# Photometric Transfer for Direct Visual Odometry

Kaiying Zhu[a], Xiaoyan Jiang[a,*], Zhijun Fang[a], Yongbin Gao[a], Hamido Fujita[b], Jenq-Neng Hwang[c]

[a]*School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, No.333 of Longteng Road, Shanghai, China*
[b]*Faculty of Software and Information Science, Iwate Prefectural University, Iwate 020-0693, Japan*
[c]*Department of Electrical and Computer Engineering, University of Washington, Box 352500, Seattle WA 98195, USA*

## Abstract

Due to efficient photometric information utilization, *direct visual odometry* (DVO) is getting widely used to estimate the ego-motion of moving cameras as well as map the environment from videos simultaneously, especially in challenging weak-texture scenarios. However, DVO suffers from brightness discrepancies since it directly utilizes intensity patterns of pixels to register frames for camera pose estimation. Most existing brightness transfer methods build a fixed transfer function which is inappropriate for successive and inconsistent brightness changes in practice. To overcome this problem, we propose a Photometric Transfer Net (PTNet) which is trained to pixel-wisely remove brightness discrepancies between two frames without ruining the context information. Photometric consistency in DVO is obtained by adjusting the source frame according to the reference frame. Since no dataset is available for training the photometric transfer model, we augment the EuRoC dataset by generating a certain number of frames with different brightness levels for each original frame by a nonlinear transformation. Afterwards, required training data containing various brightness changes and scene movements along with ground truth can be collected from the extended sequences. Evaluations on both real-world and syn-

[*] Corresponding author
*Email addresses:* `smu_zky@live.com` (Kaiying Zhu), `xiaoyan.jiang@sues.edu.cn` (Xiaoyan Jiang)

thetic datasets demonstrate the effectiveness of the proposed model. Assessment on an unseen dataset with fixed model parameters trained on another dataset proves the generalization ability of the model. Furthermore, we embed the model into DVO to preprocess input data with brightness discrepancies. Experimental results show that PTNet-based DVO achieves more robust initialization and accurate pose estimation than the original one.

*Keywords:* Photometric transfer, direct visual odometry, data augmentation, brightness discrepancy, deep learning

## 1. Introduction

The technique of visual odometry (VO), which is used to estimate the egomotion of moving cameras as well as map the environment from videos simultaneously, is essential in many applications, such as, autonomous driving, augmented reality, and robotic navigation. Robust system initialization, drift elimination for successive motion estimation, and computational efficiency are the key and challenging tasks in VO. In recent years, a series of powerful VO algorithms have been proposed [1, 2, 3, 4, 5, 6, 7] with different characteristics. These algorithms can be mainly sorted into direct methods and feature-based methods depending on the ways of utilizing visual information. Feature-based methods first extract manually designed features from each frame and then match them across frames according to feature descriptors. Afterwards, projection error can be defined as the Euclidean distance between the projected points and their correspondent points on the source frame. In contrast, direct methods skip the step of feature processing and directly use pixels to register images by finding the optimal projection of points that minimizes photometric residuals from one frame to the other. Compared with feature-based methods, direct methods can be faster and more robust in texture-less environment where the elaborately designed features are hard to be extracted.

However, since DVO utilizes original photometric information from images and follows the photometric consistency assumption, it easily suffers from bright-

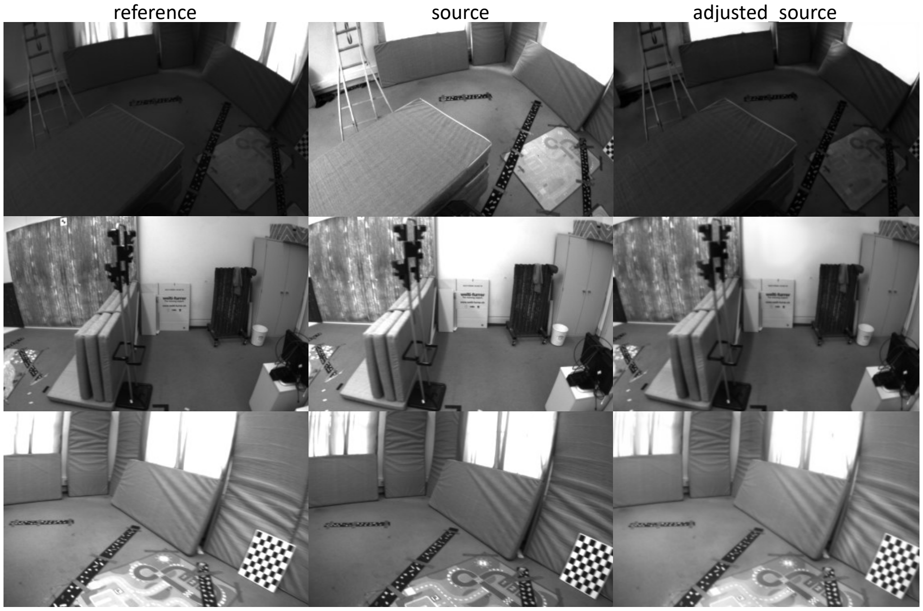

Figure 1: Brightness adjustment. In this paper, we present a CNN-based approach to pixel-wisely adjust brightness of the source frame according to the reference frame for photometric consistency. With preprocessing of input images with brightness discrepancies by the proposed model , DSO [1] achieves more robust estimation against brightness changes. From left to right are the reference images, the source images, and the adjusted source images, respectively.

ness changes, which is a critical problem. Here, photometric consistency means 2D points in different images projected from the same 3D point in the scene should keep the same pixel value. Moreover, brightness discrepancy is also common in practical scenarios due to, for example, automatic exposure change, radiance discrepancy, and dynamic lighting. Some traditional brightness transfer models either assume linear transfer relationship between frames [1] or tune brightness globally [8], neither of them are suitable for practical nonlinear cases. Other nonlinear models [9, 10] learn a fixed transfer pattern from the given samples and are also inappropriate for dynamic scenarios.

Thus, we propose a convolutional neural network (CNN), called the Photometric Transfer Net (PTNet), which pixel-wisely adjusts image brightness, to keep the brightness of sequences consistent in DVO systems. Some adjustment results are shown in Figure 1. To the best of our knowledge, we are the first to design a CNN for photometric transfer in DVO. In a VO system, the most updated camera pose is generally estimated by registering the latest frame to a reference frame on the front-end module. To match this process, our proposed model takes a pair of images including one reference image and one source image as input and the output of the model is an adjusted source image which is photometric constant with respect to the reference image.

The first challenge is designing the structure of the PTNet. So far, deep learning has proven its power in a lot of visual tasks, such as semantic segmentation [11, 12], image generation [13, 14, 15], and optical flow estimation [16, 17, 18]. These tasks require neural networks to do pixel-wise classification or regression, which have similar goals with our task. For pixel-wise image adjustment, we choose the encoder-decoder architecture [11], which has been proven effective in many other tasks, to regress an adjustment mask with the same size as the input image. For regressing pixel-wise adjustment values, the model needs to find correlations of pixels across two frames. Inspired by the optical flow estimation, we replace the original encoder with a Siamese architecture to extract multilevel features from both the source and reference images, and concatenating these features to find correlations.

Another challenge is that no sufficient training data are available. As far as we know, there is no existing dataset designed for the photometric transfer task. Since we aim to use the model to improve the performance of DVO, the training data should contain huge number of image pairs with both brightness discrepancies and scene motions. Furthermore, ground truth which contains exactly consistent context information with eliminated brightness discrepancies is also necessary but hard to obtain for supervised learning. To overcome this problem, we introduce a novel approach to create data and train the model. To be specific, we choose an sequence from the existing EuRoC dataset [19] as the base data and extend it by manually generating a certain number of images with different brightness for each original image by nonlinear transformation. Hence, we obtain more sequences containing different brightness levels. Then, required training data with corresponding ground truth can be collected from the augmented sequences and the original sequences. Moreover, collected image pairs with various brightness discrepancies and scene motions as training data ensure the generalization ability of the model.

The main contributions are summarized as follows:

- We propose the PTNet, which contains a Siamese encoder followed by multilevel correlation layers to associate multilevel features from both the source and the reference frames, and a decoder to generate an adjustment mask from coarse-to-fine. The proposed model can effectively adjust the source image according to the reference image to pixel-wisely remove brightness discrepancy.
- To obtain sufficient data for training, we present a novel approach to augment the EuRoC dataset and train the model. Generated data with various brightness discrepancies and scene motions from the augmented sequences makes the model efficiently learn the ability of brightness change elimination and ensure the generalization ability for the model.
- Experiments on both real-world and synthetic datasets prove that the model can efficiently eliminate brightness discrepancy across frames. Note that the model is only trained using part of the real-world data. Furthermore, we embed the model into DVO to preprocess the input data with brightness discrepancies and results show that DVO can achieve more robust initialization and pose estimation with our model.

The rest of the paper is structured as follows. In Sect. 2, we summarize relevant literature on dealing with brightness discrepancy problem. Before we introduce the proposed approach, we first introduce the influence of brightness discrepancy in a DVO system in Sect. 3.1. Then, we introduce our approach, including the structure of PTNet, a data augmentation method, and loss functions for training the model, in Sect. 3.2. Implementation details and experimental results are shown in Sect. 4. We conclude our work in Sect. 5.

## 2. Related Work

A VO system contains a normal architecture with a front-end for estimating camera pose initially and a back-end for jointly optimizing estimation results. Depending on the ways of utilizing data, existing algorithms can be categorized into direct methods and feature-based methods. Feature-based methods, such as ORB-SLAM [2], PTAM [20], and PL-SLAM [3], use manually designed features to register image pairs. Instead of feature extracting and matching, direct methods, such as LSD-SLAM [21] and direct sparse odometry (DSO) [1], utilize pixel intensity directly by minimizing photometric projection errors, which is more efficient on image information utilization. However, they suffer from brightness discrepancies. In the following, we summarize existing methods dealing with the brightness discrepancy problem.

**By Camera Parameter Calibration.** One weakness of DVO is that it suffers from brightness discrepancy, which is commonly caused by the change of exposure time and imaging distortion. To overcome this problem, some works aim to do photometric calibration for solving photometric inconsistence problems caused by camera parameters. [22] proposes to estimate the radiometric response function of the camera by taking images from the same scene with different exposure time. [23] uses overlapping observations of uniformly colored surface to retrieve a vignette map of the camera from large amounts of images. Both methods need to know the exposure time of cameras, which is inaccessible in many situations. [24] proposes a method to retrieve the radiometric response function from only one grayscale image. It utilizes edge regions in an image with the assumption that irradiance changes should be uniform from one side to the other side across an edge and the changes can be non-uniformly transformed because of the nonlinear radiometric response function. A similar work [25] is based on colorful images. [26] evaluates a vignetting function of camera from image regions with uniform scene in one image. [27] further estimates the radiometric response function and vignetting function simultaneously from corresponding pixels between images captured by a moving camera. [28] recovers the response and vignetting functions, exposure time of arbitrary videos by a nonlinear estimation formulation based on [29]. These methods eliminate brightness changes with an assumption that environment illumination is consistent and exact pixel values can be retrieved by calibrated parameters.

**By Robust Image Alignment.** On the other hand, some works study robust image registration methods under severe illumination situation. [30] estimates a global brightness bias by using median of photometric errors of projected points and subtracting this bias to offset the effect of brightness change. [31, 32] suggest utilizing mutual information as a metric for image registration. They optimize a displacement to minimize mutual information of images for template alignment. [33, 34, 35, 36] assume that higher order derivatives of images are illumination invariant and gradient or hessian of images can thus be used to evaluate errors. [37] proposes to use image points with local extrema of curvature to perform image registration. [5] discusses a direct visual tracking method based on normalized cross correlation (NCC), which is invariant to affine illumination changes. [38, 39] manually design illumination robust descriptors for sparse feature matching. These methods usually suffer from limited convergence basin and are less robust when scene motion is large [40].

**By Eliminating Brightness Discrepancy Directly.** Other works focus on adjusting images for photometric consistency without any calibration information. [8] corrects one image statistically to make its mean and standard deviation of pixel values similar to the reference image. [41, 42, 43] transform an image so that its histogram of pixel values matches a specified histogram. Different from previous statistics-based methods, a few methods in the literature explore the brightness transfer function to separate real illumination changes from motion-induced brightness discrepancies. [44, 10] estimate a single relative intensity transfer function from image regions where no apparent motion is present. [45] exploits physical models of time-varying brightness that consider time-dependent physical causes and changing surface orientation with respect to a directional illuminant. [46] proposes to compute the brightness transfer function via histogram specification. [9] estimates several brightness transfer functions for different regions of images and then uses principal component analysis to find one common function by merging these transfer functions with different weights. The Brightness transfer function, which gives a mapping of pixel values between two images, is commonly learned from some data samples. This makes it inapplicable for dynamic environments and its performance is subject to training data. In some dynamic scenes with continuous illumination discrepancy, some methods use a set of simple affine parameters to keep any two frames photometric consistent and the parameters are jointly optimized with camera poses during direct image registration [47, 1]. [48] uses a CNN to directly estimate affine parameters to globally adjust brightness of images. These methods cannot exactly remove the brightness discrepancies which are normally nonlinear and hard to be modeled. There also exists some novel explorations, such as transferring original frames into a new feature space which is unaffected by brightness and season changes [49]. However, the feature space is not exactly photometric consisitent for complicated scenarios.

## 3. The Proposed Approach

To deal with brightness changes of input data for DVO, we propose a CNN architecture to pixel-wisely adjust the brightness of images. The model takes both the reference frame and the source frame with brightness discrepancy as input and generates an adjusted source frame which has consistent brightness with respect to the reference frame. To gain sufficient training data containing various brightness discrepancies and camera motions, we propose a novel approach to augment the EuRoC dataset [19] and collect training data with ground truth from it. Finally, we test the model on both real-world and synthetic datasets to validate the ability of brightness discrepancy elimination and improvement for DVO. Before we introduce the details of the proposed approach, we provide a brief overview of the influence of brightness discrepancy on image registration in DVO.

### *3.1. Influence of Brightness Discrepancy on DVO*

Generally, a DVO system can be decomposed into two main components: front-end and back-end. On front-end, a new (the $j^{th}$) source frame is registered to the latest (the $i^{th}$) reference frame by directly backprojecting all image points from the reference frame to world coordinates and then projecting them to this new source frame again. This projection process can be formulated as

$$\mathbf{p}'_j = \prod\nolimits_c (\mathbf{T}_j \mathbf{T}_i^{-1} \prod\nolimits_c^{-1} (\mathbf{p}_i, d)), \tag{1}$$

with

$$\mathbf{T}_i := \begin{bmatrix} \mathbf{R}_i & \mathbf{t}_i \\ \mathbf{0} & 1 \end{bmatrix}, \ \mathbf{T}_j := \begin{bmatrix} \mathbf{R}_j & \mathbf{t}_j \\ \mathbf{0} & 1 \end{bmatrix}, \tag{2}$$

where $\prod_c(\cdot)$ denotes projecting points from 3D camera coordinates to 2D image plane coordinates; $\mathbf{p}_i \in \mathbb{R}^2$ denotes homogeneous coordinates of points in the reference image and $d \in \mathbb{R}$ denotes inverse depths of points in camera coordinates; $\mathbf{T}_i \in SE(3)$ and $\mathbf{T}_j \in SE(3)$ denote the extrinsic transformation that transfers points from world coordinates to camera coordinates of the $i^{th}$ and the $j^{th}$ frames, respectively and both can be decomposed to a rotation $\mathbf{R} \in SO(3)$ and a translation $\mathbf{t} \in \mathbb{R}^3$. Registering process is to iteratively find an optimal projection to minimize photometric error, which is defined as difference of pixel values between the points in the reference frame and corresponding projected points in the source frame. The photometric error function is defined as

$$E_{project} := \sum\nolimits_{\mathbf{p}_i \in \Omega} \left\| I_{src}[\mathbf{p}'_j] - I_{ref}[\mathbf{p}_i] \right\|_{\gamma}, \tag{3}$$

where $I_{ref}[\mathbf{p}_i]$ is the pixel value of a point $\mathbf{p}_i$ in the reference frame; $\|\cdot\|_{\gamma}$ is the Huber norm; $\Omega$ is a set of points for projection in the reference frame. To find the optimal solution, a nonlinear optimization method, e.g., the Gauss-Newton method, can be iteratively applied. Only when the points are projected close to the correct positions and the distribution of corresponding pixel values is consistent across two frames, this optimization process can be valid. *Hence, brightness discrepancy between image pairs brings more uncertainty and makes system fail to find good initial estimation.* To deal with brightness changes, we design a PTNet, which gets two frames $I_{ref}, I_{src}$ as input, to generate output $\widehat{I_{src}}$ that has consistent brightness with respect to $I_{ref}$. Then the photometric error function can be rewritten as

$$E_{project} := \sum\nolimits_{\mathbf{p}_i \in \Omega} \left\| G(I_{src}|I_{ref})[\mathbf{p}'_j] - I_{ref}[\mathbf{p}_i] \right\|_{\gamma}, \tag{4}$$

with

$$G(I_{src}|I_{ref}) := \widehat{I_{src}}, \tag{5}$$

where $G(I_{src}|I_{ref})$ denotes adjusting the brightness of $I_{src}$ according to $I_{ref}$.

To detect brightness changes across two frames rapidly, we accumulate the brightness difference between two frames as

$$E_{brightness} := |\frac{1}{n} \sum_{\mathbf{p} \in R} (I_{ref}(\mathbf{p}) - I_{src}(\mathbf{p}))|, \tag{6}$$

where $R$ is the region for accumulating pixel values which is set for robust detection; $n$ is number of pixels in the region. Even without brightness discrepancy, source frame and reference frame always have different distribution of pixel values caused by scene motions. In our experiments, we find that distribution of pixel values in the center of the image is more stable against scene motions compared with regions around the image. To avoid misjudgment of brightness change caused by scene motions, we choose a rectangle region in the center of images to detect brightness changes in our experiments and confirm it when $E_{brightness}$ exceeds a threshold $T_b$. Note that this step is just for detecting brightness discrepancy, so we do not need an accurate template alignment here.

## *3.2. **The Proposed Photometric Transfer Net (PTNet)***

### *3.2.1. **Network Architecture***

We design the PTNet based on a CNN to adjust pixel values of the source frame with respect to the reference frame. The aim is to keep pixel values constant from the reference frame to the source frame if the pixel pair is projected from the same 3D point. We choose a U-Net [11] as our backbone and the overall network architecture, as shown in Figure 2, is composed of an encoder,

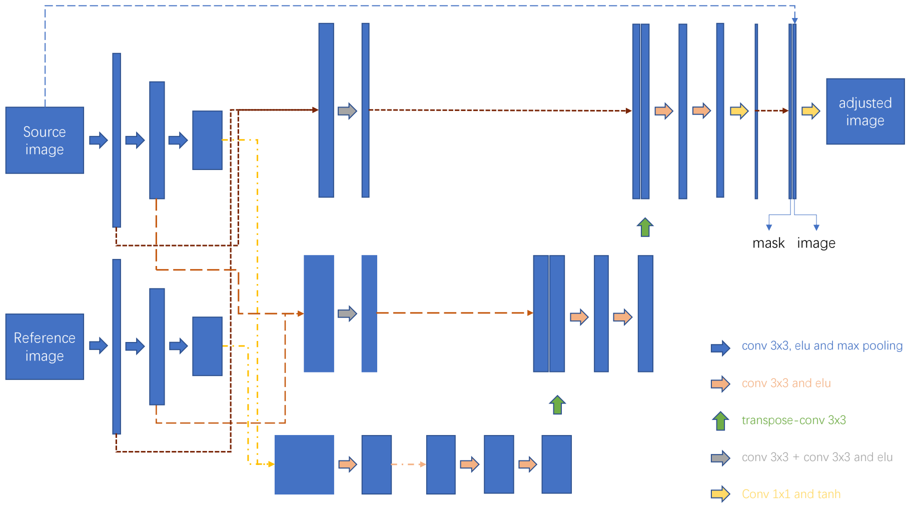

Figure 2: The architecture of the proposed PTNet. Dotted lines denote data transmission without any process. Here, the source image, the reference image, and the adjusted image are normalized to the range of [-1, 1].

correlational layers, and a decoder. In the photometric transfer task, we expect the model to associate features across two frames and regress adjustment values according to the correlation results. Inspired by the optical flow estimation, we redesign the encoder with a Siamese architecture, which is used in FlowNetCorr [16], to extract multilevel features from both the reference frame and the source frame for feature correlation. Note that, the Siamese encoder shares the weights in training and estimation. After feature extraction, we concatenate features from two frames with the same pyramid level and utilize convolutional blocks to find correlations between these features.

In the highest level of feature correlation layer, we only use one convolutional block followed by an exponential linear unit (ELU) activation function. For other two pyramid-level correlation layers, corresponding features from two frames may have a large distance in the image coordinates, which depends on the scale of scene motions. Therefore, we use two convolutional blocks without a nonlinear function between them for two lower pyramid-level correlation layers to extend the receptive field of the convolution for associating features with large distance. After feature correlation, the correlation maps are sent to the decoder to generate an adjustment mask from coarse to fine. The size of higher-level correlation maps is doubled by transposed convolutions. After each expanding, the intermediate maps are concatenated with lower-level correlation maps for fining the eventual adjustment mask. At last, the generated mask is concatenated with the source frame and pixel-wisely adjusts the source frame by a single channel convolution layer followed by the Tanh activation function. To achieve faster convergence for training, we normalize input images from [0,255] to [-1,1], and the generated mask and the outputs are also normalized by the Tanh function.

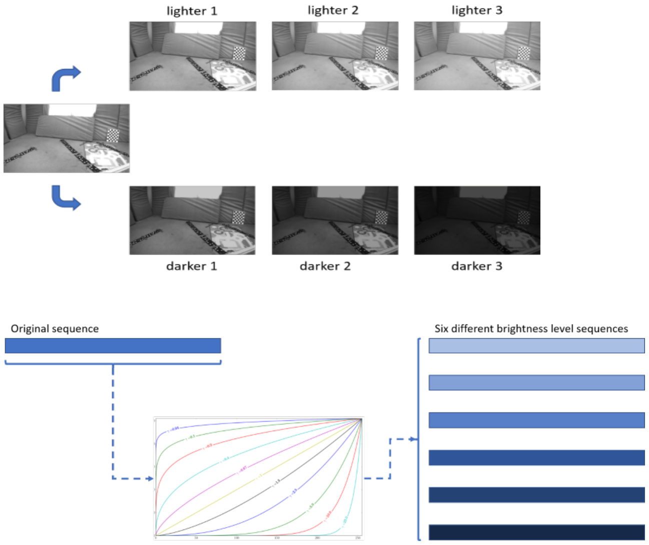

Figure 3: Process of data augmentation. Six more frames with different brightness levels are generated for each frame from the original sequence by a nonlinear transformation. Afterwards, we create six more sequences including three lighter sequences and three darker sequences.

In the encoder and decoder, the sizes of convolutional filter are all 3 x 3 and the strides are 1. Filter channels increase from low level features to high level features with 64, 128, 256 in the encoder and conversely decrease from high level to low level with 256, 128, 64 in the decoder. Each convolutional layer in the encoder follows by a max pooling with the stride of 2 to resize the feature size to be half. The size then is recovered in the decoder by transposed convolutions for two times with the filter channel of 128 and 64 from high level to low level. In feature correlation stage, a 3 x 3 convolutional layers with the stride of 1 is used for the highest level features and two 3 x 3 convolutional layers with the stride of 1, without a nonlinear function between them, are used in other two levels. Filter channel for each level feature association is the same as the encoder. All convolutional layers use the ELU nonlinear function except two convolutional layers in the correlation layers. After decoder, an expected adjustment mask is generated by a 1 x 1 convolutional layer with stride of 1 followed by a Tanh nonlinear function. Finally, the mask is concatenated with the source image and goes through a 1 x 1 convolutional layer with a Tanh nonlinear function to produce the adjusted source image.

### *3.2.2.* ***Data Augmentation and Data Preparation***

As the goal of the proposed scheme is to remove brightness discrepancy between image pairs for pose estimation in DVO, the training data should contain lots of image pairs with different brightness levels and scene movements. However as far as we know, no existing dataset is available for brightness adjustment. To overcome the difficulty, we choose one sequence from the EuRoC dataset [19], which is conventionally used to evaluate performance of VO, as the base data and augment it to create more data with various brightness levels. Specifically speaking, we manually generate six more frames with different brightness level for each original frame from the base data by the following nonlinear transformation:

$$I' = (\frac{I}{255})^{\gamma} V_{max}, \tag{7}$$

where $I$ is the original frame; $\gamma$ is used to control the trend of brightness adjustment, and $V_{max}$ is used to determine the upper bound of pixel values. By regulating $\gamma$ and $V_{max}$, we create three lighter sequences and three darker sequences as shown in Figure 3.

After data augmentation, we organize data from these sequences to compose training data including input image pairs and ground truth. To be specific, we first choose one frame in any sequence $S_k$ as a reference frame $I_k^t$, then source frame $I_l^{t+i}$ can be chosen from any sequence $S_l$, which is generated from identical image context with different brightness levels, with a sequence index difference $i$ to the reference frame. After picking the reference frame and the source frame, ground truth $I_k^{t+i}$ is automatically picked from the same sequence as the reference frame with the same sequence index as the source frame. This process is demonstrated in Figure 4.

To ensure the generalization of the trained model, we pick reference frames from all seven sequences. For each picked reference frame, we choose different time intervals to pick source frames from all seven sequences. By doing this, we achieve huge number of training data with various brightness discrepancies and scene motions. Moreover, ground truth is surely exact regardless of different discrepancies. Note we only augment first 40% frames from one sequence of the base dataset and train the model by these augmented data. In this training process, an assumption is that no brightness discrepancy exists across the base data. Unfortunately, brightness still changes a few times in the base data. To avoid picking a reference frame and the corresponding ground truth frame at the moment with changing brightness, as measured by Equation (6), we skip this reference frame if $E_{brightness}$ exceeds a threshold $T_b^*$.

### *3.2.3.* ***Loss Function***

To make the PTNet learn to correct image brightness pixel-wisely and retain the texture of scenes, we use intensity loss and gradient loss as follows:

$$L_{int} = \sum\nolimits_p \left\| I_{gt}[\mathbf{p}] - \widehat{I[\mathbf{p}]} \right\|_2, \tag{8}$$

$$L_{grad} = \sum\nolimits_p (\left\| \partial_x I_{gt}[\mathbf{p}] - \partial_x \widehat{I[\mathbf{p}]} \right\|_2 + \left\| \partial_y I_{gt}[\mathbf{p}] - \partial_y \widehat{I[\mathbf{p}]} \right\|_2), \tag{9}$$

where $I_{gt}[\mathbf{p}]$ is the pixel value at point $\mathbf{p}$ in the ground truth; $\widehat{I[\mathbf{p}]}$ is the output of the model; $\partial_x I(\cdot)$ and $\partial_y I(\cdot)$ are gradients along the horizontal direction and the vertical direction, respectively.

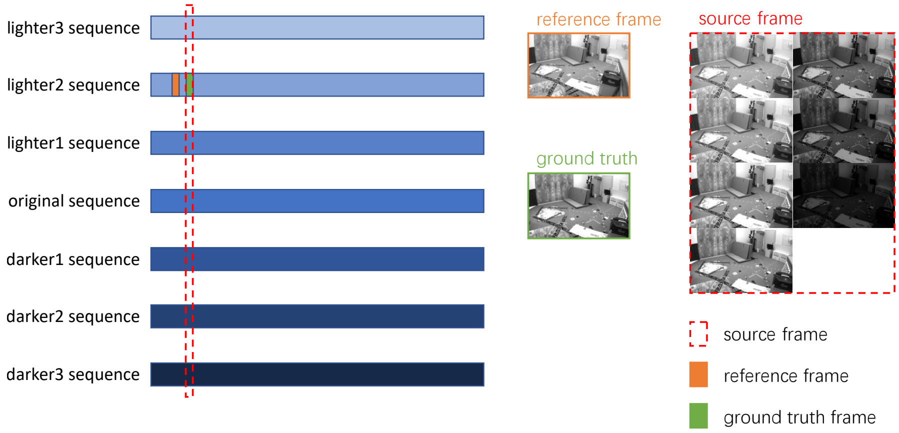

Figure 4: Example of data collection for model training. The reference frame and ground truth are chosen from the same sequence with a short time interval. For each reference frame, seven source frames are chosen from all seven sequences with the same index as the ground truth to generate seven image pairs. Note, the reference frame is chosen traversing all sequence. Before picking the reference frame, we validate that it is brightness consistent with respect to the corresponding ground truth.

To keep adjustment results smooth in poor texture regions, such as, white walls, we further utilize the smoothness loss to penalize gradients in the outputs where gradients in the source frames are small:

$$L_{smooth} = \sum\nolimits_{p} (\left\| \partial_x \widehat{I[\mathbf{p}]} \right\| e^{-\omega \partial_x I_{src}[\mathbf{p}]} + \left\| \partial_y \widehat{I[\mathbf{p}]} \right\| e^{-\omega \partial_y I_{src}[\mathbf{p}]}). \tag{10}$$

In the smoothness loss, $\omega \in \mathbb{R}^+$ is used to control the weightings for different gradient levels.

The ultimate loss function is a weighted sum of the above three loss items:

$$L_{whole} = w_i L_{int} + w_g L_{grad} + w_s L_{smooth}. \tag{11}$$

As can be seen, multiple loss items are combined to remove brightness discrepancies. First, we pixel-wisely eliminate the brightness discrepancies by the intensity loss $L_{int}$ to adjust each pixel independently and enforce the intensity distribution of the source frame to align with the distribution of the reference frame. However, independently adjusting each pixel value may ruin the structural information of the source frame when the intensity fitting is not perfect.

Therefore, the gradient loss $L_{grad}$ is used to make the model encode the relations among neighboring pixels for retaining the context of the source image. In poorly textured image regions, convolutions cannot extract valid information, which easily cause the adjustment to be irregular. To address the problem, we add the additional loss $L_{smooth}$ to make the outputs of the model to be smooth in weak-texture image regions. Different from the intensity loss and the gradient loss, the smoothness loss penalizes incorrectly adjusted regions and effectively improves the adjustment results for texture-less regions of the original source image.

## 4. Experiments

We evaluate the proposed PTNet on two challenging datasets, e.g., the EuRoC dataset [19] and the modified ICL-NUIM dataset [40]. These two datasets are designed for evaluating VO systems which contain several sequences with various brightness discrepancies. We first evaluate the performance of brightness discrepancy elimination of the model by calculating cumulative projection residuals. Then, we evaluate the performance improvement of the DVO system embedded with the proposed model by the root mean square error (RMSE) and the absolute pose error (APE) of the estimated camera poses.

### *4.1. Dataset*

The EuRoC dataset [19] presents visual-inertial data collected by a micro aerial vehicle (MAV). It contains stereo images, synchronized IMU measurements, and accurate motion ground-truth captured by a motion capture system. It includes two scenes and for each scene it has three sequences with different difficulties such as photometric variation, motion blur, etc. In our experiments, we use first 40% frames in sequence V1_01_easy to augment data and train the proposed model by these augmented data. Then we test the model on three sequences including the whole V1_01_easy sequence and two other harder sequences.

The Modified ICL-NUIM dataset [40] is a synthetic dataset, which is generated from original ICL-NUIM synthetic dataset [50]. It provides RGBD data with ground truth of camera trajectory for each frame. Comparing with original ICL dataset, it generates four more datasets with simulation of flashlight attached to the camera, local lighting variation, global lighting variation, global and local lighting variation respectively. We use the last two sequences to verify the availability of proposed model. Note we test the model on the modified ICL-NUIM dataset with fixed model parameters trained from the EuRoC dataset [19].

### *4.2.* ***Training Details***

For data augmentation, six sets of parameters $\gamma$ and $V_{max}$ for generating different brightness level sequences are 2.0, 70; 1.5, 150; 1.2, 200; 0.4, 255; 0.78, 255; 0.9, 255 respectively. During training, we select one reference frame every ten frames in turn from seven sequences and select the second frame and the fourth frame after the reference frame from all seven sequences as source frames for pairing with the reference frame. When each reference frame is selected, we check the brightness difference between the reference frame and the ground truth, and set threshold $T_b^*$ to be 6 to avoid possible photometric inconsistency between the training data and the ground truth. The batch size of training data is set to 12. The weightings of three loss items $w_i, w_g, w_s$ are set to be 10, 100, 10, respectively. For all input images, we normalize them from [0, 255] to [-1, 1] as follows

$$I_{normal} = \frac{I_{original}}{127.5} - 1, \tag{12}$$

and resize them to 376 x 240. Note, the size of outputs from the PTNet is also 376 x 240.

### *4.3.* ***Evaluation of Photometric Transfer***

To test PTNet on adjusting brightness of one frame according to a reference frame, we calculate photometric projection residuals by projecting points from reference frames to adjusted source frames according to Equation (1). We

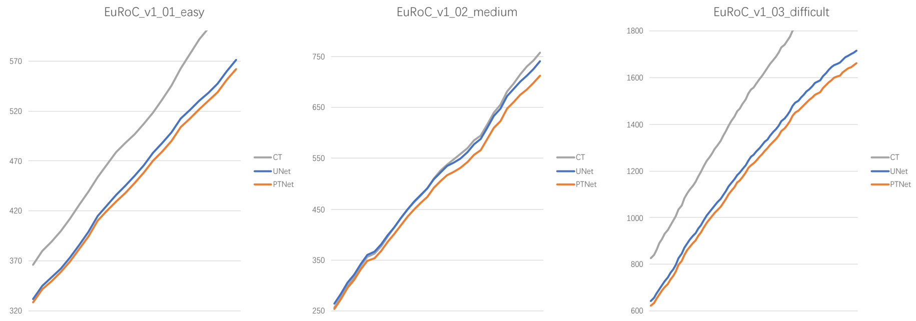

Figure 5: Cumulative projection residuals on three sequences from the EuRoC dataset [19]. We accumulate projection residuals between reference frames and the adjusted source frames when brightness discrepancies appear. Reference frames and projected points are selected by DSO [1]. CT and UNet denote Color Transfer [8] and U-Net [11], respectively. On all three sequences, our proposed PTNet achieves the best accuracy.

first compare the proposed model with two other methods, i.e. Color Transfer method [8] and the CNN model which we use as the backbone [11]. It should be noted that, instead of converting color images from RGB space to a decorrelated color space as suggested in the Color Transfer method, we use gray level images in our experiments. To the best of our knowledge, there is no publically available pretrained CNN model for the photometric transfer task, so we train the U-Net [11], which is designed for image segmentation, from scratch with the same training data and loss functions for PTNet. We show the results on three sequences from EuRoC dataset [19] in Figure 5. Since the dataset does not provide ground truth pose for each frame, we update reference frames by the latest key frame picked in DSO [1] and source frames are picked from the latest tracked frames when the difference of brightness estimated by Equation (6) exceeds the threshold $T_b = 15$. Required relative poses between reference frames and source frames and points for projection are from the estimations of DSO. Note, we ignore the projection results when the pose estimation is inaccurate. As shown in Figure 5, CNN-based methods achieve more accurate photometric transfer results and our proposed PTNet obtains the most accurate results on all three sequences. To intuitively explain the adjustment effects of

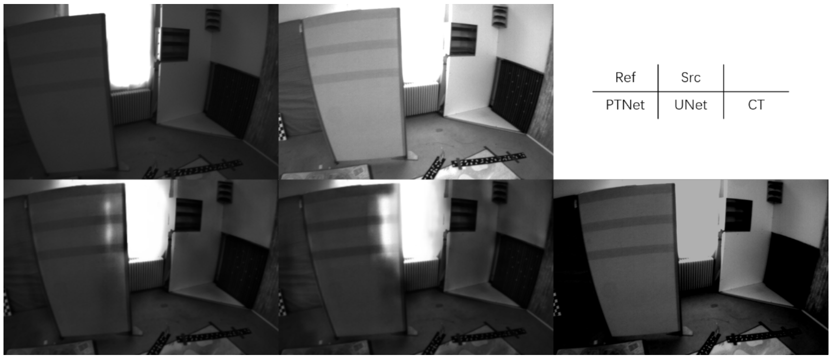

Figure 6: Comparison of adjustment results from three methods. This figure shows a challenging scenario including complicated brightness changes. Two images in the first row are the reference frame and the source frame. Three images in the second row are adjusted images by the proposed PTNet, U-Net [11], Color Transfer [8], respectively. Our model achieves visually favorable result compared with others.

three methods, we present a set of adjusted results of a challenging scenario containing complicated brightness changes in Figure 6. As we can see, brightness discrepancy has uneven distribution between the reference frame and the source frame. The Color Transfer method [8] equally adjusts the whole image and the result has higher brightness for most image regions and lower brightness for the window with overexposure compared with the reference frame. U-Net pixel-wisely adjusts the image and keeps the brightness consistent for the overexposure region. However, the result is rough in the white wall and has a large improperly adjusted regions near the overexposure region. Our proposed model achieves visually favorable adjusted results with smoother adjustment for those low-gradient regions. More adjustment results by the proposed model are shown in Figure 7.

We further discuss different adjustment results from the traditional method and the CNN-based method. The Color Transfer method [8] counts mean and variance of pixel values for both frames and then adjusts the source frame to make it achieve similar intensity distribution with respect to the reference frame by a statistical method. As a result, adjustment for extremely bright or dark

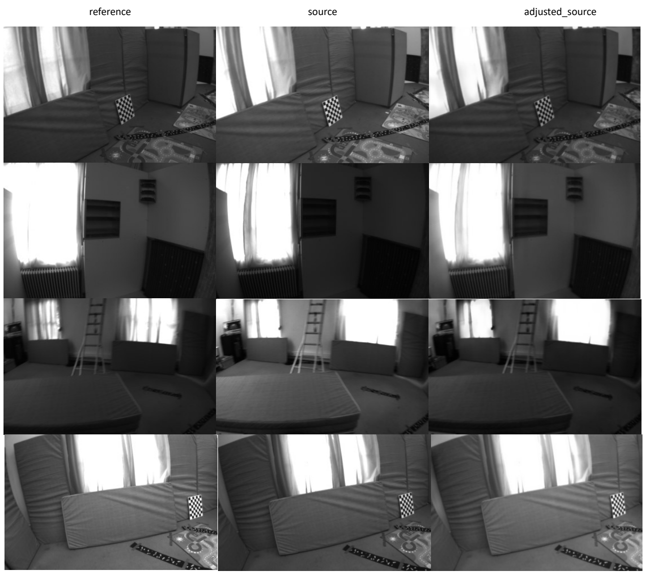

Figure 7: Examples of brightness adjustment results by the proposed PTNet on the EuRoC dataset [19]

regions in the frame affects other pixels in the frame and further causes inexact adjustment for each pixel. Our proposed model extracts multilevel features from both frames and associates them to find associated regions for regional adjustments. Moreover, low-level feature association results are facilitated with high-level feature correlations and finally achieves fine region correlations for the two frames. Therefore, our proposed model can adjust the source frame pixel-wisely and generate finer results.

To further demonstrate the effect of feature association, we manually select two sets of image pairs for adjustment by the PTNet. The adjustment results are shown in Figure 8. For the first image pair, we choose two frames with a large scene motion, containing translation and scale change. We can see that some objects disappear from the reference frame to the source frame and it causes failed feature association in the PTNet. As a result, adjustments for these regions are incorrect with black shadows. In the second image pair, there exists only a small translation. We can see that feature association works well

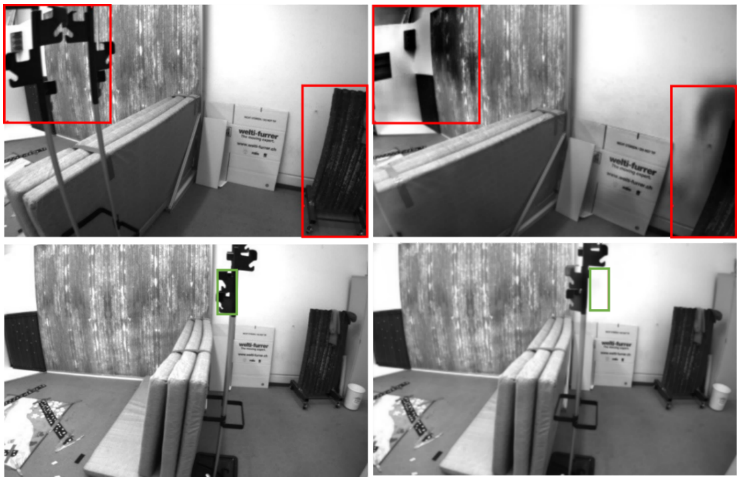

Figure 8: Examples of adjustment results by our PTNet for two image pairs with different scales of scene motion. The left side is the reference frames and the right side is the corresponding adjusted source frames. In the upper image pair with large scene motion, objects in the red box disappear from the reference frame to the source frame, which consequently cause improper adjustment. In the next image pairs with small scene motion, the object in the green box is successfully associated and incorrect adjustment is avoided.

for those small motions and the model achieves exact adjustment result. Note in VO scenario, the system always keeps a small scene motion between the latest reference frame and the new frame by updating the reference frame with a newer frame when the system detect a change of scenes which exceeds a strict threshold. Therefore, the proposed model can work well for VO systems.

In the end, we quantitatively analyze the brightness adjustment performance of the PTNet on both two datasets, i.e. the EuRoC dataset [19] and the modified ICL-NUIM dataset [40]. We compare it with several brightness adjustment methods, including Color Transfer [8], histogram matching, Exact Histogram Specificaton [43], and optimization-based affine correction method [1]. For both datasets, we manually generate additional data with uneven brightness changes. Specifically, we generate one more sequence, in which each image is generated by modifying an original image from left to right by nonlinear transformation with different parameters, from a selected original sequence to simulate the case of

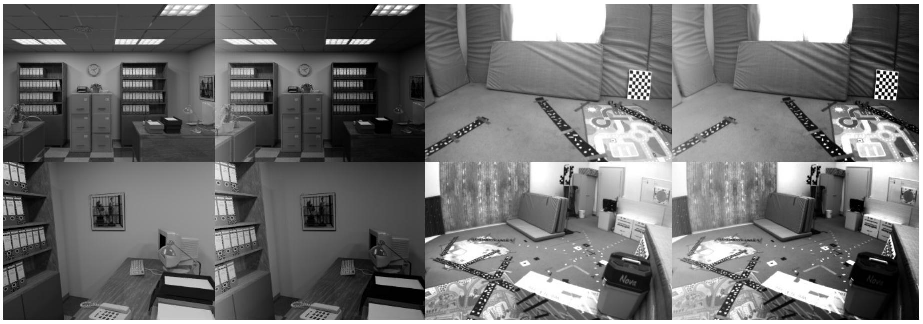

Figure 9: Example of generated images with uneven brightness changes. Left two columns are original images and generated images from the modified ICL-NUIM dataset, respectively. Accordingly, right two columns come from the EuRoC dataset. To generate uneven changes of brightness for the original data, we modify the original image from left to right by the nonlinear transformation with different parameters.

uneven brightness changes. The example of generated images from both dataset are shown in Figure 9. We evaluate the brightness adjustment accuracy by calculating intensity difference of 3 x 3 image patches and gradient difference between the reference image and the adjusted source image. The gradient is calculated by the Soble operator. On the modified ICL-NUIM dataset, we use geometric projection with provided depth maps and camera poses to project dense points with a fixed distance from the reference frame to the adjusted source frame to obtain dense point correspondences. On the EuRoC dataset, however, the depth maps and the frame-wise camera poses are not provided. Therefore, we utilize SIFT features [51], which have a certain degree of robustness of brightness change [40], with RANSAC [52] to find exact sparse corresponding points between two frames. Note, we do not evaluate the optimization-based affine correction method on the EuRoC dataset since the camera poses and point depths are unavailable. For testing on the original dataset, we traverse the sequences to find image pairs with a fixed time interval and brightness changes, which is checked by comparing the mean pixel value between the image pair. For testing on the generated data with uneven brightness changes, we traverse the original sequence with a fixed step, and pick one image from the original sequence and another image from the generated sequence with a fixed time interval. We show the results in Table 1 and Table 2.

Table 1: Mean of brightness adjustment errors on the modified ICL-NUIM dataset.

| | ethl1_global | | ethl1_global_local | | ethl1_uneven | |
|---|---|---|---|---|---|---|
| | **3x3 patch residual** | **gradient residual** | **3x3 patch residual** | **gradient residual** | **3x3 patch residual** | **gradient residual** |
| **without adjustment** | 13.4881 | 8.6711 | 13.4245 | 9.6705 | 32.1556 | **9.1353** |
| **Color Transfer[8]** | 6.2132 | 8.8163 | 8.1334 | 9.6876 | 24.5386 | 9.8273 |
| **Histogram Matching** | 5.9811 | 8.9658 | 7.8707 | 10.2944 | 21.5939 | 10.5902 |
| **EHS[43]** | **5.9530** | 8.8817 | 7.6813 | 10.1881 | 21.6916 | 10.6658 |
| **Affine[1]** | 6.2299 | **8.6016** | 8.2827 | **9.4119** | 22.3783 | 11.4134 |
| **PTNet** | 6.0637 | 9.1059 | **7.4268** | 10.4624 | **3.9427** | 11.1243 |
| | **ethl2_global** | | **ethl2_global_local** | | **ethl2_uneven** | |
| | **3x3 patch residual** | **gradient residual** | **3x3 patch residual** | **gradient residual** | **3x3 patch residual** | **gradient residual** |
| **without adjustment** | 11.6195 | 9.9472 | 11.8343 | 13.2048 | 33.1235 | **7.6229** |
| **Color Transfer[8]** | 3.8582 | 9.9122 | 7.3989 | 13.6051 | 25.6495 | 8.0411 |
| **Histogram Matching** | 3.8415 | 9.9570 | 7.5305 | 13.8738 | 21.5147 | 10.2097 |
| **EHS[43]** | **3.7934** | 9.8338 | 7.4164 | 13.7751 | 21.6255 | 10.2146 |
| **Affine[1]** | 4.2656 | **9.7698** | 8.5987 | **13.1439** | 23.2787 | 9.4632 |
| **PTNet** | 4.0719 | 10.0476 | **7.1435** | 13.7086 | **3.7623** | 9.5161 |

Table 2: Mean of brightness adjustment errors on the EuRoC dataset.

| | euroc_medium | | euroc_medium_uneven | | euroc_difficult_uneven | |
|---|---|---|---|---|---|---|
| | **3x3 patch residual** | **gradient residual** | **3x3 patch residual** | **gradient residual** | **3x3 patch residual** | **gradient residual** |
| **without adjustment** | 20.7477 | 16.0287 | 25.2880 | **15.5908** | 23.2062 | 15.1582 |
| **Color Transfer[8]** | 11.2896 | **15.7245** | 22.4723 | 16.0436 | 22.8376 | **14.9469** |
| **Histogram Matching** | **9.7349** | 17.1279 | 22.6312 | 21.7601 | 22.1101 | 20.1173 |
| **EHS[43]** | 9.7643 | 17.1281 | 22.6583 | 21.9056 | 22.2939 | 20.2907 |
| **PTNet** | 13.6937 | 19.8003 | **11.0323** | 20.1858 | **12.6388** | 17.6157 |

As shown in the table, the PTNet achieves similar birghtness adjustment performance compared with other traditional methods on the data with globally consistent brightness changes. However, these traditional methods all suffer from uneven brightness changes. On the contrary, the performance of our proposed PTNet remains intact in case of the uneven brightness changes. Note, during the model training process, the model has not seen the generated data with uneven brigthness changes. Benefiting from the pixel-wisely adjustment ability, the PTNet keeps stable performance on both global brightness changes and uneven brighness changes. However, the pixel-wisely brightness adjustment makes the PTNet lose a little accuracy on image gradient compared with other traditional methods.

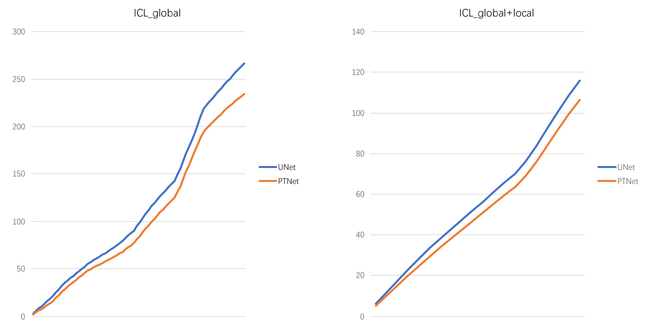

Figure 10: Cumulative projection residuals on two sequences from the modified ICL-NUIM dataset [40]. We accumulate projection residuals between reference frames and adjusted source frames when brightness discrepancy appears. We adjust source frames by U-Net [11] and our PTNet individually. Reference frames are selected traversing the sequence and source frames are selected after the reference frames with a fixed time interval. The proposed model achieves more accurate results on eliminating brightness discrepancies.

### *4.4. Generalization Ability Analysis*

From Table 1, we can see that the PTNet also work well on the synthetic sequences from the modified ICL-NUIM dataset. Note, we use the same parameters for the PTNet which is trained on EuRoC dataset [19]. Further, we compare the PTNet with the U-Net[11] on the synthetic sequences by traversing the sequence to select reference frames and select source frames after the reference frames with a fixed interval. As shown in Figure 10, PTNet also achieves better results compared with U-Net [11]. Some of the adjustment results by PTNet are shown in Figure 11.

Even testing on the unseen dataset, the proposed model still achieves favorable adjustment results. Benefited from our data augmentation, the training data contain various brightness discrepancies and scene motions which ensure the data diversity from limited base data. Abundant diversities in the training data enable the proposed model to be adaptive to all kinds of transfer functions and further allow the model to find correct adjustment value for each pixel. A CNN is good at extracting features from low-level features with simple structures to high-level features containing semantic information. These features are

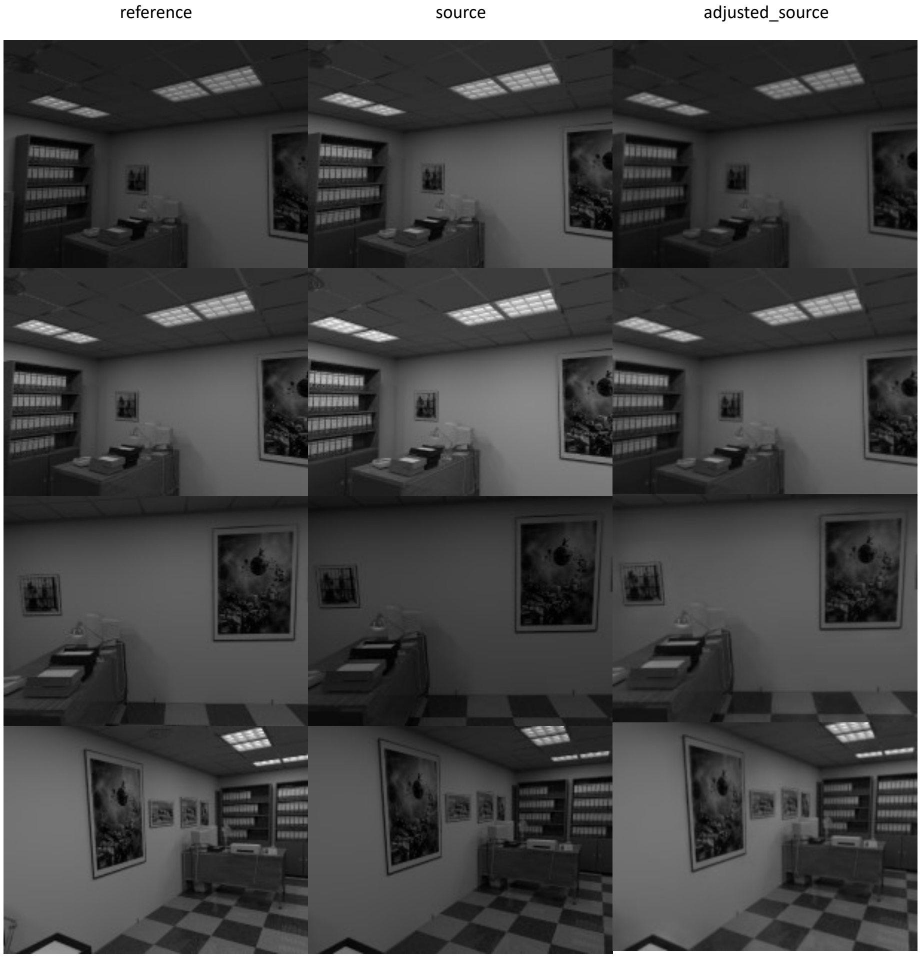

Figure 11: Examples of brightness adjustment results by our PTNet on the modified ICL-NUIM dataset [40]. In each row from left to right are the reference frame, the source frame, and the adjusted source frame. Note that the model is trained on the EuRoC dataset [19]. Even the model is trained by real world data, it achieves good adjustment resutls on the unseen synthetic dataset.

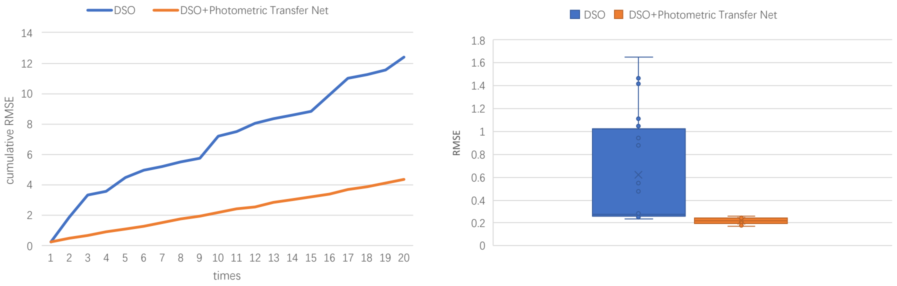

Figure 12: RMSE of estimation results of DSO [1] running on the challenging sequence V1_02_medium from EuRoC dataset [19]. We iteratively run DSO on the sequence by 20 times. We can see that with the help of proposed model, DSO achieves more accurate and robust estimation on the sequence with a lot of brightness discrepancies.

learned from the training data and can be reused for other unseen data with similar styles. Scenes in both Euroc dataset [19] and the modified ICL-NUIM dataset [40] contain abundant indoor structure features which are beneficial to feature associations and enable the model to work well on the unseen data.

### *4.5.* ***Evaluation on DVO System***

To evaluate the performance advance for DVO with data preprocessing by the PTNet, we embed the proposed model into DSO [1] and run it on the challenging sequences with brightness discrepancies from the EuRoC dataset [19]. To obtain statistically meaningful results, we run it twenty times and show the results in Figure 12. We can see that DSO performs much better with data preprocessing by PTNet and achieves more stable estimation results. In DSO, a set of affine parameters are used to globally adjust brightness of frames. These parameters are optimized with pose estimation. Results of the affine method depend on the optimization process and the selected points which should have the same brightness change. It's the reason that original DSO achieves quite different RMSE for each running results on the same sequence. While with the help of PTNet, input data is preprocessed for photometric consistency and it reduces the influence of point selection on the estimation performance

.Therefore, running DSO with the data preprocessed by the proposed model for many times achieves more stable RMSE.

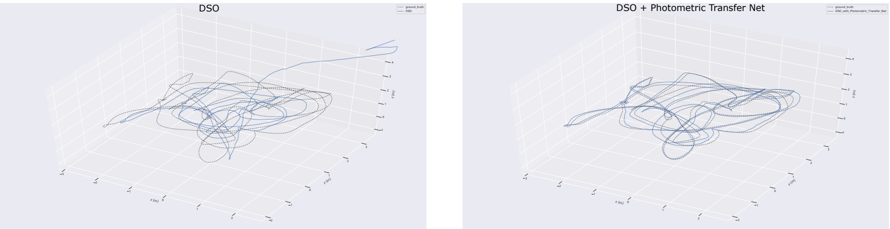

Figure 13: Estimated trajectories by DSO [1] running on the challenging sequence V1_02_medium from the EuRoC [19]. Dotted lines denote ground truth and blue lines denote the estimated trajectory. We can see that the initialization of DSO suffers from brightness discrepancy, and our PTNet improves initialization by eliminating brightness changes. Moreover, DSO achieves more accurate estimation with the help of the proposed model.

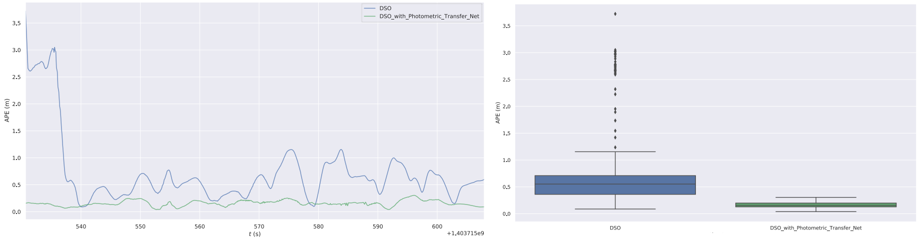

Figure 14: Absolute pose error (APE) of pose estimation by DSO [1] on the challenging sequence V1_02_medium from EuRoC [19]. We show further analysis of pose estimation to prove that the proposed model efficiently improves the estimation accuracy for DSO against brightness discrepancies.

We further choose one running result from both without-photometric-transfer experience and with-photometric-transfer experience respectively to analyze the pose estimation. Trajectories and quantitative comparison are shown in Figure 13 and Figure 14. We can see that brightness discrepancies bring much trouble to the initialization of DSO [1], which is a vital stage for any monocular VO, and the adverse influence is eliminated with the help of PTNet. In addition, input data with more consistent brightness makes the pose estimation of DSO more robust.

## 5. Conclusion

With the advantage of deep learning, we can deal with brightness discrepancy more efficiently for more robust and accurate pose estimation of DVO. In this paper, we present a CNN model which is capable to adjust the brightness of the source frame according to the reference frame in dynamic practical scenes. The adjusted source frame is pixel-wisely photometric consistent with respect to the reference frame, which satisfies the photometric consistency assumption in DVO. To train the model, we create a huge number of image pairs with different brightness levels from the EuRoC dataset by a nonlinear transformation. Experiments on real-world and synthetic dataset show that proposed approach can effectively eliminate brightness changes across frames and allow DVO to achieve better initialization and more robust pose estimation. In addition, assessment on the unseen dataset with the same model parameters trained on another dataset proves the generalization ability of the model.

## 6. Acknowledgments

The work is supported by the following projects: National Natural Science Foundation of China (NSFC) Essential project, Nr.: 61831018, U2033218; NSFC, Nr.: 61772328, 61702322.

## References

[1] J. Engel, V. Koltun, D. Cremers, Direct sparse odometry, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Vol: 40 (3) (2017) pp: 611–625.

[2] R. Mur-Artal, J. D. Tardós, Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras, IEEE Transactions on Robotics Vol: 33 (5) (2017) pp: 1255–1262.

[3] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, F. Moreno-Noguer, Pl-slam: Real-time monocular visual slam with points and lines, in: IEEE International Conference on Robotics and Automation (ICRA), 2017.

[4] T. Qin, P. Li, S. Shen, Vins-mono: A robust and versatile monocular visual-inertial state estimator, IEEE Transactions on Robotics Vol: 34 (4) (2018) pp: 1004–1020.

[5] G. G. Scandaroli, M. Meilland, R. Richa, Improving ncc-based direct visual tracking, in: European Conference on Computer Vision (ECCV), 2012.

[6] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, D. Scaramuzza, Svo: Semidirect visual odometry for monocular and multicamera systems, IEEE Transactions on Robotics Vol: 33 (2) (2016) pp: 249–265.

[7] J. Zhang, Q. Su, P. Liu, C. Xu, Z. Wang, Unlearnermc: Unsupervised learning of dense depth and camera pose using mask and cooperative loss, Knowledge-Based Systems Vol: 192 (2020) pp: 105357.

[8] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Computer Graphics and Applications Vol: 21 (5) (2001) pp: 34–41.

[9] O. Demetz, M. Stoll, S. Volz, J. Weickert, A. Bruhn, Learning brightness transfer functions for the joint recovery of illumination changes and optical flow, in: European Conference on Computer Vision (ECCV), 2014.

[10] T. D'Orazio, P. L. Mazzeo, P. Spagnolo, Color brightness transfer function evaluation for non overlapping multi camera tracking, in: ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2009.

[11] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, 2015.

[12] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[13] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), 2017.

[14] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., Conditional image generation with pixelcnn decoders, in: Advances in Neural Information Processing Systems, 2016.

[15] J. Li, F. Fang, K. Mei, G. Zhang, Multi-scale residual network for image super-resolution, in: European Conference on Computer Vision (ECCV), 2018.

[16] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: IEEE International Conference on Computer Vision (ICCV), 2015.

[17] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[18] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[19] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, R. Siegwart, The euroc micro aerial vehicle datasets, The International Journal of Robotics Research (IJRR) Vol: 35 (10) (2016) pp: 1157–1163.

[20] G. Klein, D. Murray, Parallel tracking and mapping for small ar workspaces, in: IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007.

[21] J. Engel, T. Schöps, D. Cremers, Lsd-slam: Large-scale direct monocular slam, in: European Conference on Computer Vision (ECCV), 2014.

[22] T. Mitsunaga, S. K. Nayar, Radiometric self calibration, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1999.

[23] J. Engel, V. Usenko, D. Cremers, A photometrically calibrated benchmark for monocular visual odometry, arXiv preprint arXiv:1607.02555.

[24] S. Lin, L. Zhang, Determining the radiometric response function from a single grayscale image, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[25] S. Lin, J. Gu, S. Yamazaki, H.-Y. Shum, Radiometric calibration from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004.

[26] Y. Zheng, S. Lin, C. Kambhamettu, J. Yu, S. B. Kang, Single-image vignetting correction, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Vol: 31 (12) (2008) pp: 2243–2256.

[27] S. J. Kim, M. Pollefeys, Robust radiometric calibration and vignetting correction, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Vol: 30 (4) (2008) pp: 562–576.

[28] P. Bergmann, R. Wang, D. Cremers, Online photometric calibration of auto exposure video for realtime visual odometry and slam, IEEE Robotics and Automation Letters Vol: 3 (2) (2017) pp: 627–634.

[29] D. B. Goldman, Vignette and exposure calibration and compensation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Vol: 32 (12) (2010) pp: 2276–2288.

[30] T. Gonçalves, A. I. Comport, Real-time direct tracking of color images in the presence of illumination variation, in: IEEE International Conference on Robotics and Automation (ICRA), 2011.

[31] A. Dame, E. Marchand, Second-order optimization of mutual information for real-time image registration, IEEE Transactions on Image Processing (TIP) Vol: 21 (9) (2012) pp: 4190–4203.

[32] M. Fraissinet-Tachet, M. Schmitt, Z. Wen, A. Kuijper, Multi-camera piecewise planar object tracking with mutual information, Journal of Mathematical Imaging and Vision Vol: 56 (3) (2016) pp: 591–602.

[33] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, C. Theobalt, Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration, ACM Transactions on Graphics (ToG) Vo: 36 (4) (2017) pp: 1.

[34] T. Hyun Kim, H. Seok Lee, K. Mu Lee, Optical flow via locally adaptive fusion of complementary data costs, in: IEEE International Conference on Computer Vision (ICCV), 2013.

[35] L. Xu, J. Jia, Y. Matsushita, Motion detail preserving optical flow estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Vol: 34 (9) (2011) pp: 1744–1757.

[36] N. Papenberg, A. Bruhn, T. Brox, S. Didas, J. Weickert, Highly accurate optic flow computation with theoretically justified warping, International Journal of Computer Vision (IJCV) Vol: 67 (2) (2006) pp: 141–158.

[37] M. Yokozuka, S. Oishi, S. Thompson, A. Banno, Vitamin-e: visual tracking and mapping with extremely dense feature points, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[38] C. Liu, J. Yuen, A. Torralba, Sift flow: Dense correspondence across scenes and its applications, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Vol: 33 (5) (2010) pp: 978–994.

[39] H. A. Rashwan, M. A. Mohamed, M. A. García, B. Mertsching, D. Puig, Illumination robust optical flow model based on histogram of oriented gradients, in: German Conference on Pattern Recognition (GCPR), 2013.

[40] S. Park, T. Schöps, M. Pollefeys, Illumination change robustness in direct visual slam, in: IEEE International Conference on Robotics and Automation (ICRA), 2017.

[41] C. Abbey, Fast algorithms for histogram matching: Application to texture synthesis, Journal of Electronic Imaging Vol: 9 (1) (2000) pp: 39–45.

[42] Y. Zhang, Improving the accuracy of direct histogram specification, Electronics Letters Vol: 28 (3) (1992) pp: 213–214.

[43] D. Coltuc, P. Bolon, J.-M. Chassery, Exact histogram specification, IEEE Transactions on Image Processing (TIP) Vol: 15 (5) (2006) pp: 1143–1152.

[44] D. Dederscheck, T. Müller, R. Mester, Illumination invariance for driving scene optical flow using comparagram preselection, in: IEEE Intelligent Vehicles Symposium, 2012.

[45] H. W. Haussecker, D. J. Fleet, Computing optical flow with physical models of brightness variation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) Vol: 23 (6) (2001) pp: 661–673.

[46] D. Sun, S. Roth, M. J. Black, A quantitative analysis of current practices in optical flow estimation and the principles behind them, International Journal of Computer Vision (IJCV) Vol: 106 (2) (2014) pp: 115–137.

[47] J. Engel, J. Stückler, D. Cremers, Large-scale direct slam with stereo cameras, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.

[48] N. Yang, L. v. Stumberg, R. Wang, D. Cremers, D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[49] L. von Stumberg, P. Wenzel, Q. Khan, D. Cremers, Gn-net: The gauss-newton loss for multi-weather relocalization, IEEE Robotics and Automation Letters Vol: 5 (2) (2020) pp: 890–897.

[50] A. Handa, T. Whelan, J. McDonald, A. J. Davison, A benchmark for rgb-d visual odometry, 3d reconstruction and slam, in: IEEE International Conference on Robotics and Automation (ICRA), 2014.

[51] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision (IJCV) Vol: 60 (2004) pp: 91–110.

[52] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM Vol: 24 (1981) pp: 381–395.