Unsupervised Deep Asymmetric Stereo Matching with Spatially-Adaptive Self-Similarity

Taeyong Song[1] Sunok Kim[2] Kwanghoon Sohn[3,4]
[1]Hyundai Motor Company R&D Division, [2]Korea Aerospace University,
[3]Yonsei University, [4]Korea Institute of Science and Technology (KIST)
taeyongsong@hyundai.com, sunok.kim@kau.ac.kr, khsohn@yonsei.ac.kr

Abstract

Unsupervised stereo matching has received a lot of attention since it enables the learning of disparity estimation without ground-truth data. However, most of the unsupervised stereo matching algorithms assume that the left and right images have consistent visual properties, i.e., symmetric, and easily fail when the stereo images are asymmetric. In this paper, we present a novel spatially-adaptive self-similarity (SASS) for unsupervised asymmetric stereo matching. It extends the concept of self-similarity and generates deep features that are robust to the asymmetries. The sampling patterns to calculate self-similarities are adaptively generated throughout the image regions to effectively encode diverse patterns. In order to learn the effective sampling patterns, we design a contrastive similarity loss with positive and negative weights. Consequently, SASS is further encouraged to encode asymmetry-agnostic features, while maintaining the distinctiveness for stereo correspondence. We present extensive experimental results including ablation studies and comparisons with different methods, demonstrating effectiveness of the proposed method under resolution and noise asymmetries.

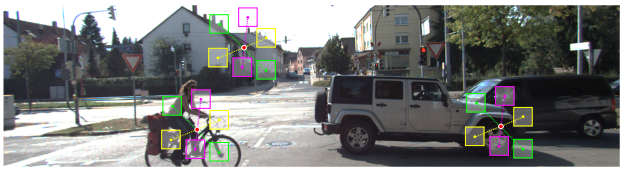
1. Introduction

Scene depth is an indispensable information in computer vision, as it can benefit numerous subsequent applications including scene recognition [5, 18], 3D scene reconstruction [22], and autonomous driving [17]. Stereo matching, which aims to find *disparities* of corresponding points in rectified left and right (stereo) images, has been widely explored since the disparity can directly converted to depth with camera calibration parameters. Recent advent of large-scale datasets and advanced hardware led the researchers to solve stereo matching with Convolutional Neural Networks (CNNs). It resulted in a number of CNN-based stereo matching algorithms that are learned in both supervised [1, 16, 19] and unsupervised manner [3, 25]. Even though the recent methods have achieved significant gain in both accuracy and speed, the existing algorithms assume that the stereo images are *symmetric*, where the stereo images have consistent visual properties in terms of brightness, resolution, noise level, modality, etc.

Recently, multi-camera systems have become more common, such as RGB-NIR cameras in Kinect, and tele-wide cameras in smartphones. Such systems usually consist of different sensors, resulting in *asymmetric* stereo images, *i.e.*, the stereo images with different visual properties. The asymmetric images are embedded into inconsistent features and make it difficult to accurately calculate the cost volume. Furthermore, the most widely adopted assumption for unsupervised stereo matching, photometric consistency, is invalid for the corresponding points in the asymmetric stereo images [2]. Consequently, the widely-used stereo matching methods assuming symmetric images [1, 3, 19] easily fail in the asymmetric scenario [15].

There have been relatively less efforts to handle stereo matching under asymmetries such as visual quality [2, 15] and spectrum [23, 31]. Several methods adopt supervised [15], or proxy-supervised [23] paradigm to solve the deep asymmetric stereo matching. However, such methods require additional active depth [15] or image [23] sensor to acquire the training label, which makes it difficult to construct the training data. In order to tackle the problem and learn the asymmetric stereo matching in an unsupervised manner, a few methods adopt feature consistency loss [2, 24]. On the other hand, several spectral-asymmetric stereo matching methods use unpaired image-to-image translation [33] algorithm to project the images into a same spectrum, followed by photometric consistency loss [14, 31]. A common approach in the unsupervised asymmetric stereo matching methods is to transfer the images into a shared space to ex-

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF2021R1A2C2006703). The work of S. Kim was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP) (NRF-2021R1C1C2005202). (Corresponding author: Kwanghoon Sohn.)



(a)

(b)

Figure 1. Self-similarity sampling patterns of (a) FCSS [10] and (b) the proposed SASS. For different pixels indicated with red circles, the sampling patterns are represented with squares, connected with dashed lines. FCSS has equivalent patterns for all pixels, while the proposed SASS generates adaptive patterns.

ploit the consistency constraint as training loss. The importance of consistent space in loss calculation for unsupervised stereo matching is further emphasized in [2].

There have been a number of researches to extract image features that are robust to different types of variations. In [21], Local Self-Similarity (LSS) descriptor has been presented based on an observation that local internal layout of self-similarity is less sensitive to photometric differences. It has demonstrated impressive robustness against large modality differences, and various derivations based on self-similarity have been formulated in handcrafted [11, 12] and deep-learning [10] frameworks, demonstrating effectiveness in cross-modal visual [11, 12] and semantic [10] correspondences. In [10, 11], in order to design self-similarity based descriptors with improved robustness and efficiency, sampling patterns are learned throughout the data. However, the learned sampling patterns are fixed for all regions as in Fig. 1(a), limiting the capability to encode robust features of varying geometries across the images.

In this paper, we present a novel Spatially-Adaptive Self-Similarity (SASS) for unsupervised stereo matching in asymmetric scenario. Motivated by the importance of the symmetry in loss calculation [2], we design a novel framework to extract asymmetry-agnostic features. We take advantage of self-similarity [21] which is robustness to domain discrepancy, and further extend it by adaptively generating the sampling patterns across the spatial locations, as illustrated in Fig. 1(b). It enables to extract asymmetry-agnostic features from the asymmetric stereo images to calculate the stereo matching loss in a symmetric space. In addition, we design a contrastive similarity loss with additional positive and negative weights to further encourage the asymmetry-agnostic property of the SASS, while preserving the discriminative capability.

The main contributions of this paper are summarized as:

- We propose a novel Spatially-Adaptive Self-Similarity (SASS) to adaptively encode asymmetry-agnostic features for unsupervised asymmetric stereo matching. The features are used to calculate the unsupervised stereo matching loss based on view consistency.
- We design a contrastive similarity loss with a novel positive and negative weighting strategy to further enhance the asymmetry-agnostic property while maintaining the discriminative capability of SASS.
- Extensive experimental results including ablation studies and comparisons with different methods demonstrate the effectiveness of the proposed method on resolution and noise asymmetries.

The rest of this paper is organized as follows: In Sec. 2, we present previous works that are related to ours. Sec. 3 explains the background and details of the proposed method. Experimental results are given in Sec. 4, followed by conclusion and future works in Sec. 5.

2. Related Works

2.1. Stereo Matching

With the advent of large-scale datasets and development of hardware performance, a number of stereo matching methods based on CNNs have been proposed. As a pioneering work, Zbontar and Lecun [27] proposed to calculate stereo matching cost with CNN. The matching cost is further processed with hand-crafted methods [8, 29] to generate the final disparity map. Mayer and Brox [16] proposed DispNet with 1D correlation layer which enables end-to-end learning of stereo matching as a regression problem. Pang et al. [19] proposed cascade residual learning framework that consists of two-stage disparity estimation, where the second stage adopts the residual learning [7] to refine the disparity map. In [9], GC-Net was proposed to learn the geometry and context using 3D convolution and differentiable soft-argmin operation. Chang et al. [1] proposed PSMNet which contains spatial pyramid pooling (SPP) [6] and 3D convolution to exploit context information and learn to regularize the cost volume. Zhang et al. [28] proposed GA-Net with semi-global and local guided aggregation layers to replace 3D convolutions for cost aggregation.

While the above methods adopt supervised learning paradigm that requires ground-truth for training, various unsupervised algorithms have also been proposed. Zhou et al. [32] proposed an unsupervised learning framework that iteratively conducts disparity prediction, confidence map estimation, training data selection, and network training. Godard at el. [3] proposed to use differentiable bilinear sampling layer to warp the stereo images and define photometric consistency loss. This method is extended in [4] by robust reprojection loss, multi-scale sampling, and an auto-masking loss. Wang et al. [25] proposed parallex attention

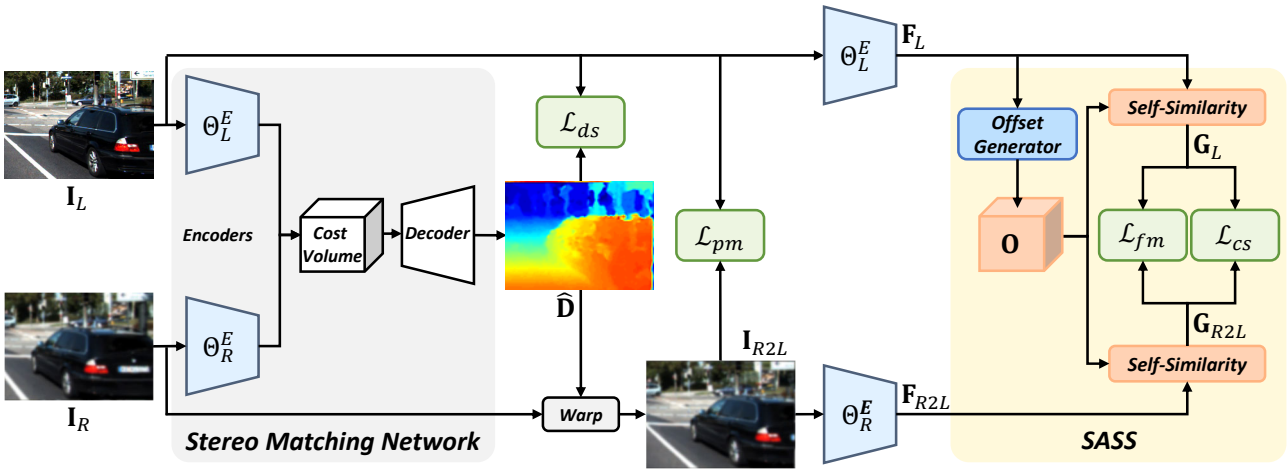


Figure 2. Illustration of our overall framework for the unsupervised asymmetric stereo matching with the proposed spatially-adaptive self-similarity.

mechanism to integrate epipolar constraints to calculate feature similarities. Peng et al. [20] proposed data grafting augmentation, self-distillation loss, and full-scale network. The above methods, both supervised and unsupervised, assume that the stereo images are symmetric.

2.2. Asymmetric Stereo Matching

There have been various methods that address different kinds of asymmetries in stereo matching. There are several methods that address asymmetry in term of light spectrum. Zhi et al. [31] presented a RGB-NIR stereo dataset captured in a driving environment, with additional information of coarse material prediction. They also presented a baseline framework that includes spectral translation, unsupervised stereo estimation, and material-aware confidence. Liang et al. [14] proposed a variant of CycleGAN [33] to translate between RGB and NIR images, followed by unsupervised stereo matching using photometric loss [3]. Walters et al. [24] employed cycle consistency that an image warped twice using left and right disparities is equivalent to its original. Recently, Tosi et al. [23] presented a novel dataset and baseline to address RGB-Multispectral stereo matching. They adopt additional RGB camera to generate pseudo ground-truth and use it as proxy to supervise the RGB-Multispectral stereo network.

Several works focused on asymmetries in visual quality of RGB images, assuming high-quality left and low-quality right images. Liu et al. [15] proposed to generate high-quality right image from the left image, then conduct stereo matching on the high-quality image pair. Even their method works on different kinds of asymmetries including resolution, noise, and rectification error, it is based on a supervised learning and heavily relies on ground-truth of both high-quality right image and disparity. More recently, Chen et al. [2] demonstrated that a key factor for reliable unsupervised learning of stereo matching is symmetry in loss calculation stage, rather than input stage. They further observed that a stereo matching encoder extracts resolution-agnostic features to some extent, and replaced photometric loss [3] with *feature-metric* loss. Motivated by [2], we further enhance the feature consistency at the loss calculation, by presenting a novel spatially-adaptive self-similarity.

2.3. Self-Similarity Based Descriptors

Local Self-Similarity (LSS) [21] is defined as an aggregation of correlations between center and surrounding patches, sampled from discretized log-polar grids. It has shown impressive robustness to domain discrepancy, and resulted in several derivations. Kim et al. [11] proposed the dense adaptive self-correlation (DASC) descriptor, which further improves robustness by randomized receptive pooling and adaptive self-correlation measure. They also learn to select the best performing patterns among the randomized receptive fields. LSS is extended into deep non-CNN architecture in [12] by building a hierarchical self-correlation surfaces. In [10], LSS is reformulated based on CNN as Fully Convolutional Self-Similarity (FCSS). Similar to DASC [11], effective sampling patterns are learned with shifting transformer module. Although the above methods show plausible results in finding correspondences between visually inconsistent images, the learned patterns in DASC [11] and FCSS [10] are fixed for all image locations, thus cannot encode the self-similarity adaptively. Consequently, it requires larger number of sampling patterns to encode diverse geometric information, leading to a computational burden.

3. Proposed Method

3.1. Background

We consider a scenario where of left and right (stereo) images $\mathbf{I}_L, \mathbf{I}_R \in \mathrm{R}^{H \times W \times C}$ are given, where H, W, C are height, width, and color channels of each image. Here, we assume that the stereo images are asymmetric. For each

pixel $\mathbf{x} = (x, y)$ in $\mathbf{I}_L$, the objective of stereo matching is to find the corresponding point $\tilde{\mathbf{x}} = (x - d, y)$ in $\mathbf{I}_R$, where d is referred to as *disparity*. A widely-used paradigm [3] for unsupervised learning of stereo matching is to warp $\mathbf{I}_R$ using the estimated disparity map $\hat{\mathbf{D}}_L$ to obtain the estimated left view $\mathbf{I}_{R2L}(x, y) = \mathbf{I}_R(x - \hat{d}, y)$, and calculate the photometric consistency loss:

$$\mathcal{L}_{pm} = \kappa(\mathbf{I}_L, \mathbf{I}_{R2L}), \tag{1}$$

where κ is a dissimilarity measure such as L_1 or SSIM loss [30]. However, in an asymmetric stereo pair, corresponding points in $\mathbf{I}_L$ and $\mathbf{I}_R$ may not have the same intensity, posing limitations to the photometric consistency loss (1).

A recent work [2] observed that a stereo matching network with asymmetric inputs generate plausible results if the photometric loss is calculated using symmetric data. In other words, a key factor for a plausible unsupervised asymmetric stereo matching is to calculate loss in a symmetric space. Based on this observation, they proposed 'feature-metric' consistency loss for unsupervised asymmetric stereo matching:

$$\mathcal{L}_{fm} = \kappa(\mathbf{G}_L, \mathbf{G}_{R2L}), \tag{2}$$

where $\mathbf{G}$ is asymmetry-agnostic feature directly extracted from stereo encoder with image $\mathbf{I}$ as the input. Motivated by the above observation [2], we aim to extract asymmetry-agnostic features using spatially-adaptive self-similarity to define symmetric loss space for unsupervised asymmetric stereo matching.

3.2. Overview

Our overall framework is illustrated in Fig. 2. We adopt a standard stereo matching network with left and right encoders $\{\Theta_L^E, \Theta_R^E\}$, a cost volume calculator, and a decoder. Given stereo image pair $\{\mathbf{I}_L, \mathbf{I}_R\}$, each image is fed into the corresponding encoder, resulting in left and right features $\{\mathbf{F}_L, \mathbf{F}_R\} \in \mathrm{R}^{H_e \times W_e \times C_e}$, with their height, width, and channel as H_e, W_e, and C_e. Then, the features are used to calculate the cost volume, which is fed into the decoder to estimate the disparity map $\hat{\mathbf{D}}_L$, aligned to $\mathbf{I}_L$. In order to train the network, the images $\mathbf{I}_L$ and $\mathbf{I}_{R2L}$ are used to calculate the photometric loss (1). They are further fed into the encoders and the proposed Spatially-Adaptive Self-Similarity (SASS) module to be embedded into the asymmetry-agnostic features $\mathbf{G}$ to calculate feature-metric loss (2). In order to learn the optimal sampling patterns that encodes asymmetry-agnostic yet distinctive features, contrastive similarity loss $\mathcal{L}_{cs}$ is applied.

3.3. Spatially-Adaptive Self-Similarity

The self-similarity feature [10, 11, 21] $\mathbf{G}(\mathbf{x})$ at pixel $\mathbf{x}$ is defined as a union of L feature values $\mathbf{G}(\mathbf{x}) = \bigcup_l G^l(\mathbf{x})$ for $l \in [1, ..., L]$, where $G^l(\mathbf{x})$ is computed as:

$$G^l(\mathbf{x}) = \max_{\dot{\mathbf{x}} \in \mathcal{N}_\mathbf{x}} \exp\left(-\frac{\mathcal{S}(P(\dot{\mathbf{x}} - \Delta\mathbf{x}_{s_l}), P(\dot{\mathbf{x}} - \Delta\mathbf{x}_{t_l}))}{\gamma}\right), \tag{3}$$

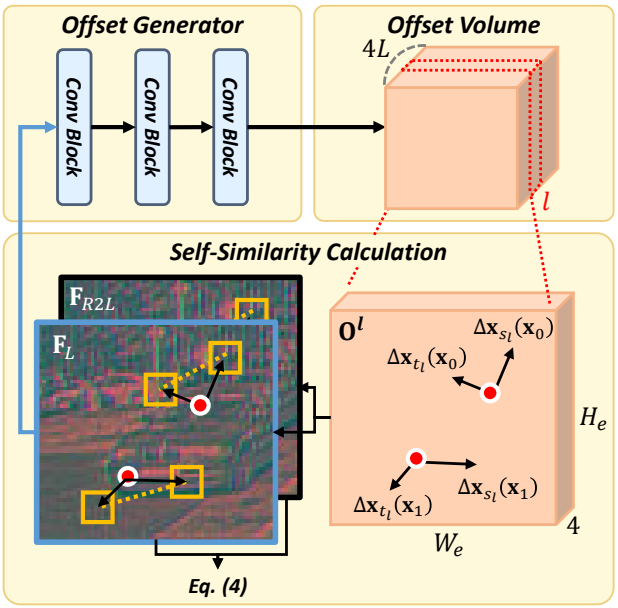


Figure 3. Pipeline of the proposed SASS. The offset generator takes $\mathbf{F}_L$ as input to generate the offset volume $\mathbf{O}$. The adaptive sampling patterns are applied to $\mathbf{F}_L$ and $\mathbf{F}_{R2L}$ to generate SASS features using (4).

where $P(\mathbf{x})$ is a patch with center $\mathbf{x}$, $\{\Delta\mathbf{x}_{s_l}, \Delta\mathbf{x}_{t_l}\}$ are the l^{th} sampling pattern offsets, and $\mathcal{S}$ is a similarity measure. The similarity measures are encoded with exponential function with bandwidth γ, and maximum operation within a support window $\mathcal{N}_\mathbf{x}$. In the previous works [10, 11], the L sampling patterns $\{\Delta\mathbf{x}_{s_l}, \Delta\mathbf{x}_{t_l}\}_{l=1}^L$ are fixed for all pixels as in Fig. 1(a).

Modification from the previous methods to the proposed SASS is straightforward. We generate sampling patterns adaptively, according to the given image and spatial location as in Fig. 1(b), so that the sampling patterns become $\{\Delta\mathbf{x}_{s_l}(\mathbf{x}), \Delta\mathbf{x}_{t_l}(\mathbf{x})\}_{l=1}^L$. To this end, we design offset generator module that consists of several convolutional blocks. The offset generator takes the left feature $\mathbf{F}_L$ as input, and generates offset volume $\mathbf{O}$, which is a concatenation of L offset maps $\{\mathbf{O}^l\}_{l=1}^L$. For the l^{th} sampling pattern, offset map $\mathbf{O}^l \in \mathbb{R}^{H_e \times W_e \times 4}$ is generated, whose first two channels correspond to the vertical and horizontal offsets for $\Delta\mathbf{x}_{s_l}(\mathbf{x})$, and last two channels correspond to those of $\Delta\mathbf{x}_{t_l}(\mathbf{x})$. Consequently, with all L offset maps concatenated, the offset volume $\mathbf{O}$ has dimensions $H_e \times W_e \times 4L$. Finally, we use $P(\mathbf{x}) = \mathbf{F}(\mathbf{x})$, and extract the SASS feature $\mathbf{G} = \bigcup_l G^l \in \mathbb{R}^{H_e \times W_e \times L}$, where G^l is obtained by:

$$G^l(\mathbf{x}) = \max_{\dot{\mathbf{x}} \in \mathcal{N}_\mathbf{x}} \exp\left(-\frac{\mathcal{S}(\mathbf{F}(\dot{\mathbf{x}} - \Delta\mathbf{x}_{s_l}(\mathbf{x})), \mathbf{F}(\dot{\mathbf{x}} - \Delta\mathbf{x}_{t_l}(\mathbf{x})))}{\gamma}\right). \tag{4}$$

Note that the subscripts L and $R2L$ are omitted for brevity. We follow the work in [10] and use simple Euclidean distance for $\mathcal{S}$. The overall procedure is depicted in Fig. 3.

3.4. Contrastive Similarity Loss

The feature-metric loss (2) has to be calculated on a feature space that is agnostic to the asymmetries, yet distinctive for correspondence matching [2]. To achieve this, we propose contrastive similarity loss to encourage the network to generate the effective SASS sampling patterns. We first define positive and negative pixels as pixels with correct and incorrect disparity estimations respectively. In order to define them without ground-truth disparity, we exploit left-right correspondence consistency [27]. To this end, we further obtain $\hat{\mathbf{D}}_R$ by performing horizontal flip and swap the stereo images to feed into the network, then flipping back the output. Then we define the correspondence error as the absolute error of the disparity estimations:

$$\mathcal{E}(\mathbf{x}) = |\hat{\mathbf{D}}_L(\mathbf{x}) - \hat{\mathbf{D}}_R(\tilde{\mathbf{x}})|, \tag{5}$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{D}}_L(\mathbf{x})$. A pixel $\mathbf{x}$ is considered to be positive if $\mathcal{E}(\mathbf{x}) \leq \tau$, and negative otherwise.

Then we define the proposed contrastive similarity loss $\mathcal{L}_{cs}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{cs} =& \frac{1}{|\Omega_p|} \sum_{\mathbf{x}\in\Omega_p} w_p(\mathbf{x})||\mathbf{G}_L(\mathbf{x}) - \mathbf{G}_{R2L}(\mathbf{x})||_2 \\ &+ \frac{1}{|\Omega_n|} \sum_{\mathbf{x}\in\Omega_n} \max(0, M - w_n(\mathbf{x})||\mathbf{G}_L(\mathbf{x}) - \mathbf{G}_{R2L}(\mathbf{x})||_2),\end{aligned} \tag{6}$$

where Ω_p and Ω_n are positive and negative pixels, with their number of pixels $|\Omega_p|$ and $|\Omega_n|$. The first term encourages the SASS features extracted from the asymmetric images to become closer at the matched pixels, so that the SASS feature becomes asymmetry-agnostic. In contrast, the second term constrains the features at non-matched pixels to become apart with a margin of M, so that the SASS feature retains discriminative property for correspondence problem.

The contrastive similarity loss is conceptually similar to contrastive correspondence loss in [10], which encourages the features at matched pixels closer and *vice versa*. We enhance the loss design by further introducing positive and negative weight terms w_p and w_n using cosine similarity. In the positive pixels, we assign higher weights as the raw encoder features $\mathbf{F}$ are dissimilar, and define $w_p = (1 - \cos(\mathbf{F}_L, \mathbf{F}_{R2L}))/2$. For the negative pixels, higher weight is assigned for similar raw features, so that the weight is defined as $w_n = (1 + \cos(\mathbf{F}_L, \mathbf{F}_{R2L}))/2$.

3.5. Total Loss

We introduce the total loss functions for training the proposed networks. We use both photometric ($\mathcal{L}_{pm}$) and feature-metric ($\mathcal{L}_{fm}$) losses, where each loss is defined as summation of L_1 and SSIM losses:

$$\begin{aligned}\mathcal{L}_{pm} =& (1 - \alpha_{pm})||\mathbf{I}_L - \mathbf{I}_{R2L}||_1 \\ &+ \alpha_{pm}(1 - SSIM(\mathbf{I}_L, \mathbf{I}_{R2L})),\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{fm} =& (1 - \alpha_{fm})||\mathbf{G}_L - \mathbf{G}_{R2L}||_1 \\ &+ \alpha_{fm}(1 - SSIM(\mathbf{G}_L, \mathbf{G}_{R2L})),\end{aligned} \tag{8}$$

where α_{pm} and α_{fm} are balancing parameters between the two terms. In addition, we use disparity smoothness loss as follows:

$$\mathcal{L}_{ds} = |\partial_x \mathbf{D}|e^{-|\partial_x \mathbf{I}_L|} + |\partial_y \mathbf{D}|e^{-|\partial_y \mathbf{I}_L|}, \tag{9}$$

where ∂_x and ∂_y are horizontal and vertical gradient operations. The total loss is a weighted summation of the above loss terms, including the contrastive similarity loss (6):

$$\mathcal{L} = \lambda_{pm}\mathcal{L}_{pm} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{cs}\mathcal{L}_{cs} + \lambda_{ds}\mathcal{L}_{ds}. \tag{10}$$

4. Experiments

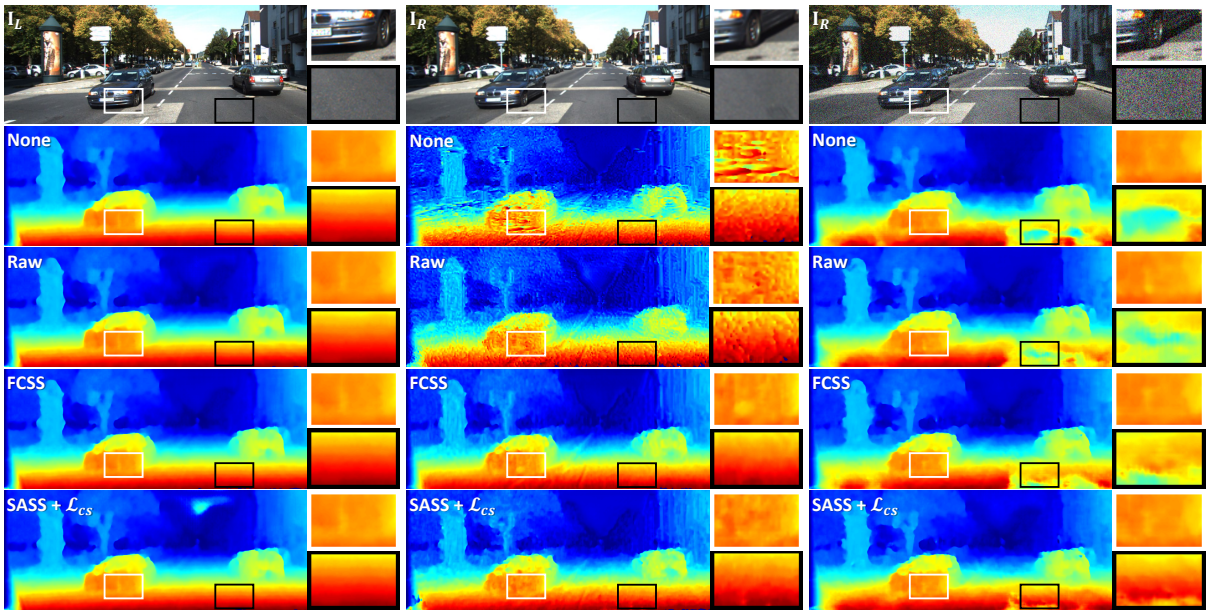
4.1. Experimental Settings

Dataset We use KITTI 2015 stereo dataset [17], which contains street view images captured from a vehicle. It consists of 200 training pairs with semi-dense ground-truth disparity labels obtained with LiDAR, and 200 testing pairs without ground-truth. As in [2], we use the original testing set to train the network, and use original training set to evaluate the performance. We first resize the images into 384×1280 resolution, and apply data augmentation including random crop of size 256×512, and random horizontal flip with left-right swap with probability of 0.5. We also apply photometric jittering of random adjustment of brightness, contrast, saturation in range $[0.8, 1.2]$, and hue in range $[-0.05, 0.05]$.

We consider asymmetries in terms of resolution and noise, assuming high-quality (high-resolution, noise-free) left and low-quality (low-resolution, noisy) right images. We generate the low-resolution image by downsampling the image by the scale s, then upscaling it to the original size using bicubic interpolation. The noisy images are generated with additive Gaussian noise with standard deviation σ. Unless specified otherwise, the experiments are conducted using asymmetry factors $s = 4$ and $\sigma = 0.15$.

Evaluation Metrics In order to quantitatively measure the stereo matching performance, we adopt widely-used metrics: end-point error (EPE) and three-pixel error (3PE). EPE is average of absolute error between the estimation and ground-truth disparities. 3PE is defined as portion of wrong estimation in term of number of pixels. For each pixel, the estimation is categorized as wrong if the absolute error is larger than 3 pixels and larger than 5% of the ground-truth value. The lower values indicate better performance for both metrics. The metrics are calculated for the valid pixels, where the ground-truth disparity is provided.

Implementation Details We adopt the '*stacked hourglass*' architecture of PSMNet [1]. During training, we use the Adam optimizer [13], with its default parameters: $\beta_1 = 0.99$ and $\beta_2 = 0.999$. We use batch size of 2, learning



(a) Symmetric (b) Resolution (c) Noise

Figure 4. Qualitative results of the ablation study experiments on (a) symmetric, (b) resolution asymmetry, and (c) noise asymmetry settings. (first row) (a) left image and (b), (c) right images, (second to fifth rows) stereo matching results of the networks of: Baseline, Raw, FCSS, and SASS + $\mathcal{L}_{cs}$.

Table 1. Quantitative results for the ablation study.

Method	End-point Error (EPE)			Three-pixel Error (3PE)		
	Symmetric	Resolution	Noise	Symmetric	Resolution	Noise
Baseline	1.887	2.652	3.614	8.90	16.22	22.63
Raw	1.891	2.574	3.128	8.92	14.83	18.58
FCSS [10]	1.882	2.329	2.919	8.87	12.33	16.22
SASS	1.884	2.221	2.727	8.90	11.69	15.76
SASS + $\mathcal{L}_{cs-}$	1.882	2.190	2.683	8.86	11.64	14.97
SASS + $\mathcal{L}_{cs}$	**1.879**	**2.183**	**2.667**	**8.83**	**11.57**	**14.84**

rate of 0.0001, and train the network for 50 epochs. We use $\gamma = 0.5$, $M = 0.5$, $\tau = 3$, and $\alpha_{pm} = \alpha_{fm} = 0.15$. We find that the features from the randomly initialized encoders rather distract the training and adopt a two-step training scheme. The first step uses the photometric loss and the loss weights are initially set as $\lambda_{pm} = 1.0, \lambda_{fm} = 0, \lambda_{cs} = 0$ and $\lambda_{ds} = 0.5$. We observe the intermediate testing performance and set $\lambda_{fm} = 1.0$ and $\lambda_{cs} = 0.2$ as the test performance converges. The experiments are conducted on a PC with a 3.60GHz CPU, and a NVIDIA TITAN RTX GPU with 24GB memory.

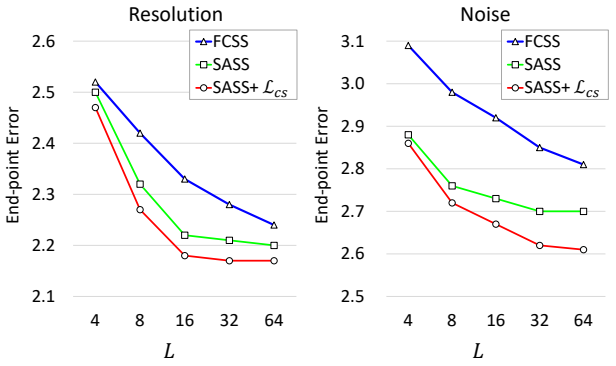
4.2. Ablation Study

Effect of the proposed components We conduct experiments to observe the effect of each proposed component. We train networks with different configurations: i) no feature-metric loss (**Baseline**), feature-metric loss calculated using ii) direct encoder feature (**Raw**), iii) **FCSS** [10] feature, and iv) the proposed **SASS** feature. Each network is trained under different asymmetric settings, including symmetric scenario, whose result can be used as reference. In this experiment, we use the number of sampling patterns $L = 16$ for FCSS and SASS. For the proposed SASS, effects of the contrastive similarity loss without ($\mathcal{L}_{cs-}$) and with the weights ($\mathcal{L}_{cs}$) are also observed.

The quantitative results are presented in Table 1. We first observe that different methods have negligible difference in stereo matching performance under the symmetric condition. The symmetric stereo pair itself provides consistent space for loss calculation, thus additional consistency constraint with the feature-metric loss shows less effect. The stereo network trained without feature-metric loss shows degraded performance under asymmetric conditions, with 2.652 EPE, 16.22 3PE for resolution, 3.614 EPE, 17.96 3PE for noise asymmetries. The performance is slightly improved as the feature-metric loss with direct encoder feature is applied (second row). We observe greatly improved performance as we adopt FCSS [10] on the feature to calculate $\mathcal{L}_{fm}$. The performance is further improved as the proposed SASS, and the contrastive similarity loss are applied. We

Table 2. End-point error results according to different number of self-similarity sampling patterns.

Method	Number of sampling patterns L				
	4	8	16	32	64
Resolution					
FCSS [10]	2.52	2.42	2.33	2.28	2.24
SASS	2.50	2.32	2.22	2.21	2.20
SASS + $\mathcal{L}_{cs}$	**2.47**	**2.27**	**2.18**	**2.17**	**2.17**
Noise					
FCSS [10]	3.09	2.98	2.92	2.85	2.81
SASS	2.88	2.76	2.73	2.70	2.70
SASS + $\mathcal{L}_{cs}$	**2.86**	**2.72**	**2.67**	**2.62**	**2.61**



(a) Resolution (b) Noise

Figure 5. End-point error in (a) resolution and (b) noise asymmetries according to different number of sampling patterns L, for different methods.

observe thebest performance with 2.183 EPE, 11.57 3PE for resolution, and 2.961 EPE, 16.38 3PE for noise asymmetries respectively, when the proposed contrastive loss with weight terms ($\mathcal{L}_{cs}$) is applied.

The qualitative results are presented in Fig. 4. We do not observe noticeable difference in the results of the different methods under the symmetric condition (1st column). With the resolution asymmetry, the baseline network generates results with artifacts, and captures visual textures as disparity discontinuity, such as lanes. On the other hand, the noise asymmetry results in incorrect disparities at flat or dark areas, which are more affected by the noise. The artifacts and degradations are slightly reduced as the raw feature-metric loss is used, and further improved as the FCSS [10] and SASS is applied. We observe that the proposed method generates the most plausible stereo matching results. We also provide the visualization of the sampling patterns of FCSS [10] and SASS in the supplementary material.

Number of sampling patterns In this experiment, we investigate the effect of number of self-similarity patterns on the stereo matching performance. To this end, we train the network using FCSS [10] and the proposed SASS with number of patterns $L = [4, 8, 16, 32, 64]$. For the SASS, we also compare the networks trained with and without the contrastive similarity loss.

Table 3. End-point error of baseline and the proposed method under different asymmetry factors.

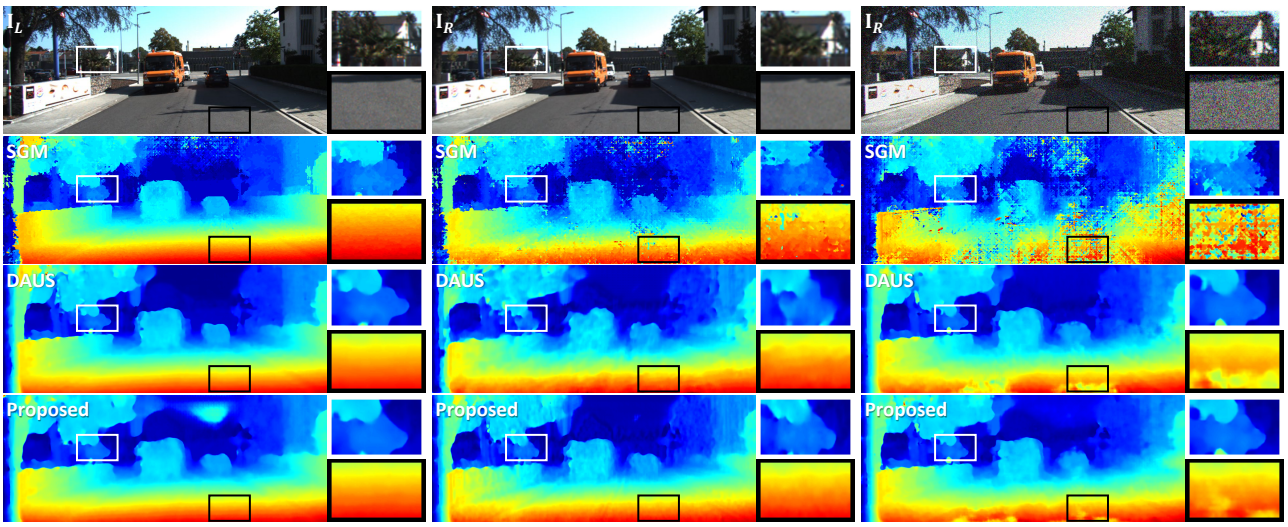
	Resolution			
s	2	4	6	8
Baseline	2.19	2.65	3.28	3.86
Proposed	**1.98**	**2.18**	**2.54**	**2.83**
	Noise			
σ	0.05	0.10	0.15	0.20
Baseline	2.12	2.57	3.61	6.24
Proposed	**1.94**	**2.13**	**2.67**	**3.33**

The quantitative results in term of EPE are presented in Table 2, and plotted in Fig. 5. First, we observe our full method (SASS + $\mathcal{L}_{cs}$) generates the best performance across different values of L. The network with FCSS [10] shows consistently decreasing EPE for larger L. In contrast, the performance of the network with SASS is nearly converged for $L \geq 16$. It can be interpreted that, the SASS adaptively generates effective patterns for each image region to encode asymmetry-agnostic features with less number of the sampling patterns. In the following experiments, we use the proposed method of SASS + $\mathcal{L}_{cs}$, with $L = 16$.

Different asymmetry factors We compare the performance of the baseline network trained without feature-metric loss, and the proposed method under different asymmetric factors for resolution and noise. We use resolution asymmetry factors $s = [2, 4, 6, 8]$, and noise asymmetry factors $\sigma = [0.05, 0.10, 0.15, 0.20]$. The EPE results are presented in Table. 3. As the asymmetry factors increase, the baseline network shows degraded performance. Especially, as σ becomes larger for noise asymmetry, the error drastically increases, as overall pixel values at the corresponding points become further inconsistent. The performance degradation is alleviated by our proposed method consistently across different factors for both resolution and noise asymmetries.

4.3. Comparison to Different Methods

We compare the stereo matching performance of different methods under asymmetric images. We conduct comparisons with semi-global matching (**SGM**) [8], and degradation-agnostic unsupervised stereo (**DAUS**) [2]. We use self-boosting stage number $k = 3$ for DAUS [2]. We also compare the scenario where the right image is restored using super-resolution or denoising [26], followed by stereo matching using SGM [8] and the baseline network. We use the released pre-trained models for the restoration method.



(a) Symmetric (b) Resolution (c) Noise

Figure 6. Qualitative results of the different stereo matching methods under (a) symmetric, (b) resolution asymmetry, and (c) noise asymmetry settings. (first row) (a) left image and (b), (c) right images, (second to fifth rows) stereo matching results of: SGM [8], DAUS [2], and the proposed method.

Table 4. Quantitative results for the comparison experiment.

Method	End-point Error (EPE)			Three-pixel Error (3PE)		
	Symmetric	Resolution	Noise	Symmetric	Resolution	Noise
SGM [8]	5.157	5.810	12.682	34.74	39.96	68.51
Restore [26] + SGM [8]	-	5.739	8.261	-	37.32	51.82
Baseline	1.887	2.652	3.614	8.90	16.22	22.63
Restore [26] + Baseline	-	2.608	3.054	-	15.16	17.04
DAUS [2]	1.887	2.423	3.071	8.90	13.62	17.21
Proposed Method	**1.879**	**2.183**	**2.667**	**8.83**	**11.57**	**14.84**

The quantitative and qualitative results are presented in Table 4 and Fig. 6. We observe the performance degradation of SGM under the asymmetries, which is much severe in the noise asymmetry. The performance is improved when image restoration, *i.e.*, super-resolution or denoising, is applied on the right image as a pre-processing. Similarly, the baseline network generates better results on the stereo pairs with restored right images. Using the feature-metric consistency, DAUS [2] and our proposed method generates plausible results. The proposed SASS and contrastive similarity loss further boost the performance, resulting in the lowest EPE and 3PE measures.

5. Conclusion and Future Works

Conclusion In this paper, we presented a novel spatially-adaptive self-similarity (SASS) for unsupervised asymmetric stereo matching. Motivated by the robustness to domain discrepancy of self-similarity, we generate adaptive self-similarity sampling patterns to effectively encode asymmetry-agnostic feature. We further designed contrastive similarity loss in order to further enhance the asymmetry-agnostic property while maintaining discriminative capability for correspondence search. Consequently, the proposed SASS generates asymmetry-agnostic features to define a symmetric loss space, enabling the plausible learning of stereo matching under asymmetric scenarios in an unsupervised manner. Experimental results on the KITTI 2015 dataset demonstrated the effectiveness of the proposed method under resolution and noise asymmetries.

Future Works Our current framework adopts two-step training scheme as described in Sec. 4.1. Relying on the photometric loss, learning in the the first stage it is limited to RGB-RGB scenario without severe shift in the image intensity, which can be caused under exposure, severe noise, and spectral asymmetries. We reserve the extension of our method to those scenarios as our future work.

Acknowledgements

This research was supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002) and the KIST Institutional Program (Project No.2E32283-23-064).

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 1, 2, 5

[2] Xihao Chen, Zhiwei Xiong, Zhen Cheng, Jiayong Peng, Yueyi Zhang, and Zheng-Jun Zha. Degradation-agnostic correspondence from resolution-asymmetric stereo. In *CVPR*, pages 12962–12971, 2022. 1, 2, 3, 4, 5, 7, 8

[3] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 1, 2, 3, 4

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019. 2

[5] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. *ECCV*, 2014. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI*, 37(9):1904–1916, 2015. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[8] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328–341, 2008. 2, 7, 8

[9] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017. 2

[10] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. FCSS: Fully convolutional self-similarity for dense semantic correspondence. In *CVPR*, pages 6560–6569, 2017. 2, 3, 4, 5, 6, 7

[11] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *CVPR*, pages 2103–2112, 2015. 2, 3, 4

[12] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Deep self-correlation descriptor for dense cross-modal correspondence. In *ECCV*, pages 679–695. Springer, 2016. 2, 3

[13] D. Kingma and J. Ba. Adam: a method for stochastic optimization. *ICLR*, 2015. 5

[14] Mingyang Liang, Xiaoyang Guo, Hongsheng Li, Xiaogang Wang, and You Song. Unsupervised cross-spectral stereo matching by learning to synthesize. In *AAAI*, volume 33, pages 8706–8713, 2019. 1, 3

[15] Yicun Liu, Jimmy Ren, Jiawei Zhang, Jianbo Liu, and Mude Lin. Visually imbalanced stereo matching. In *CVPR*, pages 2029–2038, 2020. 1, 3

[16] N. Mayer and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *CVPR*, 2016. 1, 2

[17] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. *CVPR*, 2015. 1, 5

[18] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1

[19] J. Pang, W. Sun, J. Ren, C. Yang, and Q. Yan. Cascaded residual learning: a two-stage convolutional neural network for stereo matching. *ICCV*, 2017. 1, 2

[20] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *ICCV*, pages 15560–15569, 2021. 3

[21] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *CVPR*, pages 1–8. IEEE, 2007. 2, 3, 4

[22] S. N. Shinha and M. Pollefeys. Multi-view reconstruction using photo-consistency and exact silhouette constraints: a maximum-flow formulation. *ICCV*, 2005. 1

[23] Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. RGB-Multispectral matching: Dataset, learning methodology, evaluation. In *CVPR*, pages 15958–15968, 2022. 1, 3

[24] Celyn Walters, Oscar Mendez, Mark Johnson, and Richard Bowden. There and back again: Self-supervised multispectral correspondence estimation. In *ICRA*, pages 5147–5154. IEEE, 2021. 1, 3

[25] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE TPAMI*, 2020. 1, 2

[26] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE TPAMI*, 2022. 7, 8

[27] J. Zbontar and Y. Lecun. Stereo matching by training a convolutional neural network to compare image patches. *JMLR*, 17(2):1–32, 2016. 2, 5

[28] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019. 2

[29] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. Cross-based local stereo matching using orthogonal integral images. *IEEE transactions on circuits and systems for video technology*, 19(7):1073–1079, 2009. 2

[30] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Tran. on computational imaging*, 3(1):47–57, 2016. 4

[31] Tiancheng Zhi, Bernardo R Pires, Martial Hebert, and Srinivasa G Narasimhan. Deep material-aware cross-spectral stereo matching. In *CVPR*, pages 1916–1925, 2018. 1, 3

[32] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *ICCV*, pages 1567–1575, 2017. 2

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 1, 3