifying knowledge in feature representations. Specifically, we consider **MATCHA-Light** that learns to fuse geometric and semantic features yet keeping separate descriptors for geometric and semantic matching following DIFT, **MATCHA-Light.Uni** that combines the MATCHA-Light geometric and semantic descriptors with concatenation-based merging, **MATCHA-Light.Uni.S** that further supervises MATCHA-Light.Uni with joint geometric and semantic training, as well as **MATCHA**, our final model, that combines DINOv2 with MATCHA-Light.Uni.

In Tab. 2, we show that simple concatenation-based

Method	Single Desc	Corres. Sup.	*Geometric* Aachen $AUC_{@5/10/20}(\uparrow)$	Avg(↑)	*Semantic* PF-Willow $PCK_{@0.05/0.1/0.15}(\uparrow)$	Avg(↑)	*Temporal* TapVid-Davis $PCK_{@0.05/0.1/0.15}(\uparrow)$	Avg(↑)	Average Score(↑)
MATCHA-Light	✗	GM+SM	51.4 / 60.1 / 67.1	59.5	69.0 / 90.6 / 96.2	85.3	78.7 / 86.3 / 90.2	85.1	76.6
MATCHA-Light.Uni	✓	GM+SM	50.0 / 59.0 / 66.5	58.5	60.8 / 82.8/ 90.4	78.0	78.7 / 86.3 / 90.2	85.1	73.9
MATCHA-Light.Uni.S	✓	GM+SM	49.9 / 58.4 / 65.4	57.9	36.8 / 53.0 / 62.4	50.7	79.1 / 85.9 / 89.5	84.8	64.5
MATCHA	✓	GM+SM	**51.7 / 61.0 / 68.5**	**60.4**	**70.2 / 91.3 / 97.0**	**86.2**	**87.8 / 93.5 / 95.5**	**92.3**	**79.6**

Table 2. **Ablation study on obtaining a unified feature.** We compare different ways of obtaining a unified feature. We show that simple concatenation leads to better way to keep the learned geometric and semantic representation while adding additional joint training on the concatenated feature pushes the feature to focus more on geometric matching, leading to significantly degraded semantic matching.

merging (MATCHA-Light.Uni) can effectively unify both semantic and geometric matching capabilities learned by MATCHA-Light, giving a single feature at slight decrease in matching performance across tasks. When we further finetune such unified feature with joint geometric and semantic matching supervsion, we observe significant drop in semantic matching performance. We consider such behavior is mainly caused by the imbalanced training data between geometric and semantic matching. Compared to training individual descriptors, such data limitation imposes more challenges for balancing the two tasks when training a single unified descriptor. Therefore, we finally opt for simple concatenation as our mechanism to unify different types of foundation feature representations. It turns out to be highly effective also when combing the complementary semantic knowledge learned by DINOv2 into MATCHA.

Computation analysis. We report the runtime, FLOPs, and memory usage of our models on input images with various sizes during inference in Tab. 3. Our analysis shows that the primary runtime bottleneck is diffusion feature extraction, while the overhead introduced by our feature fusion network is negligible. Our approach can benefit from advances in efficient diffusion model inference. Alternatively, MATCHA features could be distilled into a more lightweight model.

Model	Params. M	Input $H \times W$	Runtime (ms) DIFT$_{\times 2}$	DINOv2	**FusionNet**	Total	FLOPs $\times 10^{12}$	Memory GB
MATCHA	1282.8	256 × 256	156.3	9.4	6.8	172.5	2.25	7.59
		512 × 512	687.3	30	17	734.3	8.79	11.36
		1024 × 768	2687.3	141.9	101.3	2930.5	26.24	21.43
MATCHA-Light	978.4	256 × 256	161.7	0	6.5	168.2	2.03	6.46
		512 × 512	702.9	0	15.8	718.7	7.96	10.23
		1024 × 768	2715.7	0	99.9	2815.6	23.77	20.29

Table 3. Computation analysis on an NVIDIA RTX 5880 GPU.

C. Visualization

Finally, we provide visualization for different feature models through their feature similarity distribution as well as the established correspondences across different scenes. We compare MATCHA-Light and MATCHA to MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16] and the supervised DIFT (DIFT.S).

Heatmap. In Fig. 1, we visualize the heatmaps and predicted matches produced by different methods, starting from a given source point. The heatmaps represent the normalized cosine similarity between the features extracted at the source point and every pixel in the target image. DISK [17], as a local feature method, focuses primarily on local texture regions, often resulting in poor matches in scenes with repetitive structures or semantically similar content. MASt3R [6], despite being trained on a larger dataset, still exhibits similar limitations, providing suboptimal matches in these challenging scenarios. DINOv2 [10], on the other hand, excels in cases with single objects, producing sharp and localized heatmaps. However, its performance degrades in the presence of noisy backgrounds or repetitive structures, where it fails to generate accurate matches.

For DIFT.Uni, DIFT.S.Uni, and MATCHA-Light, we compute their heatmaps using concatenated geometric and semantic features. DIFT captures more low-level texture details, leading to high similarity scores in regions with repetitive patterns. DIFT.S.Uni improves upon DIFT.Uni by incorporating supervision, but it remains less robust to variations in semantic content due to its task-specific training. MATCHA-Light, with joint training and dynamic feature fusion, addresses these issues to some extent, providing more accurate matches for both repetitive textures and semantically rich content. However, as it shares the same diffusion-based features as DIFT.Uni and DIFT.S.Uni, it struggles with ambiguity in visually similar parts of the same object, such as the head and tail of an airplane.

Finally, MATCHA resolves these challenges by incorporating complementary object-level features from DINOv2. This integration significantly enhances its ability to disambiguate similar object parts and produce accurate matches even in complex scenes, making it the most robust method among the evaluated approaches.

Geometric matching. In Fig. 2 and Fig. 3, we evaluate geometric matching using relative camera pose estimation and RANSAC to identify inlier matches for both indoor and outdoor scenes. Geometric methods such as DISK [17] and MASt3R.E [6] primarily rely on low-level texture patterns, which limits their ability to handle repetitive textures

and capture high-level structures effectively. In contrast, DINOv2 [10] focuses on object-level features, capturing higher-level structures but yielding sparse matches due to its limited reliance on detailed textures. DIFT strikes a balance between low- and high-level information, yet its lack of geometric supervision reduces the number of inliers compared to its supervised counterpart, DIFT.S, MATCHA-Light, with its dynamic fusion mechanism, propagates high-level semantic knowledge to geometric features, resulting in improved inliers at both object- and patch-level. This ability is further enhanced in MATCHA, where additional features from DINOv2 are fused, leading to the highest number of inliers among the evaluated methods.

Semantic matching. As illustrated in Fig. 4, Fig. 5, Fig. 6, Fig. 7, and Fig. 8, we visualize both inliers and outliers across various objects to assess semantic matching performance. Local geometric features, such as DISK [17] and MASt3R.E, fail to establish meaningful semantic correspondences, as they primarily rely on low-level textures and patterns.

While feature models like DINOv2, DIFT, and DIFT.S demonstrate a coarse ability to capture semantic correspondences, they often struggle with utilizing low-level details for precise local discrimination, leading to inaccuracies in challenging scenarios. In contrast, MATCHA effectively integrates both geometric and semantic cues, achieving robust and accurate semantic matches even under extreme scale and viewpoint variations, outperforming other methods in these complex scenarios.

Temporal matching. We present visualizations of temporal matches in Fig. 9, Fig. 10, Fig. 11, and Fig. 12, evaluating the performance of various methods. In addition to previously discussed baselines, we include the unified DIFT feature variants, DIFT.Uni and DIFT.S.Uni, which demonstrate improved temporal matching compared to their specific geometric or semantic descriptors.

Overall, we observe that the local feature DISK performs the worst in handling highly dynamic objects, such as a jumping horse or moving bikes, due to its reliance on low-level patterns. MASt3R.E shows marginal improvement but is still outperformed by other methods. Among all approaches, MATCHA stands out as the most accurate and robust for temporal matching, effectively handling the challenges of dynamic scenes.

Failure cases. Despite its strengths, temporal matching remains a challenging task, as shown in Fig. 11 and Fig. 12. All methods struggle in scenarios where repetitive patterns in the background coincide with extreme scale and viewpoint changes caused by object motion. These limitations highlight the need for further research to improve the robustness and accuracy of temporal matching in highly dynamic and complex scenes.

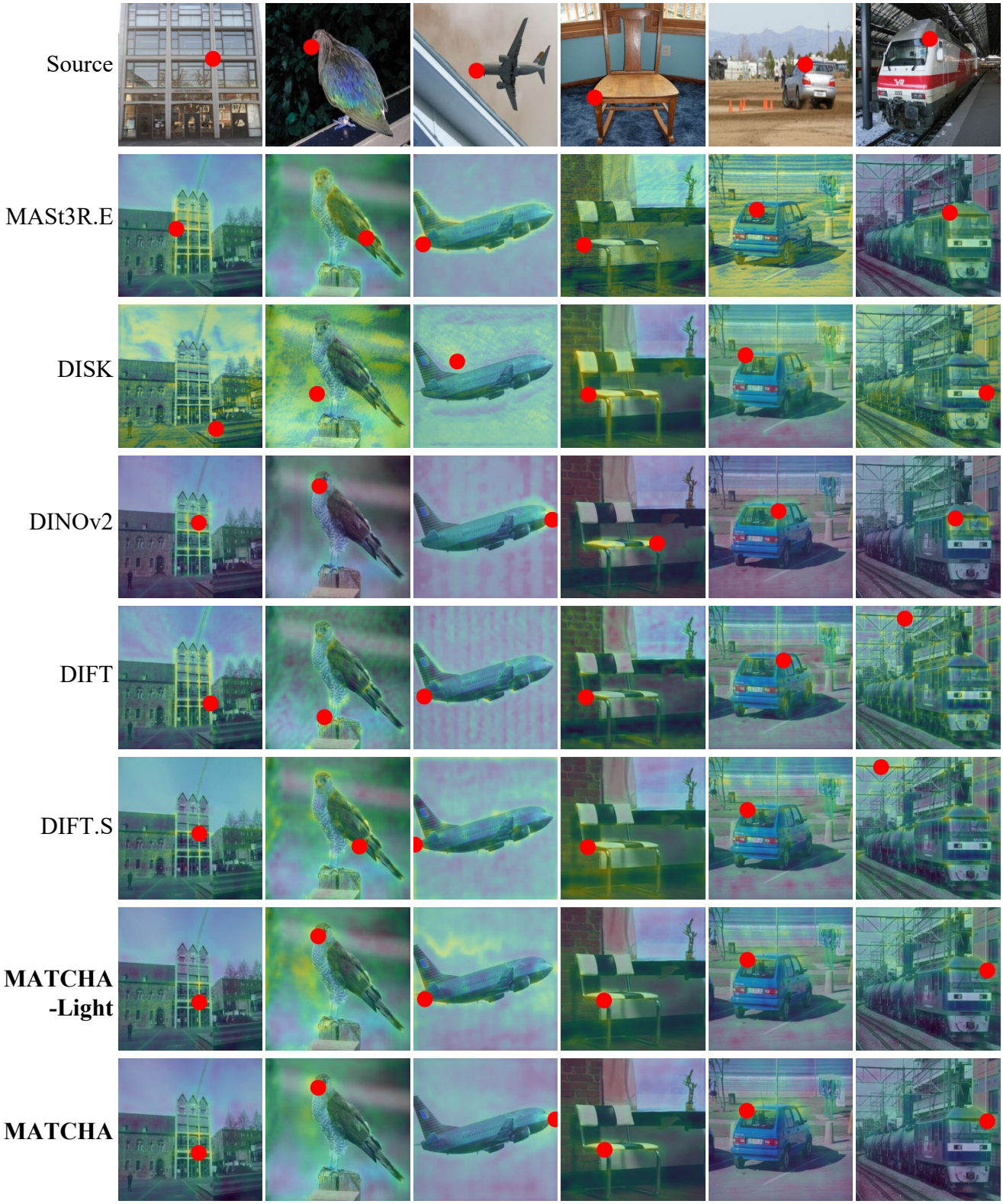

Figure 1. **Visualization of heatmap.** Given a source point (top), we visualize the heatmap and predicted matches of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

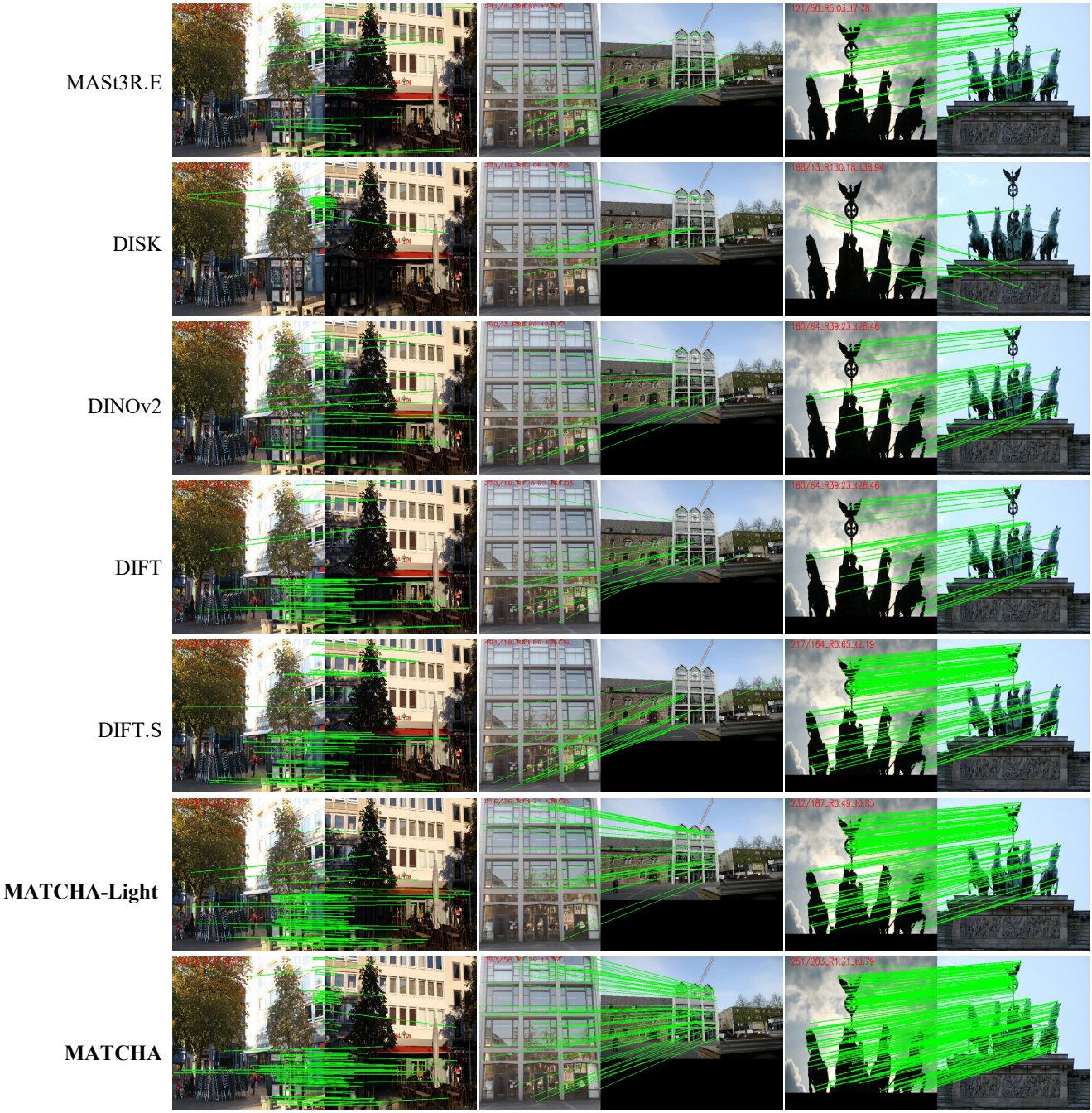

Figure 2. **Geometric matches on outdoor scenes.** We visualize the inliers after RANSAC of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA. DISK produces many inliers on local patches but is not robust to repetitive structures. MASt3R and DINOv2 focus more on structures and give close performance. DIFT works better than DINOv2 especially on regions with rich textures. With geometric supervision, DIFT.S improves the performance of DIFT. MATCHA-Light is able to find correct matches from both local patches and structures because of dynamic fusion and this ability is further enhanced by fusing features of DINOv2.

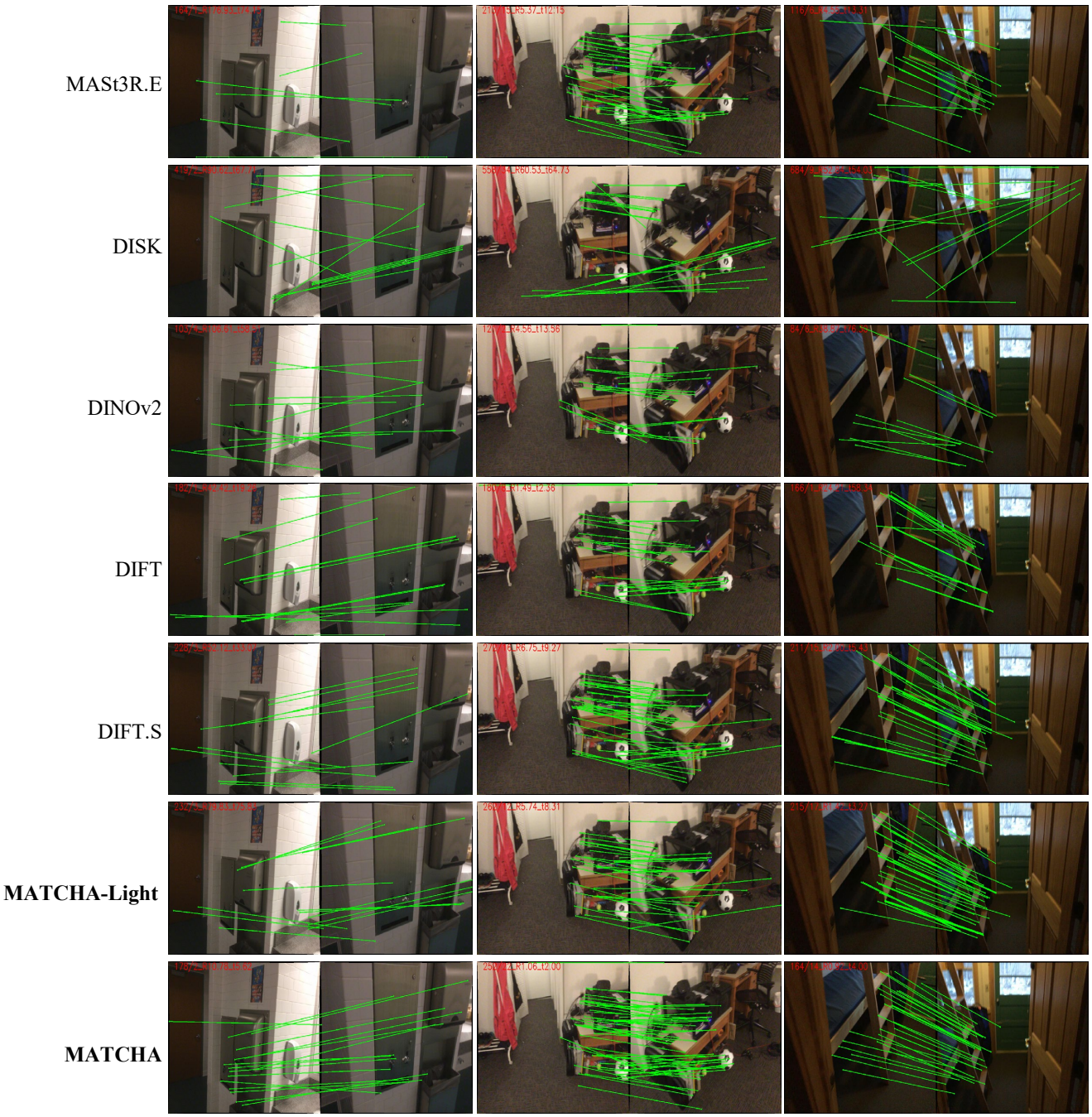

Figure 3. **Geometric matches on indoor scenes.** We visualize the inliers after RANSAC of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA. Almost all previous methods fail to find sufficient inliers on scenes with repetitive structures except MATCHA which fuses both low and high-level information. Additionally, MATCHA is able to produce more inliers in scenes with rich textures (*middle column*).

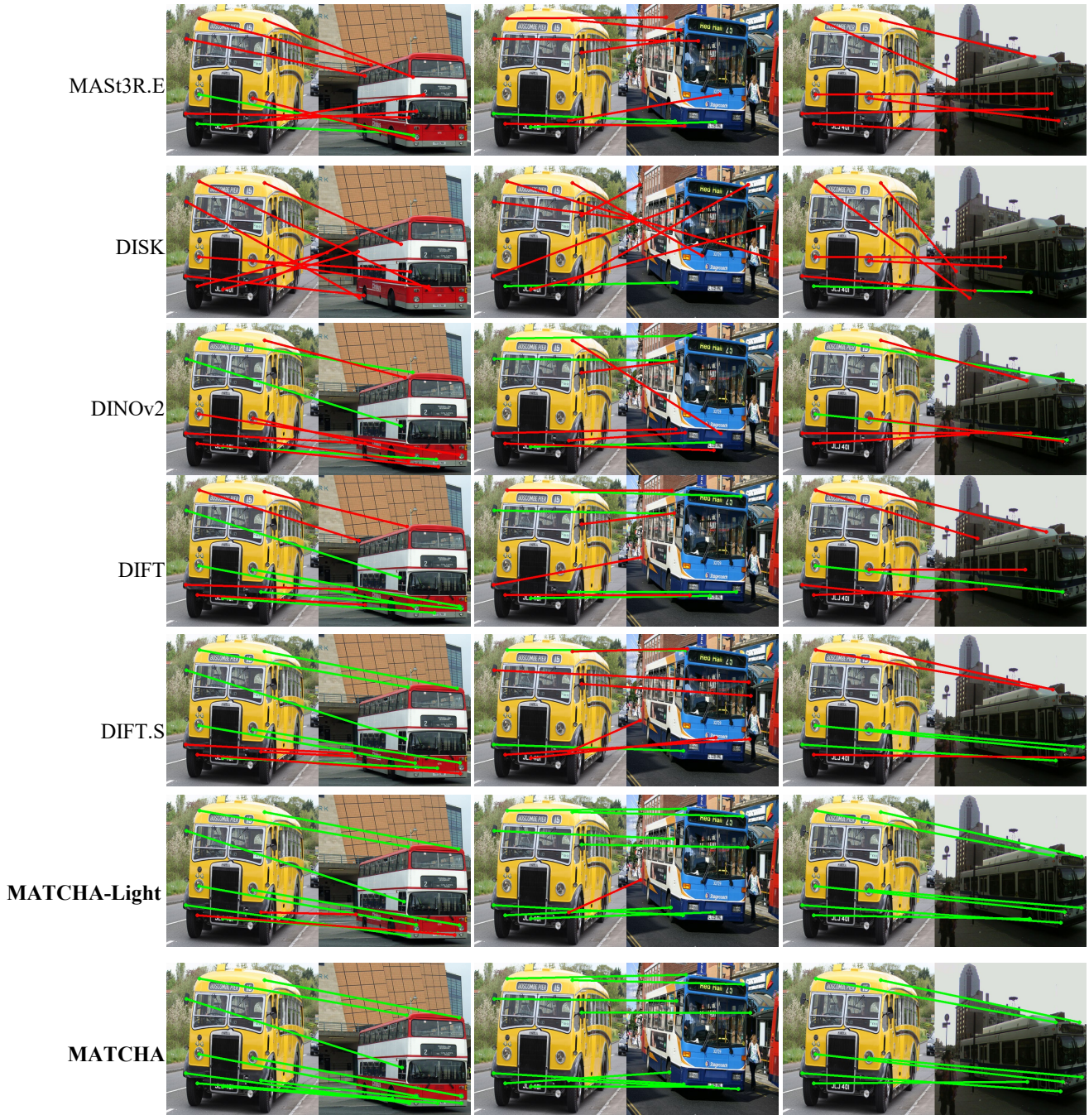

Figure 4. **Semantic matches on bus category.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

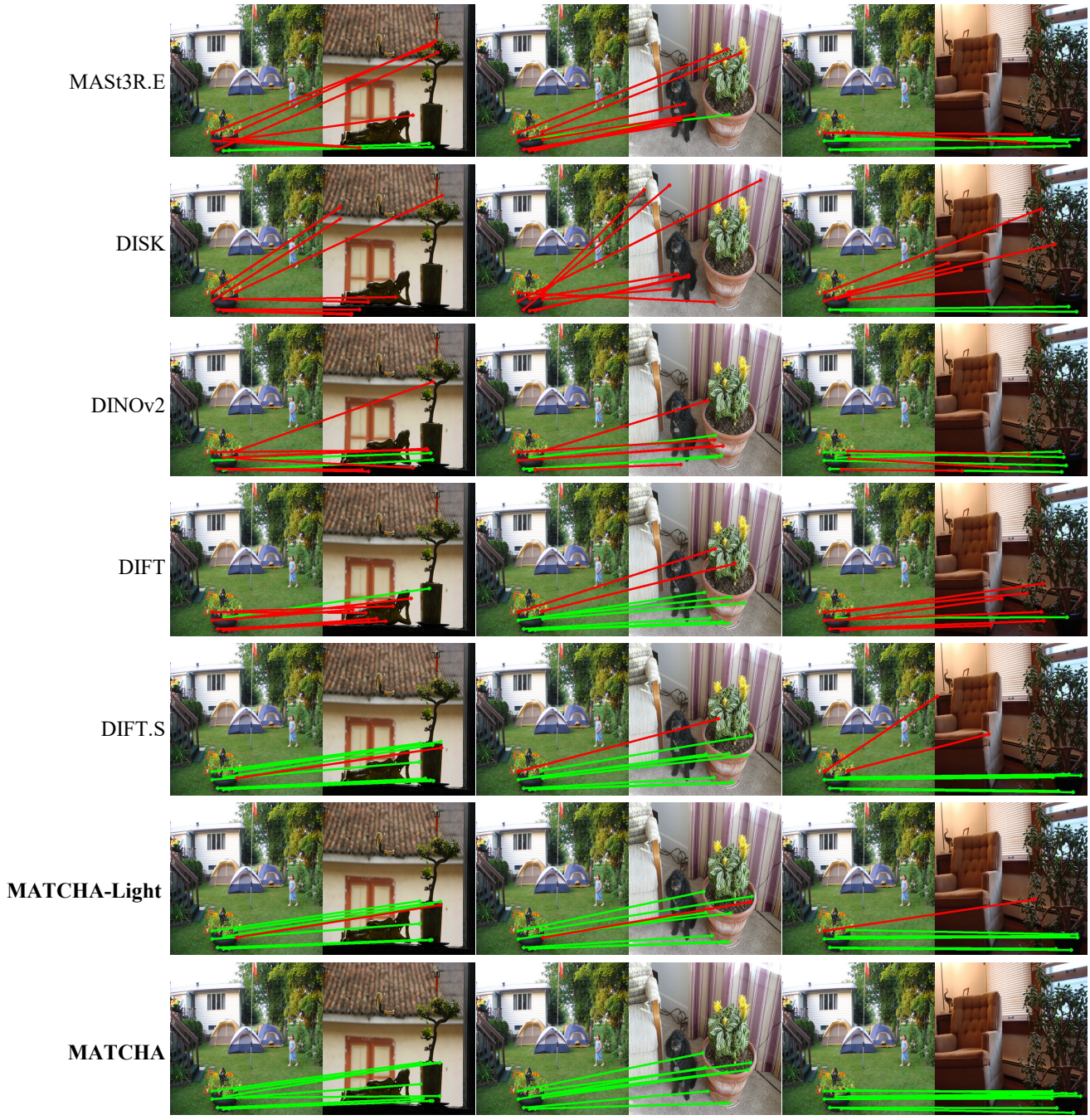

Figure 5. **Semantic matches on plant category.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

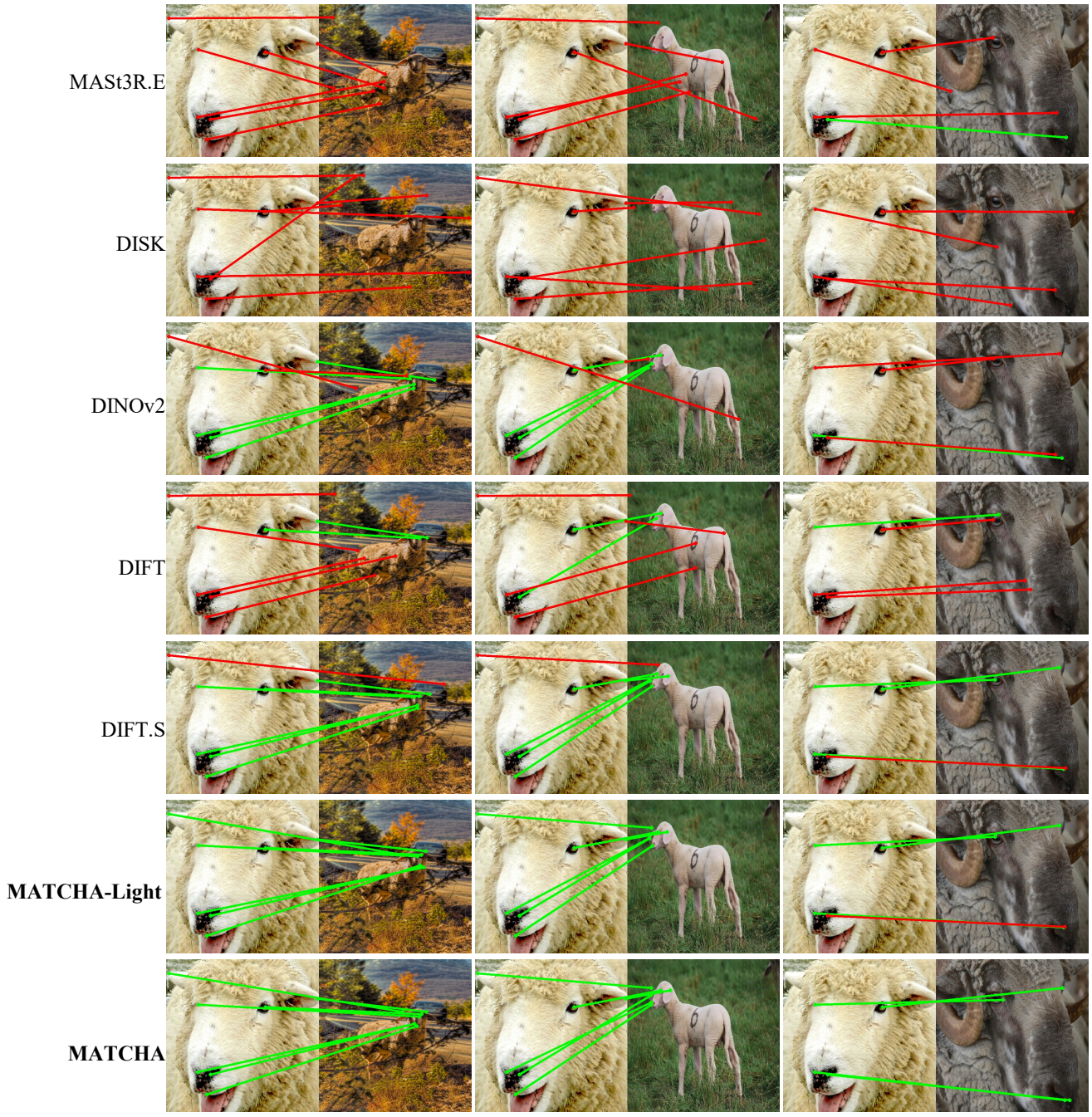

Figure 6. **Semantic matches on sheep category.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

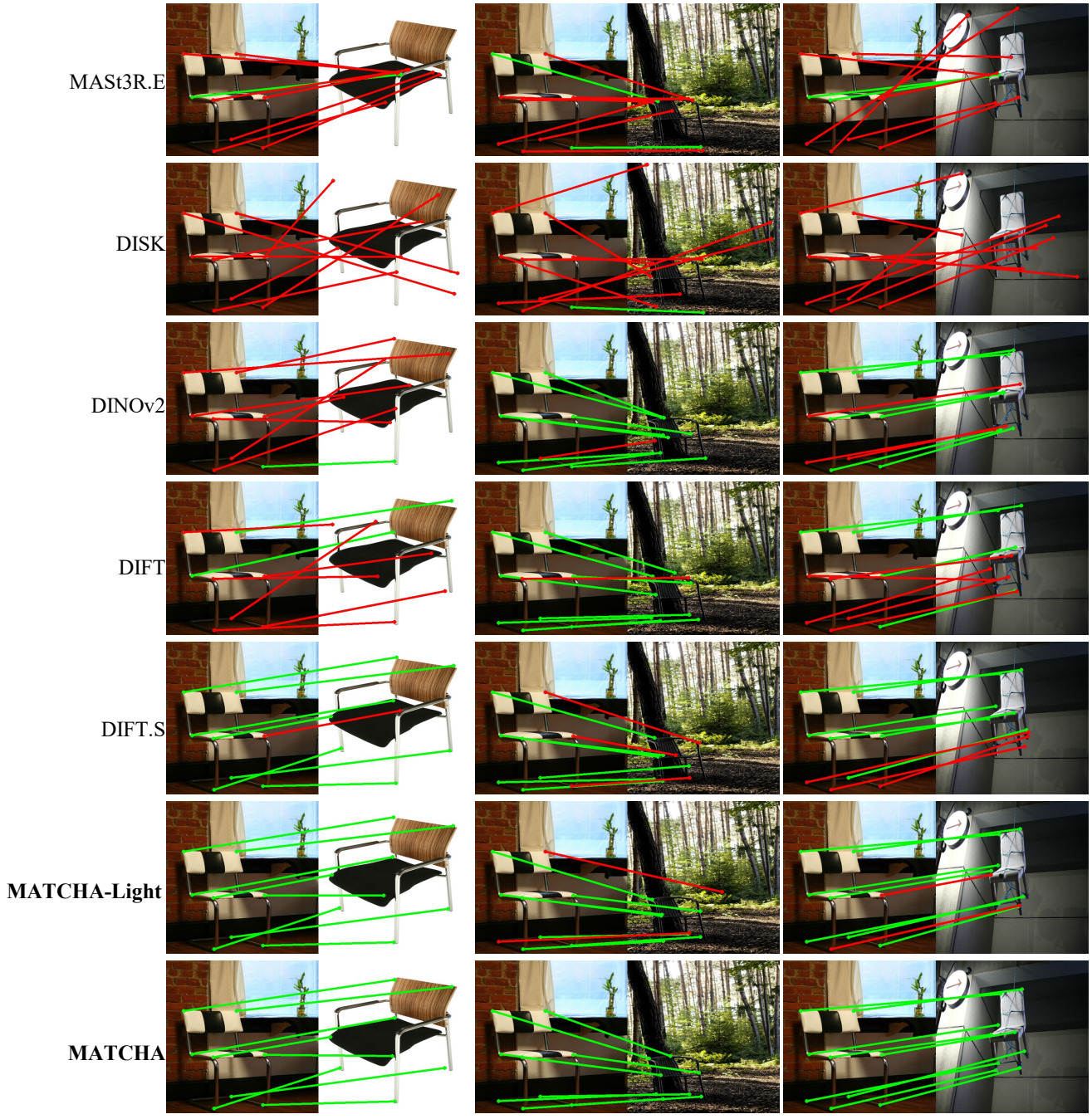

Figure 7. **Semantic matches on chair category.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

Figure 8. **Semantic matches on motorbike category.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT [16], DIFT.S (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

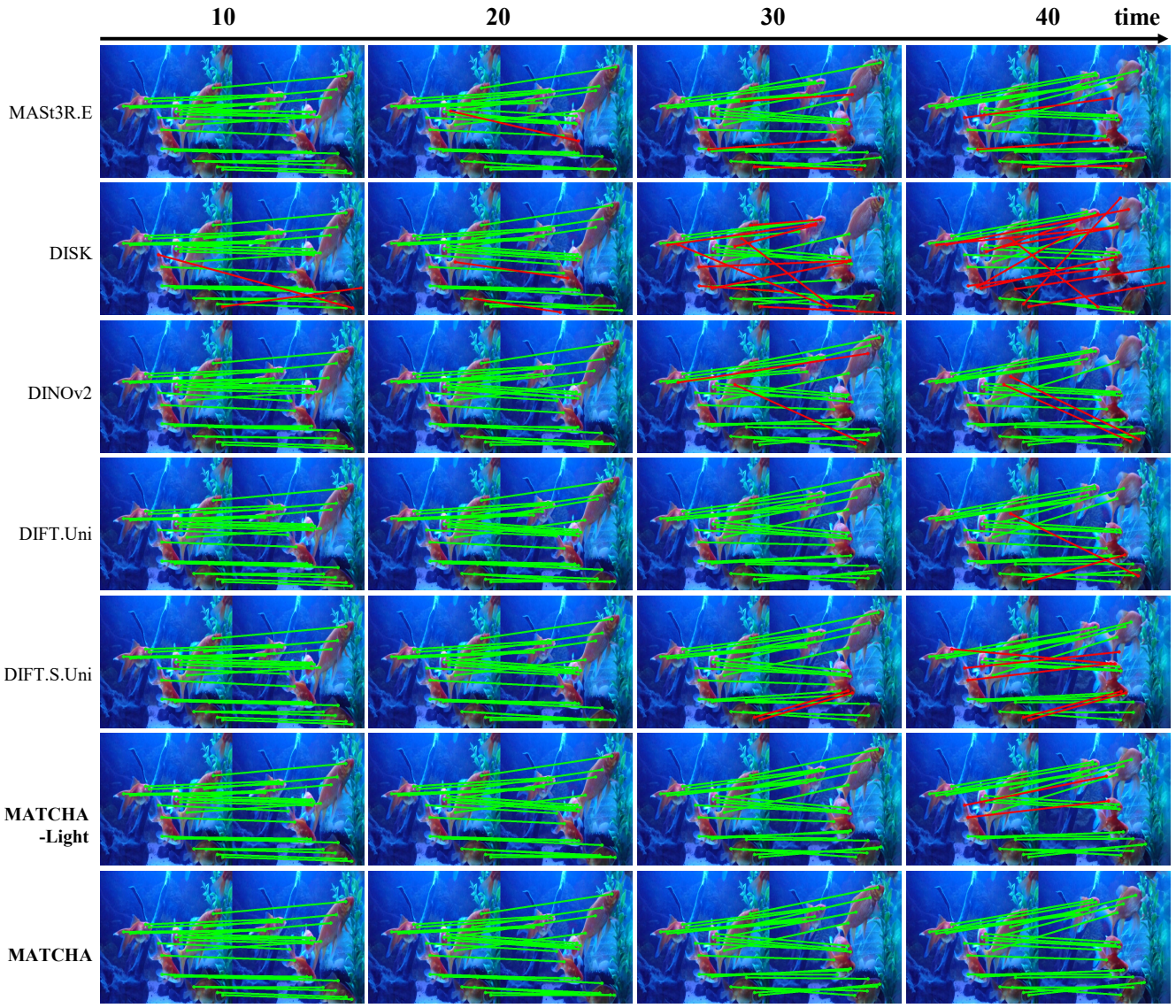

Figure 9. **Temporal matches on goldfish sequence.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT.Uni [16], DIFT.S.Uni (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

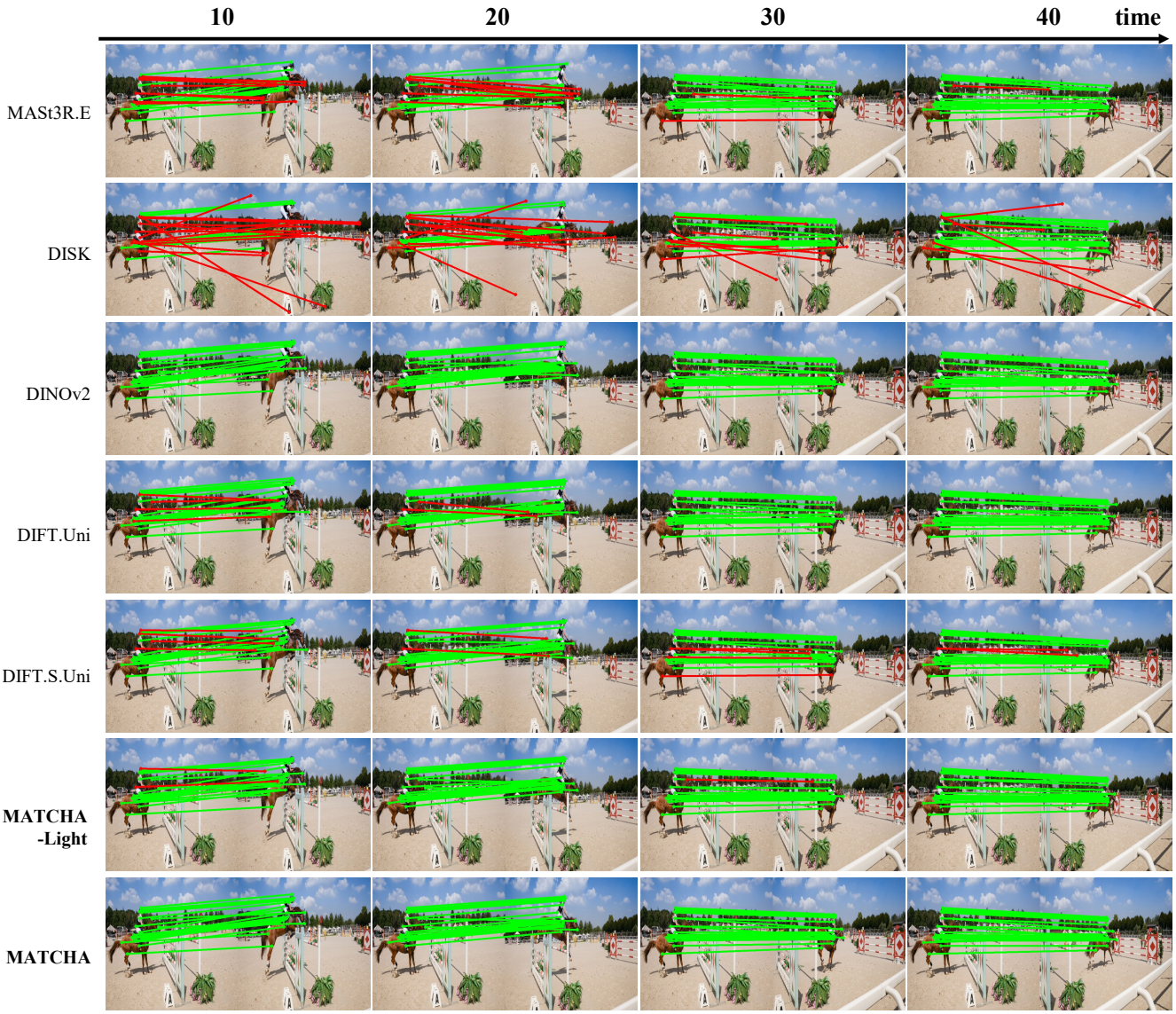

Figure 10. **Temporal matches on horsejumphigh sequence.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT.Uni [16], DIFT.S.Uni (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

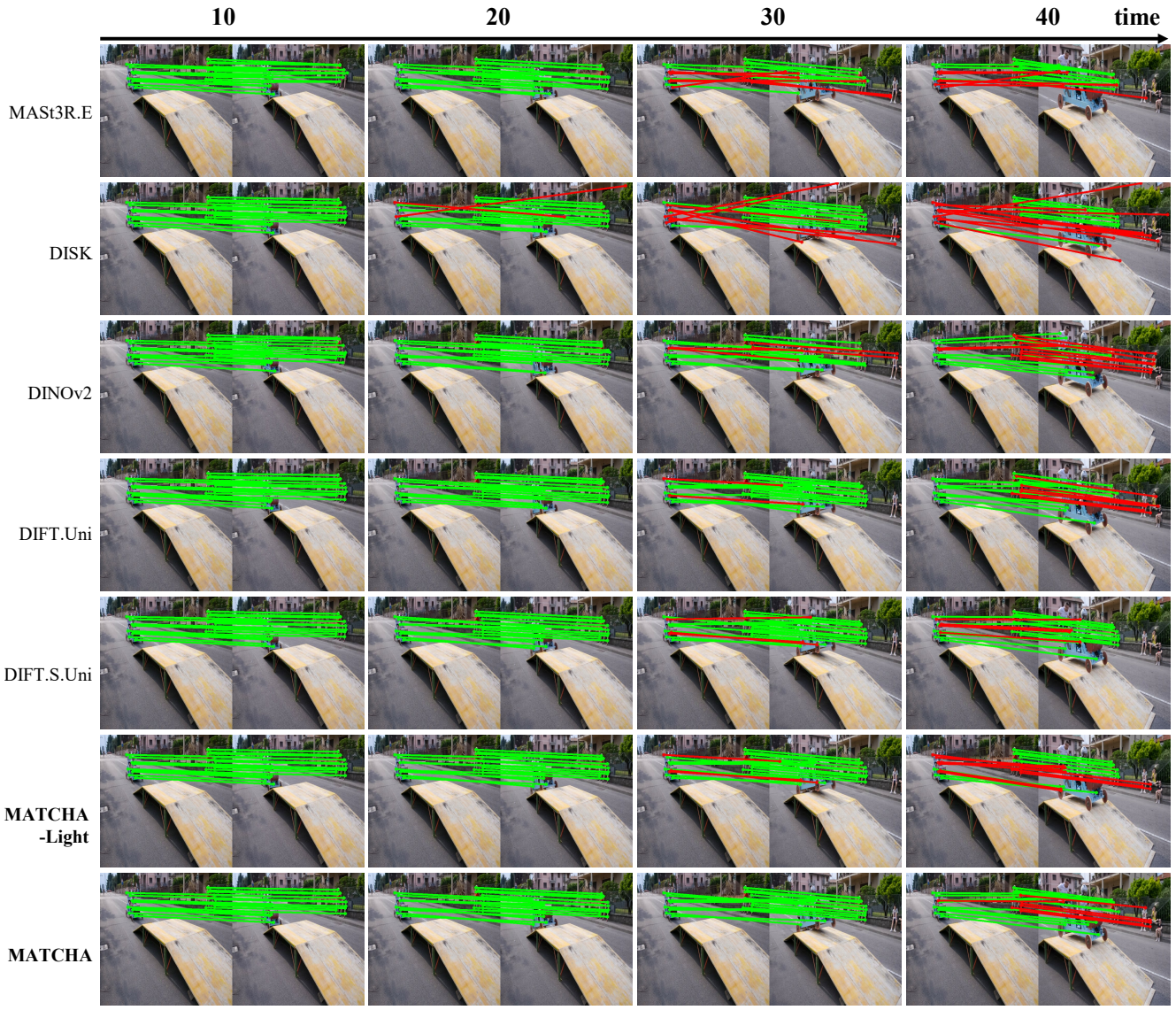

Figure 11. **Temporal matches on soapbox sequence.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT.Uni [16], DIFT.S.Uni (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

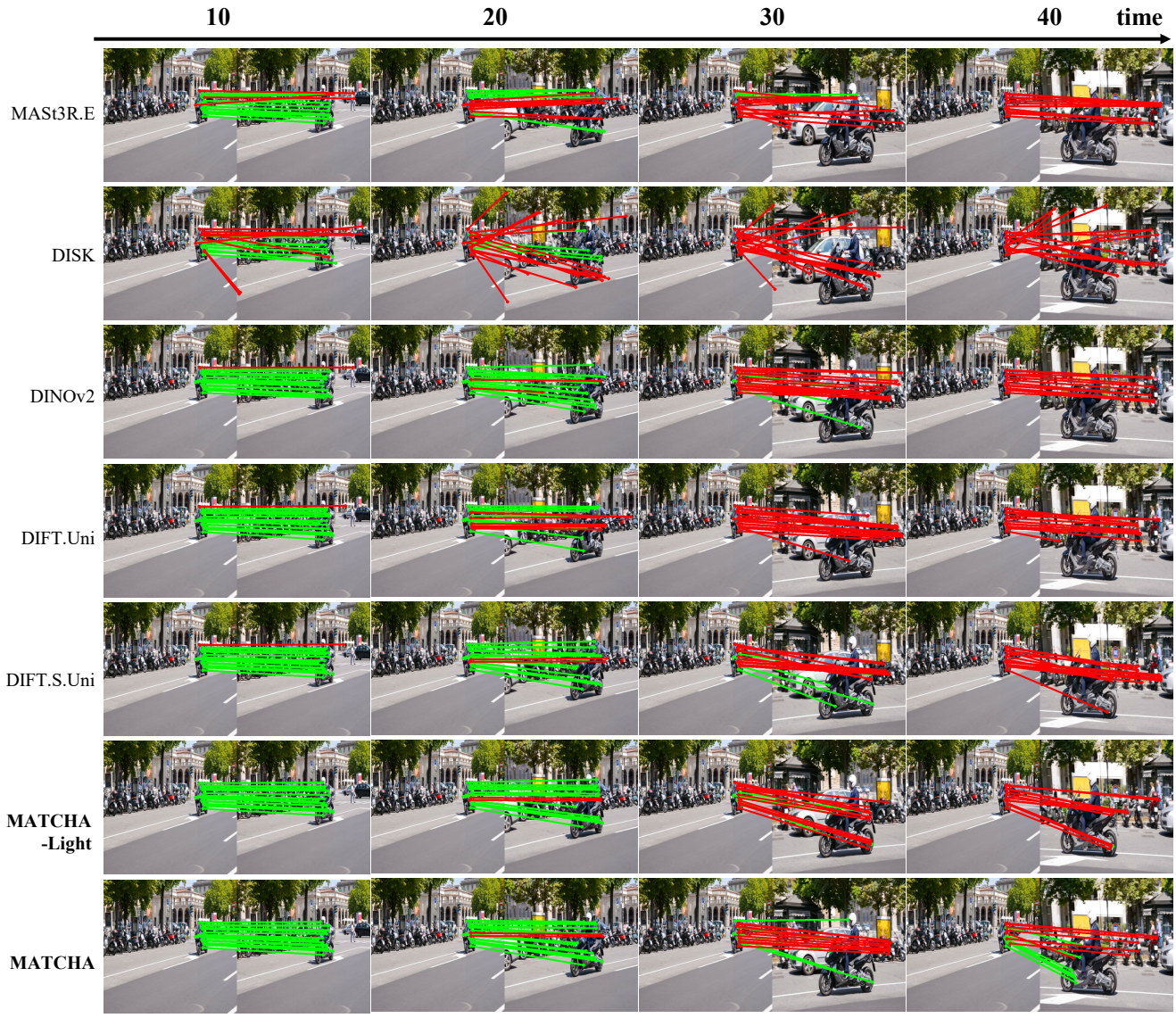

Figure 12. **Temporal matches on scooterblack sequence.** We visualize the inliers and outliers of MASt3R.E [6], DISK [17], DINOv2 [10], DIFT.Uni [16], DIFT.S.Uni (fully supervised version of DIFT), and our models MATCHA-Light and MATCHA.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[2] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1

[3] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 2

[4] 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. 1

[5] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019. 1

[6] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *European Conference on Computer Vision*, 2024. 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

[7] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1

[8] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[9] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. 1

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2

[12] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 1

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[14] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 1

[15] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1

[16] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

[17] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

[18] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 1

[19] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3076–3085, 2024. 1