A Robust Matching Network for Gradually Estimating Geometric Transformation on Remote Sensing Imagery

Dong-Geon Kim, Woo-Jeoung Nam and Seong-Whan Lee, *Fellow, IEEE*

***Abstract*— In this paper, we propose a matching network for gradually estimating the geometric transformation parameters between two aerial images taken in the same area but in different environments. To precisely matching two aerial images, there are important factors to consider such as different time, a variation of viewpoint, size, and rotation. The conventional methods for matching aerial image pairs with the large variations are extremely time-consuming process and have the limitations finding correct correspondences, because the image gradient and grayscale intensity for generating the feature descriptors are not robust to the variations. We design the network architecture as an end-to-end trainable deep neural network to reflect the characteristics of aerial images. The hierarchical structures that orderly estimate the rotation and the affine transformations make it possible to reduce the range of predictions and minimize errors caused by misalignment, resulting in more precise matching performance. Furthermore, we apply transfer learning to make the feature extraction networks more robust and suitable for the aerial image domain with the large variations. For the experiment, we apply the remote sensing image datasets from Google Earth and International Society for Photogrammetry and Remote Sensing (ISPRS). To evaluate our method quantitatively, we measure the probability of correct keypoints (PCK) metrics for objectively comparing the degree of matching. In terms of qualitative and quantitative assessment, our method demonstrates the state-of-the-art performances compared to the existing methods.**

I. INTRODUCTION

Image matching is the process to geometrically estimate the visual correspondences between two images taken in same scene but in the conditions of different sensor, viewpoint, time and weather variations. As traditional computer vision approaches, hand-crafted algorithm (such as SIFT [29, 22], SURF [20], HOG [30] and ASIFT [25]) are widely used to solve the matching tasks by computing the correspondences between two images and estimating the geometric transformation parameters. However, these methods are not robust to variation of the environment (such as time, large transformation and weather) and require lots of computational costs in high-resolution images.

As a deep neural network (DNN) has shown impressive performance in many computer vision tasks such as classification [16, 26], object detection [11], segmentation [12], group activity recognition [34], human activity prediction [31], line segment detection [35] and image generation [13], there are many attempts to overcome these limitations with DNN models [4, 7, 9, 15, 17] in an ordinary image domain. However, there are still few researches to solve the aerial image registration due to the lack of data, high-resolution, large transformation and temporal variation. Aerial images have various different characteristics, such as occlusion caused by weather, repeated patterns of the buildings, various brightness depending on time (day/night), and tilt according to taken angle. In this work, we propose a robust matching network for gradually estimating geometric transformation on the aerial image. To reflect the characteristics of aerial images, we exclude the irrelevant regions and gradually apply the transformation to focus on the adjacent regions. Furthermore, the proposed network is end-to-end trainable for learning the repetitive patterns and the environmental variations. Fig. 1 illustrates the gradual matching process of our method. Our contributions are three-fold as follows:

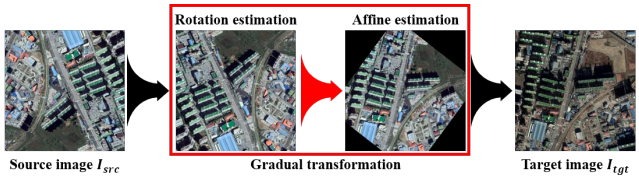

Fig. 1: An entire process for aerial image matching. Through our networks, the source image I_{src} is aligned to target image I_{tgt}.

- We propose a matching network which gradually estimates rotation and affine transformations between two aerial images. To reflect the characteristics of aerial images, we designed the end-to-end trainable network, resulting in more robustness to the variation of taken environments.
- We introduce a gradual masking method for focusing on the adjacent region during step-wise estimating process. This method makes it possible to estimate the geometric transformation parameters more precisely by reducing the regardless feature points.
- We demonstrate that the real-world matching in the aerial image domain is possible on the network trained from the synthetic generation procedure which makes a set of image pairs by applying a random affine transformation.

II. RELATED WORKS

The conventional computer vision methods for finding correspondences consist of two steps, 1) detecting feature points and 2) calculating the similarity of local descriptors

*Research was supported by DAPA (Defense Acquisition Program Administration) and ADD (Agency for Defense Development).

D.-G Kim and S.-W. Lee are with the Department of Brain and Cognitive Engineering, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea (correspondng author to provide e-mail: sw.lee@korea.ac.kr).

W.-J. Nam is with the Department of Computer Science and Engineering, Korea University, Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea.

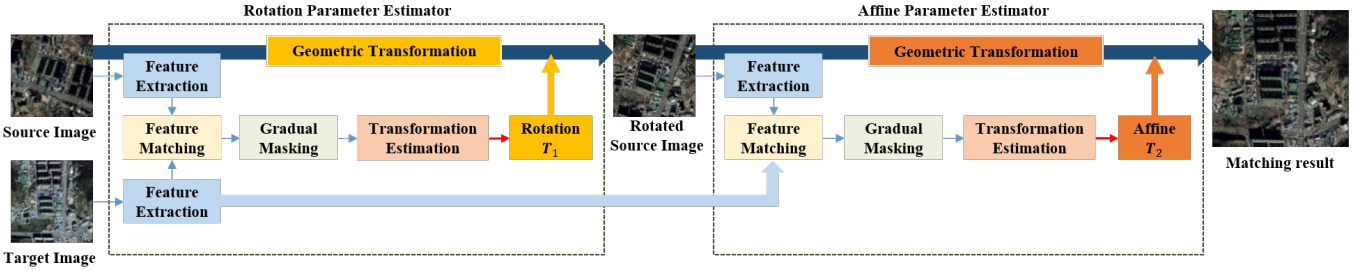

Fig. 2: Overview of the proposed method that estimates the visual correspondences between two aerial images. After passing the rotation parameter estimator, the output transformation is applied to source image, which is the input for the second estimator. Final transformation parameters are computed by composing the rotation and the residual affine transformations.

centered on these points, such as SIFT, SURF and HOG. To complement the problems of SIFT, tons of extensions have been developed [3, 19, 23, 25]. ASIFT and T. Koch *et al*. [3] use a sampling the all possible angles that transform image to find the correspondences. W. Song *et al*. [19] introduce a geometric constraints for aerial image matching. Although this approach is effective for geometrically shallow variations, it still lacks the invariance of non-linear and non-rigid deformations, resulting in limited performance.

In recent years, convolutional neural networks (CNNs) have been used to learn discriminative feature detectors and descriptors which are more invariant to appearance variations than the conventional descriptors [1, 2, 5, 6, 7, 8, 9, 10, 14, 15]. S. Kim *et al*. [1] present a descriptor, called fully convolutional self-similarity (FCSS) that is robust to an intra-class appearance variation. S. Wang *et al*. [2] introduce the network that optimally combines multi-scale feature maps. E. Simo-Serra *et al*. [6] propose a deep network using image-patch pairs. Through a comparison of various models, X. Zhang *et al*. [7] show that certain trainable models are suitable for matching. H. Altwaijry *et al*. [10] develop a binary classifier to distinguish whether the aerial image pair is matched. X. Han *et al*. [15] suggest to train a network with a positive and negative pairs at the same time. Although the aforementioned related works lead to meaningful results, they still find correspondences by extracting feature descriptor from the local patches and comparing them using an appropriate similarity metric.

Several works [4, 17, 18, 33] estimate optical flow and semantic alignment. I. Rocco *et al*. [4] propose a CNNs architecture which estimates the geometric transformation parameters. PH. Seo *et al*. [17] present an attention mechanism with offset-aware correlation (OAC) kernel. S. Jeon *et al*. [18] apply pyramid model for finding dense semantic correspondence. I. Rocco *et al*. [33] propose soft-inlier count method to eliminate outliers. Although the local constraint methods are effective only for some changes, these can not be expected to achieve accurate matching results in the aerial image field, which should be considered for the complex changes. Also, estimating the correspondence through geometric model does not guarantee a stable and accurate result.

III. Proposed Methods

In this section, we describe a novel matching network architecture for gradually estimating rotation and affine transformations between two aerial images. Fig. 2 illustrates the overview of our proposed method. The first estimator regresses the rotation parameters between the source and target aerial images. Then, the image rotated by estimated parameters is utilized as an input of second estimator. The second estimator precisely infers the affine transformation parameters between two images. Each estimator consists of four steps as illustrated in Fig. 3. We address the detailed descriptions of these steps in each section.

A. Feature Extraction

The first step for image matching is to extract features from the input images. We apply the VGG-16 [16] and the ResNet-101 [26] as a descriptor to extract robust and distinctive features. We utilize the generated feature map $f \in \mathbb{R}^{d \times h \times w}$ before passing the fully connected layer. Since the source and target images have to pass the same feature extraction process, this step has the same architecture of the siamese network [32] to share the network parameters.

B. Feature Matching

In second step, we compute the correlation map which denotes the similarity between two generated feature maps. I. Rocco *et al*. [4] first proposed this correlation map to compute the semantic similarities between two ordinary images and showed successful results in semantic alignment domain. Our step is similar to the original one, but has a difference that utilizing a pearson correlation method to correctly cope with nonlinear changes due to weather and temporal factors of the aerial images. Furthermore, ReLu function is applied in the following process to exclude elements that are not correlated with each other. The pearson correlation between two dense feature maps is obtained as follows:

$$c_{AB}(k,i,j) = \frac{(f_B(i,j) - \mu_B)^T (f_A(i_k,j_k) - \mu_A)}{\|f_B(i,j) - \mu_B\| \|f_A(i_k,j_k) - \mu_A\|} \quad (1)$$

where $c_{AB}(k,i,j)$ is the individual feature position in the $d \times h \times w$ dense feature map. $k = h(j_k - 1) + i_k$ is an assistant index for (i_k, j_k). $f_A(i_k,j_k)$ and $f_B(i,j)$ denote individual feature descriptor of f_A and f_B at the positions (i_k, j_k) and (i,j), respectively. μ_A and μ_B are the mean values of each dense feature map, f_A and f_B.

The correlation map c_{AB} has the positive and negative values that mean the relation between two features. Since there are lots of objects and repetitive patterns in aerial

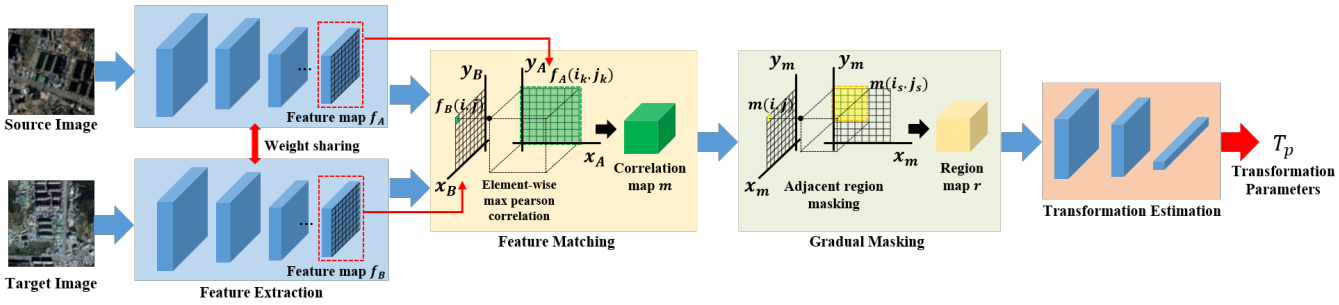

Fig. 3: The detailed structure of each parameter estimator. Each step for estimating transformation is as follow: feature extraction, feature matching, gradual masking and transformation estimation.

images, it is necessary to select the specific points instead of considering all features. Therefore, we apply the max function to the correlation map.

$$m = \max(0,\, c_{AB}) \tag{2}$$

In previous works [4, 17], applying ReLu function in the correlation map shows important role to stable the training and improve the performance. This step is also necessary in aerial image matching domain to prevent the overfitting and make the network more stable.

C. Gradual Masking

Inspired by the conventional methods [22, 28], we apply an adjacent region masking to enable more precise matching by focusing on the regions within the range. This masking separates the region to be considered in the correlation map. In the rotation estimator step, the entire region is considered because there is no information about the degree of difference between the two images. However, in the second estimator, since the source image is somewhat aligned by the estimated rotation parameters, it is possible to concentrate on the adjacent areas and make more precise estimating. Masking the adjacent region is obtained as follows:

$$r(k,i,j) = \begin{cases} m(k,i,j) & \text{if } k \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $r(k,i,j)$ and $m(k,i,j)$ are the individual feature positions in the $d \times h \times w$ region and correlation maps. $\mathcal{S} \in \{s : s = h(j_s - 1) + i_s, 1 \leq s \leq 225\}$ is a set of an assistant index for masking position (i_s, j_s) which is adjacent to (i,j). We define the masking position as $i_s \in \{x : x = i + n, n \in \{-t_p, -t_p + 1, ..., t_p\}\}$ and $j_s \in \{y : y = j + n, n \in \{-t_p, -t_p + 1, ..., t_p\}\}$. t_p is the threshold in p^{th} network, *e.g.* 1^{st} and 2^{nd} networks are the rotation and affine estimator, respectively.

D. Transformation Estimation

At the end of the stage in each estimator, the geometric transformation parameters are regressed by the transformation parameter estimator. In the rotation parameter estimator step, we apply discrete transformation instead of continuous transformation to minimize false distortion. The ranges of 1 degree to 360 degrees are divided into 8 parts by 45-degree intervals and denoted as classes. The network is trained to classify the nearest class to the actual rotation angle. Approximately estimated angle is applied to the source image and utilized as an input of the second estimator. In the affine parameter estimator step, we estimate the precise transformation with the continuous affine transformation parameters to increase the freedom of geometric transformation. The final geometric transformation parameters T_{fin} between the source and target images are computed by multiplying the rotation parameters T_1 and the affine parameters T_2.

E. Supervised Training

To train the proposed networks, we apply the cross-entropy loss and the grid distance loss in each estimator, respectively. Eq. (4) presents the cross-entropy loss function applied in first estimator.

$$L_{angle}(T_1^{GT}, T_1) = -\sum_{i=1}^{N} T_{1_i}^{GT} log(T_{1_i}) \tag{4}$$

where T_1^{GT} is the ground-truth which has 8 classes for discrete rotation angle and T_{1_i} is the output class from the first network. For the second estimator, the average grid distance loss [4] computes the distance between the transformed grids and the ground-truth.

$$L_{grid}(T_2^{GT}, T_2) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathrm{d}(T_2^{GT}(g), T_2(g))^2 \tag{5}$$

where T_2^{GT} is the ground-truth which denotes the geometric transformation parameters between source and target images. To compute the loss between the ground truth and output parameters, both are applied to a set of points in a uniform grid $\mathcal{G}$. Then, it is possible to compute the distance $\mathrm{d}(\cdot)$ between the transformed grid T_2^{GT} and T_2. The second network is trained to reduce this distance $L_{grid}(T_2^{GT}, T_2)$, resulting in reducing the difference between the the ground truth and output parameters.

IV. Experimental Results and Analysis

In this section, we describe implementation details and datasets for evaluating our proposed method. To assess our method, we compare the quantitative and qualitative results with the conventional methods: ObliqueMatching [19], ASIFT [25], WeakAlign [33], CNNGeo [4] and A2Net [17]. The final output of WeakAlign [33], CNNGeo [4] and A2Net [17] is the affine transformation parameters.

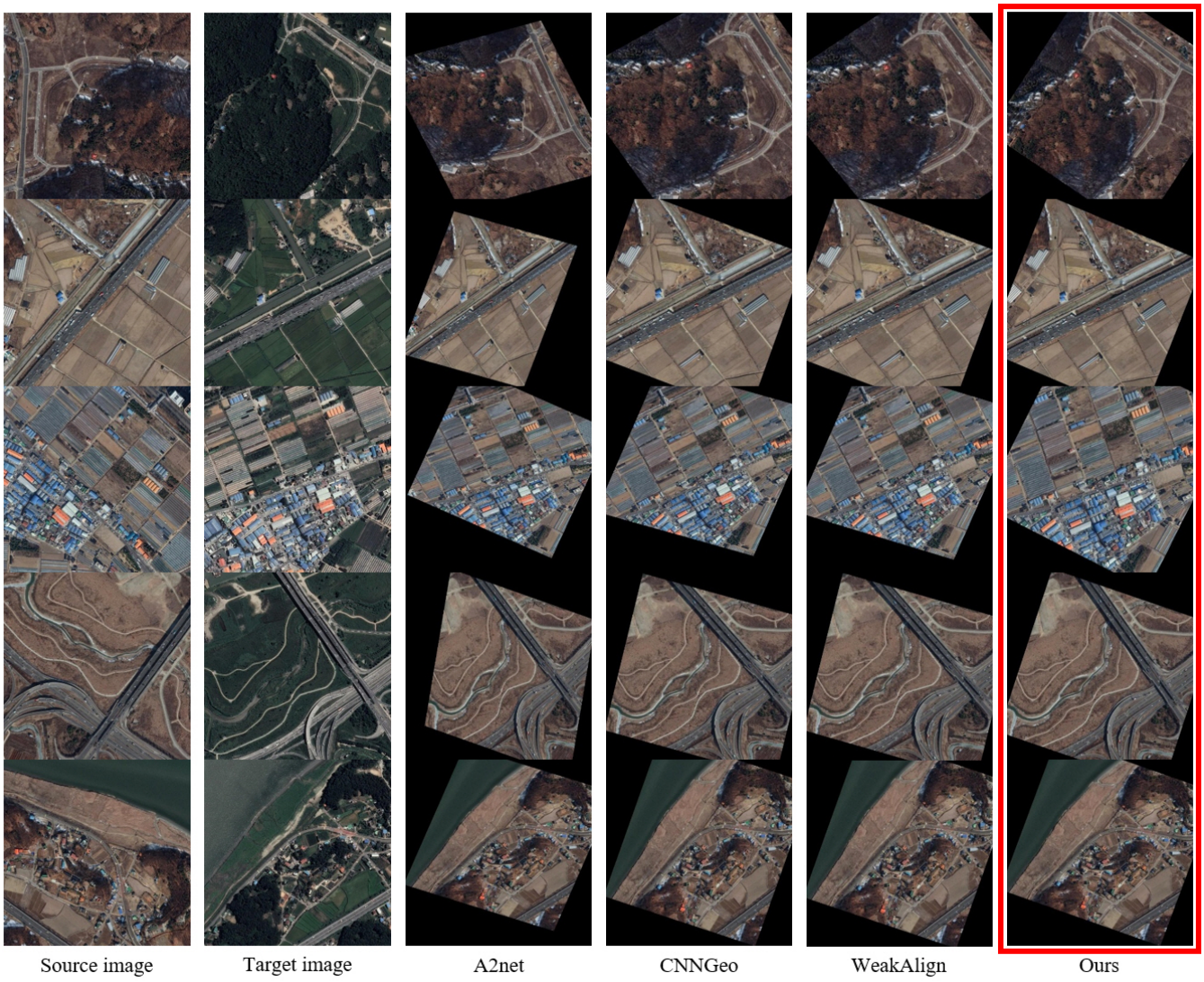

Fig. 4: The comparison of the results from our method and the conventional methods in Google Earth dataset. Each row shows a pair of images from the Google Earth dataset, other alignments and ours. Our method aligns source image to target image successfully than other methods despite of the complex variations

A. *Implementation Details*

Network architecture. For implementing our network, we use the pre-trained model for the feature extraction network learned with ImageNet [27] and fine-tune the model to extract the features suitable for the aerial image. In the experiment, we utilize the two pre-trained models: VGG-16 [16] and ResNet-101 [26].

Training details. Image pairs for training and validation are obtained from the Google-Earth. This dataset contains 10k image pairs, each of them is the same place but taken at different time. We divide them into 9k, 0.5k, and 0.5k image pairs for training, validation and test, respectively. In the case of training, 36k data are created through the data augmentation. Most of the images are taken with the Landsat 7 and 8 satellites. Since there is no aerial image dataset which provides the image pairs and geometric parameters for image registration, we generate training image pairs by transforming the target image with randomly geometric transformation parameters and utilize these parameters as the ground truth. This method makes it possible to generate training image pairs which include same area with different viewpoint. For the qualitative evaluation, we additionally utilize the real-world dataset provided by ISPRS [24]. The input images are resized to 240×240, and the threshold values $t_p = 15, 11$ are used for the gradual masking in the rotation and affine estimators, respectively. We train the network for 100 epochs each with a batch size of 16, taking about two days with a 1080ti GPU.

B. *Quantitative Evaluation*

For quantitatively assessing our model, we perform the evaluation 500 image pairs of Google Earth dataset. We follow the average probability of correct keypoints (PCK) [21]. This metric computes ratio of correctly matched points between the source and target keypoints within the threshold,

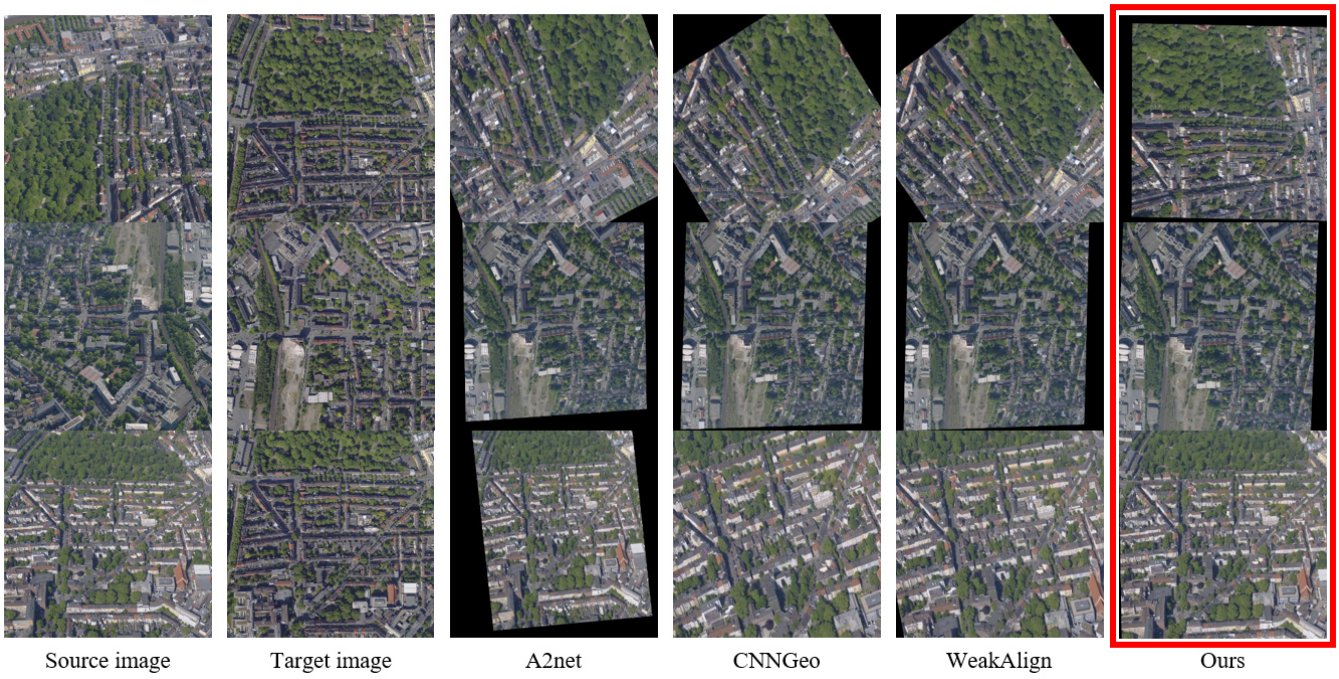

Fig. 5: The comparison of the results from our method and the conventional methods in ISPRS dataset. Each row shows a pair of images from the ISPRS dataset, other alignments and ours. When aligning the source image with different carinal directions to the target image, which is an orthoimage, our method is more stable and accurate than other methods.

TABLE I: The result of PCK evaluation metric in Google Earth dataset.

Method	PCK (%)		
	$\alpha = 0.05$	$\alpha = 0.03$	$\alpha = 0.01$
ObliqueMatching [19]	0.3	-	-
ASIFT [25]	0.5	-	-
A2Net (VGG16) [17]	60.6	36.0	5.8
CNNGeo (VGG16) [4]	64.5	37.2	7.1
WeakAlign (VGG16) [33]	64.4	37.3	7.1
A2Net (ResNet101) [17]	83.0	61.0	13.6
CNNGeo (ResNet101) [4]	83.1	65.9	16.2
WeakAlign (ResNet101) [33]	85.3	65.9	16.4
Ours (VGG16)	86.6	73.5	26.1
Ours (ResNet101)	**92.2**	**83.8**	**36.2**

$\alpha \cdot \max(h, w)$, where h and w are height and width of the input image when the transformation is applied. The PCK metric is as followed.

$$\text{PCK} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\text{d}(T_{fin_i}(p_{s_i}), p_{t_i}) < \alpha \cdot \max(h, w)] \quad (6)$$

where N is the total number of the test image pairs, T_{fin} is the final transformation parameters from our model, (p_{s_i}, p_{t_i}) is the source and target keypoints in i^{th} image pair and $\mathbb{1}[\cdot]$ is the indicator function which have the value 1 if the formula inside square brackets is satisfied and 0 otherwise.

Table I reports the matching accuracy for various aerial matching methods at the three thresholds: $\alpha = 0.05$, 0.03 and 0.01. WeakAlign [33], CNNGeo [4] and A2Net [17] are the state-of-the-art methods in semantic alignment in the ordinary image domain. Although the domain is different, they show good performance in the aerial image domain. Compared with these methods, it is clear that our method shows the state-of-the-art performance, achieving 92.2, 83.8 and 36.2 at each threshold. Furthermore, the result shows that feature extraction stage plays an important role for the aerial image registration. Similar to our intuition, ResNet-101 model with better image classification performance shows higher performance than using VGG-16 model.

C. Qualitative Results

For the qualitative evaluation, we compare the Google Earth and ISPRS datasets. ResNet-101 is used as feature extraction in both evaluations. Fig. 4 illustrates the comparison between the methods in Google Earth dataset. A2net and CNNGeo show the plausible matching results in many cases, but there is a lack of precise matching. Our method shows more precise and correct results. Fig. 5 shows qualitative results on ISPRS dataset. This dataset contains 1260 images (252 images for each direction). The direction consists of 4 cardinal and nadir directions. The dense urban areas are captured with buildings and historical facades in Dortmund, Germany. We analyze only the qualitative evaluation, because there is no geometric parameters between the source and target images in this dataset. Also, as shown in Fig. 5, the proposed method shows more accurate matching results than the conventional methods in the experiment of aligning images with various viewpoint changes based on the orthophotos. We also test the method proposed in [19] and [25], but they does not successfully match in most cases.

V. Conclusions

We proposed a novel matching network with a CNN architecture, which gradually estimates the rotation and affine transformations between aerial images. To learn the characteristics of aerial image domain, we designed the end-to-end trainable network architecture, resulting in the robust feature descriptor against the environmental variations. Furthermore, we applied the gradual masking method to focus on the adjacent region within the range in each estimator. As a result, our method shows the improved results compared to the conventional methods on Google Earth and ISPRS datasets. In the future, we will study for utilizing an unmatched-pair in the training session and focus on developing the network for precise correspondence.

References

[1] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin and K. Sohn, "FCSS: Fully Convolutional Self-Similarity for Dense Semantic Correspondence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, July 21-26, 2017.

[2] S. Wang, L. Luo, N. Zhang and J. Li, "AutoScaler: Scale-Attention Networks for Visual Correspondence," *British Machine Vision Conference*, London, UK, September 4-7, 2017.

[3] T. Koch, X. Zhuo, P. Reinartz and F. Fraundorfer, "A New Paradigm for Matching UAV and Aerial Images," *ISPRS Journal of the Photogrammetry and Remote Sensing*, Vol. 3, 2016, pp. 83-90.

[4] I. Rocco, R. Arandjelovic and J. Sivic, "Convolutional Neural Network Architecture for Geometric Matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, July 21-26, 2017.

[5] H. Altwijry, A. Veit and S. Belongie, "Learning to Detect and Match Keypoints with Deep Architectures," *British Machine Vision Conference*, York, UK, September 19-22, 2016.

[6] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua and F. Moreno-Noguer, "Discriminative Learning of Deep Convolutional Feature Point Descriptors," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, December 7-13, 2015.

[7] X. Zhang, F.X. Yu, S. Karaman and S.-F. Chang, "Learning Discriminative and Transformation Covariant Local Feature Detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, July 21-26, 2017.

[8] S. Zagoruyko and N. Komodakis, "Learning to Compare Image Patches via Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, June 8-10, 2015.

[9] V. Balntas, E. Johns, L. Tang and K. Mikolajczyk, "PN-Net: Comjoined Triple Deep Network for Learning Local Image Descriptors," *arXiv preprint arXiv:1601.05030*, 2016.

[10] H. Altwaijry, E. Trulls, J. Hays, P. Fua and S. Belongie, "Learning to Match Aerial Images with Deep Attentive Architectures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, June and July 26-1, 2016.

[11] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, Montreal, Canada, December 7-12, 2015.

[12] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, June 8-10, 2015.

[13] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, June 18-22, 2018.

[14] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo and F. Fraundorfer, "A Deep Learning Framework for Remote Sensing Image Registration," *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 145, 2018, pp 148-164.

[15] X. Han, T. Leung, Y. Jia, R. Sukthankar and A. C. Berg, "MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, June 8-10, 2015.

[16] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," in *Proceedings of the International Conference on Learning Representations*, San Diego, USA, May 7-9, 2015.

[17] P. Seo, J. Lee, D. Jung, B. Han and M. Cho, "Attentive Semantic Alignment with Offset-Aware Correlation Kernels," in *Proceedings of European Conference on Computer Vision* in, Munich, German, September 8-14, 2018.

[18] S. Jeon, S. Kim, D. Min and K. Sohn, "PARN : Pyramidal Affine Regression Networks for Dense Semantic Correspondence Estimation," in *Proceedings of European Conference on Computer Vision*, Munich, German, September 8-14, 2018.

[19] W. Song, H. Jung, I. Gwak and S.-W. Lee, "Oblique Aerial Image Matching based on Iterative Simulation and Homography Evaluation," *Pattern Recognition*, vol. 87, 2019, pp. 317-331.

[20] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Proceedings of European Conference on Computer Vision*, Graz, Austria, May 7-13, 2006.

[21] Y. Yang and D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 12, 2013, pp. 2878-2890.

[22] D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," *International Journal of Computer Vision*, vol. 60, 2004, pp. 91-110.

[23] K. Mikolajczyk and C. Schmid, "An Affine Invariant Interest Point Detector," in *Proceedings of European Conference on Computer Vision*, Copenhagen, Denmark, May 28-31, 2002.

[24] F. Nex, M. Gerke, F. Remondino, H.-J. Przybilla, M. Baumker and A. Zurhorst, "ISPRS Benchmark for Multi-Platform Photogrammetry," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 2, 2015, pp.135-142.

[25] J. Morel and G. Yu, "ASIFT: A New Framework for Fully Affine Invariant Image Comparison," *Society for Industrial and Applied Mathematics Journal on Imaging Sciences*, Vol. 2, No. 2, 2009, pp. 438-469.

[26] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, June and July 26-1, 2016.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Florida, USA, June 20-25, 2009.

[28] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communication of the ACM*, Vol. 24, No. 6, 1981, pp. 381-395.

[29] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," in *Proceedings of the IEEE International Conference on Computer Vision*, Kerkyra, Greece, September 20-25, 1999.

[30] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, USA, June 20-26, 2005.

[31] D.-G. Lee and S.-W. Lee, "Prediction of Partially Observed Human Activity based on Pre-trained Deep Representation," *Pattern Recognition*, Vol. 85, No. 1, 2019, pp. 198-206.

[32] J. Bromley, I. Guyon, Y. LeCun, E. Sckinger and R. Shah, "Signature Verification using a 'Siamese' Time Delay Neural Network," *Advances in Neural Information Processing Systems*, Denver, USA, November and December 28-3, 1994.

[33] I. Rocco, R. Arandjelovic and J. Sivic, "End-to-end Weakly-supervised Semantic Alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, June 18-22, 2018.

[34] P.-S. Kim, D.-G. Lee, S.-W. Lee, "Discriminative Context Learning with Gated Recurrent Unit for Group Activity Recognition," *Pattern Recognition*, Vol. 76, 2018, pp. 149-161.

[35] N.-G. Cho, A. Yuille, and S.-W. Lee, "A Novel Linelet-based Representation for Line Segment Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 5, 2018, pp. 1195-1208.