Supplementary Material: Robust Semi-dense Feature Matching with Language Guidance

1. Matching Module

In this work, we adopt the coarse-to-fine matching pipeline following previous work [5] in Fig. 1.

1.1. Coarse-level Matching Module

We establish coarse-level correspondences using the contextual correlation feature maps $\hat{F}_c^A$ and $\hat{F}_c^B$, which provide semantic-aware robust matching regions for subsequent subpixel-level refinement. A dense correlation between $\hat{F}_c^A$ and $\hat{F}_c^B$ is computed to form a score matrix S_c^t. Dual-softmax is then applied across both dimensions of S_c^t to estimate mutual matching probabilities, following standard practice in prior work [2, 4, 5]. Coarse matches $\{M_c\}$ are obtained by selecting matches above the score threshold τ while satisfying the mutual-nearest-neighbor (MNN) constraint.

1.2. Fine-level Matching Module

To improve efficiency, we reuse the textual refined coarse features $\tilde{F}_c^A$ and $\tilde{F}_c^B$ to derive cross-view attended fine features, avoiding additional transformation networks as used in LoFTR [4]. Specifically, $\tilde{F}_c^A$ and $\tilde{F}_c^B$ are fused with $\frac{1}{4}$ and $\frac{1}{2}$ resolution visual features through convolution and upsampling to produce fine-level features $\tilde{F}_f^A$ and $\tilde{F}_f^B$ at the original resolution. From these, we extract 8×8 discriminative patches centered at each coarse match M_c for later match refinement.

To avoid the location variance introduced by commonly used refinement-by-expectations methods in [1, 2, 4], we employ the two-stage correlation for matching refinement, which combines the strengths of mutual-nearest-neighbor (MNN) matching and subpixel refinement by expectation. The first stage aims to obtain precise pixel-level matches by densely correlating fine feature patches for each coarse correspondence M_c, followed by MNN search within the local patch score matrix S_l to select the most confident fine match. While MNN offers spatially stable correspondences, it lacks subpixel precision. To address this, the second stage further refines these pixel-level matches by computing correlations within a small 3×3 window around each fine match and applying softmax to generate a probability distribution. The final subpixel-accurate match is then derived via expectation over this distribution, combining the robustness of MNN with the precision of expectation-based refinement.

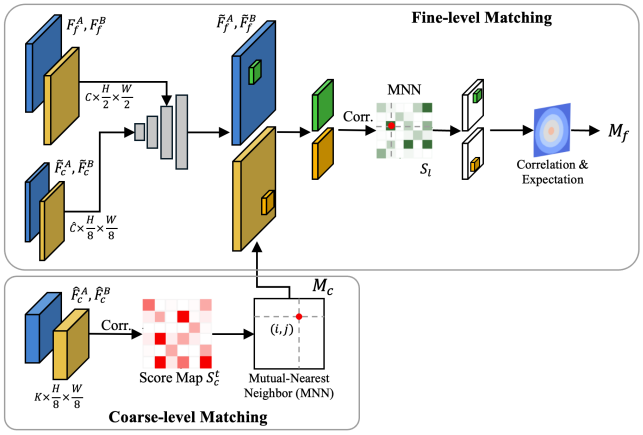

Figure 1. An overview of the coarse-level matching and fine-level matching pipeline.

2. Experiments

2.1. MegaDepth-Sync Dataset

As described in Sec. 4.2 of the main paper, we construct the MegaDepth-Sync dataset by translating all daytime images from the MegaDepth testset into nighttime images using the state-of-the-art image-to-image translation model, img2img-turbo [3]. This model adapts one-step diffusion for unpaired image translation using a CycleGAN-style framework [6]. We apply their pretrained day-to-night model to generate realistic nighttime images that preserve structural detail and high image quality, as shown in Fig. 2. Although the model occasionally introduces artificial lighting effects—especially near building edges in the sky—due to its training on the autonomous driving datasets, these artifacts do not impact the evaluation of the matching performance under day-night shifts, since they are unlikely to produce reliable keypoints and are typically excluded during feature matching. (see Fig. 2).

Figure 2. Samples of MegaDepth-Sync day-night images.

2.2. Qualitative Results

More qualitative results on the MegaDepth, ScanNet, MegaDepth-night2day and MegaDepth-night2night datasets are shown in Fig. 3.

3. Limitations and Future Works

Fig. 4 presents some failure cases on MegaDepth, Scannet, MegaDepth-dark2day and MegaDepth-dark2dark datasets. Our method exhibits failure in several scenarios: (1) when there is a large viewpoint distance between image pairs, making it difficult to obtain reliable coarse-level matches due to the loss of fine texture details from distant views; (2) when image pairs have minimal translation but share the same viewpoint angle, leading to inaccurate pose estimation by RANSAC; and (3) when significant viewpoint changes occur on low-texture surfaces, which degrades the effectiveness of semantic refinement. These issues largely originate from the heavy reliance on coarse-level matching in current coarse-to-fine pipelines, which struggles to capture fine-grained structures. While dense matching methods can address this, they require high computational costs during both training and inference. Therefore, a potential direction is to incorporate language guidance into a more computationally efficient dense matching framework in the future. Additionally, our language guidance approach leverages a closed-set tuning strategy, leaving the potential of open-vocabulary guidance unexplored for real-world scenarios.

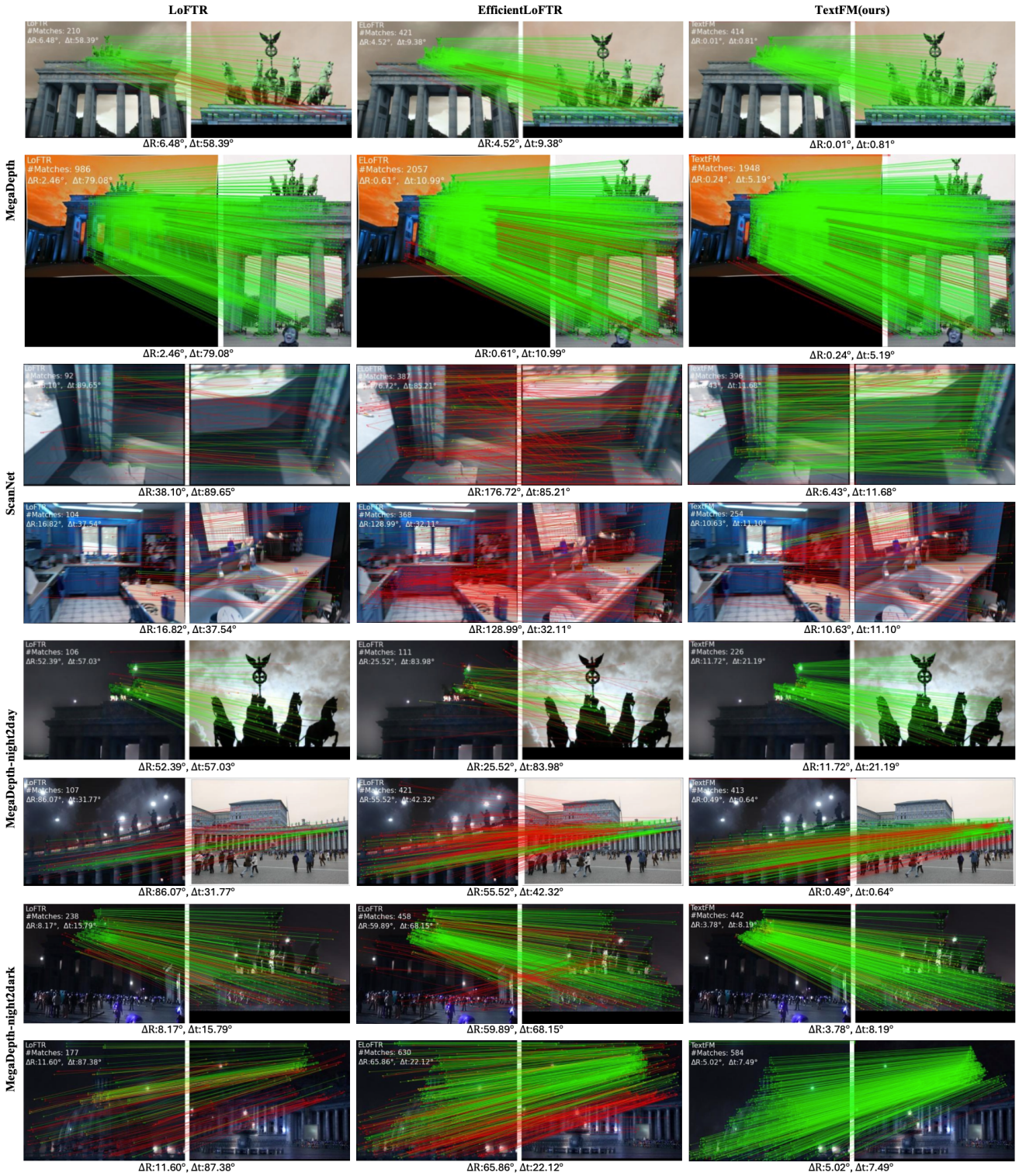

Figure 3. Qualitative comparison between our method and other coarse-to-fine methods LoFTR and EfficientLoFTR. Our method can produce a high number of accurate correspondences in challenging conditions including MegaDepth, Scannet, MegaDepth-dark2day and MegaDepth-dark2dark.

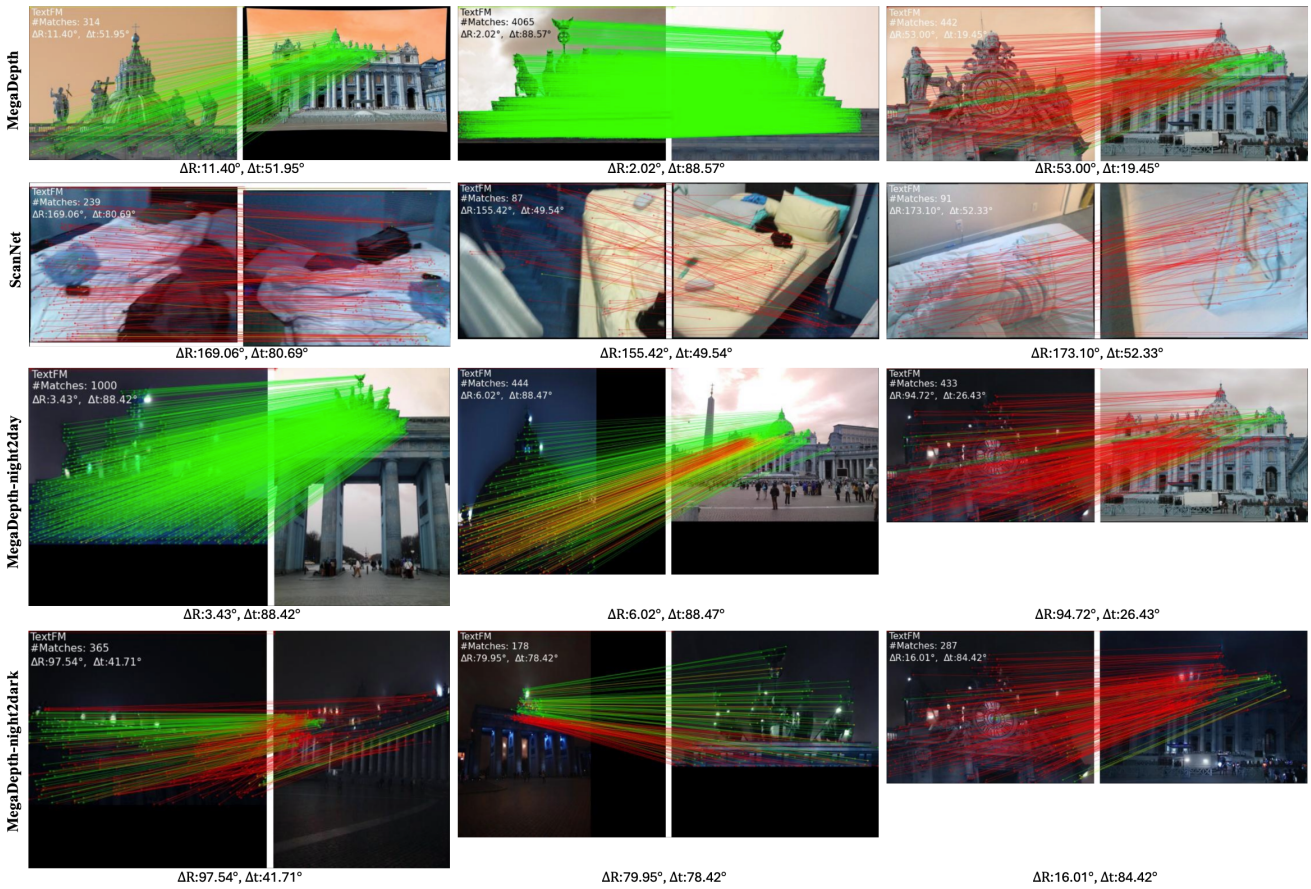

Figure 4. Qualitative visualization of failure cases in challenging conditions including MegaDepth, Scannet, MegaDepth-dark2day and MegaDepth-dark2dark.

References

[1] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European conference on computer vision*, pages 20–36. Springer, 2022. 1

[2] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2447–2455, 2023. 1

[3] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 1

[4] 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

[5] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21666–21675, 2024. 1

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1