2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

# Image-to-Image Translation for Autonomous Driving from Coarsely-Aligned Image Pairs

Youya Xia*†, Josephine Monica*‡, Wei-Lun Chao§, Bharath Hariharan†,
Kilian Q Weinberger†, Mark Campbell‡

***Abstract*— A self-driving car must be able to reliably handle adverse weather conditions (*e.g.*, snowy) to operate safely. In this paper, we investigate the idea of turning sensor inputs (*i.e.*, images) captured in an adverse condition into a benign one (*i.e.*, sunny), upon which the downstream tasks (*e.g.*, semantic segmentation) can attain high accuracy. Prior work primarily formulates this as an *unpaired* image-to-image translation problem due to the lack of paired images captured under the exact same camera poses and semantic layouts. While perfectly-aligned images are not available, one can easily obtain coarsely-paired images. For instance, many people drive the same routes daily in both good and adverse weather; thus, images captured at close-by GPS locations can form a pair. Though data from repeated traversals are unlikely to capture the same foreground objects, we posit that they provide rich contextual information to supervise the image translation model. To this end, we propose a novel training objective leveraging *coarsely*-aligned image pairs. We show that our coarsely-aligned training scheme leads to a better image translation quality and improved downstream tasks, such as semantic segmentation, monocular depth estimation, and visual localization.**

## I. Introduction

Although the development of autonomous driving perception has rapidly advanced in recent years, work has typically focused on ideal weather conditions. However, to ensure safe operation, these perception systems must also operate robustly under various adverse weather conditions (such as snowy and night) [1]. One possible solution is to train joint or specialized models to cover all conditions. However, this requires having more models or increased model size, as well as costly labeling and retraining for all tasks and conditions. Moreover, sensor signals (*e.g.*, images) in adverse weather are likely of poor quality, making some labeling tasks harder. In this paper, we investigate an alternative idea of turning images captured in adverse weather into a benign one (*i.e.*, sunny), upon which the downstream tasks are more likely to attain high accuracy, even without additional retraining.

This problem generally can be cast as an image-to-image translation task [2], training neural networks to transfer images from a source domain (*e.g.*, night) to another target domain (*e.g.*, sunny). Image-to-image translation training can typically be categorized into paired [3] or unpaired options [4], [5], [6]. Paired translations are trained with perfectly-aligned source and target image pairs, learning to map one to another. While this approach is straightforward, collecting such image pairs is challenging, if not impossible, in many application scenarios. For example, it is infeasible to obtain two driving images with the exact same semantic layouts (*e.g.*, the same set of cars and pedestrians and their placements) under different weather conditions. Thus, for many problem setups (*e.g.*, autonomous driving), existing works usually drop the notion of paired training and utilize the more challenging unpaired training, learning to map between two image distributions, usually accompanied by additional constraints (*e.g.*, cycle consistency [4]).

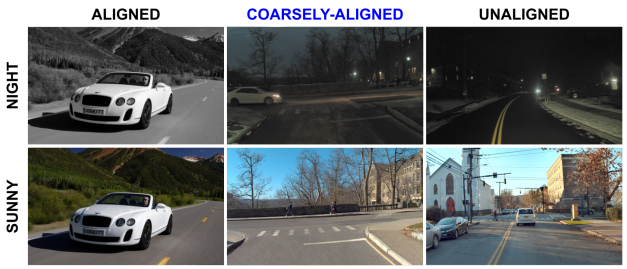


Fig. 1: Left: perfectly-aligned grayscale to color pair from Stanford car dataset [7]. Mid: coarsely-aligned pair from multiple traversals in Ithaca365 dataset [8]. Right: unaligned images [8].

In this paper, we propose Coarsely-Aligned Paired Image Translation (CAPIT), a new fashion for learning the image-to-image translation model by leveraging *coarsely-aligned* image pairs (see Figure 1). We argue that *coarsely-aligned* image pairs are much easier to obtain. For instance, many of us drive the same routes daily (*e.g.*, to and from work and school), even under different weather conditions. We can form coarsely-aligned pairs by picking images that are closest in GPS locations across traversals. While such data may not be perfectly aligned due to the shifts in camera poses or mobile objects at different times, they offer a wealth of training signals. For example, they reveal consistency and appearance changes of background objects between image domains. Moreover, as we will show, these provide sufficient (weak) supervision to train state-of-the-art image-to-image translation models without perfectly-paired data.

Our approach addresses inconsistencies across coarsely aligned images in several ways. First, as two images taken at different times are unlikely to be captured from the exact same camera pose, we propose a misalignment-tolerating L1 loss at the pixel level that first searches for the best local pixel alignment. Second, as foreground objects (*e.g.*, vehicles, pedestrians, and cyclists) may vary in terms of

* Equal contributions
† Computer Science Department, Cornell University {yx454, bh497, kqw4}@cornell.edu
‡ Mechanical and Aerospace Engineering Department, Cornell University {jm2684, mc288}@cornell.edu
§ Department of Computer Science and Engineering, the Ohio State University chao.209@osu.edu

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160815

their identities and placements, we propose to mask them and apply reconstruction loss only to unmasked regions. Furthermore, to accommodate stochastic variations (*e.g.*, due to tree swaying or shadow patterns), we propose a novel usage of the noise contrastive estimation (NCE) loss [9]. This type of loss has been used in different contexts (*e.g.*, image retrieval) to overcome significant misalignment effectively.

We validate CAPIT on the Ithaca365 dataset [8] with driving scenes through the same routes under various weather conditions (*e.g.*, sunny, snowy, and night). We show that CAPIT results in a more faithful image translation model compared to existing paired or unpaired algorithms, not only in the quality of generated images but also in the performance of downstream tasks such as semantic segmentation [10], depth estimation [11], and visual localization [12].

While we focus on autonomous driving applications, our proposed approach is general and applicable to other applications where coarsely-aligned data are easily obtainable. For example, in remote sensing [13], [14], different sensors (*e.g.*, airborne LiDAR and satellite cameras) capture different channels (*e.g.*, height, infrared, and RGB images) with different characteristics for downstream tasks. While these sensors mainly operate alone, they may have surveyed some common areas at different times. Therefore, we can pair their data by GPS locations to learn to transfer one information channel to the others.

## II. Related Work

Early image-to-image translation works focus on paired translation methods, where they apply losses that leverage the strong one-to-one correspondence between input images from the source domain and corresponding images in the target domain to guide the image translation. For example, [3] applies GAN [15] in a conditional setting as well as L1 loss between the input and paired target images. [2] further adds feature matching and perceptual loss and employs a multi-scale generator and discriminator. [16] proposes a spatially-adaptive normalization layer to improve the fidelity and alignment of translated images with the input layout. [17] combines GAN with a classic nearest-neighbor approach to generate high resolution image.

However, obtaining perfectly-aligned paired images is often challenging, if not impossible, in many cases (such as autonomous driving). Thus, the loss functions and training framework designed for paired image translation cannot be applied. Another line of work focuses on unpaired image translation, where image correspondence is not required. For example, CycleGAN [4], ToDayGAN [18], and DiscoGAN [19] use cycle-consistency loss to enforce consistency between the original image and the reconstructed image after one translation cycle. Council-GAN [20] uses a council loss to achieve similar goal. While cycle-consistency loss is a widely used constraint, ACL-GAN [5] proposes to use adversarial-consistency loss to encourage the translated image to retain important information from the source image without the need of cycle translation (*i.e.*, translated image translated back to source domain). CUT [9] proposes a patch-NCE loss that employs a contrastive learning technique [21] to maximize mutual information between input and output (translated) images. Other methods, such as MUNIT [6], DRIT [22], UNIT [23], and ForkGAN [24] focus on learning domain agnostic features to extract shared features between input and output domains.

Although unpaired image translation methods allow training without paired images, the unpaired setup leads to an under-constrained problem and inferior results compared to paired methods. Our proposed work CAPIT, Coarsely-Aligned Paired Image Translation, offers an intermediate solution by leveraging *coarsely-aligned* image pairs that are easily obtainable for autonomous driving. To our knowledge, the closest work to our proposed CAPIT is [25]. Our work CAPIT significantly differs from [25] in that, we propose a framework dedicated to learning image translation using coarsely-aligned image pairs, while the main training in [25] comes from the unpaired cycle-consistency training. [25] requires *manually* selecting *well-aligned* images to fine-tune the image translation model after the main unpaired training stage. During fine-tuning, [25] immediately treats those images with *paired* translation loss and does not address possible misalignment factors, which we show in subsection III-B and subsection IV-C can adversely degrade the resulting generative model.

While some downstream tasks may require removing only the *scattering medium* (*e.g.* snowfall and raindrop) [26] from adverse images, adapting whole street scenes (*e.g.*, removing snow from both the street and cars) can generally produce images that are closer to the target domain and benefit a variety downstream adaptation tasks that may include backgrounds, such as semantic segmentation and depth estimation. We reiterate that our contribution in leveraging coarsely-aligned image pair translation is *orthogonal* and *complimentary* to existing novel image translation methods and architectures, and can even improve many of them.

Another line of work is improving downstream tasks (*e.g.* semantic segmentation) in adverse weather by directly adapting task networks [27], [12], [28]; this allows task-specific information to be used to tailor the approach to the task. However, this often requires task labels (*e.g.*, semantic mask in the ACDC dataset [29]), which is often challenging or infeasible to annotate under adverse conditions. Adapting through image translation, on the other hand, does not need additional labels. Moreover, image translation immediately bridges the domain gap for *all* image-based downstream tasks (*e.g.*, semantic segmentation, visual localization).

## III. Proposed Approach: CAPIT

### A. Problem Setup and Baseline Algorithm

Our goal is to learn an image mapping function $G$ from a source domain $\mathcal{X}$ (*e.g.*, adverse weather) to a target domain $\mathcal{Y}$ (*e.g.*, sunny). To begin, we introduce a *paired* translation baseline, inspired by pix2pix [3]. [3] assumes that the training data are perfectly paired images $\{(x^n, y^n)\}_{n=1}^N$, $x^n, y^n \in \mathbb{R}^{h \times w}$, where $x^n \in \mathcal{X}$ corresponds to $y^n \in \mathcal{Y}$. It

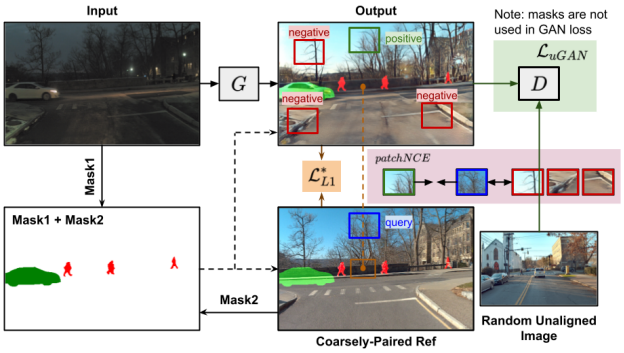


Fig. 2: CAPIT framework for training image translation model $G$ using coarsely-aligned image pairs.

introduces an L1 loss comparing translated image $G(x^n)$ to its corresponding target image $y^n$ in the pixel space:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x^n,y^n}\|G(x^n) - y^n\|_1. \quad (1)$$

[3] also employs a conditional GAN loss [15]:

$$\mathcal{L}_{\text{cGAN}}(G,D) = \mathbb{E}_{x^n,y^n}\left[\log D(y^n;x^n)\right] + \mathbb{E}_{x^n}\left[\log\left(1 - D(G(x^n);x^n)\right)\right], \quad (2)$$

where $D$ denotes a discriminator (*i.e.*, binary classifier) that tells a real image $y$ from a fake/generated one $G(x)$; the discriminator $D$ is conditioned on $x$. Through minimizing both losses with respect to $G$, we encourage the generated images to look real and similar to the target images.

### B. Learning with Coarsely-Paired Images

Obtaining paired images is often infeasible in many scenarios, including autonomous driving, thus prior works typically drop the notion of paired training and utilize the more challenging unpaired training way. Our method leverages *coarsely-aligned* images, which are much easier to obtain. Particularly, in autonomous driving, we can obtain *coarse* pairs between $x$ and $y$ if they are collected from the same area, which is common (*e.g.*, driving to work daily). The Ithaca365 dataset [8] collects images, GPS, and IMU data on the same route under different weather conditions (*e.g.*, sunny, night, and snowy). We pair images from different traversals based on the closest pose (GPS location and IMU orientation); see Figure 1-mid for a training data sample. These image pairs are *not perfectly aligned* due to potential GPS differences/errors, changes in semantic layouts at different collection times, and shifts in camera pose.

Learning with the above two loss functions (Equation 1 and Equation 2) does not explicitly account for the misalignment in the coarsely-aligned image pairs and may adversely degrade the generative model. Thus, instead of directly employing the paired translation algorithms or entirely ignoring any useful paired information in the training data by applying an unpaired algorithm, we propose CAPIT, Coarsely-Aligned Paired Image Translation, dedicated to learning image translation using *coarsely-aligned* image pairs. Specifically, we propose novel loss functions tailored to coarsely-aligned image pairs to address the aforementioned challenges. Figure 2 shows an overview of CAPIT.

*1) Foreground-Masked Loss:* Our first insight about the coarsely-aligned data is that foreground objects (*e.g.*, cars and pedestrians) change in presence and location over time. Thus, two images taken at different timestamps will have different foreground contents and layouts, even if they are taken from the same camera pose. To address this issue, we propose to mask out pixels of the foreground objects in calculating any paired loss. For instance, we calculate the L1 loss only on pixels that belong to the backgrounds in both source and corresponding target images. We denote these background pixel indices as:

$$M(x,y) = \{(i,j) | (c(x_{i,j}) \notin \text{fg}) \wedge (c(y_{i,j}) \notin \text{fg})\}, \quad (3)$$

where $c(\cdot)$ denotes the class label and fg a set of foreground classes. We can segment these objects using annotated/ground-truth masks or off-the-shelf segmentation models [30].

*2) Misalignment-Tolerating L1 Loss:* The original L1 loss penalizes differences in pixel values between the generated and target images at each pixel location, assuming that the two images are perfectly aligned. However, with coarsely-aligned data, the source and target images are unlikely from the same camera poses. Thus, a conventional L1 loss would exaggerate even with a few pixel shifts between the two images (*e.g.*, along the edges), adversely pushing the generator to fit noises. To address this problem, we propose to compare pixel $G(x)_{i,j}$ in the translated image to a set of pixels within a small rectangle window around $i,j$. Specifically, we refer to those pixel indices as

$$r(i,j) = \{(a,b) : |a-i| \le k_h \wedge |b-j| \le k_w\}, \quad (4)$$

where $k_h$ and $k_w$ are the size of the rectangular window. Given pixel location $i,j$ in the translated image, we first find (within the search window $r(i,j)$) the *best* pixel location $i',j'$ in the target image to overcome the slight pixel-to-pixel misalignment. The misalignment-tolerating L1 loss can be formulated as:

$$\mathcal{L}^{\dagger}_{L1}(G) = \mathbb{E}_{x^n,y^n}\frac{1}{hw}\sum_{i,j}\min_{i',j' \in r(i,j)}|G(x^n)_{i,j} - y^n_{i',j'}|. \quad (5)$$

Combining Equation 5 with the foreground masking defined in Equation 3, we obtain the final L1 loss:

$$\mathcal{L}^{\star}_{L1}(G) = \mathbb{E}_{x^n,y^n}\frac{1}{\text{card}(M(x,y))} \sum_{(i,j)\in M(x,y)}\min_{(i',j')\in\{r(i,j)\cap M(x,y)\}}|G(x)_{i,j} - y_{i',j'}|, \quad (6)$$

where $\text{card}(\cdot)$ is the cardinality/number of pixels.

*3) Ranking-based Loss to Overcome Large, Stochastic Variations:* The misalignment-tolerating L1 loss can effectively compensate for small pixel shifts due to slightly different camera poses. However, two images captured at different times contain inevitable *stochastic* factors that can affect the image contents and cause significant misalignment, even if they are from the same camera pose with masked-out

foreground objects. For example, a tree might be posed differently due to wind, which changes the leaves' position and shadow patterns on sidewalks. These stochastic variations are not accounted for by the previous two strategies and can significantly distort pixel values. Thus, we propose a paired loss term that is more lenient to these stochastic variations. Concretely, we build upon a key insight: although stochastic variations exist, patches of translated and corresponding target images from the same position should still be more similar (*e.g.*, similar structures and contents) than patches from a different position.

We realize this idea by *adapting* the patchNCE loss from [9]. Being an unpaired translation method, [9] computes the patchNCE loss between the source and translated images. Unlike [9], we compute the patchNCE loss between the translated and coarsely-aligned target images, leveraging our coarsely-aligned data. Let $H$ be a feature extractor that reuses the encoder of the translation model $G$, followed by a two-layer MLP; and $H_s^l$ be the extracted feature at a spatial location $s$ and layer $l$. Our patchNCE loss is:

$$\mathcal{L}^*_{\text{patchNCE}}(G,H) = \mathbb{E}_{x^n,y^n} \sum_{l=1}^{L}\sum_{s} \ell\left(H_s^l(G(x^n)), H_s^l(y^n), \{H_{s'}^l(y^n)|s' \neq s\}\right), \quad (7)$$

where $\ell$ is a cross-entropy loss to encourage the query $v = H_s^l(G(x^n))$ to match (more similar) with the positive/correct sample $v^+ = H_s^l(y^n)$ over negative samples $\{v^-\} = \{H_{s'}^l(y^n)|s' \neq s\}$:

$$\ell(v, v^+, \{v^-\}) = -\log \frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{\{v^-\}} \exp(v \cdot v^-/\tau)}. \quad (8)$$

Here, $\tau$ is a temperature constant scalar. Furthermore, we apply $M(x,y)$ to mask out the foreground objects on the patchNCE computation.

### C. *Final* CAPIT *Training Objective*

Our final CAPIT training objective combines the loss terms introduced in subsection III-B with an unpaired-GAN (uGAN) loss [4]:

$$\mathcal{L}_{\text{CAPIT}} = \mathcal{L}_{\text{uGAN}}(G,D) + \lambda_1 \mathcal{L}^\star_{L1}(G) + \lambda_2 \mathcal{L}^*_{\text{patchNCE}}(G,H), \quad (9)$$

where $\lambda_1, \lambda_2 \in \mathbb{R}^+$ are weighting coefficients. While it is natural to follow [3] to use conditional-GAN (cGAN) loss, we *empirically* find that unpaired-GAN (uGAN) loss [4] results in better performance (see subsection IV-C and subsection IV-F); thus we employ uGAN loss. We learn $G$ and $H$ to minimize the objective, while $D$ to maximize the objective. CAPIT is in theory *agnostic* to different model architectures (*i.e.*, $G$ and $D$). In this paper, we adopt a coarse-to-fine generator and a multi-scale discriminator from [2].

## IV. Experiments

First, we evaluate our *coarsely*-paired image translation approach on the Ithaca365 dataset [8] and compare our results to several baseline methods, as shown in subsection IV-B. We evaluate the quality of image translation using the FID metric [31] and several downstream tasks. In subsection IV-C, we perform an ablation study, demonstrating the key components of our approach in leveraging *coarsely*-paired dataset. In subsection IV-D, we provide an analysis on the effect of mask accuracy in foreground-masked loss computation. In subsection IV-E, we provide an analysis on the effect window size in misalignment-tolerating loss computation. Finally, in subsection IV-F we do an additional experiment on another datasets (Facades [32] and map to aerial photos [3] datasets) to justify the design choice in our GAN loss.

### A. *Datasets*

Our main experiment uses the Ithaca365 dataset [8] which contains diverse scenes (*e.g.*, urban, rural, and campus area) along a repeated 15km route. We use one traversal per weather condition (night and snowy as source domains and sunny as target domain). Each condition is split into $5208/1138$ images for training and testing, where the two splits do not have overlapping location. The images are paired between different conditions based on closest GPS poses. We use pretrained Mask R-CNN [33], [30] instance segmentation to obtain the foreground masks for our loss computation in Equation 3. For additional analysis on GAN loss (subsection IV-F), we also do experiments on additional datasets, Facades [32] and map to aerial photos [3] datasets.

### B. *Main Results*

We compare our approach to several unpaired translation models (CUT [9], ACL-GAN [5], MUNIT [6], CycleGAN [4], ForkGAN [24], and ToDayGAN [18]) to show the benefit of leveraging *coarsely*-paired images for supervision. Figure 3 shows some *qualitative* results, which demonstrate that our approach can translate an image closer to the reference image. Table I shows the *quantitative* performance of our approach compared with other baselines through different metrics and dowstream tasks (*i.e.*, semantic segmentation, monocular depth estimation, and visual localization).

*1) FID:* We assess the quality of the translated images through Frechet Inception Distance (FID) [34] to indicate distance between the distribution of translated images (fake sunny images) and of target images (real sunny images). Table I shows that our approach has lower FID than the baselines, meaning that our translated images are more similar to the real sunny images. We also compute the FID between real sunny images and original adverse weather images, *i.e.*, snowy and night, to show the performance gap in adverse conditions without any domain adaptation. Finally, the FID between two different sunny traversals is computed (last row in Table I) as a lower bound (best case) indication.

*2) Semantic Segmentation:* We evaluate the performance of image translation model through a downstream task of semantic segmentation using pixel accuracy (PixAcc) as the evaluation metric. *All* semantic segmentation experiments are evaluated using the PSPNet [10] pretrained on the MIT ADE20k dataset [35]. First, we apply the PSPNet model for semantic segmentation directly on night and snowy images. Then, to assess the quality of our image translation model,

TABLE I: Evaluation of CAPIT vs other unpaired translation methods through FID and several downstream tasks. We report the result of night/snowy to sunny translation on the Ithaca365 dataset.

| No | Method | FID↓ | PixAcc (%)↑ | $\delta_{1.25}$ (%)↑ | loc-err(m)↓ |
|---|---|---|---|---|---|
| 1. | No adaptation | 162.4 / 165.2 | 50.9 / 68.4 | 70.7 / 74.5 | 22.5 / 12.1 |
| 2. | CycleGAN | 158.3 / 136.9 | 52.8 / 64.2 | 78.3 / 73.8 | 9.3 / 5.7 |
| 3. | MUNIT | 96.3 / 162.4 | 47.4 / 58.8 | 77.0 / 67.1 | 10.8 / 7.9 |
| 4. | CUT | 94.0 / 99.2 | 53.8 / 67.8 | 79.5 / 75.8 | 9.3 / 5.0 |
| 5. | ACL-GAN | 89.4 / 101.1 | 48.0 / 68.5 | 77.0 / 73.0 | 11.2 / 6.6 |
| 6. | ToDayGAN | 83.2 / 87.8 | 57.4 / 73.5 | 80.7 / 76.4 | 8.9 / 3.4 |
| 7. | ForkGAN | 80.9 / 101.2 | 54.2 / 68.7 | 81.4 / 76.1 | 8.8 / 5.4 |
| 8. | CAPIT (ours) | **69.4** / **76.2** | **61.9** / **75.2** | **83.0** / **76.8** | **6.3** / **2.9** |
| 9. | Real sunny | 60.1 | 89.2 | 88.5 | 2.2 |

we apply the same model on the fake sunny images translated from night and snowy using our image translation model, as well as other baseline models. Table I shows that our image translation method improves the performance of semantic segmentation in adverse conditions more than all baselines.

As the Ithaca365 dataset [8] does not provide semantic segmentation labels, we generate *pseudo* ground-truth label by applying the pretrained PSPNet on the sunny *coarse* correspondence images. Additionally, when computing pixel accuracy, if a pixel has a moving foreground label in either the original adverse image or the reference sunny image, it is not included in the accuracy computation. As the pseudo ground-truth reference is approximate, we run semantic segmentation on *another* set of *coarsely-aligned* sunny images, comparing against the pseudo ground-truth to give a sense of achievable upper bound accuracy (last row in Table I).

*3) Monocular Depth Estimation:* We also evaluate the performance of image translation through monocular depth estimation downstream task. We use a standard inlier metric $\delta_{\text{threshold}}$: the percentage of pixels whose inaccuracy level $\max\left(\frac{\hat{d}_{i,j}}{d_i}, \frac{d_{i,j}}{\hat{d}_{i,j}}\right)$ are below a threshold (we use 1.25); $d_{i,j}$ and $\hat{d}_{i,j}$ are the ground-truth and predicted depth for pixel $i, j$ [11]. We train BTS [11] monocular depth estimation model on the sunny data using projected LiDAR as the sparse ground truth. We test the same depth model (trained on sunny condition) on the adverse weather condition images without domain adaptation, as well as on the fake sunny images translated using our approach and several baseline models. Finally, we evaluate the depth model on the in-domain real sunny dataset to give us an upper bound (best case) accuracy (last row in Table I). We show (Table I) the benefit of our image translation model compared to the baselines by the domain-induced performance gap that it closes.

*4) Image-based Localization:* Our final downstream task evaluation is on image-based localization, formulated as an image-retrieval problem. Specifically, given a *query* image, we find the *closest* image using SIFT matching [36] from a set of images in *library*. We evaluate based on localization error (loc-err), *i.e.* Euclidean distance from query image location to its match. We use images from a sunny traversal as the *library*, while the query images come from the adverse weather images (night and snowy). First, we directly use the adverse weather image as the *query* without any adaptation; this serves as a bottom baseline. Then, we employ image translation models to first convert the night/ snowy images to fake sunny images for query. The result in (Table I) shows that our image translation approach improves the localization error (loc-err) more than the baselines.

### C. Ablation Study

We conduct an ablation study to assess the gains from each of our proposed components and loss terms (shown in Table II). We begin with a baseline *paired* translation algorithm (Row-1) that uses conditional-GAN (cGAN) and L1 loss [3]. Comparing Row-1 and Row-2 shows that *foreground masking* consistently improves the performance. Comparing Row-2 and Row-3, we see that *unpaired-GAN (uGAN) loss* outperforms the original paired conditional-GAN (cGAN) loss. Therefore, we employ the uGAN loss instead of cGAN loss. Comparing Row-3 and Row-4, we see that using *masked patchNCE loss* between the generated image and the coarsely-aligned target image can largely overcome the stochastic misalignment and improve the performance. Finally, comparing Row-4 and Row-5 shows that accounting for the pixel misalignment in the *misalignment-tolerating L1 loss* computation can further boost the model.

TABLE II: Ablation study of loss terms for night/snowy to sunny translation on the Ithaca365 dataset.

| Loss | FID ↓ |
|---|---|
| cGAN + L1 | 114.0 / 122.1 |
| cGAN + L1 (+mask) | 106.9 / 119.8 |
| uGAN + L1 (+mask) | 94.8 / 114.8 |
| uGAN + L1 (+mask) + NCE (+mask) | 68.5 / 86.6 |
| uGAN + L1 (+mask + misalignment) + NCE (+mask) | 69.4 / 76.2 |

In short, while we have shown (subsection IV-B) the benefit of leveraging the coarsely-aligned data, *the gains are by no means trivial.* Indeed, by comparing Row-1 in Table II (baseline paired translation) to existing unpaired algorithms in Table I, we hardly see the benefit of coarsely-aligned data. However, with our proposed losses, which are dedicated to tackling the misalignment due to foreground objects, camera pose shifts, and stochastic variations, our CAPIT can leverage the valuable contextual information provided by the coarsely-paired data to excel in image translation.

### D. Effect of Mask Accuracy on Foreground-Masked Loss

Masks are used to mask out foregrounds (which are unlikely static) in the reconstruction loss computation during *training* only; they can be pre-generated prior to training. As masks are only used for training, *ground-truth masks can be used* if available. If the dataset does not provide such labels (such as in the Ithaca365 dataset), we can can apply pre-trained Mask R-CNN to each *coarsely*-aligned image pair and take the *union* as masks (Equation 3). Table III shows the FID evaluation of image translation different *confidence thresholds* are used for Mask R-CNN mask proposals acceptance. Higher threshold produces lower FP but higher FN. False positive (FP) errors mask out loss computation

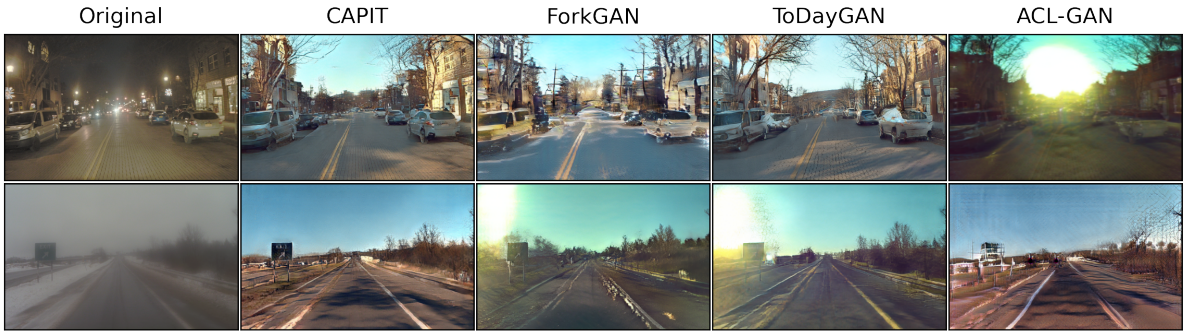


Fig. 3: Visualization results of image translation of our approach compared to several baselines.

on unnecessary area, reducing the training signal amount. False negative (FN) errors compute reconstruction loss on foreground areas that may not match in the two images, causing *noise* in the training signal. As shown, Mask R-CNN *default* threshold (0.5) produces the most accurate masks.

TABLE III: Mask confidence threshold study for night/snowy to sunny translation on the Ithaca365 dataset.

| Threshold | 0.9 | 0.7 | 0.5 | 0.3 | 0.1 |
|---|---|---|---|---|---|
| FID ↓ | 85.6 / 92.1 | 79.4 / 85.2 | 69.4 / 76.2 | 73.2 / 80.5 | 81.8 / 90.1 |

### E. Effect of Window Size on Misalignment-Tolerating Loss

Window size (in Equation 4) is studied in Table IV. If images are perfectly aligned, we can directly compare pixels of two images at the same location. However with misalignment, the window acts a search space of correspondence when comparing the reconstruction loss. A window too small cannot cover misaligned pixels, while too large over-simplifies the translation problem (*i.e.*, find similar but irrelevant pixels). We use $k_w = k_h = k = 3$ in Equation 4.

TABLE IV: Window size study for night/snowy to sunny translation on the Ithaca365 dataset.

| $k$ | 1 | 3 | 5 |
|---|---|---|---|
| FID ↓ | 75.9 / 89.7 | 69.4 / 76.2 | 73.2 / 84.3 |

### F. Additional Analysis on GAN Loss

In the previous ablation study, we find that using unpaired-GAN (uGAN) loss results in better performance than conditional-GAN (cGAN) for our *coarsely-paired* training. Inspired by this finding, we do further analysis (Table V) on different ways of using GAN loss on well-aligned datasets: Facades [32] and map-to-aerial photos [3]. Specifically, we train the Pix2pix [3] paired translation model using three different GAN losses.

Let (realA, realB) be a *paired* batch of images from domain A & B. First, in the *conditional-GAN* training, we use a conditional-discriminator following the original Pix2Pix setup [3], where every image fed into the discriminator is conditioned on the input image (realA). Second, in the *paired-GAN* training, we feed $G$(realA) and realB into the discriminator $D$. Third, in the *unpaired-GAN* training, we sample *another* batch of images realB' from domain B, and feed $G$(realA) and realB' into $D$.

Interestingly, we find that the unpaired-GAN loss combined with paired L1 results in the best performance for all three experiments. This result not only aligns with our previous finding and supports our GAN loss design, but also suggests that we should reexamine the discriminator design even with perfectly-aligned image pairs. Note that while the overall training objectives (*i.e.*, expectation over all images) of paired and unpaired GAN are the same, they lead to different training dynamics of $D$ in mini-batch training. We find this important since GAN training iterates between generator and discriminator training and is known to be less stable. Specifically, we find that *unpaired* training can better capture the overall difference between real and fake images, while *paired* training overly focuses on each paired batch.

TABLE V: Analysis on three different ways of feeding GAN discriminator loss. FID-1 is the FID for Label→Facades, PixAcc-1 is the per-pixel label accuracy for Facades→Label, and FID-2 is the FID for Map→Aerial.

| Loss | PixAcc-1 (%) ↑ | FID-1 ↓ | FID-2 ↓ |
|---|---|---|---|
| conditional-GAN + L1 | 49.2 | 165.4 | 111.1 |
| paired-GAN + L1 | 51.1 | 130.5 | 101.2 |
| unpaired-GAN + L1 | **74.6** | **124.7** | **91.4** |

## V. Conclusion

We propose an approach to leverage the *coarsely-aligned* datasets for image-to-image translation. Our proposed approach, Coarsely-Aligned Paired Image Translation (CAPIT), is built upon several key insights of the coarsely-aligned data. These include foreground masking, local re-alignment to overcome pixel shifts, and a ranking-based loss to address stochastic variation. CAPIT improves both the generated image quality and the performance of downstream tasks in autonomous driving. Furthermore, our approach is general and can be extended to other applications where coarsely-aligned data is easily available.

## Acknowledgement

This research is supported by grants from the ONR MURI (N00014-17-1-2699), the National Science Foundation NSF (IIS-1724282, TRIPODS-1740822, IIS-2107077, and IIS-2107161), the Office of Naval Research DOD (N00014-17-1-2175), the DARPA Learning with Less Labels program (HR001118S0044), and the Cornell Center for Materials Research with funding from the NSF MRSEC (DMR-1719875).

## REFERENCES


[1] Matthew Pitropov, Danson Evan Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5):681–690, 2021.

[2] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[5] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *ECCV*, 2020.

[6] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[7] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[8] Carlos Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, Wei-Lun Chao, Bharath Hariharan, Kilian Weinberger, and Mark Campbell. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

[9] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[11] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[12] Tayyab Naseer, Wolfram Burgard, and Cyrill Stachniss. Robust visual localization across seasons. *IEEE Transactions on Robotics*, 34(2):289–302, 2018.

[13] James B Campbell and Randolph H Wynne. *Introduction to remote sensing*. Guilford Press, 2011.

[14] Ralph O Dubayah and Jason B Drake. Lidar remote sensing for forestry. *Journal of forestry*, 98(6):44–46, 2000.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[16] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Pixelnn: Example-based image synthesis. *CoRR*, abs/1708.05349, 2017.

[18] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019.

[19] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017.

[20] Ori Nizan and Ayellet Tal. Breaking the cycle - colleagues are all you need. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2020.

[21] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[22] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation viadisentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020.

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[24] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. Forkgan: Seeing into the rainy night. In *European conference on computer vision*, pages 155–170. Springer, 2020.

[25] Horia Porav, Will Maddern, and Paul Newman. Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1011–1018. IEEE, 2018.

[26] Zheng Shi, Ethan Tseng, Mario Bijelic, Werner Ritter, and Felix Heide. Zeroscatter: Domain transfer for long distance imaging and vision through scattering media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3476–3486, 2021.

[27] Andreas Pfeuffer and Klaus Dietmayer. Robust semantic segmentation in adverse weather conditions by means of sensor data fusion. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2019.

[28] Aashish Sharma, Loong-Fah Cheong, Lionel Heng, and Robby T Tan. Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions. In *2020 International Conference on 3D Vision (3DV)*, pages 23–31. IEEE, 2020.

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021.

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[32] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *Proc. GCPR*, Saarbrucken, Germany, 2013.

[33] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[34] Naresh Babu Bynagari. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Asian Journal of Applied Science and Engineering*, 8:25–34, 2019.

[35] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.