# Estimating Extreme 3D Image Rotations using Cascaded Attention

Shay Dekel
Bar Ilan University
Ramat-Gan, Israel
shaydekel@gmail.com

Yosi Keller
Bar Ilan University
Ramat-Gan, Israel
yosi.keller@gmail.com

Martin Čadík
FIT, Brno University of Technology
Brno, Czech Republic
cadik@fit.vut.cz

## Abstract

*Estimating large, extreme inter-image rotations is critical for numerous computer vision domains involving images related by limited or non-overlapping fields of view. In this work, we propose an attention-based approach with a pipeline of novel algorithmic components. First, as rotation estimation pertains to image pairs, we introduce an inter-image distillation scheme using Decoders to improve embeddings. Second, whereas contemporary methods compute a 4D correlation volume (4DCV) encoding inter-image relationships, we propose an Encoder-based cross-attention approach between activation maps to compute an enhanced equivalent of the 4DCV. Finally, we present a cascaded Decoder-based technique for alternately refining the cross-attention and the rotation query. Our approach outperforms current state-of-the-art methods on extreme rotation estimation. We make our code publicly available*[1].

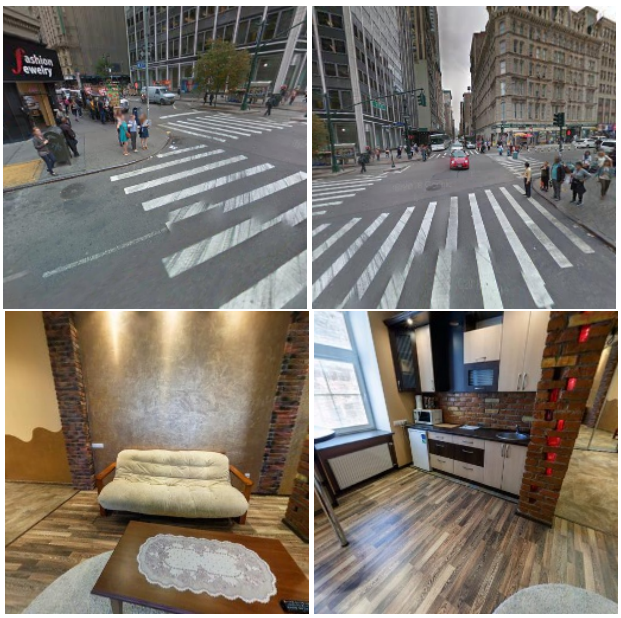

Figure 1. The estimation of extreme 3D image rotations. First row: Images pair with a small overlap. Second row: non-overlapping image pairs. The proposed scheme estimates the relative rotation between image pairs.

## 1. Introduction

Estimating the relative pose between a pair of images is a crucial task in computer vision, which is used in various applications such as indoor navigation, augmented reality, autonomous driving, 3D reconstruction [40, 44], camera localization [5, 45, 47], simultaneous localization and mapping [12, 38], and novel view synthesis [35, 42]. The current approach to image registration involves extracting features, matching them, and establishing correspondence between them. However, this approach is ineffective for input pairs with little or no overlap, making it difficult to establish sufficient feature correspondences for matching, such as in the images shown in Fig. 1.

Numerous applications [1, 32, 49] necessitate precise estimation of inter-image rotations. The prevalent approach for extreme 3D rotation estimation between images with limited or no overlap, as in Fig. 1, relates to the seminal work of Coughlan and Yuille [10]. They introduced a technique premised on linear structures within an image, primarily arising from three mutually orthogonal directions - one vertical (building walls) and two horizontal (ground pavements, roads, etc.). Similarly, "Single View Metrology" by Criminisi et al. [11] and extensions [26, 41, 61] utilize parallel image lines and corresponding vanishing points [19] for camera calibration. Furthermore, relative camera rotation can be estimated via illumination cues [2], by analyzing lighting and shadow directions.

In this work, we propose a deep-learning approach for estimating significant, extreme inter-image rotations. Unlike classical formulations [10, 11] that explicitly detect hand-crafted cues such as lines, shadows, and vanishing points, our method directly regresses the relative rotation from input images through a deep neural network. Inspired by recent successful applications of Transformers [53] in computer vision tasks including object detection [8] and im-

age recognition [24], we adapt Transformers for *multiple* tasks within the proposed pipeline shown in Fig. 2, expanding beyond previous applications of Transformers.

First, we apply Transformers-Decoders to improve the input image embeddings by distilling inter-image information between the images by cross-decoding, where each embedding uses the other's embedding as a query. This better encodes images with respect to each other. Second, a Transformer-Encoder computes a stacked multihead attention to encode *cross-attention* between the latent representations of image pairs. Thus, it improves on the 4D correlation volume (4DCV) used in prior works [15, 21, 30, 39, 51], where 4DCVs were calculated by inner products. Instead of a single layer of $N^2$ inner products as in 4DCV, the proposed Transformer-Encoder-based approach leverages multi-head attention's advanced architecture to better encode interactions between activation map entries. Third, we further improve the cross-attention encoding using a cascade of two decoders and a learnt rotation query, to jointly refine the cross-attention encoding and the rotation query. The proposed scheme is a *general-purpose* attention-based architecture for estimating attributes related to two input images such as optical flow, registration, relative pose regression, etc. This work was motivated by extreme rotation estimation, and we reserve other applications for future work, as those will require additional task-specific modifications.

Interestingly, the attention maps computed by our scheme, shown in Section 3.2, show that the Transformer-Encoder assigns high attention scores to image regions containing rotation-informative image cues, emphasizing vertical and horizontal lines. We also observe that the proposed approach can predict the rotation of non-overlapping image pairs with state-of-the-art (SOTA) accuracy. Our framework is end-to-end trainable and optimizes a regression loss. It is evaluated on three dataset benchmarks: StreetLearn [36], SUN360 [54] and InteriorNet [29], with different overlap classes in indoor and outdoor locations and under varying illumination. The experimental results in Section 4 show our model to provide state-of-the-art (SOTA) accuracy.

In summary, our contributions are as follows.

- We propose a novel scheme for estimating extreme rotations, including scenarios with minimal image overlap.
- Image embeddings are enhanced via cross-decoding, distilling inter-image information.
- A Transformer-Encoder cross-attention mechanism is proposed to encode the latent space interactions between image pairs.
- A decoder-decoder module infers relative rotation from the cross-attention encoding by learning and applying quaternionic rotation queries.
- Quantitative evaluations demonstrate favorable performance compared to state-of-the-art rotation estimation techniques on indoor and outdoor datasets.

## 2. Related Work

Our rotation estimation approach represents a specific case of the more general problem of relative pose estimation, particularly relative pose regression (RPR). The prevalent relative rotation estimation technique detects and matches 2D feature points [e.g., SIFT [34], SURF [4]] between images. For pose localization tasks [50], PnP schemes estimate the relative 3D rotation, and the query image camera pose is determined given the anchor image's 3D coordinates and pose. Other schemes for 3D rotation estimation utilized 3D Fourier transforms [22, 23], whose magnitude is invariant to translations. Recent methods apply end-to-end trainable deep networks to both images [3, 14]. Graph neural networks (GNNs) enabled multi-image RPR via aggregating localization cues across video frames [52, 55]. Neural radiance fields (NeRFs) have been explored as an alternative to traditional image or feature point storage for RPR encoding. Some schemes employ rotation-specific parametrizations, notably quaternions and Euler angles, to estimate relative 3D rotations [59]. Such parametrizations, especially quaternions, address the discontinuities intrinsic to rotation representations, attributed to the Double-Cover property. Levinson et al. [27] investigated the SVD orthogonalization approach for 3D rotation estimation via neural networks. By projecting the inferred rotation matrices onto the rotation group using SVD, they showcased that its integration supersedes conventional representations, advancing the state-of-the-art in diverse deep learning paradigms. Further, Mohlin et al. [37] introduced a neural network-based estimation of the parameters for the Fisher distribution matrix, representing the probability distributions of 3D rotations. By optimizing the negative log-likelihood loss of this distribution, they surpassed prior benchmarks in several real-world datasets. Similarly, this optimization methodology was used by Liu et al. to estimate the head pose [33]. As a noteworthy baseline, Rockwell et al. [43] devised a Vision Transformer (ViT) to approximate the eight-point algorithm for direct relative pose estimation between two images, showing competitive performance across diverse scenarios. The methods above predominantly rely on substantial overlap between input image pairs. A pronounced rotation, resulting in limited overlap, could jeopardize the accuracy of these estimations. Specifically, such techniques are ineffective for aligning non-overlapping images. Caspi and Irani [9] demonstrated the feasibility of aligning two image sequences with non-overlapping fields of view in both temporal and spatial dimensions, provided the cameras are in proximity. This alignment leverages shared temporal variations within the sequences. In a parallel vein, Shakil [46] established that multiple nonoverlapping video sequences, captured by uncalibrated video cameras, can be synchronized through inherent temporal fluctuations and inter-frame motion within the sequences.

Extending beyond mere imagery, the challenge of registering non-overlapping RGB-D scans [20, 48] serves as a notable derivative. In such cases, a holistic representation of the scene is typically deduced. In our study, we adopt the framework delineated by Cai et al. [7], where the task is to estimate the extreme relative 3D rotation from a pair of input images. Cai et al. introduced a scheme in which a CNN is used to embed the images, followed by the computation of a 4D correlation volume (4DCV) from the embeddings. An MLP is then applied to this correlation volume, optimizing it using a cross-entropy loss, resulting in state-of-the-art (SOTA) accuracy for non-overlapping images. Intrinsically, 4DCVs are an extension of Bilinear Pooling [25, 31], encoding pairwise correlations across all entities of the 2D embedding maps corresponding to the image pair. Given their encoding capability, 4DCVs have found applications in tasks necessitating long-range spatial correspondences, evident in the RAFT SOTA optical flow [51] and other optical flow models [15, 21, 56]. In a related context, 3D correlation volumes have been utilized in deep stereo matching tasks [18, 28, 30, 39], where the pixels in one image are matched with constrained spatial support in its counterpart. Diverging from these methods, our proposal emphasizes the computation of an analogous 4DCV by evaluating the cross-attention between the activation maps of the image pair through multi-head Transformer-Encoder and an associated activation mask. Specifically, this Transformer-Encoder effectively realizes the functions of multiple aggregated correlation volumes [53] via multi-head attention (MHA). Moreover, we propose an inter-image embedding distillation using Transformer-Decoders, and also improve the rotation inference using a cascaded alternating rotation decoding.

# 3. Rotation Estimation Using Cascaded Attention

The proposed methodology estimates the relative 3D rotation $\mathbf{R} \in \mathbb{R}^{3\times3}$ between input image pairs $\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{R}^{H\times W}$, outlined in Fig. 2. Siamese residual U-nets [57] with weight sharing encode inputs into activation maps $\widehat{\mathbf{I}}_1, \widehat{\mathbf{I}}_2 \in \mathbb{R}^{c\times K_1\times K_2}$, where $c$ is the number of channels and $K_1$, $K_2$ are spatial dimensions. To improve embeddings $\widehat{\mathbf{I}}_1, \widehat{\mathbf{I}}_2$ via cross-decoding, $\widehat{\mathbf{I}}_1, \widehat{\mathbf{I}}_2$ are cross-propagated into weight-sharing Transformer Decoder-0 units. Each input embedding extracts task-relevant representations $\bar{\bar{\mathbf{I}}}_1$ and $\bar{\bar{\mathbf{I}}}_2$.

To further relate the two input images, we compute the cross-attention $\hat{\mathbf{T}}$, an enhanced equivalent of the 4D correlation volume (4DCV) used in prior works [15, 21, 30, 39, 51]. Therefore, we have vectorized the rows $\bar{\bar{I}}_1$ and $\bar{\bar{I}}_2$ as two sequences $\in \mathbb{R}^{c\times K_1K_2}$, concatenated as a single tensor $T\in\ \mathbb{R}^{c\times 2K_1K_2}$, and apply a Transformer-Encoder as in Section 3.2. The rotation is decoded on the basis of cross-attention $\hat{\mathbf{T}}$ using a novel attention-based architecture that uses a cascade of two Transformer-Decoders. Initially, Transformer Decoder-1 is utilized to augment $\hat{\mathbf{T}}$ by incorporating it as a query, with guidance provided by a learnable quaternion $\bar{\mathbf{q}} \in \mathbb{R}^4$ as input. Subsequently, the output of Transformer Decoder-1, denoted as $\bar{\bar{\mathbf{T}}}$, is introduced as input to the subsequent Transformer Decoder-2, further refining the query $\bar{\mathbf{q}}$. Finally, a fully connected MLP layer is applied to predict the relative rotation encoded as quaternion $\tilde{q}$. The 3D rotation is regressed using its quaternion representation, $\mathbf{q}$, as discussed in Section 3.4.

## 3.1. Image Embedding Distillation by Cross-Decoding

Given the embeddings $\widehat{\mathbf{I}}_1, \widehat{\mathbf{I}}_2 \in \mathbb{R}^{c\times K_1\times K_2}$ of the input images, we aim to refine the embeddings by distilling the information between the input images. For that, we apply a Transformer-Decoder that is applied to each embedding, where the other embedding is used as the decoder's query. In order to transform activation maps into Transformer-compatible inputs, we follow the same sequence preparation procedure as in [8]. The activation maps $\widehat{\mathbf{I}}_1, \widehat{\mathbf{I}}_2 \in \mathbb{R}^{c\times K_1\times K_2}$ are first flattened to a sequential representation $\widehat{\mathbf{I}}_1, \widehat{\mathbf{I}}_2 \in \mathbb{R}^{c\times K_1K_2}$. Each position in the activation map is further assigned with a learned encoding to preserve the spatial information of each location. To reduce the number of parameters, two one-dimensional encodings are learned separately for the $X$,$Y$ axes. Specifically, for an activation map $\widehat{\mathbf{I}}$ we define the sets of positional embedding vectors $\mathbf{E}_u \in \mathbb{R}^{K_1\times C/2}$ and $\mathbf{E}_v \in \mathbb{R}^{K_2\times C/2}$, such that a spatial position $(i,j)$, $i \in 1..K_1$, $j \in 1..K_2$, is encoded by concatenating the two corresponding embedding vectors:

$$\mathbf{E}^{i,j}_{pos} = \begin{bmatrix}\mathbf{E}^i_u \\ \mathbf{E}^j_v\end{bmatrix} \in \mathbb{R}^C. \tag{1}$$

The processed sequence, serving as input to the Transformer is thus given by:

$$\widehat{\mathbf{I}} = \widehat{\mathbf{I}} + \mathbf{E}_\mathbf{A} \in \mathbb{R}^{K_1K_2\times C}, \tag{2}$$

where $\mathbf{E}_\mathbf{A}$ is the positional encoding of $\widehat{\mathbf{I}}$.

## 3.2. Cross-Attention Computation using a Transformer-Encoder

The cross-attention between the refinement of representations of input images $\bar{\bar{I}}_1$ and $\bar{\bar{I}}_2$ is computed using a Transformer-Encoder with $l = 2$ layers and $h = 4$ attention heads for each layer. An ablation study of this configuration is given in Section 4.6. The cross-attention maps computed by the Transformer-Encoder are an improved equivalent of the 4D correlation volumes [7, 51], encoding the interactions (inner-products) between *all* the im-

[1] https://github.com/dekelshay/AttExtremeRotation

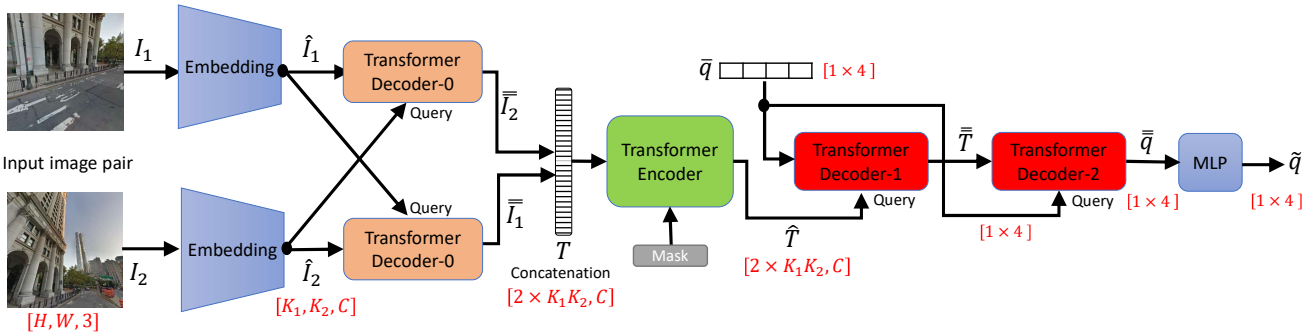


Figure 2. **The proposed architecture** utilizes weight-sharing Siamese CNNs to encode the input image pair $(I_1, I_2) \in \mathbb{R}^{H \times W}$ into feature maps $(\hat{I}_1, \hat{I}_2)$. These feature maps are then cross-decoded by the weight sharing Transformer Decoder-0 layers, cross-distilling $(\hat{I}_1, \hat{I}_2)$ into the representations $\bar{\bar{I}}_1$ and $\bar{\bar{I}}_2$. The concatenated refined embeddings $T$ are input to the Transformer-Encoder alongside an attention mask $M$ to derive the cross-attention encoding $\hat{T}$. $\hat{T}$ enters a cascade of two Transformer Decoders, where the first, Transformer Decoder-1, enhances the cross-attention as $\bar{\bar{T}}$, guided by the learned quaternion rotation query $\bar{q}$. The second, Transformer Decoder-2, encodes the rotation as $\bar{\bar{q}}$, transformed via a multilayer perceptron (MLP) to predict the relative quaternion rotation $\tilde{q}$.

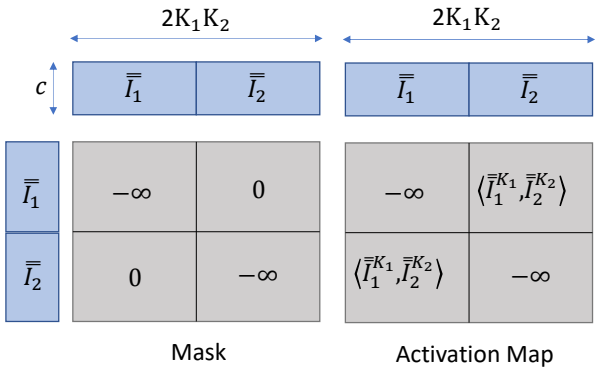


Figure 3. Computing the cross-attention using a Transformer-Encoder and the input mask, $\mathbf{M}$. The mask $\mathbf{M}$ zeros the self-attention terms, retaining only the cross-attention terms.

age cues in the activation maps. By default, a Transformer-Encoder computes the self-attention maps of the input sequence. Hence, the cross-attention $\hat{\mathbf{T}}$ of the vectorized and concatenated activation maps $T$ is computed by applying the attention mask $\mathbf{M}$ given in Eq. 3. The mask $\mathbf{M}$ nullifies the self-attention terms in the attention maps computed throughout the Transformer-Encoder, while retaining the cross-attention terms,

$$\mathbf{M} = \begin{bmatrix} -\infty & -\infty & \cdots & 0 & 0 \\ -\infty & -\infty & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & -\infty & -\infty \\ 0 & 0 & \cdots & -\infty & -\infty \end{bmatrix} \tag{3}$$

The use of the mask $\mathbf{M}$ and the corresponding structure of the attention maps is shown in Fig. 3. Any pair of image patches could hold valuable information about the overall geometric relationships in an image. The Transformer-Encoder can uncover these hints implicitly. The regions in the input images that contain rotation-related cues, explicitly or implicitly, receive higher attention scores, as seen in Section 2 of the Supplementary Materials section. This leads to a more meaningful and concise input for the distillation and the subsequent MLP layer, ultimately improving the estimation accuracy. The same as in [7], even when the image pairs are non-overlapping, the Transformer-Encoder formulation can predict the rotation using straight lines only present in a single image, the same as human cognitive capabilities. For example, in the extreme scenario of non-overlapping image pairs the roll angle can be estimated from a single image, by implicitly assuming that buildings and their edges are perpendicular to ground level. Similarly, the relative elevation angle can be estimated by assuming that the streets and pavements are parallel to the ground plane or by computing the corresponding vanishing points. Most training and test datasets in this domain depict urban scenes, adhering to these assumptions.

### 3.3. Cascaded Attention-based Decoding

Given the cross-attention tensor $\hat{\mathbf{T}}$ that encodes interrelations between the paired input images, our objective is inferring the 3D relative rotation quaternion. To achieve this, we propose an innovative cascaded decoding scheme that alternately refines both the query rotation and cross-attention. Initially, Transformer Decoder-1, enhances $\hat{\mathbf{T}}$ based on the learned quaternion $\bar{\mathbf{q}}$ to compute $\bar{\bar{\mathbf{T}}}$. Next, the refined cross-attention $\bar{\bar{\mathbf{T}}}$ is queried by $\bar{\mathbf{q}}$ to deduce the rotated encoding $\bar{\bar{\mathbf{q}}}$ using Transformer Decoder-2. This cascaded inference approach could be extended via additional dual units. However, we observed no performance gains from additional cascades. Since the decoder inputs are semantic representations, positional encodings are excluded.

### 3.4. Relative Rotation Regression

The encoded quaternion vector $\bar{\bar{\mathbf{q}}}$ is subsequently input to a Multi-Layer Perceptron (MLP) regressor, computing the quaternion output denoted $\tilde{\mathbf{q}}$. This resultant quaternion is given by $\mathbf{q} = [q_w, q_x, q_y, q_z]$. The training loss is formulated as:

$$\mathbf{L} = \mid q_0 - \tilde{\mathbf{q}}/||\tilde{\mathbf{q}}|| \mid_2, \tag{4}$$

where $q_0$ and $\tilde{q}$ are the groundtruth and predicted quaternions, respectively. Normalization ensures that the quaternion is a valid 3D rotation representation.

## 4. Experimental Results

The proposed scheme was experimentally verified by applying it to contemporary benchmark datasets with overlapping and nonoverlapping image pairs. Our experimental setup rigorously adhered to the paradigm established by Cai et al. [7], using identical datasets and image overlap categories. Utilizing their provided source code[2], to create perspective views from panoramic images, ensuring that the input images were the same in both studies, allowing fair comparisons with previous SOTA results and other contemporary schemes. For that, we also used the same Residual-Unet backbone network [58] as in [7]. Section 4.1 details the image datasets we used and their processing, according to Cai et al. [7], to derive the training and test datasets. Training details are given in Section 4.2. We compare with recent SOTA schemes listed in Section 4.3 using the geodesic error measure used in previous work [7]

$$\mathbf{E} = \arccos\left(\frac{tr(\mathbf{R}^T\mathbf{R}^*) - 1}{2}\right), \tag{5}$$

where $\mathbf{R}$ is the predicted rotation matrix and $\mathbf{R}^*$ is the groundtruth relative rotation matrix for each image pair. The experimental comparisons are reported in Section 4.4 and the attention maps are visualized in Section 2 of the Supplementary Materials to provide an intuitive interpretation of the cross-attention scores computed by the Transformer-Encoder. We studied the cross-dataset generalization properties of the proposed scheme in Section 4.5, while ablation studies of the different parameters, design choices and parameters are reported in Section 4.6.

### 4.1. Image Datasets and their Processing

We used the following datasets and train/test splits used in previous works:

**InteriorNet** [29] is a synthetic data set to understand and map interior scenes. A subset of 10,050 panoramas from 112 different houses was used, where the images of 82 houses were used for training and those of 30 houses were used for testing, respectively.

[2] https://github.com/RuojinCai/ExtremeRotation_code

**StreetLearn** [36] is an outdoor dataset consisting of approximately 140,000 panoramic views of Pittsburgh and Manhattan. We used 56K panoramic views from Manhattan, from which we randomly chose 1000 panoramic views for testing.

**SUN360** [54] is an indoor collection of high-resolution panoramas that cover a full view of $360° \times 180°$ for a variety of environmental scenes downloaded from the Internet. It also provides location category labels. We used 7K and 2K panoramas for training and testing, respectively.

As these datasets contain panoramic images, we generated 200 perspective $128 \times 128$ images by randomly cropping 200 different locations in each panoramic image. This sampling strategy ensures a consistent distribution of ground-truth image pairs with pitch resolutions spanning $[-45°, 45°]$ and yaw resolutions encompassing $[-180°, 180°]$. We estimate only the Yaw and the Pitch angles, presuming a null roll between paired images. We avoided generating textureless image pairs, that is, images that mainly contain ceilings or floors in a house or skies in an outdoor scenes, by limiting the pitch rang to $[-45°, 45°]$ for the outdoor dataset and $[-30°, 30°]$ for the indoor datasets. There is no overlap between the train and test datasets. To compare our results with prior research and to analyze the influence of camera translation on our rotation estimation approach, we partitioned the InteriorNet and StreetLearn datasets into two groups: images with and without camera translations. The non-translated images were acquired by randomly selecting pairs of cropped images from a single panorama. In contrast, datasets that include translations (known as StreetLearn-T and InteriorNet-T) were generated by randomly selecting pairs of cropped images from different panoramas, where translations are less than $3m$. However, our method was not used to estimate these translations. We evaluated our performance in overlapping and nonoverlapping pairs and use the setup of Cai et al. [7] by dividing the datasets into three overlap classes:

**Large**, contains highly overlapping pairs up to relative rotations of $45°$

**Small**, contains pairs that partially overlap with relative rotation angles $\in [45°, 90°]$

**None**, contains pairs without overlap with relative rotations $> 90°$.

### 4.2. Training Details

We use a pre-trained Residual-Unet [58] (same as in Cai et al. [7]) as a backbone to compute the feature maps of the two input images $(\hat{I}_1, \hat{I}_2) \in \mathbb{R}^{c \times k_1 \times k_2} = \mathbb{R}^{128 \times 32 \times 32}$. According to Fig. 2, subsequently, these feature maps are cross-propagated into dedicated decoder units, resulting in the refinement of representations $\bar{\bar{I}}_1$ and $\bar{\bar{I}}_2$. The refinements of the representations $\bar{\bar{I}}_1$ and $\bar{\bar{I}}_2$ were reshaped and con-

catenated along the axis of the samples to form the tensor $\mathbf{T} \in \mathbb{R}^{(2\cdot32\cdot32)\times128} = \mathbb{R}^{2048\times128}$. $\mathbf{T}$ was the input to the Transformer-Encoder, consisting of $l = 2$ layers with ReLU nonlinearity and a dropout of $p = 0.1$. Each encoder layer uses $h = 4$ MHA heads and a hidden dimension of $C_h = 768$. An ablation study of the Transformer-Encoder parameters is given in Section 4.6. The Transformer-Encoder's output $\hat{T}$ is then fed into a dual-path structure comprising two concatenated decoders. The primary decoder receives a learnt quaternion vector initialized by white Gaussian noise, $\bar{q}$, as input and the cross-attention $\hat{T}$ as a query, and produces $\bar{T}$, enhancing contextual nuances, while the subsequent decoder gets $\bar{T}$ as an input and the same empty quaternion vector, $\bar{q}$ as a query, and generates $\bar{\bar{q}}$, encapsulating pivotal rotational attributes. The two sequential attention-based decoders use $l = 2$ layers with $h = 2$ MHA heads and a hidden dimension of $C_h = 768$. Finally, the MLP regressor that computes the quaternion representation for the regression loss in Eq. 4. The MLP regressor contains two fully connected layers. Throughout all experiments, the model is optimized using an Adam optimizer with an initial learning rate of $\lambda = 5e-4$, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-10}$, and a batch size of 20. Our model is implemented in PyTorch, it is end-to-end trainable, and all experiments were performed on an 8GB NVIDIA GeForce GTX 2080 GPU.

### 4.3. Comparative baselines

In line with Cai et al. [7], we compare our method with contemporary schemes using the datasets in Section 4.1:

**A SIFT-based approach [6].** A method for matching SIFT features [34] using RANSAC [17] in image pairs of the same panorama, and estimating the relative rotation matrix using Homography equations or the Essential matrix.

**CNN-based methods [13].** Deep learning schemes that detect and encode local image features using SuperPointNet [13] and D2-Net [16].

**Self-supervised interest point [59].** A scheme by Zhou et al. [59] (Reg6D) that applies a CNN to approximate the mappings between various rotation representations and fits continuous 5D and 6D rotation representations, instead of the commonly used Euler and quaternion representations.

**Extreme rotation estimation [7].** A deep learning technique to estimate the relative 3D rotation of image pairs in an extreme setting [7] where the images have little or no overlap. They proposed a network that automatically learns implicit visual cues by computing a 4D correlation volume.

**Attention-based methods.** We also compare to recent work by Rockwell et al. [43] (8PointVit) using a Vision Transformer (ViT) to estimate the relative pose. Although Rockwell et al. achieve competitive results in multiple settings, their approach is less suited for extreme view changes.

### 4.4. Experimental comparisons

The results of the comparison of the proposed scheme with baselines and SOTA schemes are reported in Table 1. We report the mean and median of the geodesic error given in Eq. 5 and the percentage of image pairs whose estimated relative rotation error was less than $10°$. We compared the accuracy of our proposed model to the schemes detailed in Section 4.3. The proposed approach is shown to be accurate for both indoor and outdoor scenes and significantly outperforms the baseline schemes in all overlap categories. For nonoverlapping pairs, correspondence-based methods such as SIFT [6], SuperPointNet [13], Reg6D [59] and 8PointViT [43] failed to provide any estimates, as they require feature correspondence. The DenseCorrVol approach [7] provides accurate results in extreme cases, but our approach outperforms it. Qualitative experimental results are given in Section 1 of the Supplementary Materials section. The qualitative results of the rotation estimation are shown in Fig. 4 for the StreetLearn and SUN360 datasets, for the large, small and nonoverlapping cases. We show the full panoramas, the footprints of the cropped images that were used as inputs for the proposed scheme and the footprint of the estimated image crop based on the estimated rotation. In all cases, we achieve high estimation accuracy.

### 4.5. Cross-Dataset Generalization

The cross-dataset generalization properties of our approach were evaluated using the Holicity dataset [60]. The Manhattan dataset was used to train the models, while the London dataset was used for testing. The test images were divided into three overlap classes according to Section 4.1. We compared the generalization of our approach with Cai et al.'s [7]. The results, in Table 2, show that our approach outperformed Cai et al in all overlap classes.

### 4.6. Ablation Study

**Transformer-Encoder parameters.** Table 3 summarizes multiple Transformer-Encoder configurations for each overlap category. The expressive power of the Transformer-Encoder depends on the number of heads and layers. The more are used, the better the expressive power. However, using an excessive number might lead to overfitting, and the optimal constellation, in terms of accuracy in Table 3, is given by $h = 4$, $l = 2$. In particular, this constellation is a sweetspot so that increasing the number of heads or layers results in reduced accuracy.

**Backbone ablation.** In Table 4, we examine the depth of the Residual-Unet [58] backbone by altering the number of residual blocks it contains. Increasing the number of residual blocks enhances the backbone's expressive capability, but an excessively deep architecture may result in overfitting. We found that using three residual blocks is the optimal choice, which aligns with our original decision.

| Overlap | Method | InteriorNet | | | InteriorNet-T | | | SUN360 | | | StreetLearn | | | StreetLearn-T | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Avg(°↓) | Med(°↓) | 10°(%↑) | Avg(°↓) | Med(°↓) | 10°(%↑) | Avg(°↓) | Med(°↓) | 10°(%↑) | Avg(°↓) | Med(°↓) | 10°(%↑) | Avg(°↓) | Med(°↓) | 10°(%↑) |
| Large | SIFT* [34] | 6.09 | 4.00 | 84.86 | 7.78 | 2.95 | 55.52 | 5.46 | 3.88 | 93.10 | 5.84 | 3.16 | 91.18 | 18.86 | 3.13 | 22.37 |
| | SuperPoint* [13] | 5.40 | 3.53 | 87.10 | 5.46 | 2.79 | 65.97 | 4.69 | 3.18 | 92.12 | 6.23 | 3.61 | 91.18 | 6.38 | 1.79 | 16.45 |
| | Reg6D [59] | 9.05 | 5.90 | 68.49 | 17.00 | 11.95 | 41.79 | 16.51 | 12.43 | 40.39 | 11.70 | 8.87 | 58.24 | 36.71 | 24.79 | 23.03 |
| | DenseCorrVo [7] | 1.53 | 1.10 | 99.26 | 2.89 | 1.10 | 97.61 | 1.00 | 0.94 | **100.00** | 1.19 | 1.02 | 99.41 | 9.12 | 2.91 | 87.50 |
| | 8PointViT [43] | 0.48 | 0.40 | **100.00** | 2.90 | 1.83 | 97.91 | - | - | - | 0.62 | 0.52 | **100.00** | 4.08 | 2.43 | **90.13** |
| | Ours | **0.43** | **0.38** | 99.65 | **1.75** | **0.95** | **98.8** | **0.85** | **0.45** | 99.95 | **0.58** | **0.48** | 99.31 | **3.88** | **1.69** | 87.20 |
| Small | SIFT* [34] | 24.18 | 8.57 | 39.73 | 18.16 | 10.01 | 18.52 | 13.71 | 6.33 | 56.77 | 16.22 | 7.35 | 55.81 | 38.78 | 13.81 | 5.68 |
| | SuperPoint* [13] | 16.72 | 8.43 | 21.58 | 11.61 | 5.82 | 11.73 | 17.63 | 7.70 | 26.69 | 19.29 | 7.60 | 24.58 | 6.80 | 6.85 | 0.95 |
| | Reg6D [59] | 25.71 | 15.56 | 33.56 | 42.93 | 28.92 | 23.15 | 42.55 | 32.11 | 9.40 | 24.77 | 15.11 | 30.56 | 46.61 | 34.33 | 13.88 |
| | DenseCorrVol [7] | 6.45 | 1.61 | 95.89 | 10.24 | 1.38 | 89.81 | 3.09 | 1.41 | 98.50 | 2.32 | 1.41 | 98.67 | 13.04 | 3.49 | 84.23 |
| | 8PointViT [43] | 1.84 | 0.94 | 99.32 | 4.48 | 2.38 | 96.30 | - | - | - | 1.46 | 1.09 | **100.00** | 9.19 | 3.25 | 87.7 |
| | Ours | **1.55** | **0.872** | **99.85** | **4.25** | **0.777** | **97.55** | **2.109** | **0.831** | **98.99** | **1.21** | **0.718** | 99.122 | **7.48** | **1.8666** | **88.996** |
| None | SIFT* [34] | 109.30 | 92.86 | 0.00 | 93.79 | 113.86 | 0.00 | 127.61 | 129.07 | 0.00 | 83.49 | 90.00 | 0.38 | 85.90 | 106.84 | 0.38 |
| | SuperPoint* [13] | 120.28 | 120.28 | 0.00 | – | – | 0.00 | 149.80 | 165.24 | 0.00 | – | – | 0.00 | – | – | 0.00 |
| | Reg6D [59] | 48.36 | 32.93 | 10.82 | 60.91 | 51.26 | 11.14 | 64.74 | 56.55 | 3.77 | 28.48 | 18.86 | 24.39 | 49.23 | 35.66 | 11.86 |
| | 8PointViT [43] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | DenseCorrVol [7] | 37.69 | 3.15 | 61.97 | 49.44 | 4.17 | 58.36 | 34.92 | 4.43 | 61.39 | 5.77 | 1.53 | **96.41** | 30.98 | 3.50 | **72.69** |
| | Ours | **35.13** | **2.814** | **65.20** | **45.32** | **4.05** | **59.56** | **32.46** | **4.19** | **63.17** | **5.33** | **1.20** | 96.22 | **28.13** | **3.25** | 72.43 |

Table 1. Relative rotation estimation results. We utilized the InteriorNet, SUN360, and StreetLearn datasets and show the average and median of geodesic errors. We also present the percentage of image pairs with relative rotation error below 10°, for the overlap categories in Section 4.1. The gray numbers indicate errors exceeding 50%. The asterisk $*$ signifies that the mean and median errors did not lead to pose estimation, and their calculations are performed only on successful image pairs.

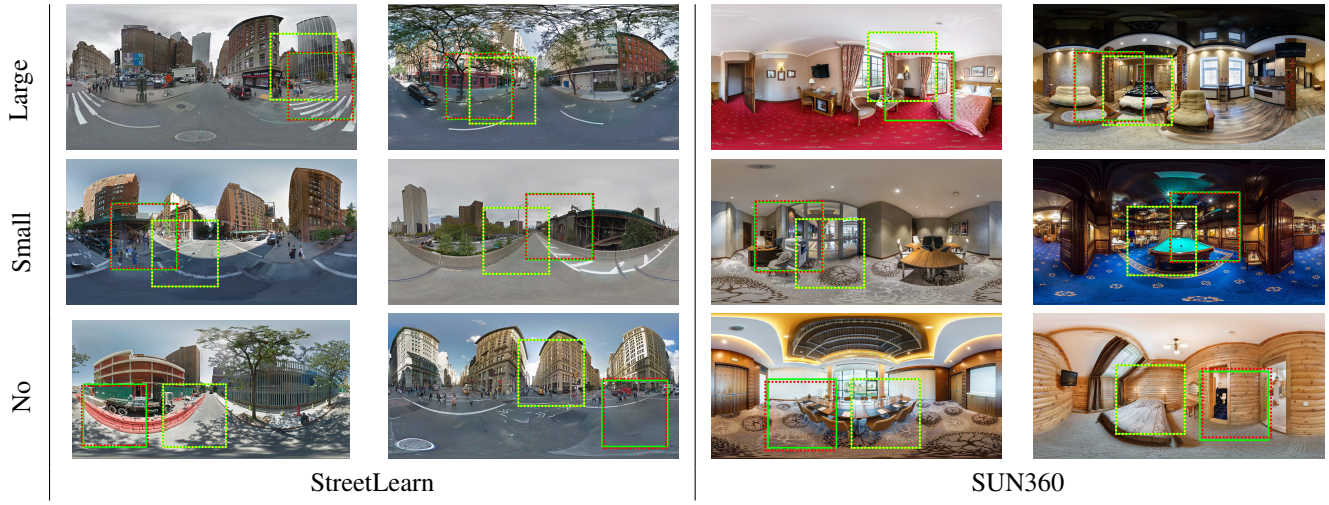


Figure 4. Rotation estimation results. The panoramic and cropped groundtruth images are marked with green and yellow dots. The predicted footprint of one of the cropped images is marked by the red-dotted line. The first row shows the results of the matching of images with large overlaps. The second and last rows show the matching of small overlap and non-overlapping images.

| Overlap | Ours [°] | Cai et al. [7] [°] |
|---|---|---|
| Large | **9.55** | 11.23 |
| Small | **16.33** | 20.87 |
| None | **38.48** | 40.82 |

Table 2. Cross-dataset generalization. We trained the models on the Manhattan dataset and tested them on the London dataset. The average geodesic error is reported.

**3D rotation encoding and training losses.** The ablations of different rotation encodings and their corresponding training losses were evaluated and are presented in Table 5. The evaluation was performed by applying the Residual-Unet backbone [58] and a proposed Transformer-Encoder-based cross-attention method to image pairs from the Stree-Learn dataset with large overlaps. For the discrete formulation, in line with Cai et al. [7], the pitch and yaw angles were discretized into 360 bins $\in [-180°, 180°]$, and a cross-entropy loss was used to train the network. These results were compared to those obtained using the $L_2$ regression loss, as described in Eq. 4 in our scheme. The results in Table 5 show that our $L_2$ regression outperforms the discrete Euler angle approach proposed by Cai et al. [7].

**Architecture Ablations** To evaluate the proposed architecture and assess the contribution of each proposed com-

| Overlap | Heads | Layers | Rotational error Avg [°] | Med [°] | 10°[%] |
|---|---|---|---|---|---|
| Large | 1 | 1 | 1.21 | 0.97 | 99.1 |
| | 4 | 1 | 0.82 | 0.65 | 99.1 |
| | 2 | 2 | 0.87 | 0.71 | 99.12 |
| | 4 | 2 | **0.58** | **0.48** | **99.31** |
| | 4 | 4 | 0.65 | 0.59 | 99.2 |
| Small | 1 | 1 | 7.13 | 3.26 | 94.99 |
| | 4 | 1 | 6.42 | 2.46 | 95.32 |
| | 2 | 2 | 5.44 | 2.05 | 96.55 |
| | 4 | 2 | **1.21** | **0.718** | **99.122** |
| | 4 | 4 | 4.61 | 0.98 | 95.43 |
| None | 1 | 1 | 7.43 | 2.88 | 91.55 |
| | 4 | 1 | 6.15 | 2.55 | 92.55 |
| | 2 | 2 | 6.23 | 3.02 | 91.35 |
| | 4 | 2 | **5.33** | **1.20** | **96.22** |
| | 4 | 4 | 5.78 | 1.87 | 95.22 |

Table 3. Ablation of the Transformer-Encoder parameters using the StreeLearn dataset. For each overlap class, there is an optimal configuration that balances the Transformer-Encoder's expressive power and overfitting.

| Backbone | Layers | Rotational error Avg [°] | Med [°] | 10°[%] |
|---|---|---|---|---|
| 1 | Conv(k=7, s=2, d=64)<br>1× Residual blocks<br>Conv(k=3, s=1, d=512)<br>Conv(k=3, s=1, d=256) | 2.95 | 1.53 | 94.13 |
| 2 | Conv(k=7, s=2, d=64)<br>2× Residual blocks<br>Conv(k=3, s=1, d=512)<br>Conv(k=3, s=1, d=256) | 1.75 | 1.05 | 96.13 |
| 3 | Conv(k=7, s=2, d=64)<br>**3 × Residual blocks**<br>Conv(k=3, s=1, d=512)<br>Conv(k=3, s=1, d=256) | **1.21** | **0.7818** | **99.122** |
| 4 | Conv(k=7, s=2, d=64)<br>4× Residual blocks<br>Conv(k=3, s=1, d=512)<br>Conv(k=3, s=1, d=256) | 1.45 | 0.86 | 97.12 |

Table 4. Backbone Ablation. We evaluate the depth of the Residual-Unet backbone network [58], used in our scheme, by changing the number of residual blocks.

ponent, the use of the different Transformer Decoders in particular, we conducted a series of experiments employing various architectural variations of the proposed architecture introduced in Section 3 and Fig. 2. The results are shown in Table 6 and the corresponding architectures are shown in the

| Representation | Loss function | Rotational error Avg [°] | Med [°] | 10°[%] |
|---|---|---|---|---|
| Quaternions | $L_2$ Regression | **1.21** | **0.78** | **99.122** |
| Euler angles | Cross-Entropy | 2.37 | 1.46 | 98.13 |
| Euler angles | $L_2$ Regression | 2.22 | 1.32 | 98.99 |

Table 5. Ablation of 3D rotation encoding and training losses. We compare the 3D rotations encodings by Euler angles and quaternions in discrete and continuous domains and the corresponding training losses.

Supplementary Materials. In each experiment, we used a particular partial configuration of the proposed Transformer Decoders and evaluated the resulting estimation error using the StreetLearn dataset with large overlaps between the input images. The results in Table 6 show that the proposed configuration outperforms all other configurations. In particular, configuration #1 shows that using the sequential attention-based decoders, TD1 and TD2 improves the accuracy significantly. The cross-decoding by TD0 provides additional, but not as significant improvement.

| | TD0 | TD1 | TD2 | Avg [°] | Med [°] | 10°[%] |
|---|---|---|---|---|---|---|
| 0 | + | + | + | **0.58** | **0.48** | **99.3** |
| 1 | - | + | + | 0.98 | 0.81 | 95.15 |
| 2 | - | - | - | 3.35 | 2.44 | 88.16 |
| 3 | - | + | - | 1.76 | 1.55 | 93.12 |
| 4 | + | + | - | 0.86 | 0.72 | 95.64 |
| 5 | + | - | - | 1.97 | 1.65 | 92.82 |

Table 6. Architectural ablation study. We compare the estimation accuracy of different configurations of Transformers Decoders (TDs). The corresponding architectures are shown in the Supplementary Materials, and the first configuration is shown in Section 3 and Fig. 2.

## 5. Conclusion

We present a novel formulation for estimating the relative rotation between a pair of images. In particular, we study the estimation of rotations between images with small and no overlap. We propose an attention-based approach using a Transformer-Encoder to calculate the cross-attention between image pair embedding maps, which outperforms the previous use of 4D correlation volumes [7, 51] and a decoder-decoder mechanism to estimate the output quaternion. Our framework can be trained end-to-end and optimizes a regression loss. It has been experimentally shown to outperform previous SOTA schemes [7] on multiple datasets used in contemporary work. In particular, for the challenging small and nonoverlapping cases.

## References


[1] Hadar Averbuch-Elor and Daniel Cohen-Or. Ringit: Ring-ordering casual photos of a temporal event. *ACM Trans. Graph.*, 34(3), 2015. 1

[2] Alexandru O. Balan, Michael J. Black, Horst Haussecker, and Leonid Sigal. Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 1

[3] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *European Conference on Computer Vision (ECCV)*, September 2018. 2

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006. 2

[5] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6684–6692, 2017. 1

[6] Matthew Brown, Richard I Hartley, and David Nistér. Minimal solutions for panoramic stitching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 6

[7] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 4, 5, 6, 7, 8

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, pages 213–229, Cham, 2020. Springer International Publishing. 1, 3

[9] Yaron Caspi and Michal Irani. Aligning non-overlapping sequences. *International Journal of Computer Vision*, 48(1):39–51, 2002. 2

[10] J.M. Coughlan and A.L. Yuille. Manhattan world: compass direction from a single image by bayesian inference. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 941–947 vol.2, 1999. 1

[11] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. In *IEEE International Conference on Computer Vision (ICCV)*, pages 434–441. IEEE Computer Society, 1999. 1

[12] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007. 1

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2018. 6, 7

[14] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera re-localization. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 2, 3

[16] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[17] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 6

[18] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2020. 3

[19] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1

[20] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, and Konrad Schindler Andreas Wieser. Predator: Registration of 3D point clouds with low overlap. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[21] Di Jia, Kai Wang, ShunLi Luo, TianYu Liu, and Ying Liu. Braft: Recurrent all-pairs field transforms for optical flow based on correlation blocks. *IEEE Signal Processing Letters*, 28:1575–1579, 2021. 2, 3

[22] Y. Keller, A. Averbuch, and Y. Shkolnisky. Algebraically accurate volume registration using euler's theorem and the 3D pseudopolar FFT. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 795–800 vol. 2, 2005. 2

[23] Y. Keller, Y. Shkolnisky, and A. Averbuch. Volume registration using the 3D pseudopolar fourier transform. *IEEE Transactions on Signal Processing*, 54(11):4323–4331, 2006. 2

[24] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[25] S. Kong and C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7025–7034, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 3

[26] Jinwoo Lee, Minhyuk Sung, Hyunjoon Lee, and Junho Kim. Neural geometric parser for single image camera calibration. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer*

*Vision (ECCV)*, pages 541–557, Cham, 2020. Springer International Publishing. 1

[27] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of SVD for deep rotation estimation. *Advances in Neural Information Processing Systems (NIPS)*, 33, 2020. 2

[28] Minhua Li, Qingling Chang, Yuhan Wang, Xinglin Liu, Shiting Xu, and Yan Cui. Stereo matching with multiscale hybrid cost volume. *IEEE Access*, 10:100128–100136, 2022. 3

[29] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference*, 2018. 2, 5

[30] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 3

[31] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[32] Chenxi Liu, Alex Schwing, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Rent3d: Floor-plan priors for monocular layout estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[33] Hai Liu, Shuai Fang, Zhaoli Zhang, Duantengchuan Li, Ke Lin, and Jiazhang Wang. Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Transactions on Multimedia*, 24:2449–2460, 2022. 2

[34] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2, 6, 7

[35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[36] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The StreetLearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019. 2, 5

[37] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. *Advances in Neural Information Processing Systems (NIPS)*, 33, 2020. 2

[38] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

[39] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3283–3291, 2019. 2, 3

[40] Onur Ozyesil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *arXiv preprint arXiv:1701.08493*, 2017. 1

[41] Yiming Qian and James H. Elder. A reliable online method for joint estimation of focal length and camera rotation. In *European Conference on Computer Vision (ECCV)*, page 249–265, Berlin, Heidelberg, 2022. Springer-Verlag. 1

[42] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[43] Chris Rockwell, Justin Johnson, and David F. Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *International Conference on 3D Vision (3DV)*, 2022. 2, 6, 7

[44] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[45] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6896–6906, 2018. 1

[46] O. Shakil. An efficient video alignment approach for non-overlapping sequences with free camera movement. In *ICASSP*, volume 2, pages II–II, 2006. 2

[47] Yoli Shavit and Yosi Keller. Camera pose auto-encoders for improving pose regression. In *European Conference on Computer Vision (ECCV)*, pages 140–157, Cham, 2022. Springer Nature Switzerland. 1

[48] Che Sun, Yunde Jia, Yi Guo, and Yuwei Wu. Global-aware registration of less-overlap rgb-d scans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6357–6366, June 2022. 3

[49] Dekel (Basha) T., Moses Y., and Avidan S. Photo sequencing. *International Journal of Computer Vision*, 110(3):275 – 289, 2014. Cited by: 8. 1

[50] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7199–7209, 2018. 2

[51] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, page 402–419, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 3, 8

[52] M. Turkoglu, E. Brachmann, K. Schindler, G. J. Brostow, and A. Monszpart. Visual camera re-localization using graph neural networks and relative pose supervision. In *International Conference on 3D Vision (3DV)*, pages 145–155, Los Alamitos, CA, USA, dec 2021. 2

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 30. Curran Associates, Inc., 2017. 1, 3

[54] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic

place representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2695–2702. IEEE, 2012. 2, 5

[55] Fei Xue, Xin Wu, Shaojun Cai, and Junqiu Wang. Learning multi-view camera relocalization with graph neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11372–11381, 2020. 2

[56] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 32. Curran Associates, Inc., 2019. 3

[57] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15:749–753, 2018. 3

[58] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 5, 6, 7, 8

[59] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 2, 6, 7

[60] Yichao Zhou, Jingwei Huang, Xili Dai, Linjie Luo, Zhili Chen, and Yi Ma. HoliCity: A city-scale data platform for learning holistic 3D structures. 2020. arXiv:2008.03286 [cs.CV]. 6

[61] Rui Zhu, Xingyi Yang, Yannick Hold-Geoffroy, Federico Perazzi, Jonathan Eisenmann, Kalyan Sunkavalli, and Manmohan Chandraker. Single view metrology in the wild. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, pages 316–333, Cham, 2020. Springer International Publishing. 1