

# Deep Phase Correlation for End-to-End Heterogeneous Sensor Measurements Matching

Zexi Chen Xuecheng Xu Yue Wang\* Rong Xiong

Control Science and Engineering of Zhejiang University  
{chenzexi, xuechengxu, ywang24, rxiong}@zju.edu.cn

\* is the corresponding author

**Abstract:** The crucial step for localization is to match the current observation to the map. When the two sensor modalities are significantly different, matching becomes challenging. In this paper, we present an end-to-end deep phase correlation network (DPCN) to match heterogeneous sensor measurements. In DPCN, the primary component is a differentiable correlation-based estimator that back-propagates the pose error to learnable feature extractors, which addresses the problem that there are no direct common features for supervision. Also, it eliminates the exhaustive evaluation in some previous methods, improving efficiency. With the interpretable modeling, the network is light-weighted and promising for better generalization. We evaluate the system on both the simulation data and Aero-Ground Dataset which consists of heterogeneous sensor images and aerial images acquired by satellites or aerial robots. The results show that our method is able to match the heterogeneous sensor measurements, outperforming the comparative traditional phase correlation and other learning-based methods. [Code is available here.](#)

**Keywords:** Image Matching, Heterogeneous Sensor Modality, Deep Learning

## 1 Introduction

Localization is one of the most fundamental problem for mobile robots. With a decade of research, localization given the measurement and the map built by the same sensor is relatively mature. But for matching measurements from heterogeneous sensor modalities remains an open problem. This problem is practical considering the effort to build a map. We would like the map to be sharable by multiple robot users, even equipped with heterogeneous sensors. Note the great progress in visual inertial navigation and mapping [1], the solution space to localization can be reduced to 3 or 4, namely the translation and the heading angle. By exploiting birds-eye view measurements, the problem is re-formulated as an image matching problem with warping space built on  $\text{SIM}(2)$ . In this paper, we focus on this scenarios with birds-eye view images acquired by ground vehicles, satellites and UAV with vision and LiDAR, as shown in Figure 1.

Previous researches on homogeneous image matching can be divided into two categories: those rely on point features correspondences to localize in specific setups [2, 3, 4, 5], those apply dense correlation methods to find the best pose candidate in solution space [6, 7, 8, 9, 10]. However, all these approaches do not perform well when heterogeneous measurements are given. As for heterogeneous image matching, [11, 12] utilize hand-craft features to localize LiDAR against satellite maps. These frameworks rely heavily on the coverage and optimality of the hand-crafted features over the variations in real environment. Learning-based methods proved to have certain generalization ability are intuitively appropriate for heterogeneous image matching. [13, 14] learn the embeddings for heterogeneous observations and exhaustively search for the optimal pose in the discrete solution space, which is more interpretable than regressing the pose by an end-to-end network as in [15, 16]. However, the former suffers from low efficiency and limited pose range due to the exhaustive evaluation on the large pose space, constraining the application with known scale.

By revealing the problems encountered by prior works, one can think of an ideal matcher that can obtain the solution without exhaustive evaluation and also have good interpretability and generalization. In our work, we set to propose such a learnable matcher, of which the essence is a differentiable phase correlation. Phase correlation [6] is a similarity-based matcher that performs well for inputs

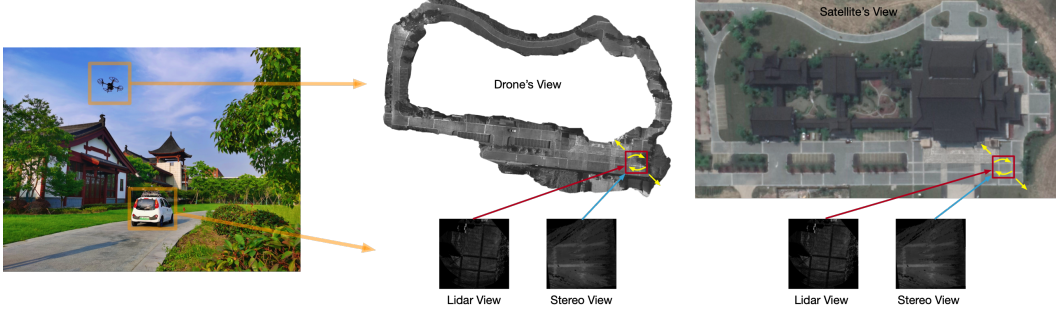


Figure 1: A typical scenario for localization based on matching measurements from heterogeneous sensors. **Left:** Aerial-Ground cooperation; **middle:** matching observations from LiDAR or stereo cameras of the ground robot to drone’s global map; **right:** matching observations from LiDAR or stereo cameras of the ground robot to satellite’s global map.

with the same modality but only tolerate small high frequency noise. We modify the phase correlation into a differentiable manner and embed it into our end-to-end framework. This architecture allows our system to find optimal feature extractors with respect to the resultant pose of image matching. Specifically, we adopt the conventional phase correlation pipeline proposed by [6] and explicitly endow the Discrete Fourier Transform (DFT) layer, log-polar transformation layer (LPT), and differentiable correlation layer (DC) with differentiability and thus make it trainable for our end-to-end matching network as shown in Figure 2. Our experiments show the robustness and efficiency of the propose method on matching heterogeneous sensor measurements.

## 2 Deep Phase Correlation Network

With known gravity direction, the relative pose between the two observed birds-eye view images ( $O_1, O_2$ ) can be simplified as similarity transform  $S$ :

$$S = \begin{pmatrix} s\mathbf{R}_\theta & \mathbf{t} \\ 0 & 1 \end{pmatrix} \in \text{SIM}(2) \quad (1)$$

where  $s \in \mathbb{R}^+$  is the scale,  $\mathbf{R}_\theta \in \text{SO}(2)$  is the rotation matrix generated by the heading angle  $\theta$ , and  $\mathbf{t} \in \mathbb{R}^2$  is the translation vector. In general, the two images are disturbed by illumination, shadow and occlusion, or even acquired by heterogeneous sensors. For example,  $O_{p1}$  is acquired by a birds-eye camera of a UAV in the morning, while  $O_{p2}$  is a local elevation map constructed by the ground robot with a LiDAR. To address this issue, a general process is to extract features from the two images, and estimate the relative pose using the features instead of the original sensor measurements. By applying the template matching to the features, we derive the optimal feature extractors  $f_1$  and  $f_2$  for the two images respectively by solving

$$\min_{f_1, f_2} \|S^* - \arg \max_S H(f_1(O_1), T\{f_2(O_2), S\})\| \quad (2)$$

where  $S^*$  is the ground truth,  $H$  is a scoring function to measure the similarity, and  $T\{\cdot, S\}$  is to transform  $\cdot$  by relative pose  $S$ . Inner product is often regarded as the scorer function. Then we have

$$\min_{f_1, f_2} \|S^* - \arg \max_S f_1(O_1) \cdot T\{f_2(O_2), S\}\| \quad (3)$$

There are two problems. First, an exhaustive evaluation is required for all elements in  $S$ , which is extremely time consuming considering the 4 dimensional space  $\text{SIM}(2)$ . Second,  $\arg \max$  is not differentiable, thus (3) is hard to be optimized. If  $f_1$  and  $f_2$  are hand-crafted processes, the optimization is even harder. In this paper, we set to differentiate (3) and eliminate the exhaustive evaluation to find efficient data-driven feature extractors  $f_1$  and  $f_2$ .

### 2.1 Decoupled correlation based pose estimator

**Cross-correlation:** We begin with the known scale and rotation  $\{\theta^*, s^*\}$ , reducing the unknown parameters to 2. Denoting  $F_1 = f_1(O_1)$  and  $F_2 = T\{f_2(O_2), \{\theta^*, s^*\}\}$ , we have

$$\min_{f_1, f_2} \|\mathbf{t}^* - \arg \max_{\mathbf{t}} \sum_x F_1(x) F_2(x + \mathbf{t})\| \quad (4)$$

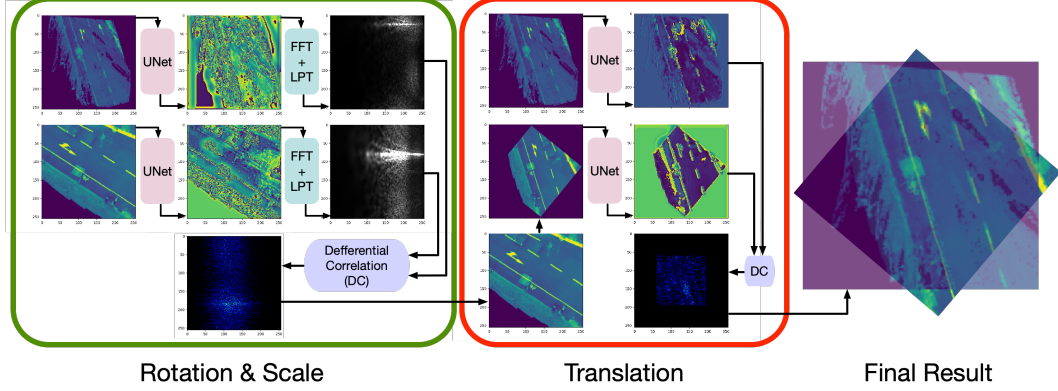


Figure 2: The architecture of proposed Deep Phase Correlation Network (DPCN), which consists of 4 U-Net based feature extractors and differentiable phase correlation for decoupled scale, rotation and translation estimation.

where  $x$  is a position in the feature. Note the term in the  $\arg \max$  is the cross-correlation function parameterized by  $t$  between  $F_1$  and  $F_2$ , which can be evaluated very efficiently using convolution.

**Translation invariance:** In general, if we achieve scale and rotation by assuming known translation, then the problem is chicken-and-egg. We introduce a representation which is invariant to translation, but variant to scale and rotation. Thus we can ignore the translation when solving for scale and rotation. Specifically, we refer to the magnitude of frequency spectrum  $\Omega$  of  $O_1$  and  $O_2$ . According to the property of Fourier transform, we have

$$|\mathfrak{F}\{T\{f_2(O_2), \{\theta, s\}\}\}| = |\mathfrak{F}\{T\{f_2(O_2), S\}\}| \quad (5)$$

where  $\mathfrak{F}$  is the fourier transform. It means that only rotation and scale have effects on the magnitude of frequency spectrum.

$$\min_{f_1, f_2} \|\{\theta^*, s^*\} - \arg \max_{\theta, s} \Omega_1 \cdot \Omega_2(\{\theta, s\})\| \quad (6)$$

**Cartesian to log-polar:** We decouple rotation and scale from relative pose using the frequency spectrum. However, exhaustive evaluation is still required for rotation and scale. By looking into the frequency domain, we have

$$\Omega_1(\psi) = \Omega_2(s^* \mathbf{R}_{\theta^*} \psi) \quad (7)$$

where  $\psi$  is a position in the frequency spectrum. By representing it in the polar coordinates, we have

$$\Omega_1(\alpha, \beta) = \Omega_2(s^* \alpha, \beta + \theta^*) \quad (8)$$

where  $\alpha = |\psi|$  and  $\beta = \angle \psi$ . To deal with the scale, we further apply the logarithm to the position

$$\Omega_1(\log \alpha, \beta) = \Omega_2(\log \alpha + \log s^*, \beta + \theta^*) \quad (9)$$

Now we finally arrive at the correlation form with respect to the scale and rotation, which eliminates all exhaustive evaluation in the whole pose estimator.

$$\min_{f_1, f_2} \|\{\log s^*, \theta^*\} - \arg \max_{\{\log s, \theta\}} \sum_{\{\log \alpha, \beta\}} \Omega_1(\log \alpha, \beta) \Omega_2(\log \alpha + \log s, \beta + \theta)\| \quad (10)$$

When there is no feature extractor, or equivalently, the original sensor measurements are fed, the process is called phase correlation, estimating the relative pose in  $\mathbb{S}\mathbb{I}\mathbb{M}(2)$  very efficiently. However, when there is variation between the pair of inputs, the feature extractor is indispensable.

## 2.2 End-to-end learnable feature extractor

**Expectation as differentiable  $\arg \max$ :** To find the feature extractor which is optimal with respect to the training data, we have to approximate the  $\arg \max$  in (4) and (10) with differentiable function. We connect the cross-correlation to a softmax function, which maps the input to a discrete probability density function  $p(\cdot)$ . Set translation part as example, the cross-correlation in (4) is

$$p(\mathbf{t}) = \frac{\exp(-\sum_x F_1(x)F_2(x + \mathbf{t}))}{\sum_{\mathbf{t}} \exp(-\sum_x F_1(x)F_2(x + \mathbf{t}))} \quad (11)$$

By keeping the features positive, we do not need to care about the negative correlation. Still referring to the translation part, given the probability function, we derive the expectation as the optimal translation estimator, and substitute it into (4), yielding the differentiable form

$$\min_{f_1, f_2} \|\mathbf{t}^* - \sum_{\mathbf{t}} \mathbf{t} p(\mathbf{t})\| \quad (12)$$

For rotation and scale (10), the same expectation estimator can be applied to approximate the non-differentiable  $\arg \max$  based estimator. When learning, we assign a temperature coefficient to the softmax to tune the range of feature input, which accelerates the convergence, but does not make difference in theoretic derivation. The whole pose estimator can be regarded as a differentiable phase correlation (DPC) with back-propagated gradients to enforce the learning of feature extractor.

**Deep feature extractor:** The conventional phase correlation [6] utilizes high pass filters to suppress high frequency random noise of two inputs, which can be seen as a feature extractor. For more distinct variation between the pair of inputs, one high pass filter is far from sufficient. Considering that there is no common feature to directly supervise the feature extractor, we utilize the end-to-end learning to address the problem.

We adopt U-Net for feature extraction, aiming at learning the common features of the two images implicitly. We construct 4 separate U-Nets with the input and output size of  $256 \times 256$  respectively for the template image and the source image in the rotation phase and the translation phase, shown in Figure 2. Each U-Net is constructed with 4 down-sampling encoder layers and 4 up-sampling decoder layers to extract features. As the training progresses, the parameters of four U-Nets are tuned. Note that this network is light-weighted so that it could be efficient enough for real-time execution. Combining the feature extractor and the DPC, we name the whole network as DPCN.

### 2.3 Data preparation for learning

One question is that why we supervise DPCN on the estimated pose, but not the correlation matrix with a one-peak matrix centering at the correct position, thus the expectation based estimator can be eliminated. We argue that enforcing the correlation matrix to be one-peak is over-supervision. In theory of phase correlation, the position corresponding to the maximum correlation, does not necessarily mean zero correlation for the others, which can be explained by the resonance in physics. However, when the correlation map is passed through softmax estimator, the temperature coefficient suppresses the non-maximal part in the correlation matrix to be close to 0, resulting in a normalized probability density function. In addition to the pose supervision as shown in (12), we also supervise the probability density function of translation, and scale/rotation, to be one-peak. Still refer to the translation as example, we do it by applying the KL-divergence:

$$\min_{f_1, f_2} \delta_{\mathbf{t}^*}(t) p(\mathbf{t}) \quad (13)$$

where  $\delta_{\mathbf{t}^*}(t)$  is a normalized one-peak function centering at  $\mathbf{t}^*$ . In practice, we slightly expand the one-peak function with Gaussian smooth. Theoretically, the resultant distribution for some cases can be multi-modal, e.g. the repetitive local environment. However, given massive training data with similar input and different output, the optimal solution to (13) is a multi-modal probability function, i.e. multi-modal  $p(\mathbf{t})$ . The two loss (13) and (12) are mixed by weight. For translation part, multi-modal result occurs more often, so we increase the loss of (13) in the total loss for translation phase.

## 3 Experimental Results

In this section, we explicitly evaluate the performance of our approach. By utilizing phase correlation with DFT and log-polar transform over the learned representation of two images, we are able to fully estimate the rotation and scale transformation of the two heterogeneous images, and eventually able to estimate the 4-DoF relative pose  $S$ .

**DATASET & Metrics:** Our approach is evaluated both on randomly generated simulation dataset and on real-world dataset "Aero-Ground Dataset"[17] which contains several different image pairs shown as follows:

- **12d:** "LiDAR Local Map" to "Drone's Birds-eye Camera View";
- **12sat:** "LiDAR Local Map" to "Satellite Map";

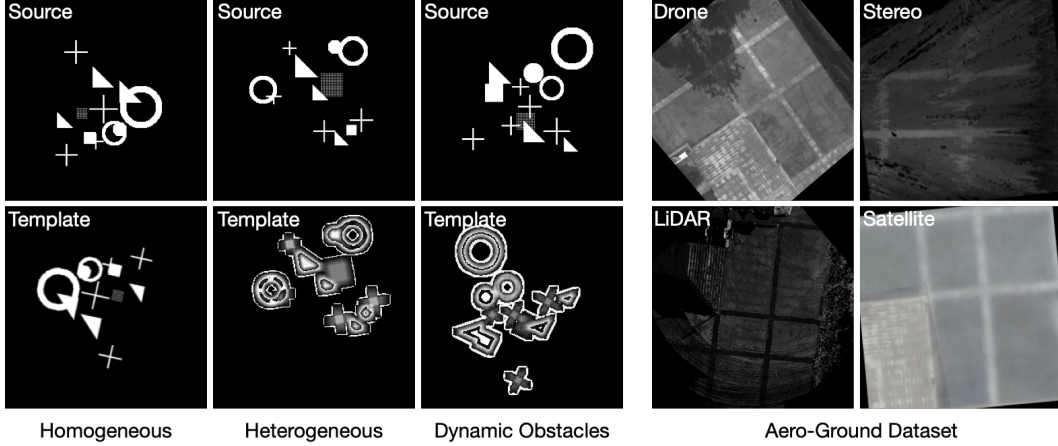


Figure 3: **Simulation dataset(left)** containing “Homogeneous”, “Heterogeneous” and “Dynamic Obstacles” sets. **Aero-Ground Dataset(right)** containing “drone’s view”, “LiDAR intensity”, “stereo” and “satellite”

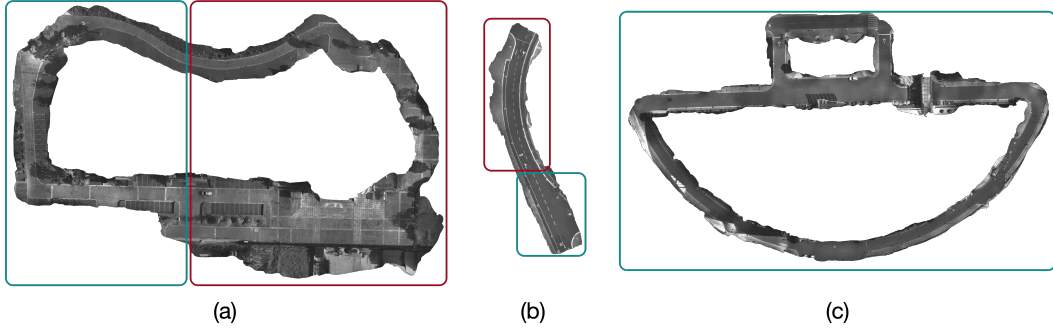


Figure 4: Demonstration of birds-eye views of the different scenes in Aero-Ground dataset. The experiments are carried out on location (a) and (b) separately in which the model is trained on images pairs generated inside **red** areas and validated on images pairs generated inside **blue** area. The generalization experiment is carried out with estimating poses of images inside location (c) with models trained on (a) and (b).

- **s2d**: “Stereo Local Map” to “Drone’s Birds-eye Camera View”;
- **s2sat**: “Stereo Local Map” to “Satellite Map”.

On the simulation dataset, we evaluate our work on homogeneous, heterogeneous, and heterogeneous with dynamic obstacles images pairs whereas on the Aero-Ground Dataset, we evaluate our work on the application of cooperative SLAM system between ground mobile robots, the MAV and the Satellite. The demonstration of both datasets is shown in Figure 3 and 4. Finally, we applied our method to Monte Carlo Localization to prove the real-time capability and robustness of our method. In all datasets, we constrained translations of both  $x$  and  $y$ , rotation changes and scale changes in the range of  $[-50, 50]$  pixels,  $[0, \pi]$  and  $[0.8, 1.2]$  respectively with images shapes of  $256 \times 256$ .

For evaluating the accuracy and error rate of estimation, we consider “Accuracy in Units” and Mean Square Error(MSE) of the estimated result and the ground truth as the mean indicator:

$$Acc_{v_t} = \frac{1}{N} \cdot \sum_{i=1}^N \mathbf{Count}\{|S_i - G_i| \leq t\} \times 100\% \quad (14)$$

$$E = \frac{1}{N} \cdot \sum_{i=1}^N (S_i - G_i)^2 \quad (15)$$

where  $v$  being the output type( $x$ ,  $y$ , rotation, and scale),  $t$  being the threshold of accuracy (pixel for translation, degree for rotation and multiplier for scale),  $N$  being the amount of image pairs,  $S_i$  being the estimated result of the  $i$ th image pair and  $G_i$  being the corresponding ground truth. “Accuracy in Units” is calculated as the percentage of estimation with an error lower than the threshold.

**COMPARATIVE METHODS:** Benchmarks in the experiments include conventional PHASE CORRELATION[6], deep learning based QATM[5], DAM[18], RELATIVE POSE REGRESSION(RPR),

and DENSE SEARCH(DS). **PHASE CORRELATION** is the baseline for registering two homogeneous images and the pipeline of which is also partly adopted in our approach. We select it as a benchmark for evaluating the performance of our approach in matching homogeneous images. **QATM** is a representative work in image matching applying deep learning to learn features for matching and NMS for matching selection. It could handles translation displacement with high accuracy and therefore we select it as the benchmark for evaluating translation estimations in heterogeneous images. Unfortunately, the author of QATM only provided a pretrained model without a training script so that we could only evaluate its performance with the provided model. **DAM** trains and learns affine transformations including translation, rotation and scale changes, and therefore we trained DAM on the same dataset and compare our method with it on the four aspects. **RELATIVE POSE REGRESSION(RPR)** is adopted by multiple methods[15, 16] in pose regression by outputting desired classes of estimation through several fully-connected layers without an analytical solutions. Finally, **DENSE SEARCH(DS)** is the methodology adopted by [10] which violently rotates images to some discrete angles to estimate the relative shifts. We trained DS in a smaller rotation range [0,15] due to its discrete property and time consumption. It is selected as a benchmark to demonstrate the advantage of estimation continuity and swiftness of our approach.

### 3.1 Simulation Dataset

In order to verify the feasibility of our approach under several different situations and the generalization capability of the fully differentiable methodology, we conduct various of experiments on the simulation dataset. The experiments on a pair of homogeneous images is conducted to verify the equivalence to the conventional phase correlation when dealing with images of same styles. Moreover, with experiments on a pair of heterogeneous images conducted, we show the unique capability of our approach with estimating the 4-DoF pose  $S$  of two images with drastic style changes. Finally, we introduce dynamic obstacles to the heterogeneous image pairs to demonstrate the robustness of our approach for being insensitive to dynamic obstacles.

Table 1: Results of simulation dataset. We choose the threshold error of  $5\ pixels$  for translation,  $1^\circ$  for rotation and  $0.2\times$  for scale. More thresholds are elaborated in Appendix C

Benchmarks	Exp.	$E_x$	$Acc_{x_s}\%$	$E_y$	$Acc_{y_s}\%$	$E_r$	$Acc_{r_1}\%$	$E_s$	$Acc_{s_{0.2}}\%$	Runtime(ms)
<u>IMAGE REGISTRATION</u>	1	<b>0.6635</b>	<b>99.2</b>	<b>0.9231</b>	<b>99.5</b>	<b>0.0663</b>	<b>99.7</b>	<b>0.0710</b>	<b>98.9</b>	141.4
	2	1774.1592	52.3	3233.8133	33.2	145.8561	72.3	0.1992	<b>97.6</b>	138.8
	3	2319.5537	49.2	2945.3017	42.6	121.5026	67.9	<b>0.1218</b>	<b>96.7</b>	137.0
<u>QATM</u>	1	15.4820	95.4	8.9192	96.3	\	\	\	\	108.3
	2	2999.3710	31.4	4286.4810	26.4	\	\	\	\	108.9
	3	3651.4691	25.6	4901.7201	21.5	\	\	\	\	109.1
<u>DAM</u>	1	53.7597	90.6	28.6825	95.9	19.2243	81.7	0.1452	90.5	111.7
	2	<b>2.5816</b>	<b>98.4</b>	<b>4.6234</b>	<b>97.9</b>	<b>28.9341</b>	<b>80.8</b>	0.1432	90.9	114.2
	3	46.1165	71.3	89.6835	68.2	<b>36.5608</b>	<b>77.8</b>	0.3625	87.3	110.4
<u>RPR</u>	1	8.6754	96.9	2.2201	97.3	16.2723	90.2	0.0805	95.2	<b>64.1</b>
	2	22.3634	39.9	32.8334	41.2	97.8517	78.3	<b>0.1367</b>	96.7	<b>64.7</b>
	3	14.6842	51.1	11.1322	56.8	101.3329	76.8	0.1846	96.1	<b>63.5</b>
<u>DS</u>	1	7.4092	85.1	11.1108	79.3	26.2656	33.4	\	\	304.5
	2	5.4970	86.6	7.3333	86.5	31.3891	23.9	\	\	301.3
	3	<b>6.2850</b>	<b>83.9</b>	<b>7.8176</b>	<b>79.6</b>	25.6643	27.1	\	\	301.4
<u>DPCN</u>	1	<b>0.1031</b>	<b>100</b>	<b>0.2162</b>	<b>100</b>	<b>0.0528</b>	<b>100</b>	<b>0.0522</b>	<b>100</b>	<b>71.9</b>
	2	<b>0.0073</b>	<b>100</b>	<b>0.0172</b>	<b>100</b>	<b>0.0397</b>	<b>100</b>	<b>0.0642</b>	<b>100</b>	<b>72.1</b>
	3	<b>0.0761</b>	<b>100</b>	<b>0.4671</b>	<b>100</b>	<b>0.0913</b>	<b>100</b>	<b>0.0013</b>	<b>100</b>	<b>72.1</b>

Note: *Exp.1*, *Exp.2* and *Exp.3* is conducted on ‘‘Homogeneous’’, ‘‘Heterogeneous’’, and ‘‘Dynamic Obstacles’’ sets respectively. **Red** is the best performance and **blue** is the secondary.

Experimental results on the simulation dataset are shown in Table 1. The results show that for the homogeneous pairs of images, our method maintains an equivalent performance with the conventional phase correlation pipeline in accuracy with faster speed and outperformed the rest of the baselines. In the heterogeneous dataset and dynamic obstacle dataset, our approach outperformed any existing baselines in accuracy and is only a bit slower than the Relative Pose Regression.

The key to our approach is applying a fully optimal differentiable DFT with log-polar transformation to back-propagate the supervised error to train a multi-channel feature-based representation which is optimized for phase correlation to estimate the relative pose. It outperformed the conventional phase correlation method applying high pass filter with the learned representation when dealing with images that is not simply a cutout to one another. The learned representation, the result of the DFT, log-polar transformation, and the correlation map shown in Figure 6 indicate that the network is

interpretable: by supervising end-to-end, the network is finally able to predict the transformation of two heterogeneous images with their U-Net outputs of convergent features. Note that the simulation dataset is randomly generated to reduce overfitting.

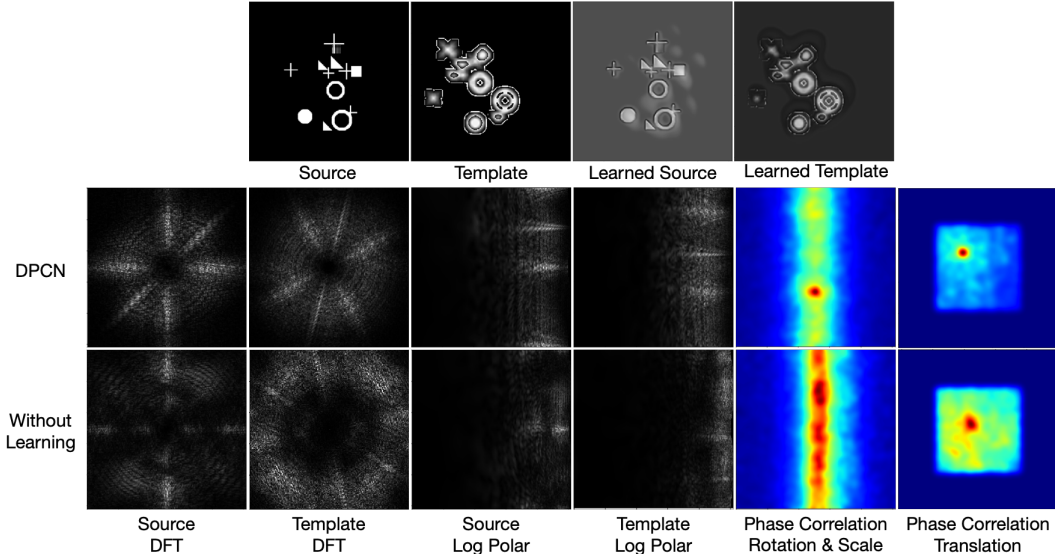


Figure 5: Workflow visualization showing that with codecs applied and trained, the network is able to reduce the impact of noise and style changes while conventional phase correlation without feature learning fails to recognize the relations of two images.

Table 2 and elaborations in Appendix C verify the generalizing ability of our approach. For generalization experiments, the model is trained on the “Heterogeneous” set and is evaluated on two other sets in simulation datasets. The results show that with models not specifically trained, it can still maintain a high rate of accuracy in all 4-DoF.

Table 2: Results of generalization experiments with simulation dataset.

Exp.	$E_x$	$Acc_{x_5}(\%)$	$E_y$	$Acc_{y_5}(\%)$	$E_{rot}$	$Acc_{rot_1}(\%)$	$E_{scale}$	$Acc_{scale_{0.2}}(\%)$
Homogeneous	0.4121	100	0.705	100	0.0432	100	0.015	100
Dynamic Obstacle	0.0276	100	0.105	100	0.039	100	0.003	100

### 3.2 Aero-Ground Dataset

To evaluate the applicability in the real world of our approach, we conduct experiments on Aero-Ground Dataset to match images across sensors. Few baseline support 4-DoF(or more) pose estimations across sensors and therefore we relax the condition by providing rotation and scale ground truth as initials to several benchmark for comparison, including the DAM with scaling initialized and QATM with both rotation and scale initialized. We train and validate our models in two different scenes (Figure 4(a) and (b)) separately and generalize them in the third scene(Figure 4(c)).

Table 3 shows validation results on scene (a) and (b) with the threshold error of 10 *pixels* for translation,  $1^\circ$  for rotation and  $0.2\times$  for scale. Acknowledging that each ground-image is generated with a scale of 0.1 *meters per pixel(MPP)* in the Aero-Ground Dataset, the threshold error of 10 *pixels* for translation could be transformed to 1 *meter* in the real world. The results prove that when estimating 4-DoF poses across real-world sensors, our approach is the first to finish this job with an accuracy of at least 89.22% when considering error lower than 1 *meter*. Even when we relax the conditions and provide initials to the rest of the benchmarks, our approach still outperforms the rest of them in 2-DoF(QATM) and 3-DoF(DAM). Additional demonstrations of the experiment comparing with conventional phase correlation is shown in appendix B.

To evaluate the generalizing capability of our approach, we conduct experiments on scene (c) with DPCN models trained on scene(a) and (b) and DAM model trained on scene(c). The results shown

Table 3: Results of the scene (a) and (b). More thresholds are elaborated in Appendix C

Scene	Benchmarks	Exp.	$E_x$	$Acc_{x_{10}}$ (%)	$E_y$	$Acc_{y_{10}}$ (%)	$E_{rot}$	$Acc_{rot_1}$ (%)	$E_{scale}$	$Acc_{scale_{0.2}}$ (%)	Runtime(ms)
(a)	QATM	l2sat	6533.8164	21.5	4082.8421	57.1	\	\	\	\	109.2
		l2d	8168.9537	34.9	4160.1207	21.9	\	\	\	\	108.8
		s2sat	8268.8350	24.3	7135.4707	19.5	\	\	\	\	108.9
		s2d	5288.4170	33.6	5309.1103	31.7	\	\	\	\	109.7
	DAM	l2sat	507.1945	55.4	208.8668	70.8	44.2139	37.8	\	\	110.6
		l2d	690.1782	39.4	301.1191	66.8	96.5603	22.5	\	\	117.3
		s2sat	740.6881	35.2	732.4164	33.6	105.1678	24.1	\	\	114.4
		s2d	536.5027	51.5	616.4043	43.9	68.1288	33.9	\	\	114.2
	DPCN (Ours)	l2sat	40.5561	96.9	4.8175	98.0	0.1172	99.2	0.00345	95.5	74.39
		l2d	15.53	98.2	6.4531	94.0	0.0412	99.2	0.0122	94.2	74.63
		s2sat	65.373	90.9	15.5920	97.8	0.1078	97.4	0.0055	93.7	75.38
		s2d	327.31	91.3	14.493	92.6	0.2274	99.3	0.0070	93.5	73.44
<b>Benchmarks</b>											
(b)	QATM	l2d	3603.8648	37.5	4018.3337	35.9	\	\	\	\	109.2
		s2d	2808.5308	36.3	2878.3589	31.0	\	\	\	\	108.5
		l2d	972.8225	30.1	588.4123	42.2	61.3341	35.1	\	\	113.9
		s2d	633.2790	40.9	484.3626	49.6	85.3438	27.4	\	\	116.5
	DPCN (Ours)	l2d	8.0043	96.2	102.359	89.2	0.0059	99.7	0.0005	99.7	75.34
		s2d	88.7428	91.6	61.0860	90.6	0.7634	99.4	0.0035	95.0	75.96

in Figure 6 and Table 4(Appendix C) prove that with the types of source sensors given and fixed, our approach is capable of estimating poses regardless of scene changes and illumination changes with similar accuracy and still outperformed DAM which is specifically trained on (c). Therefore, the robustness of our approach in the real-world application is well documented.

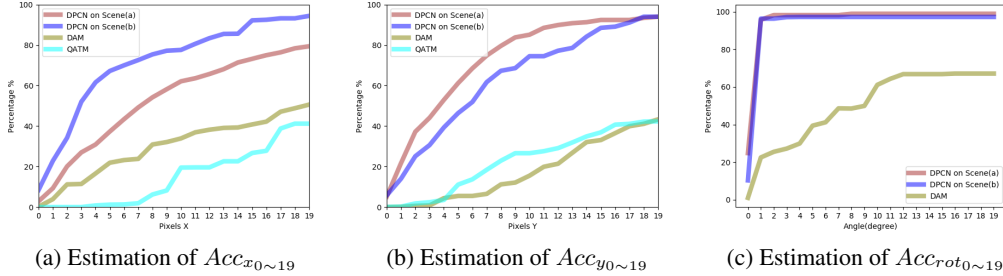


Figure 6: Transformation estimation in generalization on scene (c).

### 3.3 Application in satellite map based localization

In this section, we demonstrate the application in localization by introducing our approach to Monte Carlo Localization. It proves that with corresponding maps as the output, our model is capable of real-time air-ground localization, e.g. scene(a). A demonstration is shown in Figure 7, where the green dashed line is the odometry estimated by “VINS” through a stereo camera, the red dashed line is generated by MCL by matching 4-DoF poses of LiDAR intensity map and satellite ,ap, and the yellow line is the ground truth. As the cumulative error of the odometry gradually increases, the corrective effect of our method is sufficiently demonstrated.



Figure 7: Application in Monte Carlo Localization.

## 4 Conclusion

We present an approach for precise multi-sensor pose matching which greatly fascinates multi-agent collaborative exploration. We achieve this by training pairs of individual U-Nets with end-to-end poses to learn representations of heterogeneous images to be recognized by differentiable phase correlation(DFT+LPT+Correlation). We show that by training the network end-to-end with a fully differentiable pipeline, the network is easy and fast to be trained, precise in matching and capable of running in real-time. We also show that with every estimation analytical, the network is completely interpretable and has the capability of generalization.



## Acknowledgments

This project is supported and funded by the National Key R&D Program of China(2018AAA0102703), the National Nature Science Foundation of China(61903332), and the Science and Technology on Space Intelligent Control Laboratory(Grant No.HTKJ2019KL502002). We give sincere appreciations to all the reviewers.

## References

- [1] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings - IEEE International Conference on Robotics and Automation*, 2007. ISBN 1424406021. doi:10.1109/ROBOT.2007.364024.
- [2] M. Noda, T. Takahashi, D. Deguchi, I. Ide, H. Murase, Y. Kojima, and T. Naito. Vehicle ego-localization by matching in-vehicle camera images to an aerial image. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6469 LNCS(PART 2):163–173, 2011. ISSN 03029743. doi:10.1007/978-3-642-22819-3\_17.
- [3] T. Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015. ISBN 9781467369640. doi:10.1109/CVPR.2015.7299135.
- [4] D. Barnes, W. Maddern, G. Pascoe, and I. Posner. Driven to Distraction: Self-Supervised Distractor Learning for Robust Monocular Visual Odometry in Urban Environments. nov 2017. URL <http://arxiv.org/abs/1711.06623>.
- [5] J. Cheng, Y. Wu, W. Abdalmageed, and P. Natarajan. QATM: Quality-aware template matching for deep learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:11545–11554, 2019. ISSN 10636919. doi:10.1109/CVPR.2019.01182.
- [6] B. Srinivasa Reddy and B. N. Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 1996. ISSN 10577149. doi:10.1109/83.506761.
- [7] R. Kaslin, P. Fankhauser, E. Stumm, Z. Taylor, E. Mueggler, J. Delmerico, D. Scaramuzza, R. Siegwart, and M. Hutter. Collaborative localization of aerial and ground robots through elevation maps. *SSRR 2016 - International Symposium on Safety, Security and Rescue Robotics*, pages 284–290, 2016. doi:10.1109/SSRR.2016.7784317.
- [8] I. Talmi, R. Mechrez, and L. Zelnik-Manor. Template matching with deformable diversity similarity. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. ISBN 9781538604571. doi:10.1109/CVPR.2017.144.
- [9] J. Kim, J. Kim, S. Choi, M. A. Hasan, and C. Kim. Robust template matching using scale-adaptive deep convolutional features. *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, 2018-Febru (December):708–711, 2018. doi:10.1109/APSIPA.2017.8282124.
- [10] D. Barnes, R. Weston, and I. Posner. Masking by Moving: Learning Distraction-Free Radar Odometry from Pose Information. (CoRL), 2019. URL <http://arxiv.org/abs/1909.03752>.
- [11] R. Kummerle, B. Steder, C. Dornhege, A. Kleiner, G. Grisetti, and W. Burgard. Large scale graph-based SLAM using aerial images as prior information. *Robotics: Science and Systems*, 5:297–304, 2010. ISSN 2330765X. doi:10.15607/rss.2009.v.038.
- [12] P. Ruchti, B. Steder, M. Ruhnke, and W. Burgard. Localization on OpenStreetMap data using a 3D laser scanner. In *Proceedings - IEEE International Conference on Robotics and Automation*, 2015. doi:10.1109/ICRA.2015.7139932.

- [13] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song. L3-net: Towards learning based lidar localization for autonomous driving. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. ISBN 9781728132938. doi:10.1109/CVPR.2019.00655.
- [14] T. Y. Tang, D. De Martini, D. Barnes, and P. Newman. RSL-Net: Localising in Satellite Images from a Radar on the Ground. *IEEE Robotics and Automation Letters*, 5(2):1087–1094, 2020. ISSN 23773766. doi:10.1109/LRA.2020.2965907.
- [15] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. ISBN 9781467383912. doi:10.1109/ICCV.2015.336.
- [16] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. ISBN 9781538604571. doi:10.1109/CVPR.2017.694.
- [17] ZJU-Lab. Aero-ground dataset, 2020. URL <https://github.com/ZJU-Robotics-Lab/OpenDataSet>.
- [18] J. H. Park, W. J. Nam, and S. W. Lee. A two-stream symmetric network with bidirectional ensemble for aerial image matching. *Remote Sensing*, 2020. ISSN 20724292. doi:10.3390/rs12030465.
- [19] Y. Pan, X. Xu, Y. Wang, X. Ding, and R. Xiong. GPU accelerated real-time traversability mapping. In *IEEE International Conference on Robotics and Biomimetics, ROBIO 2019*, 2019. ISBN 9781728163215. doi:10.1109/ROBIO49542.2019.8961816.

## A Network Structure and Experimental Setup

### A.1 Network Structure

We train the DPCN network with the input of pairs of heterogeneous images and the ground truth of their relative pose. There are two phases here in the training. **Phase 1:** Rotation and scale changes are trained and estimated in this phase. The pair of images go through a pair of U-Nets and their outputs go through the DFT layer for translation reduction. The Log Polar Transform remaps the DFT’s output to the log polar coordinate so that rotation and scale variances are then shown in columns and rows. Afterward, phase correlation is able to estimate the rotation and scale with a relative power spectrum as the output. Finally, we supervised the U-Net using Cross-Entropy Loss of both rotation, scale, and their ground truth. **Phase 2:** Translation of the image is trained and estimated in this phase. The source image of the image pair is rotated and scaled by the result in phase 1 and the new image pair(original template image and the new source image) go through a new pair of codec. The translation is then estimated by phase correlation and therefore the codec is supervised using Cross-Entropy Loss and L1 Loss of estimated transformation and their ground truth.

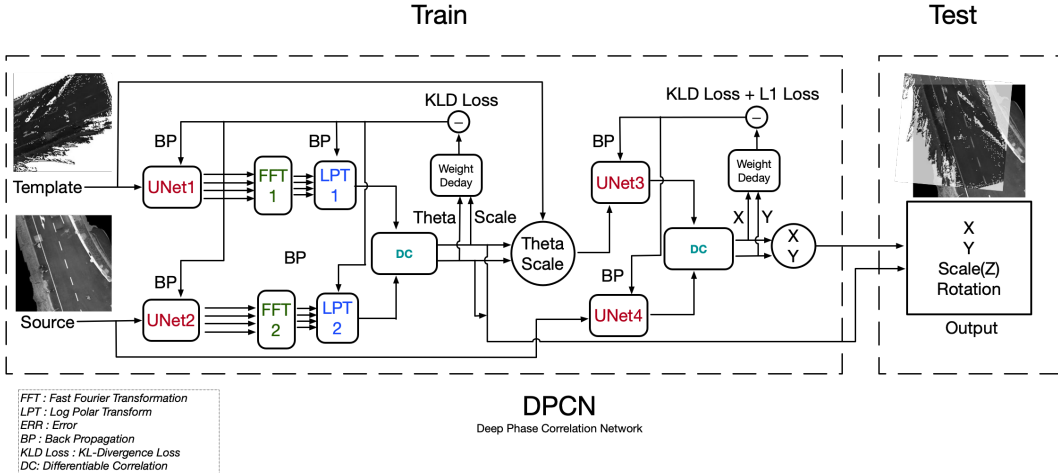


Figure 8: The network structure of DPCN

### A.2 Experimental Setup

**Hardware:** All models for the full method and their comparison models were trained on a single server with CPU of Intel i9-9900X @ 3.5GHz x 20 and GPU of RTX 2080ti x 4. The validation experiments are all conducted on a single desktop computer(AMD Ryzen 3700X @ 3.7GHz x8) with the GPU of an RTX 2060super. The Aero-Ground Dataset is record by a ground robot and a DJI M100 drone. The ground robot is equipped with three 32-wire Velodyne LiDAR, four pairs of stereo cameras by Flir and one RTK dGPS provided by QX. The DJI drone records videos by one pair of Intel Realsense D435i stereo camera with one facing forward and one downward.

**Software:** The birds-eye global map from the drone used in the training and validation is constructed by full-licensed software METASHAPE with video clips from the Aero-Ground dataset. The local map of both LiDAR intensity style and stereo style from the ground robot is constructed by ELEVATION MAP[19]. All of the maps constructed above and the satellite map obtained from GOOGLE MAPS have the resolution of 0.1 MPP.

## B Additional Demonstration

Demonstrations of comparing DPCN and conventional phase correlation on real-world dataset Aero-Ground Dataset is shown in Figure 9 and Figure 10. They show that when matching heterogeneous images from different sensors, the trainable feature extractors in DPCN play important roles in outperforming the conventional phase correlation. However, the core of the entire DPCN is the differentiable DFT, LPT and phase correlation that could back propagate losses and eventually train the feature extractors. They also prove that by learning features through the supervision of the end-to-end poses, the approach is capable of reducing hollows, rejecting noises and assimilate different styles to estimate the 4-DoF relative pose.

### B.1 DPCN Results

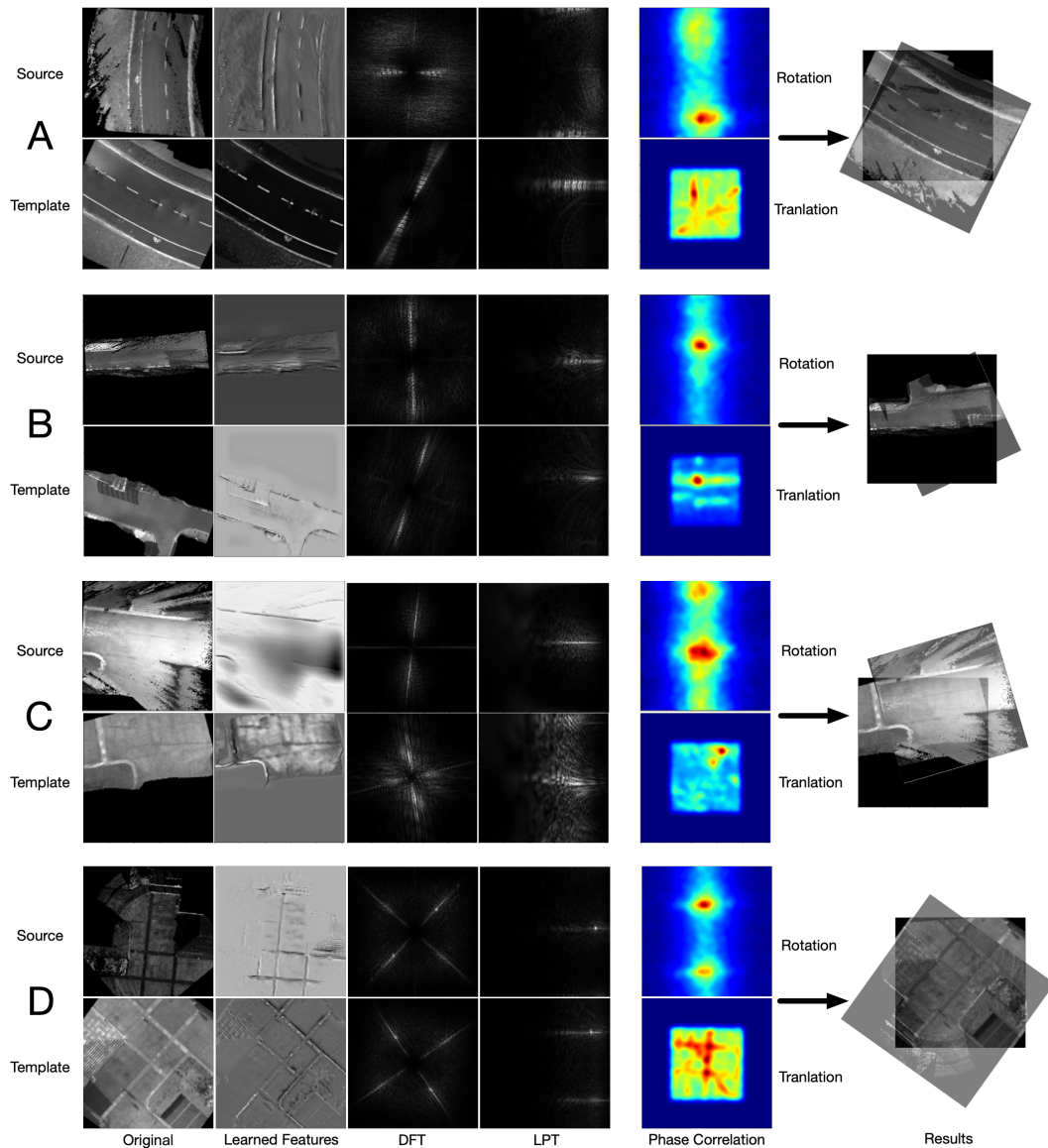


Figure 9: Additional four demonstrations matching heterogeneous images pairs from Aero-Ground Dataset. The respective comparisons are shown in subsection B.2.

## B.2 Conventional Phase Correlation Results

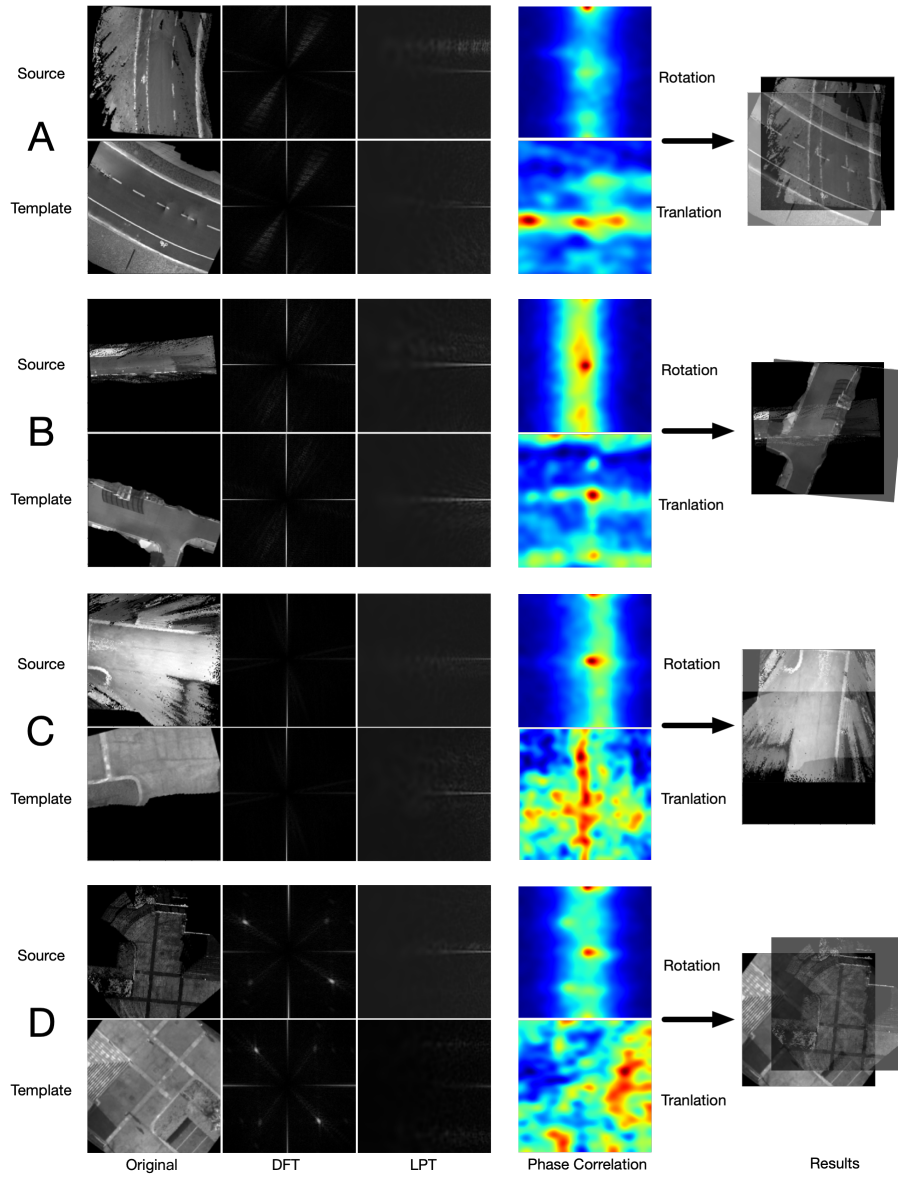
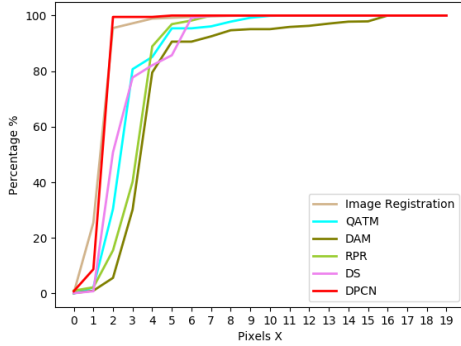


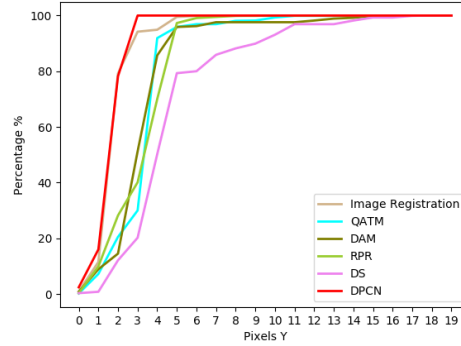
Figure 10: Comparison using conventional phase correlation to match heterogeneous images pairs from subsection B.1.

## C Elaboration on Estimation

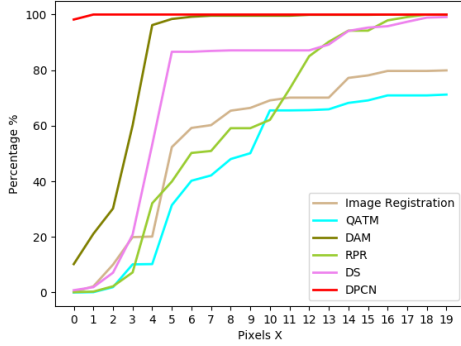
The threshold of estimation in experiments in 3 is elaborated in this section by the means of graphs. Figure 11 shows the  $Acc_{0\ to\ 19}$  of translation estimation in simulation dataset, figure 12 shows the  $Acc_{0\ to\ 19}$  of translation estimation in Aero-Ground dataset, Figure 13 shows the  $Acc_{0\ to\ 19}$  of translation estimation in simulation dataset, and Table 4 show the exact error of generalization experiments on Aero-Ground Dataset.



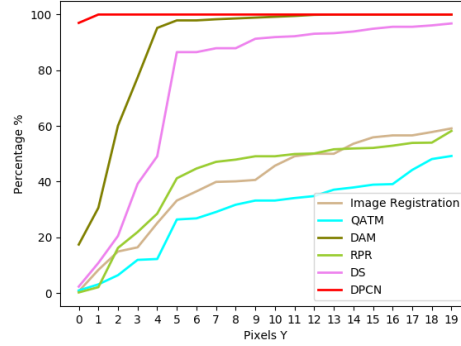
(a) Estimation of  $x$  in Homogeneous dataset



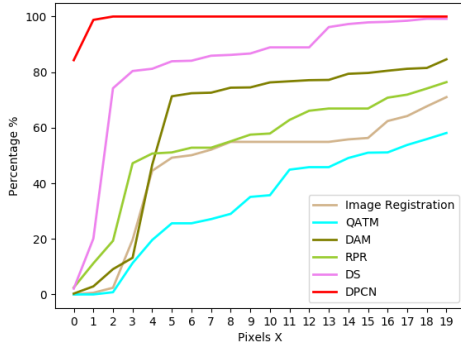
(b) Estimation of  $y$  in Homogeneous dataset



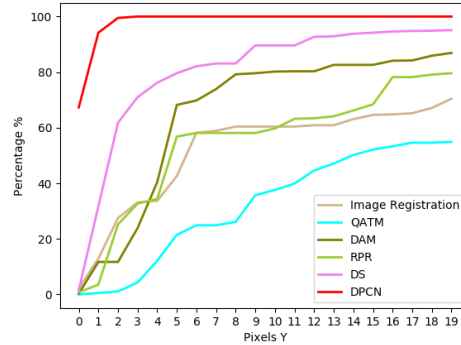
(c) Estimation of  $x$  in Heterogeneous dataset



(d) Estimation of  $y$  in Heterogeneous dataset

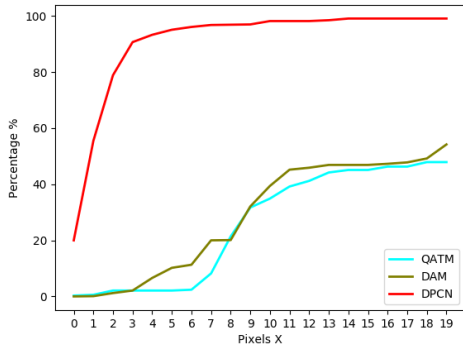


(e) Estimation of  $x$  in Dynamic Obstacle dataset

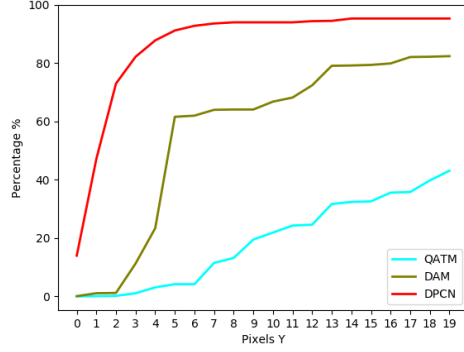


(f) Estimation of  $y$  in Dynamic Obstacle dataset

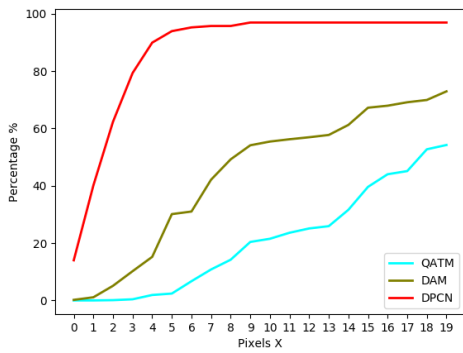
Figure 11:  $Acc_{0\ to\ 19}$  of translation estimation in simulation dataset.



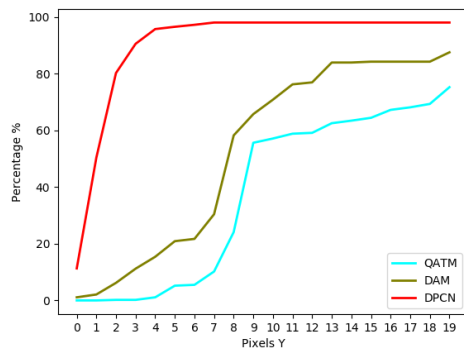
(a) Estimation of  $x$  in “LiDAR to Drone” scene(a)



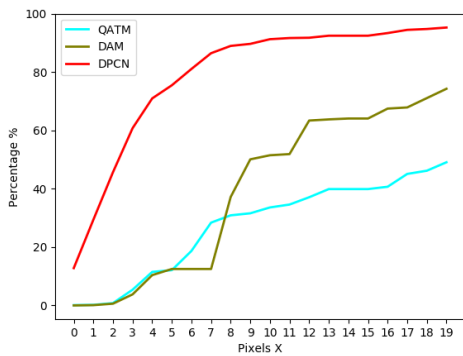
(b) Estimation of  $y$  in “LiDAR to Drone” scene(a)



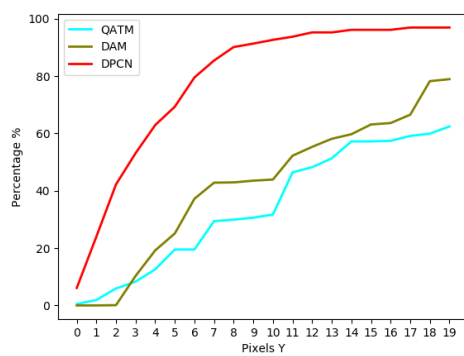
(c) Estimation of  $x$  in “LiDAR to Satellite” scene(a)



(d) Estimation of  $y$  in “LiDAR to Satellite” scene(a)

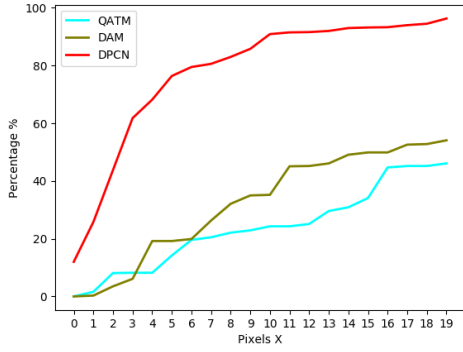


(e) Estimation of  $x$  in “Stereo to Drone” scene(a)

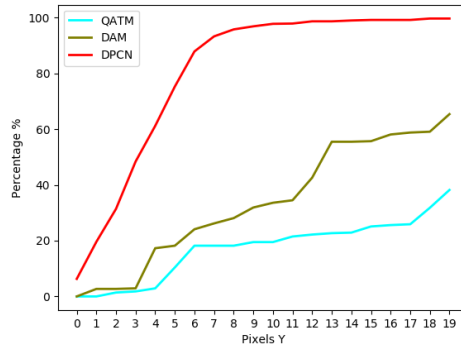


(f) Estimation of  $y$  in “Stereo to Drone” scene(a)

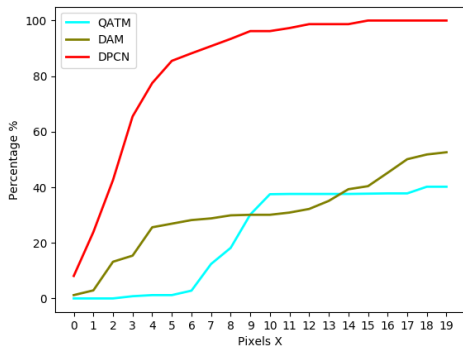
Figure 12:  $Acc_0$  to  $19$  of translation estimation in Aero-Ground dataset.



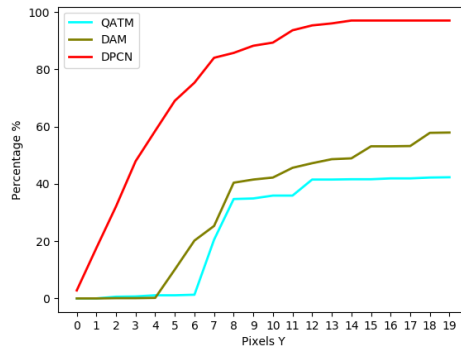
(g) Estimation of  $x$  in "Stereo to Satellite" scene(a)



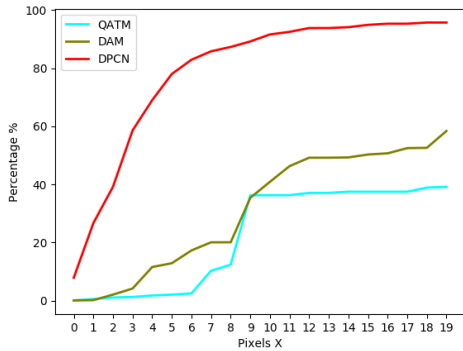
(h) Estimation of  $y$  in "Stereo to Satellite" scene(a)



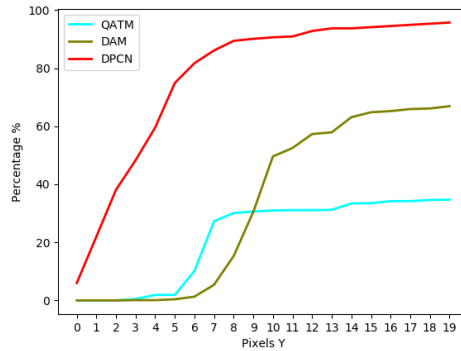
(i) Estimation of  $x$  in "LiDAR to Drone" scene(b)



(j) Estimation of  $y$  in "LiDAR to Drone" scene(b)



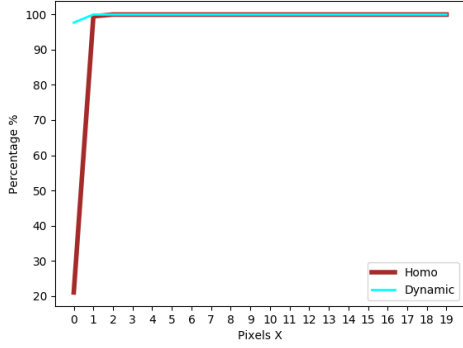
(k) Estimation of  $x$  in "Stereo to Drone" scene(b)



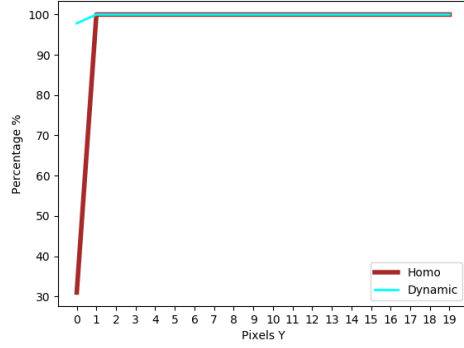
(l) Estimation of  $y$  in "Stereo to Drone" scene(b)

Figure 12:  $Acc_0$  to 19 of translation estimation in Aero-Ground dataset.





(a) Estimation of  $x$  in simulation generalization



(b) Estimation of  $y$  in simulation generalization

Figure 13:  $Acc_0$  to 19 of translation estimation in generalization.

Table 4: Results of generalization experiments with Aero-Ground Dataset. Experiments are conducted with the input type of stereo camera and drone’s birds-eye (AKA “s2d” in Table 3), therefore, the model applied in these experiments are trained on the “s2d” dataset in scene (a) and (b). For generalization, we choose the threshold error of 15 *pixels* for translation,  $1^\circ$  for rotation and  $0.2\times$  for scale. More threshold is elaborated in Figure 6

Model	$E_x$	$Acc_{x_{15}}(\%)$	$E_y$	$Acc_{y_{15}}(\%)$	$E_{rot}$	$Acc_{rot_1}(\%)$	$E_{scale}$	$Acc_{scale_{0.2}}(\%)$
DPCN in (a)	232.4638	73.2	29.9253	92.4	89.0943	95.7	0.0084	95.0
DPCN in (b)	31.2071	92.2	138.5449	88.5	2.8793	96.3	0.0153	93.3
DAM	602.8490	40.8	720.9244	33.1	88.7239	22.5	0.0153	93.3
QATM	3922.6715	26.7	1103.6291	36.9	\	\	\	\