2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SURE: Semi-dense Uncertainty-REfined Feature Matching

Sicheng Li[1,3], Zaiwang Gu[2], Jie Zhang[3], Qing Guo[4], Xudong Jiang[1] and Jun Cheng[2*]

***Abstract*— Establishing reliable image correspondences is essential for many robotic vision problems. However, existing methods often struggle in challenging scenarios with large viewpoint changes or textureless regions, where incorrect correspondences may still receive high similarity scores. This is mainly because conventional models rely solely on feature similarity, lacking an explicit mechanism to estimate the reliability of predicted matches, leading to overconfident errors. To address this issue, we propose SURE, a Semi-dense Uncertainty-REfined matching framework that jointly predicts correspondences and their confidence by modeling both aleatoric and epistemic uncertainties. Our approach introduces a novel evidential head for trustworthy coordinate regression, along with a lightweight spatial fusion module that enhances local feature precision with minimal overhead. We evaluated our method on multiple standard benchmarks, where it consistently outperforms existing state-of-the-art semi-dense matching models in both accuracy and efficiency. our code will be available on https://github.com/LSC-ALAN/SURE.**

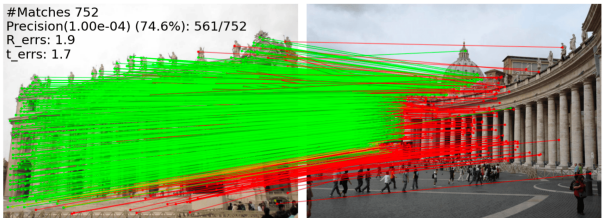
I. INTRODUCTION

Feature matching aims to establish visual correspondences between two images and serves as a cornerstone in many robotic vision applications, including Structure-from-Motion (SfM) [1], visual localization [2], 3D reconstruction [3], and SLAM [4], [5]. However, achieving accurate and reliable feature matching remains challenging in real-world scenarios due to factors such as repetitive textures, occlusions, and large viewpoint, scale or modality changes [6], [7], [8].

Recent methods such as LoFTR [9] abandon explicit keypoint detection and adopt a coarse-to-fine framework. Using transformer [10] as feature extractor, they compute patch-wise similarities for coarse matching, followed by fine-level refinement at the pixel level. Building on this idea, MatchFormer [11] incorporates multiscale features to improve robustness across varying resolutions, while E-LoFTR [12] introduces more efficient modules to optimize fine-level matching.

Despite their success, most of the existing methods still suffer from two critical limitations. First, most existing methods estimate confidence purely based on feature similarity, without explicitly modeling the intrinsic trustworthiness of the prediction itself. This becomes problematic in challenging scenarios such as large viewpoint changes or textureless regions, where incorrect matches often receive high similarity scores, and thus cannot be effectively filtered. In Structure-from-Motion or SLAM systems, unreliable matches can severely affect pose estimation and 3D reconstruction [13], making the ability to assess match reliability just as important as finding the matches. Second, many existing models prioritize accuracy at the cost of efficiency, relying on large architectures and highly complex computations, which limit their applicability in real-time or resource-constrained scenarios.

(a) SURE



(b) E-LoFTR [12]

Fig. 1: Comparison of E-LoFTR and our method on the MegaDepth dataset. Lines highlighted in green and red correspond to points with an epipolar error less than or greater than 10^{-4} respectively.

To address these issues, we propose SURE, a Semi-dense Uncertainty-Refined matching framework that jointly estimates correspondences and their associated uncertainties. At the core of SURE is trustworthy regression, which employs evidential learning[14] to predict both aleatoric and epistemic uncertainties along with the correspondence coordinates. Instead of computing similarity across high-resolution 2D windows, SURE estimates two sets of 1D heatmaps representing the marginal distributions of the coordinates x and y. These 1D heatmaps are treated as probabilistic evidence, enabling the model to jointly infer precise coordinates and associated uncertainties in a principled and efficient manner. This formulation not only reduces computational complexity significantly, but also yields a principled confidence estimate

*Corresponding author: Jun Cheng (cheng_jun@a-star.edu.sg).
[1]S. Li and X. Jiang are with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore.
[2]J. Cheng and Z. Gu are with the Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research (A*STAR), Singapore.
[3]J. Zhang and S. Li are with Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore.
[4]Q. Guo is with the College of Computer Science, Nankai University, Tianjin, China.
This research is supported by the National Research Foundation Singapore under the AI Singapore Programme (NO. AISG4-GC-2023-008-1B) and A*STAR under its MTC Programmatic Funds (GranNo. M23L7b0021).

for each match. To further enhance fine-level prediction, we introduce a lightweight spatial fusion strategy that effectively preserves detailed low-level spatial information, thereby improving accuracy without incurring significant computational overhead. As illustrated in Figure 1, our method yields a higher proportion of geometrically consistent correspondences, with noticeably fewer mismatches.

We evaluated our method on several widely used benchmarks, including MegaDepth [15] and ScanNet [16], where SURE consistently outperforms existing state-of-the-art semi-dense matching models in both accuracy and efficiency. Our approach demonstrates not only strong generalization across scenes, but also the ability to produce trustworthy confidence scores for match filtering and downstream tasks.

Our main contributions are summarized as follows.

- We propose SURE, a novel semi-dense matching framework that integrates correspondence prediction with uncertainty estimation.
- We introduce an evidential regression head that jointly models aleatoric and epistemic uncertainties, providing reliable confidence scores for match evaluation.
- We propose a spatial fusion module that refines local features by incorporating hierarchical spatial information and enhancing structural details.
- We demonstrate that our method surpasses previous state-of-the-art approaches such as E-LoFTR in both accuracy and efficiency on standard benchmarks.

II. RELATED WORKS

A. Feature Matching

1) Sparse Matching: Sparse feature matching methods identify and describe a limited set of keypoints. Early approaches such as SIFT [17] and ORB [18] rely on handcrafted descriptors with nearest-neighbor matching. With deep learning, methods like R2D2 [19] and SuperPoint [20] learn robust detectors and descriptors using convolutional networks. Extensions include semantic-aware SLAM frameworks that enhance correspondence and loop closure [21]. SuperGlue [22] models relationships between features via graph neural networks, while LightGlue [23] improves efficiency. Descriptor distillation [24] learns compact representations through teacher–student training, and LiftFeat [25] integrates pseudo-3D geometric cues to improve robustness. Despite these advances, sparse methods remain limited in low-texture or repetitive regions due to unreliable keypoint detection.

2) Dense Matching: Dense matchers aim to predict correspondences across nearly all image pixels. Early approaches such as DGC-Net [26] and DRC-Net [27] rely on 4D correlation volumes to explore full matching spaces, while PDC-Net [28] improves performance through progressive refinement with deformable convolutions. More recent methods, including DKM [29], formulate dense matching probabilistically and predict confidence maps to filter unreliable matches. RoMa [30] further leverages frozen DINOv2 [31] features with a dedicated convolutional decoder for refinement. Despite strong accuracy and alignment, dense methods remain computationally expensive due to high-resolution processing and large feature volumes.

3) Semi-Dense Matching: Semi-dense methods aim to strike a balance between matching density and computational cost. LoFTR [9] pioneers a coarse-to-fine strategy, first building coarse-level correspondences on a downsampled grid, then refines them using fine-grained feature patches. This approach enables more complete coverage than sparse techniques while avoiding the cost of full-resolution matching. MatchFormer [11] leverages a hierarchical transformer to jointly model global context. TopicFM [32] introduces topic-based modules to leverage semantic context for more robust matching, and E-LoFTR [12] addresses efficiency limitation by compressing local features and refining matches in a more compact representation. However, challenges remain in ensuring reliable fine-level alignment without high overhead. Our method follows this semi-dense paradigm to maintain wide coverage and efficiency, but differs by integrating uncertainty-aware predictions for more reliable refinement.

B. Uncertainty Estimation

Recent advances in deep learning have led to various uncertainty modeling techniques. Bayesian neural networks [33] and deep ensembles [34] model epistemic uncertainty by treating weights as distributions or training multiple models, but are often computationally expensive. Monte Carlo Dropout [35] offers a simpler approximation via stochastic forward passes. Deep Evidential Learning [36], [37] avoids sampling by using deterministic evidence-based inference. For regression, Deep Evidential Regression [14] extends this framework by modeling outputs with a normal-inverse gamma prior, allowing joint estimation of aleatoric and epistemic uncertainties in a single pass.

In correspondence estimation, uncertainty modeling remains to be explored. Evidential learning has recently been applied to other tasks such as stereo matching [38] and vision language calibration [39], but its potential in feature matching has not yet been fully explored. To address this gap, we introduce an evidential formulation that jointly estimates correspondence offsets and quantifies both aleatoric and epistemic uncertainties in a principled manner.

III. METHOD

This section presents an overview of the proposed SURE framework. As depicted in Figure 2, the architecture consists of four main components: a CNN backbone for feature extraction, a coarse matching module to generate initial correspondence candidates via global context-aware features, a lightweight spatial fusion module that integrates multi-scale details to enrich local features, and a trustworthy regression stage incorporating an evidential head that predicts precise matches along with aleatoric and epistemic uncertainty estimates. This design enables SURE to provide accurate, reliable, and efficient semi-dense matching by combining hierarchical feature representations and principled uncertainty modeling.

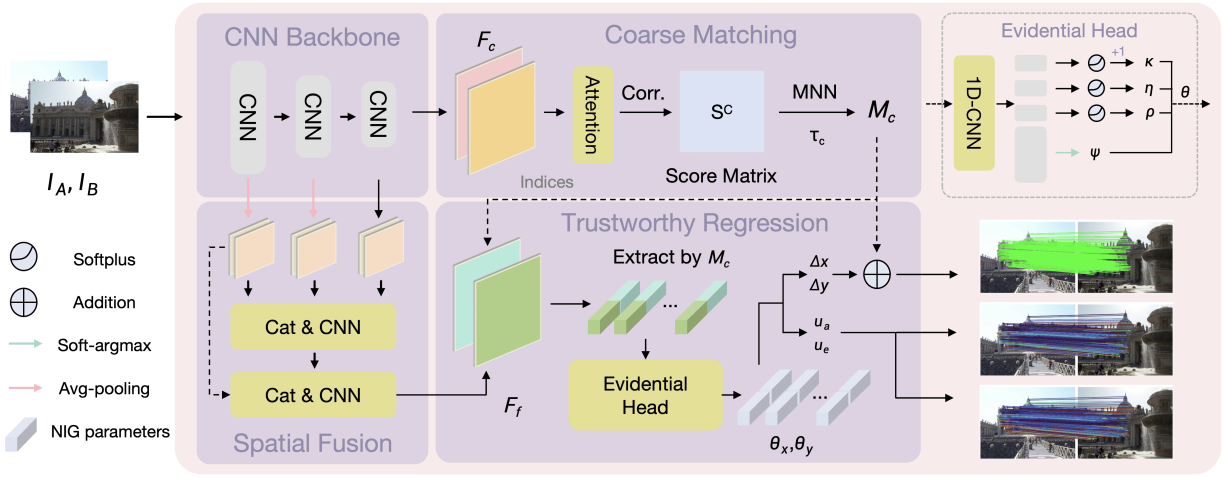


Fig. 2: Overview of the proposed SURE framework. (1) A backbone extracts coarse features F_c and a Spatial Fusion Module provides fine features F_f. (2) The coarse matching module produces initial correspondences M_c, which are used to sample F_f for fine-level refinement. (3) Trustworthy Regression produces precise offsets $(\Delta x, \Delta y)$ along with uncertainty estimates. Specifically, an Evidential Head predicts the parameters $(\psi, \eta, \kappa, \rho)$ of a Normal-Inverse-Gamma distribution, which jointly encode the offset and its associated aleatoric and epistemic uncertainties.

A. Feature Extraction

To extract hierarchical visual representations, we utilize a single-branch compact backbone with RepVGG [40]. Given a pair of input images $I^A, I^B \in \mathbb{R}^{C\times H\times W}$, the network generates multilevel features, where C, H, and W refer to the number of channels, height, and width, respectively. The deepest layer produces coarse descriptors $F_c^A, F_c^B \in \mathbb{R}^{C_c\times H_c\times W_c}$, where $H_c = \frac{H}{8}$, $W_c = \frac{W}{8}$, and $C_c = 256$, taking advantage of broad contextual patterns.

B. Coarse Matching

Following prior work [12], we apply self and cross attention to the coarse-level features F_c^AandF_c^B, yielding the enhanced representations $\hat{F}_c^A$ and $\hat{F}_c^B$. We establish initial correspondences by computing a bidirectional similarity matrix followed by confidence-based filtering.

We first calculate the coarse similarity matrix $S^c \in \mathbb{R}^{H_cW_c\times H_cW_c}$ using the inner product scaled by temperature:

$$S^c_{i,j} = \frac{1}{\tau}\cdot\langle \hat{F}_c^A(i), \hat{F}_c^B(j)\rangle, \tag{1}$$

where τ is a fixed scalar that controls the sharpness of the distribution.

To infer the matching confidence in both directions, we perform softmax normalization along rows and columns, respectively:

$$P^{A\to B}_{i,j} = \frac{\exp(S^c_{i,j})}{\sum_k \exp(S^c_{i,k})}, P^{B\to A}_{i,j} = \frac{\exp(S^c_{i,j})}{\sum_k \exp(S^c_{k,j})}. \tag{2}$$

Following common practice, we adopt mutual nearest neighbor (MNN) filtering [23], selecting correspondences that are bidirectional maxima and exceed a confidence threshold τ_c. The resulting set of coarse matches, denoted as $\mathcal{M}_c$, serves as candidate correspondences for fine-level refinement.

C. Spatial Fusion Module

Unlike traditional FPN-style designs in feature matching [9], which upsample low-resolution features through top-down fusion, we adopt a streamlined multi-scale fusion strategy optimized for fine stage regression. While prior methods often restore features to full resolution for dense refinement, leading to high computational cost, our method uniformly aligns all features to a fixed resolution $\frac{1}{8}$ for the final regression. This avoids costly cropping and maintains efficiency. To retain structural detail, we incorporate a high-resolution enhancement path inspired by HRNet [41], enriching the fused features with both spatial precision and semantic depth.

Specifically, we extract intermediate features from the backbone at three spatial scales: $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$ of the input resolution. These features are denoted as $\{F_{(s)}\}_{s\in\{2,4,8\}}$, where $F_{(s)} \in \mathbb{R}^{C^{(s)}\times\frac{H}{s}\times\frac{W}{s}}$ corresponds to the feature map at scale $\frac{1}{s}$. To unify the dimensions of the channels, each $F_{(s)}$ is projected onto a common embedding space of size $C_f = 256$ using a 1×1 convolution.

The projected features are then downsampled through adaptive average pooling to match the spatial resolution of the $\frac{1}{8}$ scale, resulting in aligned features $\left\{\hat{F}_{\frac{1}{2}}, \hat{F}_{\frac{1}{4}}, \hat{F}_{\frac{1}{8}}\right\} \in \mathbb{R}^{C_f\times\frac{H}{8}\times\frac{W}{8}}$. These are concatenated along the channel dimension and fused through a 1×1 convolution followed by batch normalization and ReLU activation, producing the fused representation F_{fused}.

To further preserve high-frequency details, we introduce a residual path from the original $F_{\frac{1}{2}}$ scale. It is processed

through a separate 1×1 convolution and pooled to the same $\frac{1}{8}$ resolution. The final fine-level feature F_f is computed as:

$$F_f = F_{\text{fused}} + \text{Pool}\left(\text{Conv}_{1\times1}(\hat{F}_{\frac{1}{2}})\right). \tag{3}$$

The resulting fused representation $F_f \in \mathbb{R}^{C_f \times H_{\frac{1}{8}} \times W_{\frac{1}{8}}}$ retains both the semantic context and local structural details, which benefits downstream fine-level refinement and uncertainty estimation.

D. Trustworthy Regression

1) Probabilistic Modeling With Uncertainty: From an evidential perspective, each offset z is sampled from a Gaussian distribution with unknown parameters ξ and υ^2. These parameters are further modeled by placing a normal prior in ξ and an inverse gamma prior in υ^2.

$$z \sim \mathcal{N}(\xi, \upsilon^2), \xi \sim \mathcal{N}(\psi, \upsilon^2\eta^{-1}), \upsilon^2 \sim \mathcal{IG}(\kappa, \rho), \tag{4}$$

where $\mathcal{IG}$ denotes the inverse gamma distribution, $\psi \in \mathbb{R}$, $\eta > 0$, $\kappa > 1$, and $\rho > 0$. Assuming the independence between the mean and the variance, the joint posterior $q(\xi, \upsilon^2)$ follows a normal-inverse-gamma (NIG) distribution, parameterized by $\theta = (\psi, \eta, \kappa, \rho)$:

$$q(\xi, \upsilon^2) = \mathcal{NIG}(\psi, \eta, \kappa, \rho). \tag{5}$$

The expected prediction $\hat{z}$, along with its aleatoric uncertainty u_a^z and epistemic uncertainty u_e^z, is derived from the offset distribution, whose mean represents the predicted offset, and variance encodes both types of uncertainty, following the formulation in [14].

$$\hat{z} = \psi, u_a^z = \frac{\rho}{\kappa - 1}, u_e^z = \frac{\rho}{\eta(\kappa - 1)}. \tag{6}$$

To train this evidential model, we minimize a loss composed of two terms: a negative log-evidence term derived from the NIG distribution.

$$\begin{aligned}\mathcal{L}^{\text{evi}}(\mathbf{w}) = &\frac{1}{2}\log\left(\frac{\pi}{\eta}\right) - \kappa\log(\Theta) \\ &+ \left(\kappa + \frac{1}{2}\right)\log\left((\mathbf{y}^* - \psi)^2\eta + \Theta\right) \\ &+ \log\left(\frac{\Gamma(\kappa)}{\Gamma(\kappa + \frac{1}{2})}\right),\end{aligned} \tag{7}$$

where $\Theta = 2\rho(1 + \eta)$ and $\mathbf{w}$ denotes the network-estimated parameters.

A regularization term is introduced to penalize incorrect predictions with high confidence:

$$\mathcal{L}^{\text{reg}}(\mathbf{w}) = |\mathbf{y}^* - \psi| \cdot \Phi, \quad \text{with} \quad \Phi = 2\eta + \kappa, \tag{8}$$

where $\mathbf{y}^*$ is the ground truth label of offsets, and Φ quantifies total evidence.

The total uncertainty-aware loss is computed over all N predictions as:

$$\mathcal{L}_f(\mathbf{w}) = \frac{1}{N}\sum_{j=1}^{N}\left(\mathcal{L}_j^{\text{evi}}(\mathbf{w}) + \zeta\mathcal{L}_j^{\text{reg}}(\mathbf{w})\right), \tag{9}$$

where $\zeta > 0$ is the regularization coefficient.

2) Evidential Regression Head: Inspired by coordinate classification approaches [42], we design a lightweight 1D regression head that decouples offset estimation along the x and y axes, allowing for efficient and stable sub-pixel refinement.

Given a set of coarse matches, we first extract their corresponding fine-level descriptors from F_f^A and F_f^B, and concatenate them to form a fused feature tensor $\hat{F}_f \in \mathbb{R}^{M\times 2d}$, where M denotes the number of coarse matches and d is the dimensionality of each feature vector. This tensor serves as the input to two independent regression branches for each axis.

The vector $\hat{F}_f$ is subsequently processed by two independent 1-D convolutional regression heads, one for each axis. Although both heads take the same input $\hat{F}_f$, they have separate parameters and output structures.

Each head produces a $(N + 3)$-dimensional output vector that directly parameterizes a Normal-Inverse-Gamma (NIG) distribution. Specifically, the first N elements are spatial logits, from which a soft-argmax operation computes the expected value $\psi \in [-0.5, 0.5]$, which also serves as the predicted offset z along the corresponding axis. The remaining three values are transformed through soft plus activation to obtain the additional NIG parameters: η, κ, and ρ, which together characterize the predictive uncertainty of the model. This design allows the model to jointly learn both the prediction and its associated confidence in a fully differentiable and probabilistic manner.

3) Uncertainty Filtering: Given the predicted parameters of the NIG distribution, the aleatoric uncertainty u_a, and the epistemic uncertainty u_e along both axes can be calculated using Eq.(6). For each matched pair, we obtain (u_a^x, u_a^y) and (u_e^x, u_e^y) from the respective regression heads. The final aleatoric and epistemic uncertainties are obtained by averaging the values across axes:

$$u_a = \frac{u_a^x + u_a^y}{2}, \quad u_e = \frac{u_e^x + u_e^y}{2}. \tag{10}$$

We then filter the predicted matches based on these uncertainty estimates. Specifically, we set thresholds τ_a and τ_e, and discard any predictions whose $u_a > \tau_a$ or $u_e > \tau_e$. These thresholds are chosen as quantiles to retain a desired percentage of the most confident predictions.

E. Loss

We supervise the matching process at both the coarse and fine levels to ensure robust global localization and subpixel precision.

1) Coarse-Level Loss: Following previous work [43], we adopt a dual-direction focal loss to supervise the coarse matching probability matrices $P^{A\to B}$ and $P^{B\to A}$, using ground truth correspondences P_c^{gt} generated by warping the grid centers via depth and camera pose.

$$\mathcal{L}_c = \text{FL}(P_c^{gt}, P^{A\to B}) + \text{FL}(P_c^{gt}, P^{B\to A}), \tag{11}$$

where FL denotes the focal loss [44] defined as:

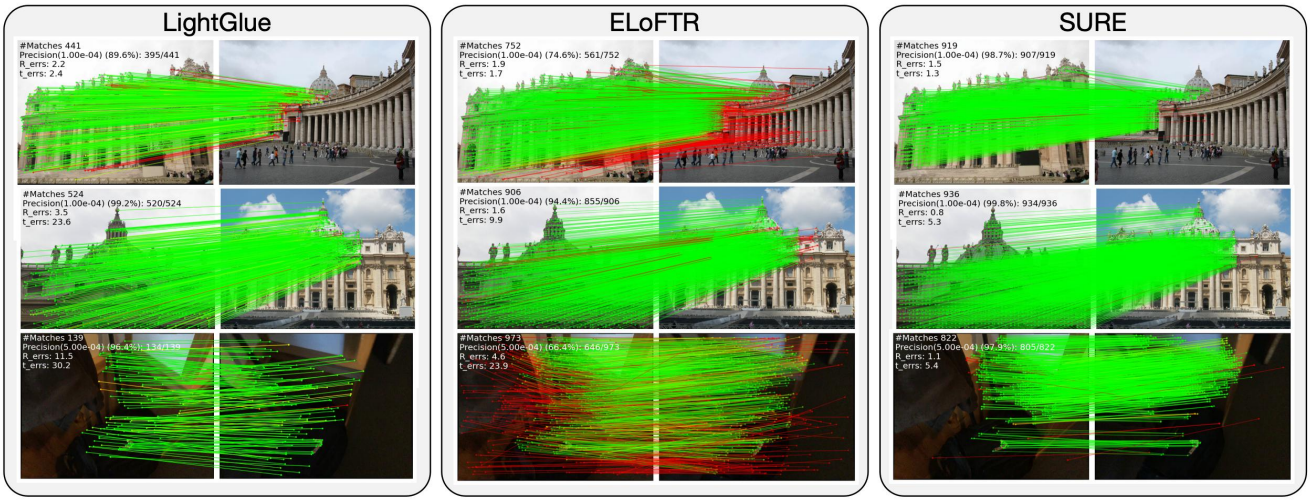
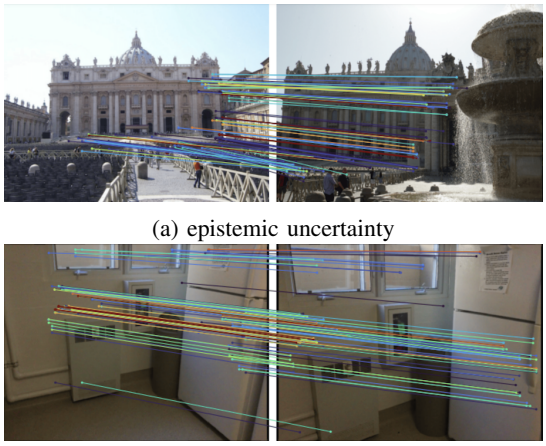


Fig. 3: Qualitative comparisons of SURE against Light Glue and E-LoFTR on both indoor and outdoor scenes. SURE achieves a higher number of correct matches and reduces mismatches, demonstrating robustness in low-texture areas as well as under significant viewpoint and lighting variations. Red regions denote points with epipolar error exceeding 5×10^{-4} for indoor and 1×10^{-4} for outdoor scenes.



(a) epistemic uncertainty

(b) aleatoric uncertainty

Fig. 4: We selected large viewpoint changes and weak-texture scenarios. Among 2048 correspondences, the 50 pairs with the highest model uncertainty and data uncertainty were chosen. The lighter the line color, the higher the uncertainty.

$$\mathrm{FL}(p) = -\alpha(1-p)^{\gamma}\log p, \tag{12}$$

and α, γ are the weighting and focusing parameters, respectively.

2) Total Loss: The final training objective is a weighted sum of the two terms:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_f(\mathcal{L}_f(\mathbf{w}^x) + \mathcal{L}_f(\mathbf{w}^y)), \tag{13}$$

where λ_c and λ_f balance the contributions of the coarse and fine level supervision. $\mathbf{w}^x$ and $\mathbf{w}^y$ denote the parameters of the fine-level regression heads along the horizontal x and vertical y axes.

IV. EXPERIMENTS

A. Implementation Details

Our model is trained on the MegaDepth dataset [15]. We adopt the standard training and testing splits provided by [12] to ensure fair comparisons with existing methods. The entire pipeline is trained end-to-end using the AdamW optimizer, with an initial learning rate of 2×10^{-3} and a weight decay of 1×10^{-4}. In our implementation, the fine-level loss weight ζ is fixed at 1, while the focal loss parameters γ and α are set to 2 and 0.25, respectively. The total loss is formulated as a weighted sum of the objectives of the coarse level and the fine level, with respective weights set to $\lambda_c = 1.0$ and $\lambda_f = 0.25$. We train the model for 30 epochs using 4 NVIDIA 3090 GPUs, with a batch size of 16 and an input image resolution of 832×832. The bin number N set to 16. The temperature parameter τ used in softmax normalization is fixed to 0.1 throughout training. The trustworthiness regression network consists of two 1D convolutional layers. During inference, the uncertainty thresholds for filtering, τ_a and τ_e, are both set to 0.95.

B. Relative Pose Estimation

1) Dataset: Following prior protocols [9], we evaluate on 1500 image pairs each from ScanNet [16] (indoor) and MegaDepth [15] (outdoor). All ScanNet images are resized to 640×480, while MegaDepth images are uniformly scaled to 832×832 for consistency

2) Baselines: We compare with three categories of methods: (1) sparse keypoint-based matchers including SuperPoint [20], SuperGlue [22], LightGlue [23], and XFeat [45]; (2) semi-dense matchers such as RCM [46], LoFTR [9], MatchFormer [11], AspanFormer [47], E-LoFTR [12],and JAMMA [43]; (3) dense matcher ROMA [30] and DKM [29].

TABLE I: Results of Relative Pose Estimation on ScanNet and MegaDepth Dataset. We report the pose estimation AUC under three error thresholds, along with the overall runtime efficiency of the models.

Category	Method	ScanNet			MegaDepth			Time(ms)
		AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	
Sparse	SP + SG *CVPR'20*	16.2	32.8	49.7	57.6	72.6	83.5	96.9
	SP + LG *ICCV'23*	14.8	30.8	47.5	58.8	73.6	84.1	84.2
	XFeat *CVPR'24*	16.7	32.6	47.8	44.2	58.2	69.2	14.2
Dense	DKM *CVPR'23*	26.6	47.1	64.1	67.3	79.7	88.1	554.2
	ROMA *CVPR'24*	28.9	50.4	68.3	68.5	80.6	88.8	824.9
Semi-Dense	LoFTR *CVPR'21*	16.9	33.6	50.6	62.1	75.5	84.9	117.5
	MatchFormer *ACCV'22*	15.8	32.0	48.0	62.0	75.6	84.9	156.0
	ASpanFormer *ECCV'22*	19.6	37.7	54.4	62.6	76.1	85.7	155.7
	RCM *ECCV'24*	17.3	34.6	52.1	58.3	72.8	83.5	93.0
	E-LoFTR *CVPR'24*	19.2	37.0	53.6	63.7	77.0	86.4	69.6
	JamMa *CVPR'25*	15.1	31.6	48.5	64.1	77.4	86.5	**59.9**
	SURE (Proposed)	**20.3**	**38.6**	**55.3**	**64.7**	**77.7**	**86.8**	62.8

3) Evaluation Protocol: Following previous work [43], we report the area under the curve (AUC) of the pose error at thresholds of $5°$, $10°$, and $20°$. We use LO-RANSAC to estimate the essential matrix for MegaDepth, and RANSAC for ScanNet. All methods use a RANSAC inlier threshold of 0.5. Furthermore, we measure the matching runtime on the MegaDepth dataset using a single NVIDIA 4090 GPU to assess efficiency.

4) Results: As shown in Tab. I, our method achieves state-of-the-art performance among sparse and semi-dense matchers on both MegaDepth and ScanNet. It surpasses recent semi-dense approaches such as E-LoFTR, ASpanFormer and JamMa across all AUC thresholds, while maintaining high computational efficiency. Compared to dense matchers like DKM and RoMa, our model achieves a more favorable balance between accuracy and speed, making it more suitable for real-time applications. As illustrated in Fig. 3, we visualize the correspondences along with their estimated uncertainties, demonstrating that our model produces cleaner and more reliable matches overall.

As further evidence, Fig. 4 highlights how our uncertainty modeling behaves in challenging scenarios. Out of 2048 predicted correspondences, we visualize the 100 pairs with the highest uncertainty. We observe that epistemic uncertainty tends to localize occluded areas under large viewpoint changes, while aleatoric uncertainty concentrates on weak-texture regions. Since these areas are also prone to incorrect matches, the results confirm that our uncertainty estimation is both meaningful and effective. By filtering or down-weighting such matches, our method not only improves robustness but also prevents error propagation to downstream tasks.

C. Homography Estimation

1) Dataset: We conduct evaluation on the HPatches benchmark [48], which has 108 planar image sequences.

2) Baselines: We compare our method against recent sparse and semi-dense approaches. Sparse matchers include SuperPoint [20], R2D2 [19], DISK [49], SuperGlue [22], and LightGlue [23]. Semi-dense matchers include DRC-Net [27], Patch2Pix [50], LoFTR [9], ASpanFormer [47], and E-LoFTR [12].

TABLE II: Homography estimation results on the HPatches dataset. We report the AUC of corner reprojection error.

Category	Method	AUC		
		@3px	@5px	@10px
Sparse	SP + NN	41.6	55.8	71.7
	R2D2 + NN	50.6	63.9	76.8
	DISK + NN	52.3	64.9	78.9
	SP + SG	53.9	68.3	81.7
	SP + LG	54.2	68.3	81.5
Semi-Dense	DRC-Net	50.6	56.2	68.3
	Patch2Pix	59.3	70.6	81.2
	LoFTR	65.9	75.6	84.6
	ASpanFormer	**67.4**	76.9	85.6
	E-LoFTR	66.5	76.4	85.5
	SURE(ours)	67.0	**77.2**	**86.1**

3) Metric: Following the protocol in [9], we use the top 1000 matches for homography estimation and apply vanilla RANSAC. The AUC of the projection error is calculated at thresholds 3, 5, and 10 pixels. All images are resized so that the shorter side is 480 pixels.

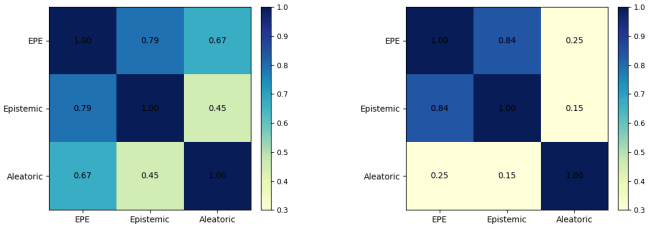
4) Results: As shown in Tab. II, SURE achieves the best AUC at both 5px and 10px, demonstrating a strong coarse-level localization. While slightly behind others at stricter thresholds, it is mainly due to the use of more complex fine matching modules in existing semi-dense methods. Nonetheless, SURE strikes a favorable balance between accuracy and efficiency.

D. Ablation Study

We evaluate key design choices via ablation studies (Tab. III). All models are trained for 30 epochs at 832 resolution and evaluated using the protocol in Sec. IV-B. AUC@10° is reported on MegaDepth and ScanNet, with runtime measured on MegaDepth. *I*: We start with E-LoFTR

TABLE III: Ablation study on MegaDepth and ScanNet datasets. We report AUC@10° and inference time. Each component is examined to assess its individual and combined impact on performance.

Setup ↓	AUC@10° ↑		Time (ms) ↓
	ScanNet	MegaDepth	
I (Baseline): E-LoFTR	37.0	77.0	69.6
II: Direct Regression	35.9	75.0	58.3
III: Coord. Regression	36.3	75.8	59.0
IV: Spatial Fusion	37.3	76.7	62.5
V: KL Loss	37.2	76.5	62.5
VI: Evidential Head	38.2	77.3	62.7
VII (Ours): Filtering	**38.6**	**77.7**	62.8



(a) Megadepth (b) Scannet

Fig. 5: Uncertainty analysis. (a) and (b) are the heat maps of the Spearman rank correlation analysis between the EPE and the uncertainties.

as our baseline. *II*: Replacing the fine matching module with a simple 2-layer 1-D CNN regression head that directly predicts (x, y) coordinates using L2 loss results in a noticeable drop in AUC to (35.9% / 75.0%), along with a reduced inference time of -11.3 ms. *III*: Decomposing the regression into two 1-D CNN heads for x and y slightly improves performance to (36.3% / 75.8%) at 59.0 ms, suggesting marginal stability benefits from axis-wise modeling. *IV*: Introducing our spatial fusion module improves generalization, boosting performance to (37.3% / 76.7%) with a moderate increase in time to 62.5 ms. This highlights the value of integrating fine-level structural cues. *V*: Applying a KL loss following SimCC [42] does not offer further gains, resulting in (37.2% / 76.5%). This suggests that our proposed uncertainty loss better captures spatial confidence. *VI*: Replacing the L2 loss with our evidential formulation improves the performance to (38.2% / 77.3%), with negligible time cost (62.7 ms), confirming the advantage of uncertainty-aware regression. *VII*: Filtering unreliable matches using predicted uncertainty yields the best result (38.6% / 77.7%), with a minimal time increase to 62.8 ms. Compared to E-LoFTR, our method delivers both higher accuracy and lower latency, highlighting its superior balance between effectiveness and efficiency.

E. Uncertainty Analysis

To evaluate uncertainty quality, we compute the Spearman correlation between end-point error (EPE) and two types of uncertainties: aleatoric (data) uncertainty and epistemic (model) uncertainty, as shown in Fig. 5. In MegaDepth, model uncertainty correlates more strongly with EPE 0.79 than data uncertainty 0.67, indicating its superior alignment with actual prediction error. The modest correlation between the two 0.45 suggests they capture complementary aspects—data uncertainty reflects input noise, while model uncertainty measures confidence due to limited knowledge. In ScanNet, model uncertainty remains reliable with a higher correlation 0.84, whereas data uncertainty drops to 0.25, likely due to domain shift. This highlights the robustness of epistemic uncertainty across datasets, and the sensitivity of aleatoric uncertainty to scene-specific variations. These results confirm that our uncertainty estimates provide reliable indicators of prediction confidence.

V. CONCLUSIONS

We present a novel matching framework that integrates uncertainty-aware learning with an efficient spatial fusion strategy to enhance the accuracy of the correspondence. By jointly modeling confidence and structure, our method effectively captures both semantic context and spatial precision. Extensive experiments on challenging benchmarks demonstrate superior performance, and ablation studies confirm the contribution of each component. These results highlight the importance of incorporating probabilistic reasoning and fine-grained feature cues in semi-dense feature matching.

REFERENCES

[1] P. Lindenberger, P.-E. Sarlin, V. Larsson, and M. Pollefeys, "Pixel-perfect structure-from-motion with featuremetric refinement," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5987–5997.

[2] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 716–12 725.

[3] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm, "Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset)," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3287–3295.

[4] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[5] C.-M. Chung, Y.-C. Tseng, Y.-C. Hsu, X.-Q. Shi, Y.-H. Hua, J.-F. Yeh, W.-C. Chen, Y.-T. Chen, and W. H. Hsu, "Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9400–9406.

[6] Y. Liu, Z. Sun, B. Yu, Y. Zhao, B. Du, Y. Xu, and J. Cheng, "Mifnet: Learning modality-invariant features for generalizable multimodal image matching," *IEEE Transactions on Image Processing*, vol. 34, pp. 3593–3608, 2025.

[7] W. Liu, W. Zhou, J. Liu, P. Hu, J. Cheng, J. Han, and W. Lin, "Modality-aware feature matching: A comprehensive review of single- and cross-modality techniques," *arXiv preprint arXiv:2507.22791*, 2025.

[8] J. Zeng, Z. Gu, W. Liu, L. Cai, and J. Cheng, "Uncertainty aware interest point detection and description," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 2144–2153.

[9] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Matchformer: Interleaving attention in transformers for feature matching," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 2746–2762.

[12] Y. Wang, X. He, S. Peng, D. Tan, and X. 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*, 2024, pp. 21 666–21 675.

[13] F. Engelmann, K. Rematas, B. Leibe, and V. Ferrari, "From points to multi-object 3d reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4588–4597.

[14] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[15] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2041–2050.

[16] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[19] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," *Advances in neural information processing systems*, vol. 32, 2019.

[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[21] O. Ilter, I. Armeni, M. Pollefeys, and D. Barath, "Semantically guided feature matching for visual slam," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 12 013–12 019.

[22] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[23] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 17 627–17 638.

[24] X. Guo, J. Hu, H. Bao, and G. Zhang, "Descriptor distillation for efficient multi-robot slam," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6210–6216.

[25] Y. Liu, W. Lai, Z. Zhao, Y. Xiong, J. Zhu, J. Cheng, and Y. Xu, "Liftfeat: 3d geometry-aware local feature matching," *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11 714–11 720, 2025.

[26] I. Melekhov, A. Tiulpin, T. Sattler, M. Pollefeys, E. Rahtu, and J. Kannala, "Dgc-net: Dense geometric correspondence network," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1034–1042.

[27] X. Li, K. Han, S. Li, and V. Prisacariu, "Dual-resolution correspondence networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 346–17 357, 2020.

[28] P. Truong, M. Danelljan, L. Van Gool, and R. Timofte, "Learning accurate dense correspondences and when to trust them," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5714–5724.

[29] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg, "Dkm: Dense kernelized feature matching for geometry estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 765–17 775.

[30] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 790–19 800.

[31] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[32] K. T. Giang, S. Song, and S. Jo, "Topicfm: Robust and interpretable topic-assisted feature matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2447–2455.

[33] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[34] Y. Wen, D. Tran, and J. Ba, "Batchensemble: an alternative approach to efficient ensemble and lifelong learning," in *International Conference on Learning Representations*, 2020.

[35] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[36] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[37] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[38] J. Lou, W. Liu, Z. Chen, F. Liu, and J. Cheng, "Elfnet: Evidential local-global fusion for stereo matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 784–17 793.

[39] C. Oh, H. Lim, M. Kim, D. Han, S. Yun, J. Choo, A. Hauptmann, Z.-Q. Cheng, and K. Song, "Towards calibrated robust fine-tuning of vision-language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 12 677–12 707, 2024.

[40] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 733–13 742.

[41] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[42] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, and S.-T. Xia, "Simcc: A simple coordinate classification perspective for human pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 89–106.

[43] X. Lu and S. Du, "Jamma: Ultra-lightweight local feature matching with joint mamba," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 934–14 943.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[45] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, "Xfeat: Accelerated features for lightweight image matching," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 2682–2691.

[46] X. Lu and S. Du, "Raising the ceiling: Conflict-free local feature matching with dynamic view switching," in *European Conference on Computer Vision*, 2025, pp. 256–273.

[47] H. Chen, Z. Luo, L. Zhou, Y. Tian, M. Zhen, T. Fang, D. Mckinnon, Y. Tsin, and L. Quan, "Aspanformer: Detector-free image matching with adaptive span transformer," in *European conference on computer vision*. Springer, 2022, pp. 20–36.

[48] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5173–5182.

[49] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," *Advances in neural information processing systems*, vol. 33, pp. 14 254–14 265, 2020.

[50] Q. Zhou, T. Sattler, and L. Leal-Taixe, "Patch2pix: Epipolar-guided pixel-level correspondences," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4669–4678.