TWO-PHASE MULTIMODAL IMAGE FUSION USING CONVOLUTIONAL NEURAL NETWORKS

Kushal Kusram, Shane Transue, and Min-Hyung Choi

Department of Computer Science and Engineering
University of Colorado Denver

ABSTRACT

The fusion of multiple imaging modalities presents an important contribution to machine vision, but remains an ongoing challenge due to the limitations in traditional calibration methods that perform a single, global alignment. For depth and thermal imaging devices, sensor and lens intrinsics (FOV, resolution, etc.) may vary considerably, making per-pixel fusion accuracy difficult. In this paper, we present *AccuFusion*, a two-phase non-linear registration method to fuse multimodal images at a per-pixel level to obtain an efficient and accurate image registration. The two phases: the Coarse Fusion Network (CRN) and Refining Fusion Network (RFN), are designed to learn a robust image-space fusion that provides a non-linear mapping for accurate alignment. By employing the refinement process, we obtain per-pixel displacements to minimize local alignment errors and observe an increase of 18% in average accuracy over global registration.

Index Terms— Nonlinear Image Registration, Depth-Thermal Fusion, Two-phase CNN, Multimodal Imaging

1. INTRODUCTION

The registration of multimodal images represents the process of aligning and integrating multiple image streams into a composite image. It has been well-studied for combining thermal and 3D imaging data [1, 2, 3]. This *fusion* process provides multimodal sensor data that is crucial due to its significance in facial authentication [3], autonomous vehicles [4], remote sensing [5], medical imaging [6], and 6D-SLAM environmental reconstruction[7]. Numerous multimodal image registration methods have been proposed, with approaches including stereoscopic calibration and real-time feature correlation [8]. These common registration methods involve the following steps: feature detection, template matching, and transformation estimation [1]. However, determining accurate correspondence becomes a hurdle due to the challenges involved in detecting accurate and reliable common features [9, 10] given the inconsistent edges within each image type, aggravated by wide discrepancies in FOV and resolution. To address this, we integrate the efficiency of global rigid registration with localized non-linear displacements into a two-stage convolutional fusion network. This enables us to leverage the benefits of *scene independence* and efficiency of rigid methods with the improved accuracy of localized deformations provided by non-linear registrations.

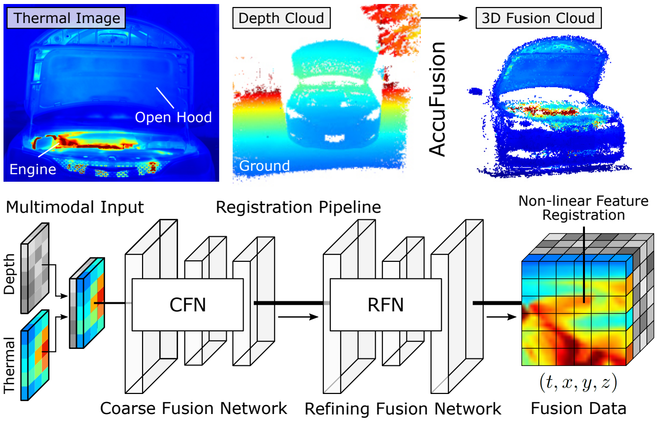

Fig. 1. Two-stage deep learning solution for non-linear image registration for real-time thermal-depth fusion.

Work in multimodal registration has largely focused on adapting stereoscopic calibration through a generated homography transformation T using various depth-thermal template designs [2]. Towards improving this level of accuracy, we build on recent advances that have shifted to utilizing *scene dependent* features with *image-space* or pixel-to-pixel correspondence [11] and extend these approaches to provide an accurate dense transformation space $\vec{T}(i, j)$ with displacements for each pixel (i, j). These pixel-level displacements can improve accuracy by optimizing the global error during training. The problem impeding this approach is that non-linear feature alignment is an expensive run-time process. We must identify how we can store dense image-space displacements in a model that can process real-time data streams.

In our solution, we formulate a two-stage model that performs a *global preconditioning* transformation followed by *localized deformations* that perform per-pixel transformations. We achieve this by creating a non-linear registration with localized deformations that integrates depth and thermal data into spatial point-cloud data as shown in Fig. 1. By performing dense image-space alignments performed at a pixel-level, our method provides an efficient method for multimodal fusion, even for device pairs with large differences in Field-of-Views (FOVs) and resolutions.

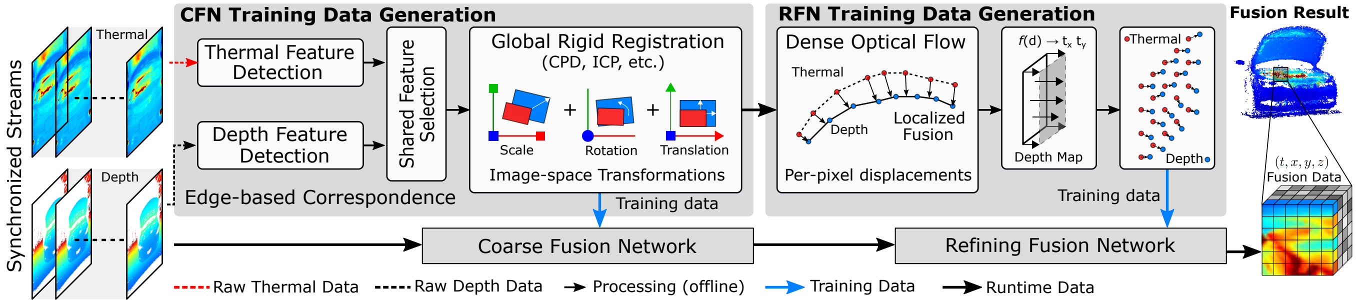

Fig. 2. Overall architecture: Dual streams (depth-thermal) are provided to our two-phase network. Training is performed to identify the optimal alignment between edge features, forming a model for per-pixel displacements for non-rigid alignments.

2. PROPOSED METHOD: ACCUFUSION

We propose a new form of multimodal fusion through deep learning. We combine two key phases: (1) a global rigid alignment through a preconditioning network and (2) a pixel-to-pixel non-rigid alignment network that predicts localized alignments that can reduce the overall fusion error. These correspond to the Coarse Fusion Network (CFN) and Refining Fusion Network (RFN). Our solution assumes the provision of depth and thermal images are available for training. The run-time execution of the networks generates a displacement field $\vec{T}_{i,j}$ as an output that can be applied to the input thermal stream to perform a refined alignment to the depth stream.

2.1. Coarse Fusion Network (CFN)

The CRN serves as a *preconditioning* alignment towards accurate fusion by aligning depth I_d with thermal I_t images at a global level. The network architecture consists of five convolution layers followed by three dense layers, eventually predicting the rotation, scale and translation of T. We are using CNN variants of conventional registration approaches [12, 13] based on *edge-correspondence* to generate training sets. We represent I_d and I_t as gray-scale images I_d^g and I_t^g respectively to perform $Edge(I_d^g, I_t^g)$[14, 15, 16] reducing I_d and I_t to edge maps I_d^e and I_t^e where $I_{d,t}^e : U \to [0,1]^c$. Thus, when c = 1 and $\forall U = 1$ we consider I_t^e and I_t^e as point-set representation p_d and p_t respectively to perform a rigid alignment $Rigid(p_d, p_t)$ [17, 18, 19] to determine the estimate transformation matrix T. The depth image I_d and corresponding T form input and expected training data for CRN. During run-time, the CFN replaces conventional approaches as a preconditioning step to the RFN as a coarse registration by applying the predicted T to align I_d with I_t image.

2.2. Refining Fusion Network (RFN)

The RFN serves as the second stage used to predict a per-pixel dense displacement field $\vec{T}(i,j)$ to perform localized non-linear deformations to provide optimal alignment of depth data with thermal data. The RFN architecture is subdivided into identical networks. Each network consists of 3 convolutional layers, a dense layer and 3 deconvolutional layers. The job of convolutional layers combined with the dense layer is to learn and embed local deformations which is processed further by deconvolutional layers to upscale to match the required resolution of predicted field $\vec{T}(i,j)$. We utilize an optical flow estimation algorithm [20] during training to determine the local deformations. We merge output from these identical networks by $Concat(t_x, t_y)$. We apply $OpticalFlow(I_d^e, I_t^e)$ to estimate the non-linear deformations $\vec{T}$ = $(t_x, t_y) \in R^{(i \times j)}$ where the t_x, t_y are valid for $U = 1$ of I_d^e tracing back to pixels in I_d representing edges. We establish $map(I_d, \vec{T})$ to establish a relationship between depth pixels $d \in (I_d^e \cap I_d)$ and per-pixel displacements (t_x, t_y). The linear relationship $\vec{T} = \hat{\beta}_0 I_d + \hat{\epsilon}_i$ generates this map. The I_d and $\vec{T}$ form the input and expected data during training. After sufficiently saturating the training domain, RFN during run-time predicts $\vec{T}$ to be applied to I_d from CFN. The $\vec{T}$ aligns pixels at a localized level refining the output of CFN to improve registration accuracy. We generate displacement error using *Hausdorff distance* [21], a well-known technique in image space to measure the degree of dissimilarity between two images or point sets. The goal is to minimize the distance as much as possible. We compute $\mathcal{H}_{CFN}(I_d^e, I_t^e)$ and $\mathcal{H}_{RFN}(I_d^e, I_t^e)$ where given two point sets $I_d^e = \{d_1, d_2, d_3 ... d_p\}$ and $I_t^e = \{t_1, t_2, t_3 ... t_q\}$ we define $\mathcal{H}(I_d^e, I_t^e) = max(h(I_d^e, I_t^e), h(I_t^e, I_d^e))$ where $h(I_d^e, I_t^e) = \max\limits_{d_i \in I_d}(\min\limits_{t_i \in I_t} \|d_i - t_j\|)$ and $h(I_t^e, I_d^e) = \max\limits_{t_i \in I_t}(\min\limits_{d_i \in I_d} \|t_i - d_j\|)$ where $\|.\|$ is the Euclidean distance between two points. A lower Hausdorff value translates to good accuracy and vice-versa. Thus, with RFN we aim for $\mathcal{H}_{RFN} \leq \mathcal{H}_{CFN}$ through:

$$\arg\min_{i,j} \vec{T} = ||\mathcal{H}_{RFN}(I_d^e, I_t^e)||^2 \tag{1}$$

Where the field $\vec{T}(i,j)$ stores the displacements that generate the minimized $\mathcal{H}(I_d^e, I_t^e)$ edge alignment error. A key understanding of this approach is that it only generates a displacement vector $\vec{t}$ for each edge pixel as a function of input depth, computed using the depth lookup for the optical flow shown in Figure 2. To account for displacements for all pixels, we populate the training set with edges contained in the input images for short-range distances $(d \leq 8.0[m])$. For long-range distances, we leverage a regression model of the map for image-space displacement vectors to train the RFN.

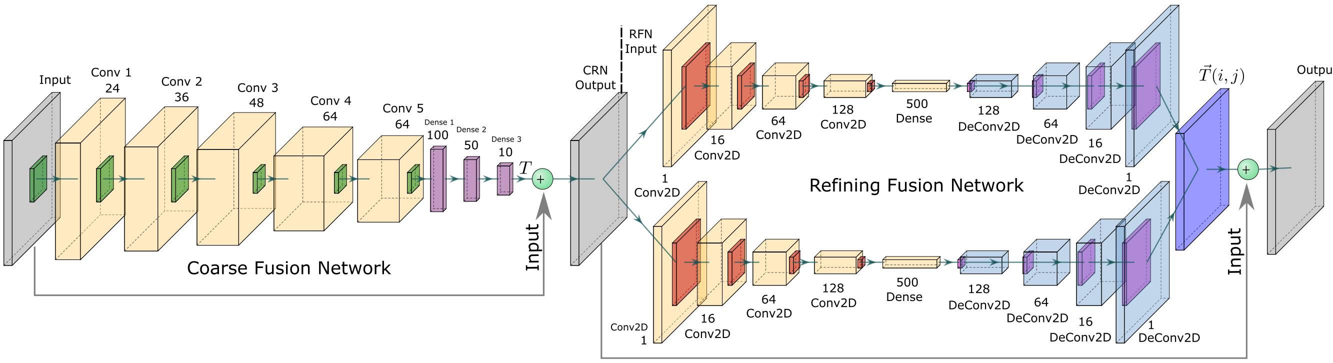

Fig. 3. The two-phase network architecture: Coarse Fusion Network (left) and Refining Fusion Network (right)

3. IMPLEMENTATION DETAILS

At its core, our approach is a deep learning technique enabled through the use of Convolutional Neural Networks (CNNs). The method performs image-space operations to identify the alignment required to fuse *edge-features* common to both modalities. To generate the required training datasets, we use a real-world modular template to generate edges used to train multimodal correspondence. This is implemented in Keras [22] with the TensorFlow [23] backend.

3.1. Template-based Correspondence

Image feature registration based on homography [2], or on edge-correspondence mapping can be significantly simplified through the use of a calibration template. While there are template-free methods [11], we aim to provide the most accurate training data possible during the calibration process. There are three classes of templates for depth-thermal fusion: passive, active, and human-guided. Passive templates are simple to construct and require no power. This method can be as simple as holes in foam board [2]. Problems with passive templates include: low heat signature, poor edge detection, require correct environmental setup. Active templates are heated by an external power source in a fixed pattern [2]. Human-guided templates could also be used [24] which requires no construction. For best *edge accuracy*, we created an active template where thermoelectric plates are combined into a uniform heat distribution through copper-plated aluminium with a matte surface, as shown in Figure 4.

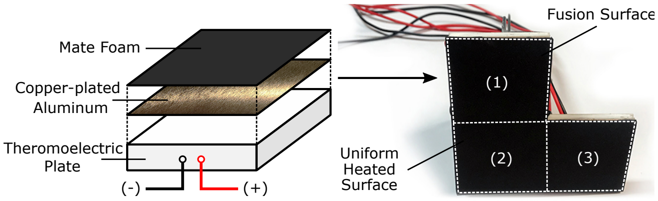

Fig. 4. Templates for edge-based correspondence. We use an *active template* to train the distance interval 0.5[m] - 8.0[m].

3.2. Training and Network Configuration

The generation of our training dataset is defined by collecting a sequence of synchronized depth and thermal images over a spatial distance of 0.5[m] - 8.0[m]. This dataset provides the edge templates used to perform the accurate alignment process to identify the correspondence between edges in each image modality. This correspondence results in a training dataset of edge-mapped consisting of recorded sequences containing synchronized template video collected at 30[Hz].

The two networks by design function independently. Thus, they undergo training separately. The dataset distribution for both remains 60% train, 20% test, and 20% validation. For hyperparameters, the CFN is configured to learn at 0.0001, with the Adam optimizer [25], the kernels initialized as glorot uniform and have valid padding. The first three layers have 5x5 kernel and followed by layers having 3x3 kernel size. The configured network uses the ReLU activation function and Mean Squared Error (MSE) to calculate the loss during training. The Refining Fusion Network is configured to learn at 0.001, Adam optimizer, a batch size of 16, and an epoch size of 100. The upsampling layers have the nearest interpolation as their configuration, while the transpose layers have valid padding, glolot uniform as kernel initialization. All the layers use LeakyReLu as their activation function with an alpha of 0.3. The network was trained using Euclidean distance as a metric of loss between the predicted dense vector field and generated ground truth dense vector field. This implementation aims to provide an efficient fusion algorithm that can be executed on the GPU for real-time systems. Our mobile system configuration is an Intel i7-10750H, a 6-core processor coupled with Nvidia GeForce RTX 2070 Max-Q, and 32GB of memory.

4. EXPERIMENTAL RESULTS

We evaluate our localized fusion approach by streaming depth and thermal data from different experimental scenes and measure the Hausdorff distance as the sum of the pixel error calculated. We have multiple experiment scenarios including: (1) running vehicle, (2) face profile, (3) hand gesture, and (4) vehicle with open-hood showing the engine compartment.

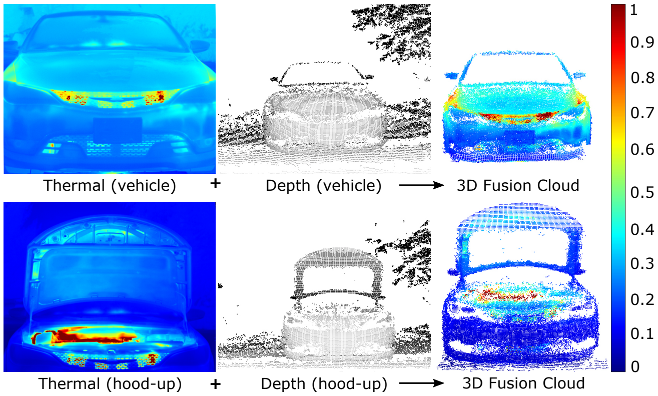

Fig. 5. Resulting fusion of the thermal (left) and depth (center) data used to generate the 3D fusion of a stationary vehicle.

Since the premise of the RFN is to provide localized transformation $\vec{T}(i,j)$ to improve accuracy, we demonstrate our method using low-resolution devices with a wide FOV disparity in challenging real-world scenarios. In our initial results, we demonstrate the generation of the 3D point cloud generated from a running vehicle, as shown in Figure 5. The FOV difference and limited resolution makes accurate fusion particularly challenging. We evaluate the accuracy using Hausdorff distance on the low resolution data and demonstrate the versatility and robustness of different scenes to quantitatively compare the alignment generated by the CFN with the RFN as shown in Figure 7. For other datasets, including body movement and the facial profile, maintaining key edges is required for precise applications such as pedestrian tracking, gesture recognition, and thermal-depth biometrics. Figure 6 illustrates the fusion result for identifying unique body features commonly used in body-tracking and facial recognition. This provides the ability to eliminate key features as shown in Figure 6 that may hamper recognition or the overall accuracy in body movement or gesture tracking.

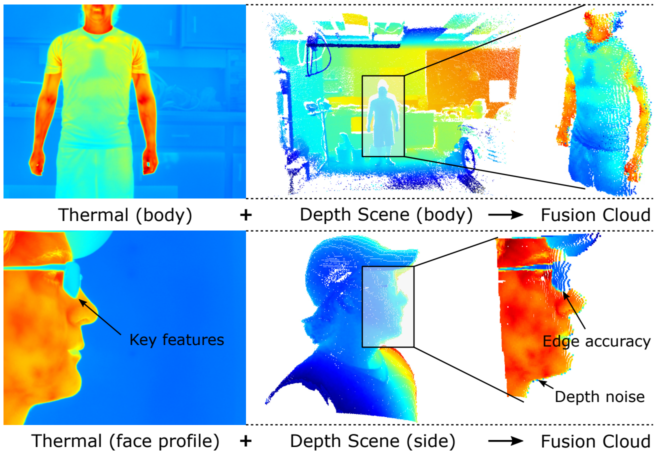

Fig. 6. Resulting fusion for body movement and profile capture. Unique multimodal mappings can provide inter-modal recognition, tracking, or biometrics based on localized detail improvements provide by the our refined alignment network.

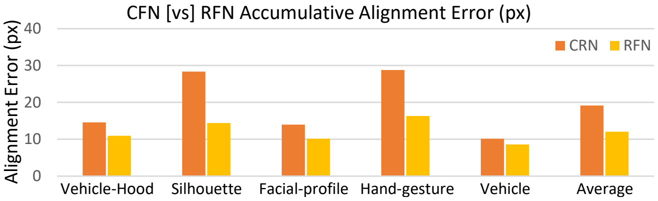

Fig. 7. Alignment error measured as the accumulative distance in pixels (px) in the Hausdorff distance metric for five selected experimental scenes. The average graph illustrates the improvements facilitated by RFN with an overall 18% improvement in accuracy over CRN ranging from 9 to 28%.

5. DISCUSSION

Performing image-space fusion provides a reliable method for identifying the best alignment between observable features that can be mapped in a trained CNN model. Our approach offers a *hybrid* method that performs the estimation using a depth lookup map and a rigid global alignment, followed by a non-linear refinement step. Based on the initial estimate of this alignment, the refining network can improve the alignment between prominent features and localized mismatches. This method works well when the alignment may be difficult, including instances where the devices have large differences in intrinsic characteristics such as the FOVs and image resolutions. In our experimental setup, the FOV for depth and thermal devices are 70° and 21.5° respectively. This illustrates that even in this extreme case with limited image resolution, our method can be used to improve overall accuracy. The accuracy of the method is dictated by a combination of the selected edge detection algorithm, training datasets, RFN formulation, and hyper-parameters. While the accuracy can be improved using this method, it still requires a template during the training dataset generation. For the input of these images, we assume that the synchronized images contain minimal distortion. We leverage data logistics reduction by utilizing CPU and GPU processing for unifying the CFN and RFN in the deep learning space. The improved accuracy level is key to enabling new multimodal-related applications, including 3D environmental reconstruction, facial authentication, and medical imaging, where accuracy is paramount.

6. CONCLUSION

Most multimodal fusion techniques primarily rely on stereographic homography or context-dependent features, and iterative point-set registration techniques to determine global correspondence at run-time. Although we have seen recent work relying on machine learning methods, with this paper, we manage to reproduce coarse fusion with run-time efficiency and a step further in image fusion by performing local alignment, improving the accuracy by 18%. This provides pixel-level 3D thermographic images accurate to edge-level that benefit a wide variety of fusion applications.

7. REFERENCES

[1] Barbara Zitova and Jan Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[2] Johannes Rangel, Samuel Soldan, and A Kroll, "3d thermal imaging: Fusion of thermography and depth cameras," in *International Conference on Quantitative InfraRed Thermography*, 2014, vol. 3.

[3] Marc Oliu Simón, Ciprian Corneanu, Kamal Nasrollahi, Olegs Nikisins, Sergio Escalera, Yunlian Sun, Haiqing Li, Zhenan Sun, Thomas B Moeslund, and Modris Greitans, "Improved rgb-dt based face recognition," *Iet Biometrics*, vol. 5, no. 4, pp. 297–303, 2016.

[4] Shoaib Azam et al., "Data fusion of lidar and thermal camera for autonomous driving," in *Applied Industrial Optics 2019*. 2019, p. T2A.5, Optical Society of America.

[5] Behnood Rasti and Pedram Ghamisi, "Remote sensing image classification using subspace sensor fusion," *Information Fusion*, vol. 64, pp. 121–130, 2020.

[6] Changmiao Wang et al., "Lung nodule classification using deep feature fusion in chest radiography," *Computerized Medical Imaging and Graphics*, vol. 57, pp. 10–18, 2017.

[7] Dibyendu Ghosh, Biswajit Samanta, and Debashish Chakravarty, "Multi sensor data fusion for 6d pose estimation and 3d underground mine mapping using autonomous mobile robot," *International Journal of Image and Data Fusion*, vol. 8, no. 2, pp. 173–187, 2017.

[8] Florian Walch et al., "Image-based localization using lstms for structured feature correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 627–637.

[9] Ehab Salahat and Murad Qasaimeh, "Recent advances in features extraction and description algorithms: A comprehensive survey," in *2017 IEEE international conference on industrial technology (ICIT)*. IEEE, 2017, pp. 1059–1063.

[10] Seyed Muhammad Hossein Mousavi, Vyacheslav Lyashenko, and Surya Prasath, "Analysis of a robust edge detection system in different color spaces using color and depth images," , vol. 43, no. 4, 2019.

[11] Ignacio Rocco Spremolla, Michel Antunes, Djamila Aouada, and Björn E Ottersten, "Rgb-d and thermal sensor fusion-application in person tracking.," in *VISIGRAPP (3: VISAPP)*, 2016, pp. 612–619.

[12] Sayan Nag, "Image registration techniques: A survey," *arXiv preprint arXiv:1712.07540*, 2017.

[13] Renbo Xia, Jibin Zhao, and Yunpeng Liu, "A robust feature-based registration method of multimodal image using phase congruency and coherent point drift," in *MIPPR 2013: Pattern Recognition and Computer Vision*. International Society for Optics and Photonics, 2013, vol. 8919, p. 891903.

[14] Saket Bhardwaj and Ajay Mittal, "A survey on various edge detector techniques," *Procedia Technology*, vol. 4, pp. 220–226, 2012.

[15] Saining Xie and Zhuowen Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[16] John Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, , no. 6, pp. 679–698, 1986.

[17] Fang Wang and Zijian Zhao, "A survey of iterative closest point algorithm," in *2017 Chinese Automation Congress (CAC)*. IEEE, 2017, pp. 4395–4399.

[18] Baraka Maiseli, Yanfeng Gu, and Huijun Gao, "Recent developments and trends in point set registration methods," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 95–106, 2017.

[19] Andriy Myronenko and Xubo Song, "Point set registration: Coherent point drift," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[20] Alexey Dosovitskiy et al., "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[21] Dejun Zhang, Fazhi He, Soonhung Han, Lu Zou, Yiqi Wu, and Yilin Chen, "An efficient approach to directly compute the exact hausdorff distance for 3d point sets," *Integrated Computer-Aided Engineering*, vol. 24, no. 3, pp. 261–277, 2017.

[22] Francois Chollet et al., "Keras (python)," 2015.

[23] Martín Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[24] Jake T Lussier and Sebastian Thrun, "Automatic calibration of rgbd and thermal cameras," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 451–458.

[25] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.