Supplementary Materials
Patch Your Matcher: Correspondence-Aware Image-to-Image Translation Unlocks Cross-Modal Matching via Single-Modality Priors

Anton Frolov
Bauhaus-Universität Weimar
{firstname}.{lastname}@uni-weimar.de

Volker Rodehorst
Bauhaus-Universität Weimar
{firstname}.{lastname}@uni-weimar.de

1. Limitations

Synthetic training data. Our models are trained exclusively on synthetically generated imagery (MD-Syn). We observe that some of such samples are not physically plausible or not faithful to the real target modalities (*e.g.*, hallucinated infrared structure; overly sparse or overly dense event frames), which is particularly harmful for I2I translation. Future work should pursue higher-fidelity training data and more realistic evaluation scenarios such as cross-modal SfM, sensor relocalization, and SLAM.

Architecture and hyperparameters. Our designs for generator and discriminator take inspiration in prior work (*i.e.*, DP-GAN-like generator and ELoFTR-style transformer module), however we do not perform a systematic architecture or hyperparameter search. Multi-objective optimization (accuracy vs. FLOPs/latency/parameters), automated hyperparameter optimization and neural architecture search could discover Pareto-superior trade-offs and produce lighter models for resource-constrained settings.

Hard-sample mining. During training, average discriminator scores spread as the generator learns to align easy pairs, implicitly yielding per-pair difficulty estimates. We currently do not exploit this signal sufficiently. Curriculum learning, uncertainty-aware weighting, or explicit hard-negative mining could focus learning on challenging correspondences and improve overall robustness.

Many-to-one multi-modal translation. MINIMA [4] matcher variants demonstrate the added benefit of multi-modal training setting. While more computationally intensive, such training could yield significantly higher robustness and generalizability. Our PYM architecture can support such conditioning, but a rigorous study is beyond the present scope and orthogonal to our core contribution. Nonetheless, this remains a viable direction of future work.

Table 1. Performance and efficiency overview. Latency for batch size of 1, on AMD EPYC 7532 + 1 x NVIDIA A100 GPU.

Variant	#Params (M)	FLOPs (G)	VRAM, GB	Latency, ms
PYM	55.50	499	3.67	17.24
LG	13.15	357	4.19	35.93
LG+PYM	73.94	856	7.56	53.22
EloFTR	15.05	477	3.01	40.53
EloFTR+PYM	75.84	976	6.67	57.65
RoMA	111.29	3,581	34.54	265.83
RoMA+PYM	172.08	4,080	38.21	286.98

2. Performance and efficiency

While specialist PYM models converge faster than generalist MINIMA, I2I translation layers bear a significant overhead cost, that also partially persists during inference. In table Tab. 1, we provide summary statements for the parameter counts, FLOPs, forward/backward pass size, and inference times on an Nvidia A100 GPU. We consider scenarios of sparse [3] and semi-dense [7] matching with and without translation, as well as translation-only scenario. Notably, despite no dedicated optimizations, and the significant overhead costs, the resulting performance is real-time.

3. Additional qualitative evaluations

To enable a more descriptive comparison against CycleGAN [8] and Pix2Pix [2], we provide qualitative side-by-side comparisons for MDSyn modalities (see Fig. 1,Fig. 2,Fig. 3,Fig. 4).

To diagnose failure modes of ELoFTR$_{\text{PYM}}$, we present qualitative examples arranged by increasing precision for each of the evaluated datasets: MDSyn [4] IR-G (Fig. 5), MDSyn [4] D-G (Fig. 6), MDSyn [4] N-G (Fig. 7), MDSyn [4] E-G (Fig. 8), METU-VisTIR [5] IR-G (Fig. 9), DIODE [6] D-G (Fig. 10), DSEC [1] E-G (Fig. 11).

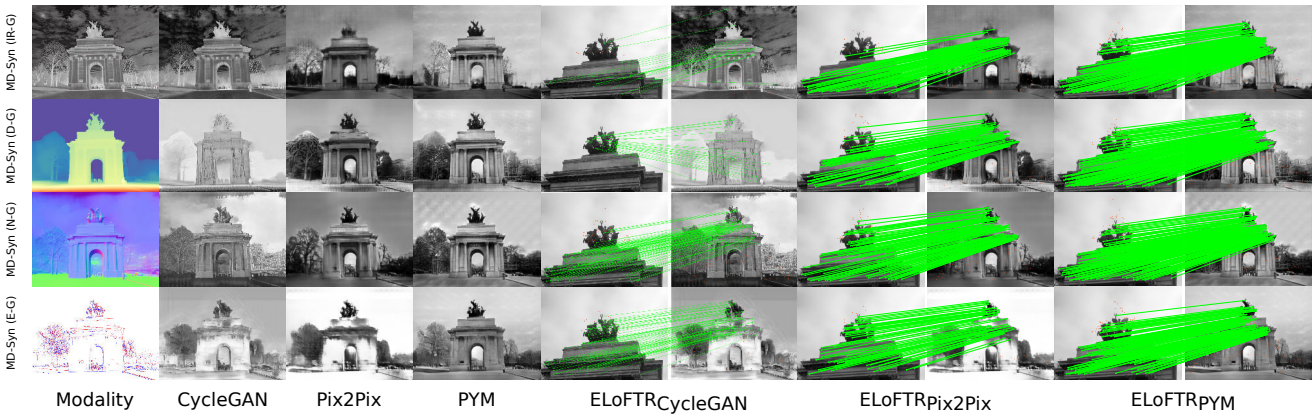

Figure 1. Matching results on MD-Syn. Sample 1.

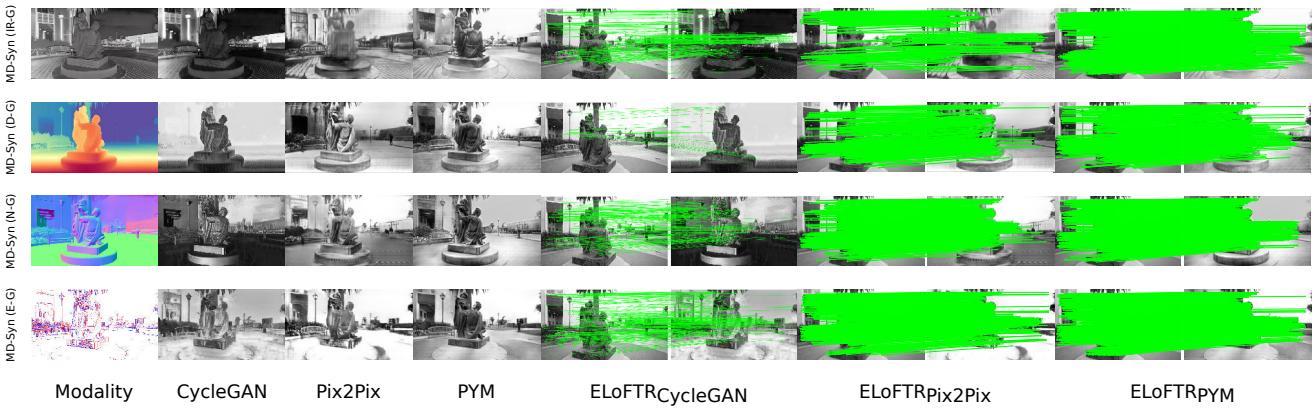

Figure 2. Matching results on MD-Syn. Sample 2.

References

[1] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6:4947–4954, 2021. 1

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[3] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. pages 17581–17592, Paris, France, 2023. IEEE. 1

[4] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. Minima: Modality invariant image matching, 2024. 1

[5] Önder Tuzcuoglu, Aybora Köksal, Bugra Sofu, Sinan Kalkan, and A. Aydin Alatan. Xoftr: Cross-modal feature matching transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 4275–4286. IEEE, 2024. 1

[6] Igor Vasiljevic, Nicholas I. Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor depth dataset. *CoRR*, abs/1908.00463, 2019. 1

[7] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21666–21675. IEEE, 2024. 1

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017. 1

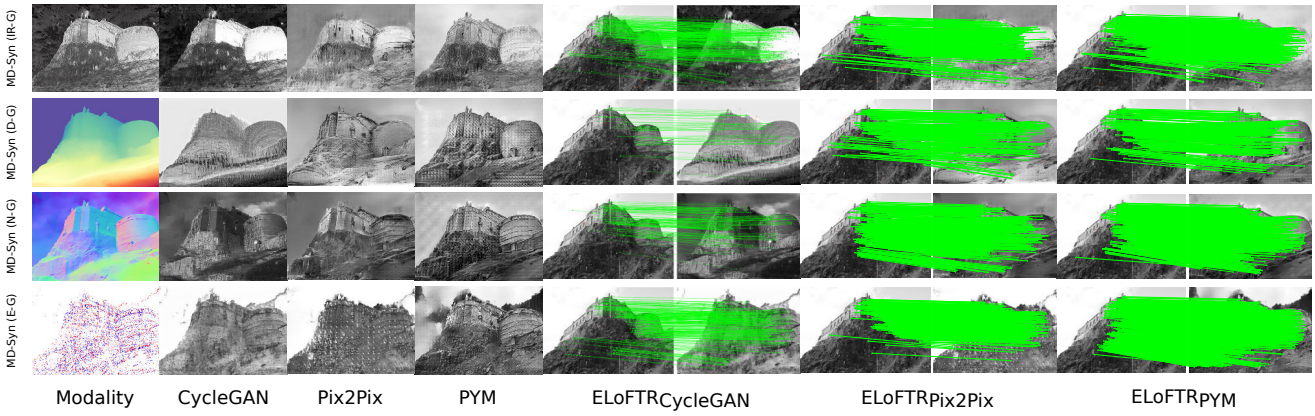

Figure 3. Matching results on MD-Syn. Sample 3.

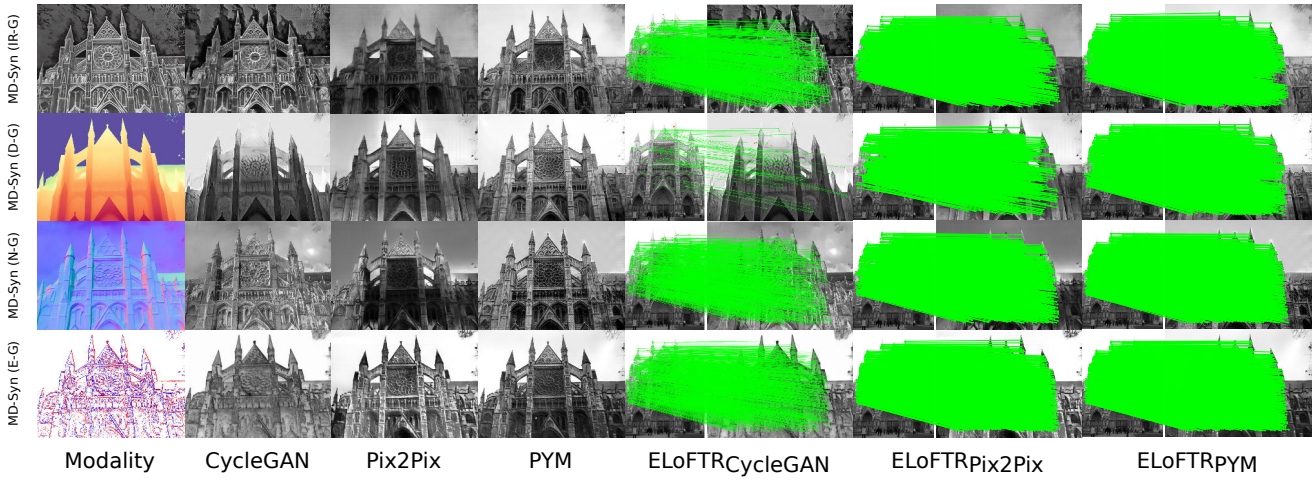

Figure 4. Matching results on MD-Syn. Sample 4.

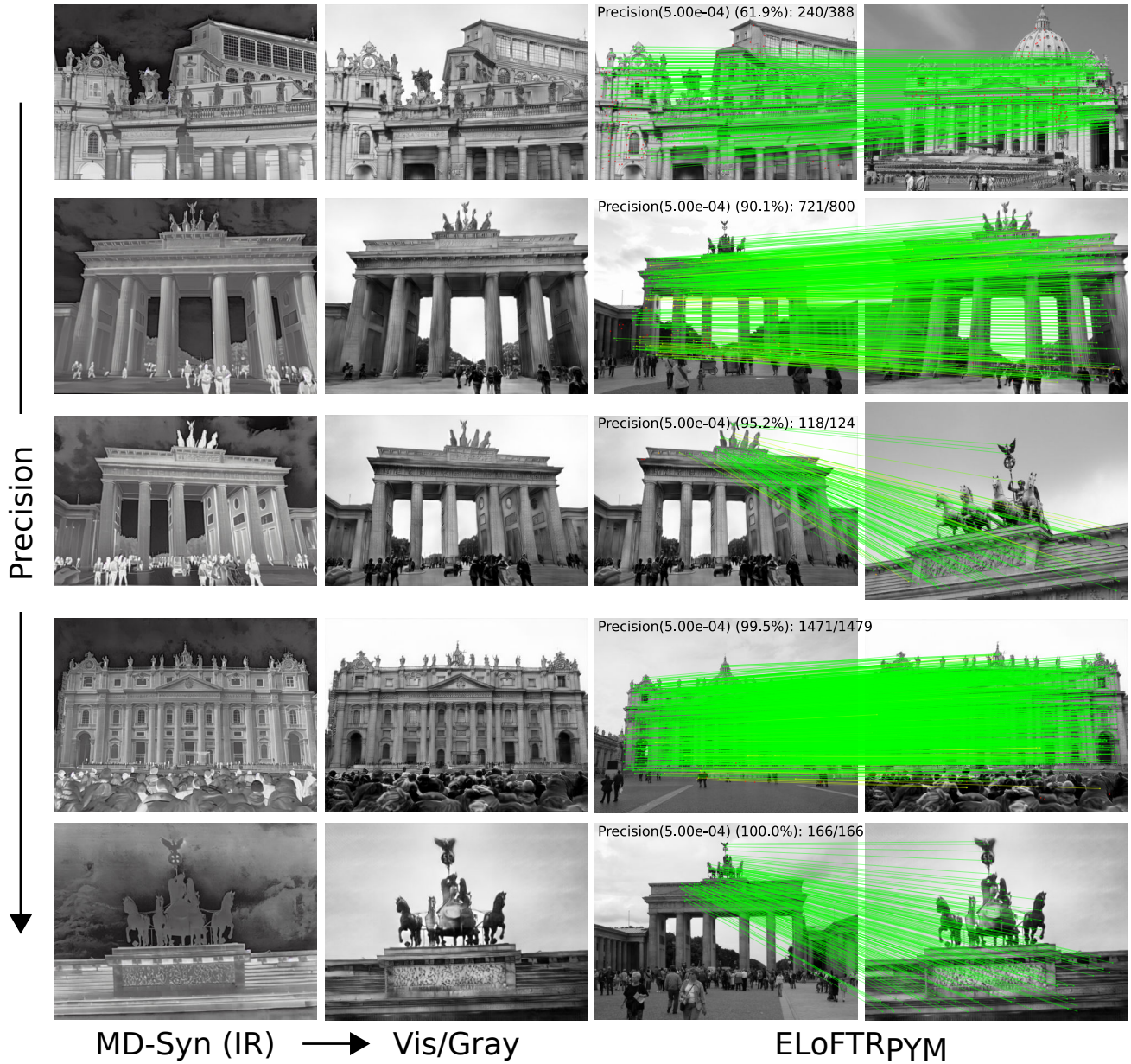

Figure 5. Matching results on MD-Syn (IR-G) with ELoFTR$_{\text{PYM}}$. From 61.9% to 100.0% prec @ 5×10^{-4}.

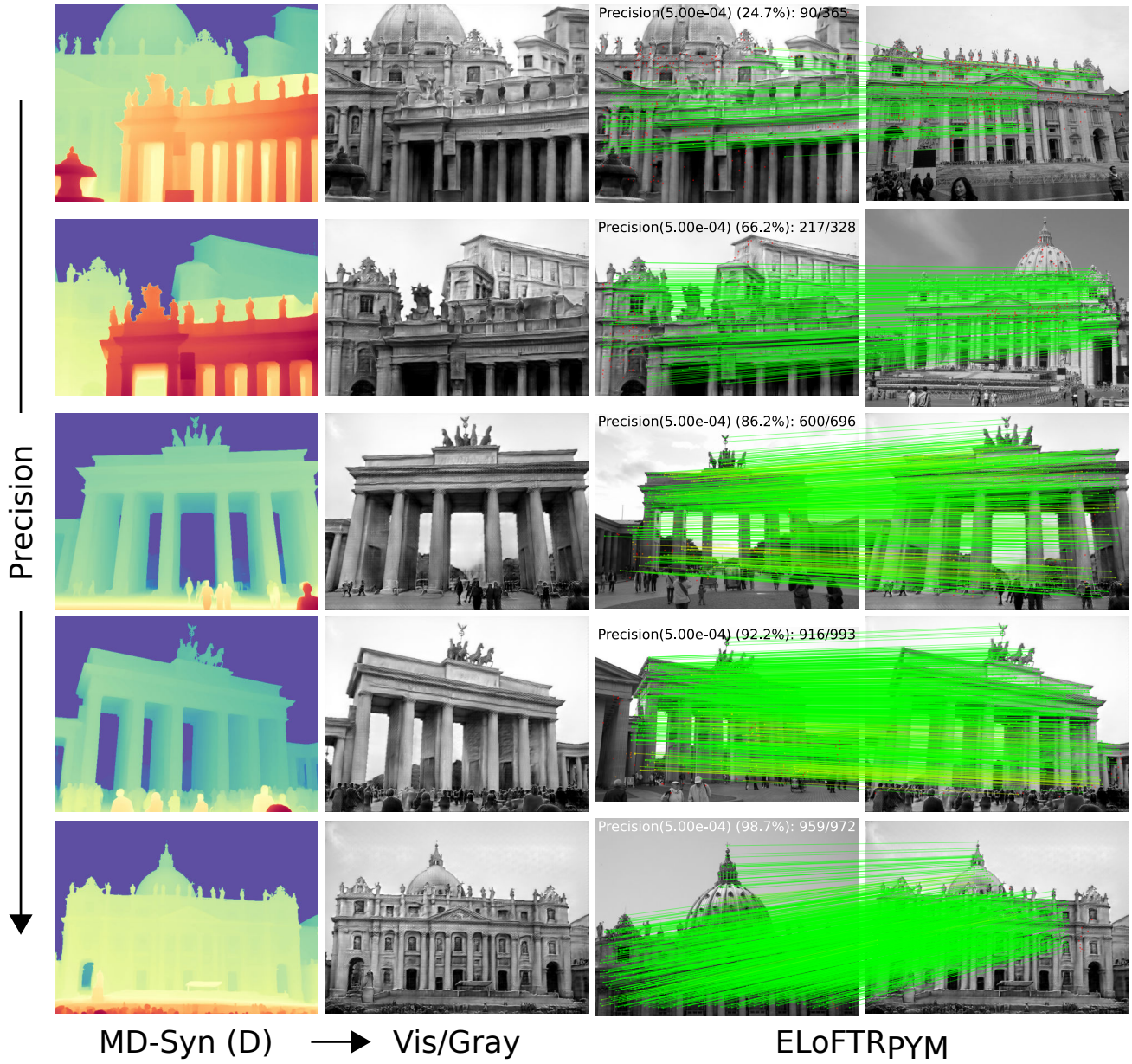

Figure 6. Matching results on MD-Syn (D-G) with ELoFTR$_{\text{PYM}}$. From 24.7% to 98.7% prec @ 5×10^{-4}.

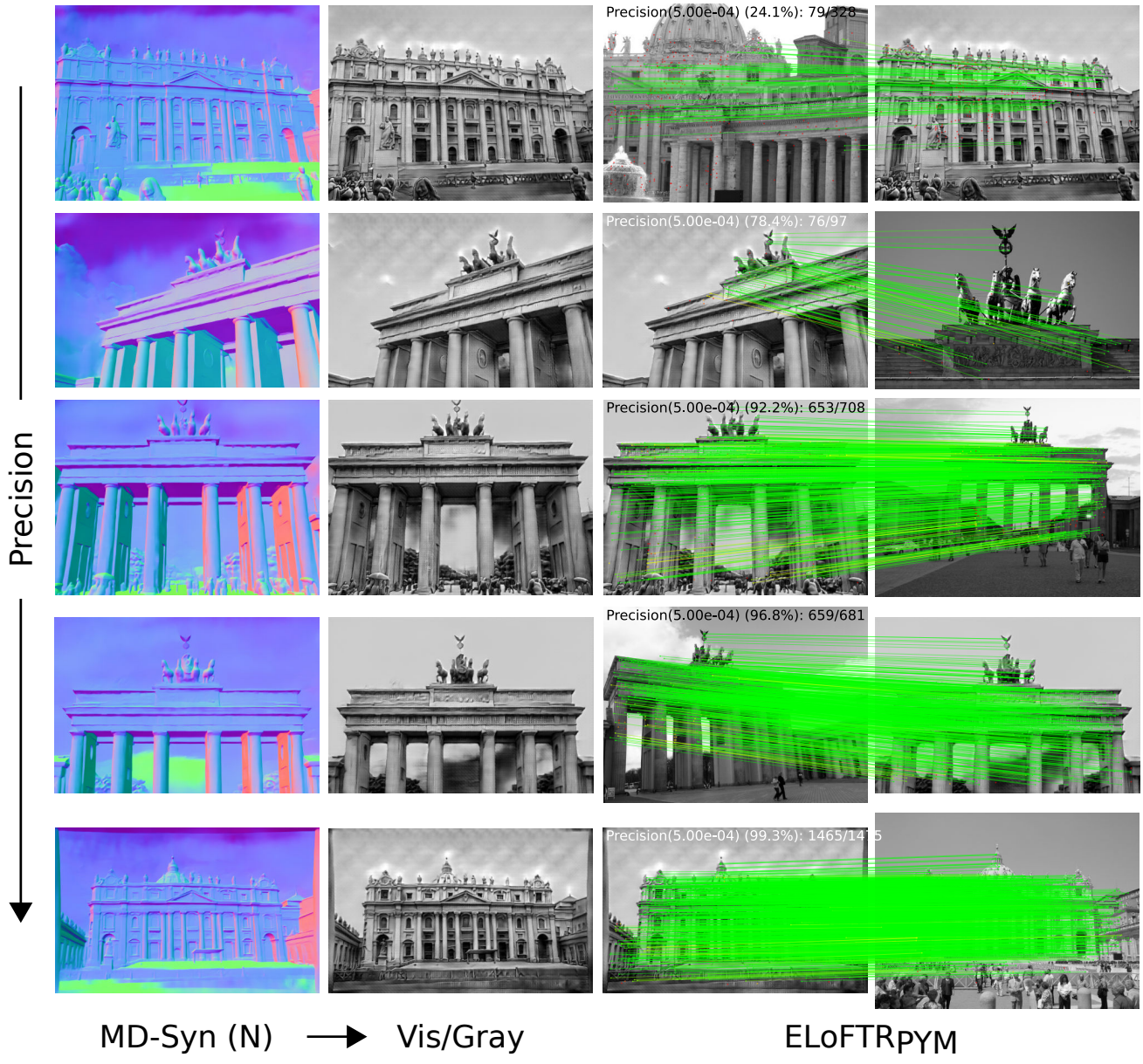

Figure 7. Matching results on MD-Syn (N-G) with ELoFTR$_{\text{PYM}}$. From 24.1% to 99.3% prec @ 5×10^{-4}.

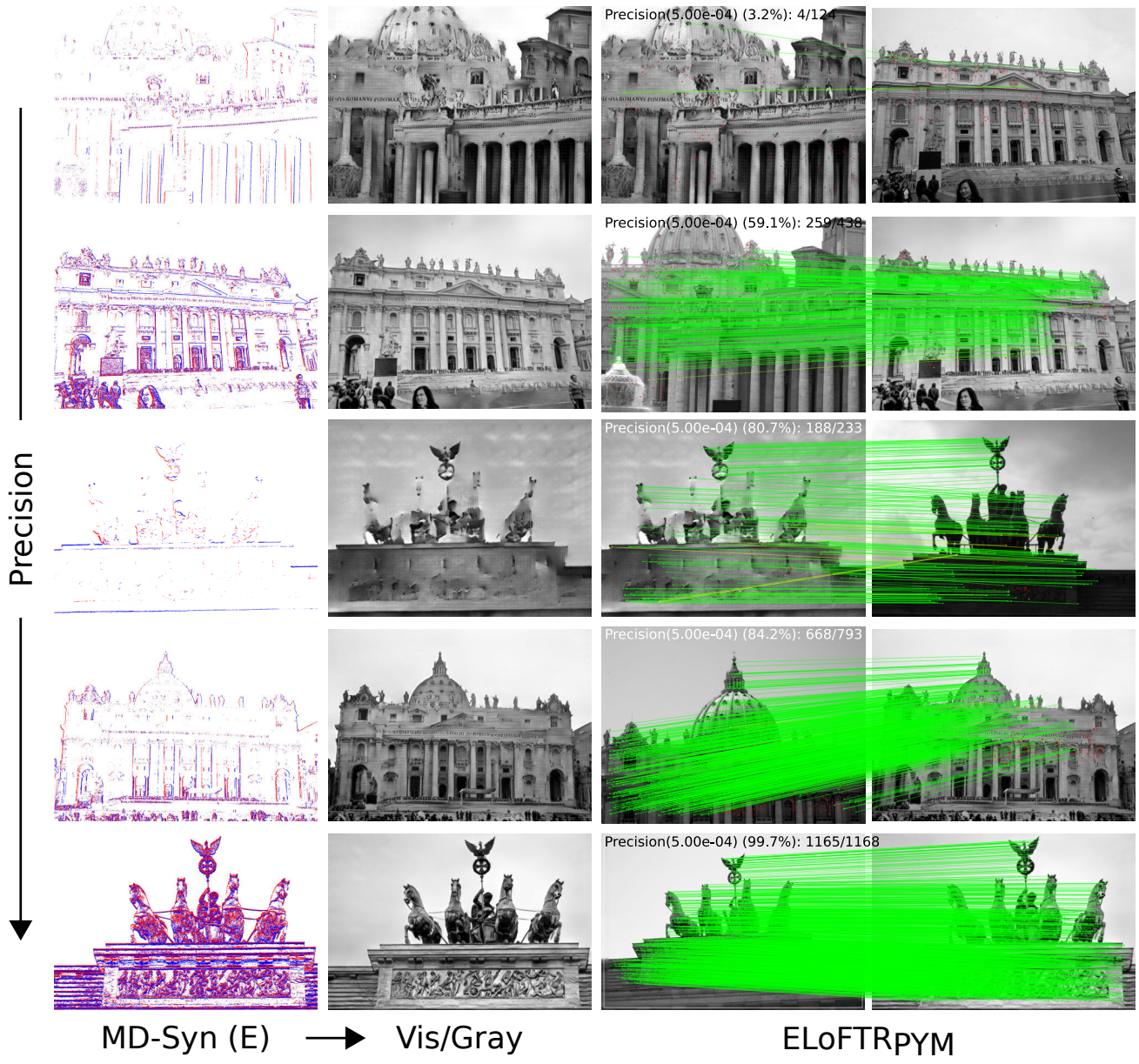

Figure 8. Matching results on MD-Syn (E-G) with ELoFTR$_{PYM}$. From 3.2% to 99.7% prec @ 5×10^{-4}.

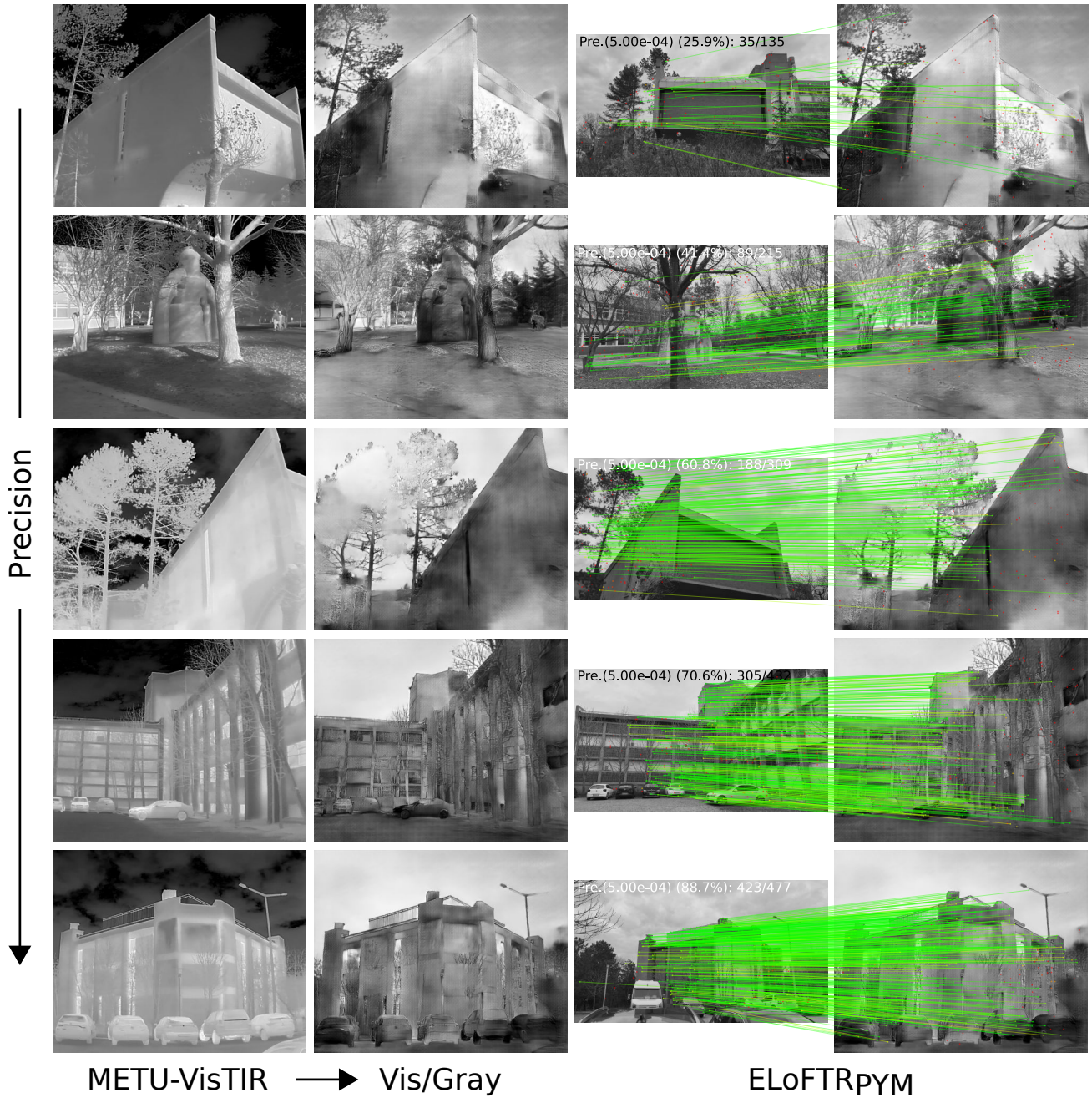

Figure 9. Matching results on METU-VisTIR (IR-G) with ELoFTR$_{\text{PYM}}$. From 25.9% to 88.7% prec @ 5×10^{-4}.

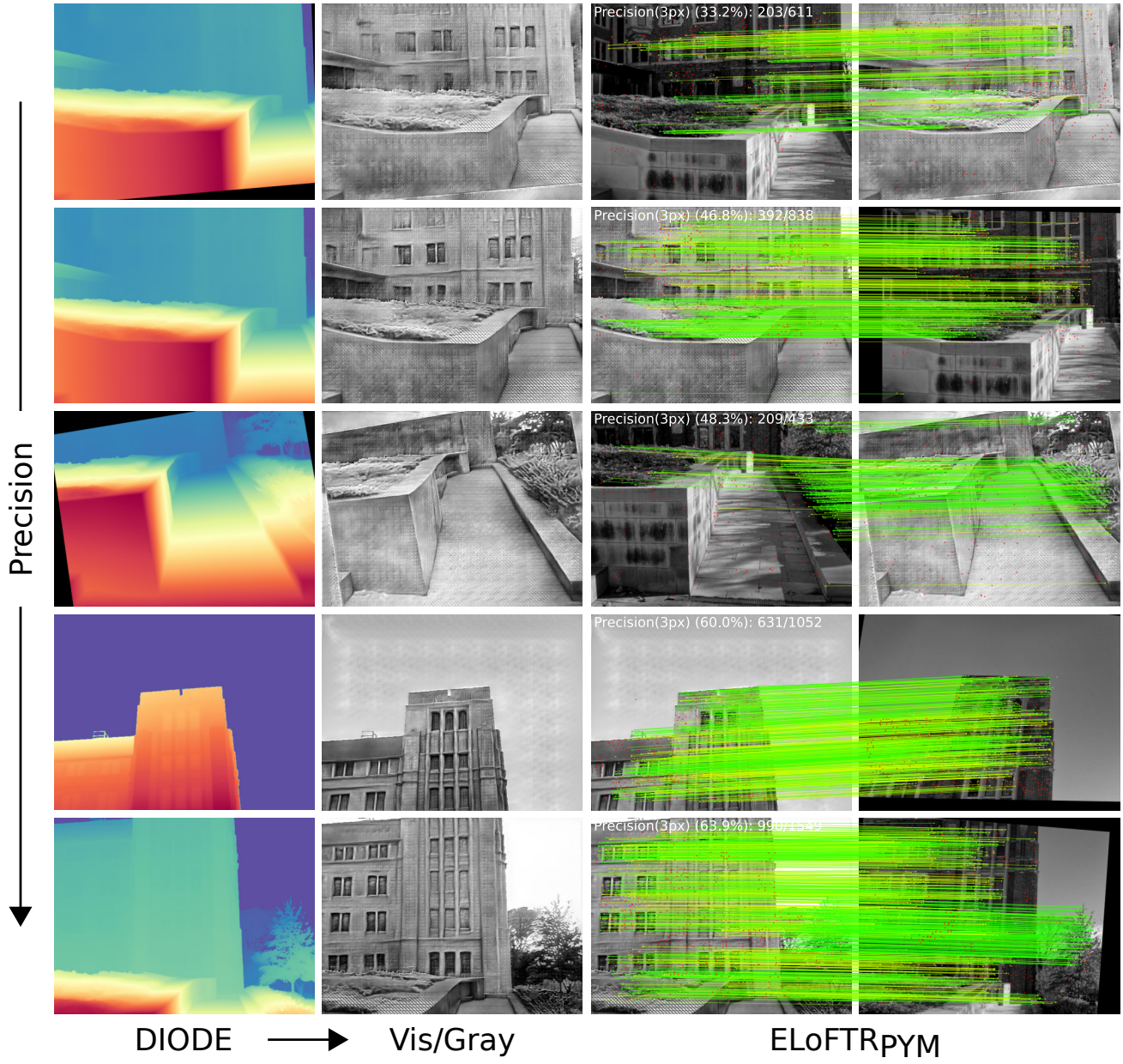

Figure 10. Matching results on DIODE (D-G) with ELoFTR$_{\text{PYM}}$. From 33.2% to 63.9% prec @ 3px.

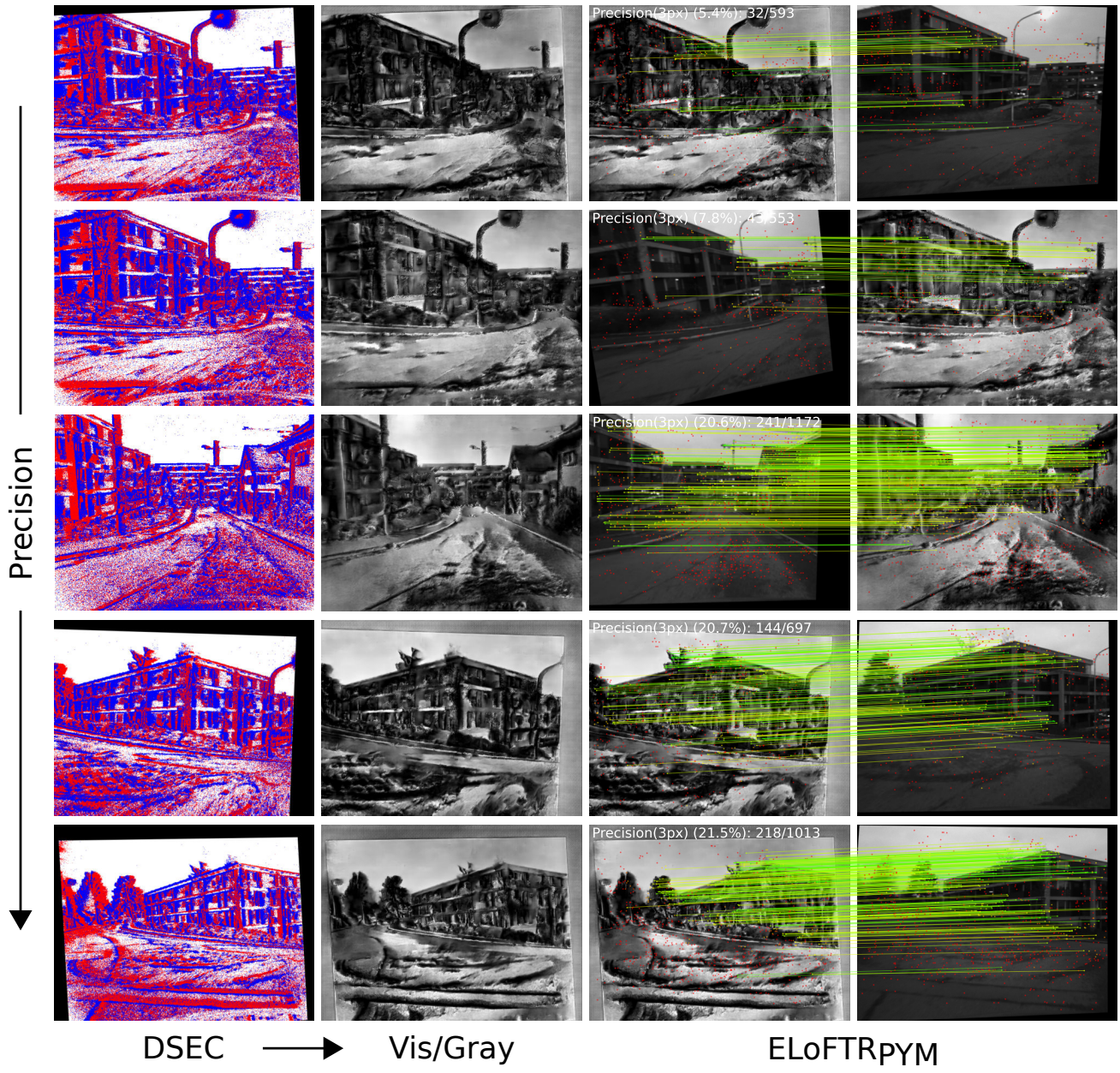

Figure 11. Matching results on DSEC (E-G) with ELoFTR$_{PYM}$. From 5.4% to 21.5% prec @ 3px.