Article

A Time Sequence Images Matching Method Based on the Siamese Network

Bo Tao ¹, Licheng Huang ^{2,*}, Haoyi Zhao ², Gongfa Li ¹ and Xiliang Tong ²

¹ Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China; taoboq@wust.edu.cn (B.T.); ligongfa@wust.edu.cn (G.L.)

² Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, China; liuying3025@wust.edu.cn (H.Z.); tongxiliang@wust.edu.cn (X.T.)

* Correspondence: 18759117912@163.com

Abstract: The similar analysis of time sequence images to achieve image matching is a foundation of tasks in dynamic environments, such as multi-object tracking and dynamic gesture recognition. Therefore, we propose a matching method of time sequence images based on the Siamese network. Inspired by comparative learning, two different comparative parts are designed and embedded in the network. The first part makes a comparison between the input image pairs to generate the correlation matrix. The second part compares the correlation matrix, which is the output of the first comparison part, with a template, in order to calculate the similarity. The improved loss function is used to constrain the image matching and similarity calculation. After experimental verification, we found that it not only performs better, but also has some ability to estimate the camera pose.

Keywords: similarity; image pair; the Siamese network; correlation matrix; comparison

Citation: Tao, B.; Huang, L.; Zhao, H.; Li, G.; Tong, X. A Time Sequence Images Matching Method Based on the Siamese Network. *Sensors* **2021**, *21*, 5900. <https://doi.org/10.3390/s21175900>

Academic Editor: Chen Chen

Received: 27 July 2021

Accepted: 31 August 2021

Published: 2 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Judging the relationship of an image pair is a common issue in the field of computer vision. It has an important effect in SfM (structure from motion), image retrieval, pose estimation, and stereo match. Image matching [1–3] refers to judging the relationship between image pairs by identifying the same or similar scenes, objects, shapes, and semantics of two images. Image matching is a basic vision technology, and the result of image matching directly influences the effect of SLAM (simultaneous localization and mapping), 3D reconstruction, and scene understanding. Statistically, above 40% of vision perception tasks rely on precise and effective image matching. These applications include computer vision, pattern recognition, military security, medical diagnosis, and so on.

With the rapid development of image matching technology, many approaches based on the Siamese network [4,5] are proposed. Due to its unique architecture, the Siamese network performs much better compared to other networks in face recognition [6,7], instance segmentation [8], object tracking [9–12], and so on. It is considered as a promising network in image matching. In this paper, the raw data are time sequence images. We make random samples in sequence images to obtain an image pair. A label is constructed by the time interval of images in an image pair. The value range of the label is between 0 and 1. Details of the label construction are presented in Section 4. A label indicates the similarity of images in a pair. We can judge whether the image pair is matched by similarity. The method can easily and conveniently construct a Siamese network dataset.

In conclusion, our major contributions are as follows:

1. Propose a Siamese network for image matching;
2. Propose a convenient and rapid approach to making the dataset;
3. Verify the effect of the Siamese network in image matching.

This paper is organized as follows. Section 1 introduces the application, background, and significance of image matching. Section 2 introduces the related work about the Siamese network and image matching. Section 3 describes our proposed method and architecture in detail. Section 4 shows the details of the experiment. In Section 5, conclusions are presented and future work is discussed.

2. Related Works

Image matching is a classical issue in the image process. In the past 10 years, CNNs (convolutional neural networks) have achieved great progress in image classification, object detection, semantic segmentation, and other fields with a single image. However, most CNNs find it difficult to establish data association with multiple images. Classical CNN architecture is unsuitable for image matching. At present, the major obstacles in image matching are the effects of environmental changes. These changes are caused by factors such as the viewpoint, scale, and illumination. Some traditional algorithms were proposed to deal with these obstacles. There are also some typical algorithms, such as SIFT [13], SURF [14], ORB [15], and the color histogram algorithm [16]. These traditional algorithms use different features to achieve image matching. Based on traditional ones, some improved algorithms were also proposed, such as CSIFT [17], BRISK [18], ORB [19], FREAK [20], LBP [21], and stereo key-point matching [22]. The improved algorithms mentioned are simplified in feature representation in order to improve the overall computational efficiency. However, the traditional algorithms depend on the design and extraction of handcraft features, which are time-consuming, susceptible to environmental interference, less robust, and lacking generalization in practice.

The Siamese network is different from classical CNN architecture. It takes dual inputs and has two feature extractions that share weights. The Siamese network can establish the data association between dual inputs by some fusion operations in the output of feature extractions, such as element subtraction, weighted summation, and so on. The Siamese network was first proposed by LeCun et al. [23]. LeCun applied the Siamese network to handwriting signature verification and achieved the best performance in that time. Then, Hinton et al. first combined it and deep learning to face recognition. With the popularity of deep learning, more and more researchers have combined deep learning and the Siamese network. Luca et al. [24] replaced full-connection layers with convolutional layers to propose FC-Siamese (fully convolutional Siamese) networks. Forian et al. proposed FaceNet [25]. Held et al. [26] explored the effect of the Siamese network in few-shot learning. Some researchers have made some significant works for the Siamese network in representation learning [27,28], and the Siamese network also became an important direction in self-supervised learning [29–31].

3. Method

In image matching, the input image pair has two attributes: similarity and matching. The matching attribute has 3 states to select: pixel matching, semantic matching, or mismatching. The similarity has 2 states to select: similarity or dissimilarity. We defined images in inputs as the original image and the matching images, respectively. Pixel matching requires that pixels in the original image can be found in the matched image with the same intensity and geometric relationship. Semantic matching refers to the establishment of a correspondence domain between two images based on semantic consistency [32]. In other words, semantic matching relies on the semantic information of the pixels rather than the physical information (intensity and geometric relationships). The physical information of a pixel is susceptible to environmental changes, but the semantic information is relatively stable. Therefore, pixel matching is only suitable for image matching in static environments, whereas semantic matching is suitable for image matching in both dynamic and static environments. Mismatching means that pixels in the original image and the matching image are different in both physical and semantic domain. Similarity is our final aim. In Figure 1, pixel matching and mismatching can directly judge between similar or

dissimilar, which is indicated by solid lines. By contrast, semantic matching finds it hard to discriminate between similar or dissimilar, which is indicated by dotted lines.

Figure 1. Relationship between match attribution and similarity attribution.

In fact, pixel matching is a special case in semantic matching. Due to the fact that motion brings translation and rotation to the object, the match image is wrapped in pixels. When matching image is wrapped with camera motion, pixel matching degenerates to semantic matching. Semantic matching has semantic invariance, which is robust. However, the relationship between semantic matching and similarity attribute is not formulated. The reasons are as follows:

- Semantic information is hard to measure;
- Boundary between similar and dissimilar pairs is hard to define.

The most widely used criterion is Euclidean distance. Compared to the Euclidean distance, the probability distribution greatly reduces the influence of local features for image matching. The probability reflects the motion trend in physical and consistency in mathematics. Thus, it retains local features and global features in high dimension space. For the boundary, we designed a neural network to search in a high dimension and output a value to describe similarity attribute.

In this paper, we built 16 anchors uniformly in each image of input pairs. Full convolutional network was used as a feature extraction. Feature extraction was applied to inputs to obtain feature maps, which are composed of 16 anchors. Each anchor in the feature map represents a local feature in the inputs. We called it a patch. Due to the complexity of the distribution, 16 anchors are few. It cannot construct a precious explicit probability distribution.

We turned to construct a 16×16 correlation matrix. It describes the correlation between each patch in the original image and all patches in the matching image. Then non-linear transformation operated on correlation matrix to obtain correlation distribution φ . When the original image and matching image are the same image, the correlation matrix represents a unique form: identity matrix, which remains the same formulation in correlation distribution. We defined the correlation distribution with the attribute as label distribution ζ . In general cases, the original image and the matching image had some differences, which are reflected in φ and lead to φ not being equal to ζ . Therefore, differences, which are negatively related to the similarity, between ζ and φ , are good criterion.

3.1. The Siamese Network

The Siamese network's attributes indicate that it can eliminate random noise and find correlation. For each input, the Siamese network has an instance of feature extraction, and instances share the same weight. In parallel architecture networks, multiple inputs are fed to the parallel blocks. Parameters in parallel blocks are different. Thus, parallel architecture reserves the attribute of random and enforces inputs correlation. Random is harmful, which introduces noise to the system. In series architecture network, single input is required and only one data flow is supported. During training, all inputs are considered

independent and identically distributed. Thus, series architecture is stable without random noise. Single input means that it is hard to find the correlation in dataset.

As Table 1 shows, the Siamese network combines pros of parallel and series architecture using hybrid architecture. The front part constructed a parallel network with two feature extraction instances that are shared weights. The latter part of it is a series network to evaluate the similarity.

Table 1. Pros and cons of Siamese network.

	Parallel Network	Series Network	Siamese Network
Pros	Keep correlation	Stability	Eliminate random noise
Cons	Introduce random noise	Ignore correlation	Keep correlation Correlation hard to model

The Siamese network can avoid mutual interference of inputs and maintain the structural similarity of the input data. In the AI (artificial intelligence) fields, the Siamese network belongs to comparative learning or metric learning. The base principle for comparative learning is boundary measurement. It needs one input as a template, which contains much prior knowledge. Another input is compared to the template. The network can find right boundaries to distinguish different attributes by comparison. Based on that, the network learned the representation.

For that purpose, the dataset is a key. Fitting dataset makes network learned latent representation and improper dataset leads to bad parameters. In the dataset, positive samples were the image pairs with a similarity, and negative samples were the image pairs with enough difference. Negative samples are more than three times as large as positive samples in order to keep data distribution in a suitable domain.

Correlation matrix and correlation distribution φ needed to meet the following conditions:

1. When the input images are the same, φ is equal to ζ ;
2. The stronger the correlation of the corresponding patch, the greater the probability value;
3. In probability, the value range of each element in φ is between 0 and 1 (including 0 and 1), and the elements in each row accumulate to 1;
4. The more similar the input image pairs are, the more similar φ is to ζ .

As for condition 1, the output of the feature extractor in the Siamese network was subtracted to generate a correlation matrix. Rows and columns in matrix represent the position coordinates of the patch in the original image and the matching image, respectively. The element's value in matrix represents the correlation. The stronger the correlation, the lower the value. When the original image and matching image are equal, the correlation is the strongest, and the value is 0. In order to satisfy condition 2, the Gaussian function Formula (1) was used to transform the output space and enhance nonlinearity. In order to satisfy condition 3, the softmax function was used, as in Formula (2).

As shown in Figure 2, along the column direction of the correlation matrix, each row in the correlation matrix was normalized. Since normalization can transform rows into probability distributions, the correlation matrix becomes a correlation distribution φ , which is composed of a series of probability distributions. Each row in φ represents a classification result. The input of this classification is a patch of the original image, and categories are all patches in the matching image.

The variable x in the Gaussian function represents each element in the correlation matrix, which is linear. The parameter σ in the Gaussian function is related to the scale of the input and affects distribution concentration. A lower σ can strengthen the positive correlation and make unimodality. A higher σ can strengthen the negative correlation and

make multimodality. The express $\exp(x_i)$ in Formula (2) is equal to Formula (1); thus x_i is the same as $-\left(\frac{x}{2\sigma}\right)^2$. The parameter k is the number of columns. As for Figure 2, k is 16.

$$\text{Gaussian}(x) = \exp\left(-\left(\frac{x}{2\sigma}\right)^2\right) \quad (1)$$

$$\text{Gaussian}(x) = \exp(x_i) \quad (2)$$

φ can reflect the local correlation between the original image and the matching image. It cannot solve the global correlation between the original image and the matching image directly. The relationship between similarity and φ is more complicated. S and φ are negatively correlated in value, and the structure relationship between S and φ is closely related. We decided to use conditional probability to describe the relationship, which is non-linear, as in the Formula (3) shown.

$$P(S|\varphi, \zeta) = \text{Similar_network}(D(\varphi, \zeta)) \quad (3)$$

The Similar network is a neural network for solving the conditional probability $P(S|\varphi, \zeta)$. The Similar network is designed as a simple fully connected network. $D(\varphi, \zeta)$ denotes the result of the improve KL (Kullback–Leibler) divergence, which is discussed in next section. The activation function uses the tanh function with bias of 0. The tanh function is shown in Formula (4). The variable x represents the output of fully connected network. We took the absolute value of the tanh function as the similarity to ensure that the value range of the similarity is same to the value range of the label.

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \quad (4)$$

The overall architecture of the network is shown in Figure 3. In our architecture, there were two different comparison parts. The first part was to compare the original image with matching images. The second part was to compare correlation distribution φ with label distribution ζ . The first part aimed to find correlation of inputs and generate the correlation distribution φ . The second part aimed to calculate the similarity by correlation.

Figure 2. The label distribution ζ and the correlation distribution φ . The intensity of the color indicates the strength of the correlation; the strongest is 1, and the weakest is 0. The ordinate indicates the position of each patch in the original image, the abscissa indicates the position of each patch in the matching image, and the row indicates the correlation between a certain patch in the original image and all patches in the matching image. If the original image and matching image are equal, each row is unimodal. Rows being multimodal means that the patch is ambiguous and hard to match.

Figure 3. Network architecture.

3.2. Improved KL Divergence

In Siamese network, the most classic contrastive loss function is Info NCE (Info Noise-Contrastive Estimation) loss function [33]. For similar or dissimilar inputs, the Info NCE loss function can guide the network to reduce or increase differences, respectively. Therefore, the key of the Info NCE loss function lies in the semantic information difference measurement. Some classic measurement functions, such as Euclidean distance and cosine distance, are naive. There is a lack of comparison of the overall difference in the input, and it is susceptible to the influence of some special features, which leads to larger deviations. This paper proposes an improved KL divergence as the similarity measurement function to compare the similarity of the inputs.

For probability distribution, KL divergence is typically used to measure the differences between the two probability distributions, as in Formula (5). However, KL divergence is sensitive to the distribution $p(x)$, but insensitive to the distribution $q(x)$. In distribution $p(x)$, the part near the peak region can be exactly estimated, and the probability of the part that is far from the peak region approaches 0. The position on the distribution $p(x)$ represents the space coordinates of the corresponding patch. According to object motion, the closer the space positions are to patch, the stronger the similarity is. Ignoring the spatial coordinates increases the error.

$$D_{KL}(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx \quad (5)$$

Due to being unimodal, the influence of the spatial coordinates is concentrated near the peak, and the value in non-peak area is zero. Therefore, we use coordinates to weigh the difference value, and only the value near the single peak of the distribution φ and ζ are affected, whereas the other positions are masked. Similarity between the distributions is measured by integration. Since the coordinates on the image have different directivities, we weigh φ according to different directions. The Manhattan distance is used to calculate the total distance, as shown in Formula (6).

$$D(\varphi, \zeta) = \int x \ln \frac{\varphi(x, y)}{\zeta(x, y)} dx + \int y \ln \frac{\varphi(x, y)}{\zeta(x, y)} dy \quad (6)$$

Parameters x, y represent the coordinate $\varphi(x, y)$, and $\zeta(x, y)$ represents the probability that the x -th part in the original image matches the y -th part in the matching image, based on the distribution φ and the distribution ζ , respectively. The lower the D scores, the closer ζ is to φ , and the higher the D scores, the larger the difference between ζ and φ .

3.3. The Channel Attention

Channel attention mechanism refers to applying attention mechanism on channel dimension. The channel in feature map represents various semantic information. While each channel is independent, it increases difficulty to feature fusion. Channel attention can promote the quality of feature fusion, producing more advanced features, which are important factors in the image matching process. It learns channel weight distribution, which contains channel correlation. Each channel multiply corresponding weights in order to make feature fusion.

In this paper, we adopted SE (squeeze and excitation) architecture to achieve channel attention mechanism [34]. The SE unit could focus on important features and ignore unimportant features. A SE unit is shown in Figure 4. It has a series of convolution layers and a squeeze layer. The squeeze layer, which includes global average pooling and global maximization pooling, generates weight.

Figure 4. SE unit. W represents the vector width. H represents the vector height. C represents the number of vector channel. The input x is convolved to obtain the feature u . The weight is generated by compressing u into the channel dimension through the squeeze process. Finally, the weighted feature y is obtained by multiplying the weight by the corresponding channel.

3.4. The Design of Loss Function

In general, loss function in the Siamese network is characterized by high cohesion and low coupling. Loss function can maximize inter-class differences and minimize intra-class differences. In this paper, we designed a loss function, which is shown in Formula (7)

$$\text{loss} = \text{BCE}(l, P(S|\varphi, \zeta)) + L \quad (7)$$

Our designed loss function consisted of two parts. The first part was the classic BCE (binary cross-entropy) function, which is widely used in binary classification tasks. The BCE function requires l and $P(S|\varphi, \zeta)$ as inputs, where l is a label that decides whether the original image matches the matching image ($l = 1$) or not ($l = 0$), and $P(S|\varphi, \zeta)$ is defined in Formula (3). Since $P(S|\varphi, \zeta)$ is the result of Similar network, the BCE function is mainly trained on the parameters of Similar network. The second part is a contrastive loss function, which is shown in Formula (8):

$$L = (1 - l)D + l\{\max(\text{margin} - D, 0)\} \quad (8)$$

where D is a function defined in Formula (6), the parameter l is same as the l in BCE function, and margin represents the boundary of differences. Only when the difference value is within a certain range can the loss be optimized. If the difference is out of range, a constant is set and no optimization occurs. In the experiment, the margin was set to 1. The L required label and correlation distribution φ for learning. Since φ are results of feature extraction, φ focused on feature representation, and L mainly trained the parameters in the feature extractor.

4. Experiment

In experiment, the dataset was some color image videos from KITTI [35], and the input image was resized to 64 pixels on the horizon and vertical. Positive and negative

samples were constructed by chronological order. The image pairs with time interval less than n were taken as positive samples and labeled as 0. The ones greater than n were taken as negative samples and labeled as 1. The selection of hyper parameter n was related to the correlation of image pair: If n too small, not enough correlation would be learned; otherwise, unnecessary errors will be introduced if n is too large. Fewer positive samples could not learn correlation precisely. Fewer negative samples introduced unnecessary errors. According to the instructions given by KITTI dataset, one interval was 0.1 s on average. We decided to set n to 5. When n was greater than 5, the change caused by camera motion was apparent and the correlation of image pairs was low.

By means of sampling, a dataset was constructed. The dataset contained positive and negative sample pairs. The positive sample pairs were guaranteed to account for approximately 25% of the dataset, and the negative sample pairs were far greater than the positive sample pairs. In the Siamese network, feature extraction adopted the full convolutional network architecture and the similar network adopted 1 layer of full connection network. The network architecture is shown in Table 2. Due to the fact that the padding in the convolution process can affect the scale invariance, all the convolution layers had no padding.

Table 2. Network architecture.

Feature Extraction	Similar Network
Input $64 \times 64 \times 3$	Input 16×16
7×7 conv 16 BN ReLU valid	Flatten
5×5 conv 32 BN ReLU valid	1 Dense Tanh
3×3 conv 256 BN ReLU valid	
3×3 conv 128 BN ReLU valid	
3×3 conv 128 BN ReLU valid	
2×2 maxpool	
3×3 conv 256 BN ReLU valid	
3×3 conv 512 BN ReLU valid	
3×3 conv 256 BN ReLU valid	
3×3 conv 256 BN ReLU valid	
3×3 conv 512 BN ReLU valid	
3×3 conv 256 BN ReLU valid	
3×3 conv 256 BN ReLU valid	

In the experiment, the model was trained on NVIDIA GTX 1080 Ti GPU with 12 GB memory. The inference time for single image was 10 ms, and the total parameter size was 20 M. The accuracy was 89.2%. We compared a variety of networks with different architectures, as shown in Table 3.

Table 3. Result of different networks.

Name	Accuracy (%)
Siamese	84.2
Siamese + improved KL divergence	88.6
Siamese + SE unit + improved KL divergence	89.2

We compared the effects of improving the KL divergence and SE unit. As for the SE unit, it was found that, after using SE structure in the network training process, the value of loss function decreased more stably and the convergence speed was faster. With the improved KL divergence, the loss value could be maintained at a lower level and had better generalization performance. In conclusion, SE unit and improved KL divergence can improve the matching accuracy and processing speed in the Siamese network.

We compared the performance of our method, triple Siamese network, and sHybirdNet, in different scene datasets. These scenes included street, urban, and highway. Different scenes have different characteristics. The results are shown in Table 4.

- Highways have fewer features and different features are highly similar (e.g., lanes);
- Cities have more features and some features have random motion (e.g., pedestrians);
- Streets have more distinguishable features and most of the features have small position changes in the image (e.g., buildings).

Table 4. Comparison of different approach.

Dataset Scene	Siamese + SE Unit + Improved KL Divergence (Our Method)	Triple Siamese Network [36]	sHybirdNet [37]
Streets	89.2	82.5	75.2
Cities	78.2	74.6	65.4
Highways	69.9	68.2	54.2

Through experiments, we found that all networks work better in street scenes and worse in highway scenes. Our proposed method performs much better than sHybirdNet in all scenes, with an improvement greater than 10%. Compared to triple network, our proposed method also has some advantages.

As shown in Figure 5, the correlation of the input image pairs is manifested in the structure and intensity. The intensity indicates the similarity of each patch in image pairs. In the higher similarity of image pairs, there are more bright patches in φ , which indicates that most patches in the original image can find the corresponding matching patches in the matching images. The more similar the image pairs are, the more patches with high correlation in φ . The structure of the φ represents the position relationship of each block in the original image and the matching image. The higher the similarity of image pairs, the closer φ and ζ are in structure.

Figure 5. Correlation distribution φ and input image: top, φ with identical input images, and the output from the similar network is 1×10^{-6} ; middle, φ with similar input images (difference between 5 frames), and the similar network output is 0.653; bottom, φ with completely different input images (difference between 50 frames), and the similarity is 0.998.

We can obtain more information from the structure of φ , such as in Figure 6. When the camera rotates to the right or left, the corresponding distribution can shift to the right or left without breaking the structure as much as possible. The structure and intensity of φ can be analyzed by the similar network, which is sensitive to changes in φ . Therefore, the motion changes between the original image and the matching image can be further inferred through the similar network.

Figure 6. Correlation distribution φ in the case of camera rotation: top, the change in the φ when the camera is rotated to the right; bottom, the change in the φ when the camera rotates to the left.

5. Conclusions

In this paper, an image matching method based on the Siamese network is proposed. We use probability distribution to model the input. The network can distinguish the similarity and difference in input through a comparison. Based on the method, the network outputs the similarity between images and determines the matching relationship between images. Based on the basic Siamese network, an improved KL divergence is used to evaluate the distribution difference, which comprehensively considers the influence of the position. It also introduces the SE unit in order to accelerate the convergence and keep it stable in the training process. It uses convolution to divide the image into multiple image patches for matching, which is different from directly matching the whole image. By comparing label distribution ζ , it keeps the consistency and global constraint on the matching results. In addition, the correlation distribution φ can estimate the basic motion of the camera.

In this paper, the input is resized to 64×64 . With a larger input scale and more complex feature extractor, fine features can be extracted; it also generates more patches, which increases the difficulty of matching. The similar network is constructed by a simple fully connected layer. More complex network structures can measure the similarity in different metrics. The correlation matrix is obtained by subtraction. In addition, a more accurate correlation matrix can be obtained by more advanced methods, such as the graph network method.

In future work, we will use a larger scale input, combined with the principle of target detection, to screen blocks in order to obtain interested blocks, so as to greatly reduce the complexity caused by the increase in input scale.

Author Contributions: Conceptualization, B.T. and L.H.; methodology, L.H.; software, L.H.; validation, B.T., L.H. and H.Z.; formal analysis, L.H.; investigation, H.Z.; resources, G.L.; data curation, B.T.; writing—original draft preparation, L.H.; writing—review and editing, B.T.; visualization, X.T.; supervision, G.L.; project administration, G.L.; funding acquisition, B.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 51505349, National Natural Science Foundation of China, grant number 51575407, and Hubei Provincial Department of Education, grant number D20201106.

Data Availability Statement: Data sharing is not applicable to this article.

Acknowledgments: This project is supported by National Natural Science Foundation of China (No. 51505349, No. 51575407) and Hubei Provincial Department of Education (D20201106). We gratefully acknowledge the support of NSF.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Wei, Y. Research on image matching based on CNN feature. *Harbin Inst. Technol.* **2017**. [\[CrossRef\]](#)
- Zitova, B.; Flusser, J. Image registration methods: A survey. *Image Vis. Comput.* **2003**, *21*, 977–1000. [\[CrossRef\]](#)
- Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; Yan, J. Image matching from handcrafted to deep features: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 23–79. [\[CrossRef\]](#)
- Henry, H.Y.; Liu, J.; Sun, H.; Wang, Z.; Zhang, H. GetNet: Get Target Area for Image Pairing. In Proceedings of the 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), Dunedin, New Zealand, 2–4 December 2019; pp. 1–6.
- Chicco, D. Siamese neural networks: An overview. *Artif. Neural Netw.* **2021**, *2190*, 73–94.
- Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010.
- Song, L.; Gong, D.; Li, Z.; Liu, C.; Liu, W. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 773–782.
- Luo, S.; Li, C. Few-shot instance segmentation based on double similarity Siamese network. *J. Wuhan Univ. Sci. Technol.* **2020**, *43*, 59–66.
- He, A.; Luo, C.; Tian, X.; Zeng, W. A twofold siamese network for real-time object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4834–4843.
- Guo, Q.; Feng, W.; Zhou, C.; Huang, R.; Wan, L.; Wang, S. Learning dynamic siamese network for visual object tracking. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 1763–1771.
- Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the European Conference on Computer Vision, Munich, The Germany, 8–14 September 2018; pp. 101–117.
- Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. Hight performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
- Lowe, D.G. Distinctive Image Feature from Scale-Invariant Keypoints(SIFT). *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [\[CrossRef\]](#)
- Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [\[CrossRef\]](#)
- Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
- Shen, D. Image registration by local histogram matching. *Pattern Recognit.* **2007**, *40*, 1161–1172. [\[CrossRef\]](#)
- Abdel-Hakim, A.E.; Farag, A.A. CSIFT: A SIFT descriptor with color invariant characteristics. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1978–1983.
- Leutenegger, S.; Chli, M.; Siegwart, R.Y. BRISK: Binary robust invariant scalable keypoints. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2548–2555.
- Yu, L.; Yu, Z.; Gong, Y. An improved ORB algorithm of extracting and matching. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2015**, *8*, 117–126. [\[CrossRef\]](#)
- Alahi, A.; Ortiz, R.; Vandergheynst, P. Freak: Fast retina keypoint. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 510–517.
- Guo, Z.; Zhang, L.; Zhang, D. Rotation invariant texture classification using LBP variance (LBPV) with global matching. *Pattern Recognit.* **2010**, *43*, 706–719. [\[CrossRef\]](#)
- Le, A.V.; Won, C.S. Key-point based stereo matching and its application to interpolations. *Multidimens. Syst. Signal Process.* **2017**, *28*, 265–280. [\[CrossRef\]](#)

23. Bromley, J.; Bentz, J.W.; Bottou, L.; Guyon, I.; LeCun, Y.; Moore, C.; Säckinger, E.; Shah, R. Signature verification using a “Siamese” time delay neural network. *Adv. Neural Inf. Process. Syst.* **1993**, *6*, 737–744. [\[CrossRef\]](#)
24. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 850–865.
25. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
26. Held, D.; Thrun, S.; Savarese, S. Learning to track at 100 fps with deep regression networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 749–765.
27. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H. End-to-end representation learning for correlation filter based tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2805–2813.
28. Chen, X.; He, K. Exploring Simple Siamese Representation Learning. *arXiv* **2020**, arXiv:2011.10566.
29. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 16–18 June 2020; pp. 9729–9738.
30. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv* **2021**, arXiv:2103.03230.
31. Grill, J.B.; Strub, F.; Althé, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent: A new approach to self supervised learning. *arXiv* **2020**, arXiv:2006.07733.
32. Xu, H.; Chen, X.; Cai, H.; Wang, Y.; Liang, H.; Li, H. Semantic Matching Based on Semantic Segmentation and Neighborhood Consensus. *Appl. Sci.* **2021**, *11*, 4648. [\[CrossRef\]](#)
33. Oord, A.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [\[CrossRef\]](#)
36. Banerjee, D.; Kyrarini, M.; Kim, W.H. Image-Label Recovery on Fashion Data Using Image Similarity from Triple Siamese Network. *Technologies* **2021**, *9*, 10. [\[CrossRef\]](#)
37. Melekhov, I.; Kannala, J.; Rahtu, E. Siamese network features for image matching. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 378–383.